Improving Spatial Resolution of Low Dosage CBCTs through Neural Networks

2025 Research Aid Awards (RAA)

Dr John Nelson

johnn@unc.edu
O: 203-482-0728

FollowUp Form

Award Information

In an attempt to make things a little easier for the reviewer who will read this report, please consider these two questions before this is sent for review:

- *Is this an example of your very best work, in that it provides sufficient explanation and justification, and is something otherwise worthy of publication? (We do publish the Final Report on our website, so this does need to be complete and polished.)*
- *Does this Final Report provide the level of detail, etc. that you would expect, if you were the reviewer?*

Title of Project:*
Improving Spatial Resolution of Low Dosage CBCTs through Neural Networks

Award Type
Research Aid Award (RAA)

Period of AAOF Support
July 1, 2025 through June 30, 2026

Institution
University of North Carolina

Names of principal advisor(s) / mentor(s), co-investigator(s) and consultant(s)
Laura Jacox, Kelly Mitchell, Don Tyndall

Amount of Funding
$6,000.00

Abstract

(add specific directions for each type here)

ABSTRACT & AIMS: Cone beam computed tomography (CBCT) imaging accurately represents 3D volumetric structures for detecting caries and guiding implant placement, but at a higher radiation dose compared to conventional 2D X-rays. In orthodontics, CBCTs are used for visualization of impacted teeth, asymmetric growth, and orthognathic surgical planning. Some orthodontists are utilizing CBCTs for standard-of-care diagnostic imaging, with the caveat of radiation. Although low-dose CBCT scans can be acquired, the reduced spatial resolution coupled with high noise limits these scans' diagnostic value, and traditional upscaling interpolation techniques are often ineffective.

Machine learning has shown neural networks (NNs) capable of complex pattern identification with great promise for imaging applications. To develop an approach to provide low radiation, high resolution CBCT data, we are proposing the application of NNs to upscale low-dose CBCT scans. After CBCT signal acquisition, reconstruction algorithms (i.e., back projection, iterative reconstruction) average data points into adjacent voxels, causing loss of spatial resolution. Decomposition or separation of these merged signals can then be approximated through training of a NN to increase resolution. Deep NNs consist of multiple layer types (i.e. convolutional, encoder-decoder) with each type exceling at different attributes (i.e. spatial reasoning, sequence-to-sequence). Multiple layers can be sequenced to further improve predictive ability. Using input and prediction datasets fed into widely used forms of NN training (in gradient descent via backpropagation), the parameters (or weights) of a NN can be optimized to make more accurate predictions. This NN could then accurately upscale lower resolution CBCTs.

Our objective is to evaluate the accuracy of NNs in upscaling low-dose CBCT scans.

Our hypothesis is that NNs can significantly reduce the radiation burden associated with CBCT scans and improve spatial resolution of CBCT scans. To this end, we propose:

AIM 1: Determine an optimal NN architecture for upscaling 3D CBCT data. Leading NN structures will be evaluated for their ability to upscale low-dose, low spatial resolution (LSR) CBCT data; resulting upscaled low-spatial resolution (ULSR) scans will be compared against high spatial resolution (HSR) scan data. First, baseline voxel variation (BVV) of the CBCT machine will be determined by 20 successive HSR scans in one position. An ingestion pipeline with drop-in design for multiple network architectures will be trained and evaluated on successive LSR and HSR scans of five phantom skulls at 20 rotations and translations. The dataset will be divided into 80:20 train-validation folds with data augmentation techniques applied to each train pair. Loss functions during training will optimize the NN using mean-squared-error (MSE) and maximum voxel error between each LSR and HSR training pair. Lowest validation MSE will identify the best network. A t-test will compare BVV to MSE as a measure of error from the upscaling algorithm.

AIM 2: Quantify lower limits of CBCT scanning parameters for accurate upscaling.

Consecutive HSR and LSR scan pairs will be taken varying mA, kVp, and voxel size at machine specific increments. Each scan setting will have 10 scan pairs performed at varied positions. The ULSR volume will be generated by each NN. MSE and maximum voxel error will be computed between ULSR and HSR volumes. A t-test will be performed to compare the BVV to each scanning parameter group ($p < 0.05$), with no difference indicating accurate upscaling.

Respond to the following questions:

Detailed results and inferences:*

If the work has been published, please attach a pdf of manuscript below by clicking "Upload a file".

OR

Use the text box below to describe in detail the results of your study. The intent is to share the knowledge you

have generated with the AAOF and orthodontic community specifically and other who may benefit from your study. Table, Figures, Statistical Analysis, and interpretation of results should also be attached by clicking "Upload a file".

JDR-26-0639_Proof.pdf
Manuscript was submitted for publication but has not been accepted yet.

Were the original, specific aims of the proposal realized?*

AIM 1 evolved to capture the optimization of a U-Net architecture (a type of convolutional network commonly used in image upsampling) and the error function was derived from a hybrid discriminative power - capturing multiple domains of image differences versus just raw voxel values.

AIM 2 involved training multiple networks around medium to high enhancement tasks and low to high enhancement tasks. The data used for training was strictly phantom data showing promising enhancement results on holdout subjects. A small held-out human sample also showed enhancement improvements from a phantom-only trained network, thereby supporting a strong case of transfer learning across phantom to human domains.

Larger sample sizes and across multiple scanners is now needed to further generalize these findings.

Were the results published?*

No

Have the results of this proposal been presented?*

Yes

To what extent have you used, or how do you intend to use, AAOF funding to further your career?*

The AAOF's current support of funding has helped realize applying my background in software engineering to solve modern orthodontic and dental problems. By mitigating technological costs, I was able to pursue this line of research to help conclude a feasibility study around using neural networks to enhance the spatial resolution of CBCT images. This is not only a benefit to us as providers but more importantly a benefit to our patients. The funding helped to support patient participants, acquire phantoms used in the study, and support the computation resources needed to generate a neural network.

Accounting: Were there any leftover funds?*

If "yes", enter your best estimate and work with your grants manager to finalize financial reports and send refund payable to: AAOF
Attn: George
401 N. Lindbergh Blvd.
St. Louis, MO. 63141-7839

If "no", enter zero.

$0.00

Not Published

Are there plans to publish? If not, why not?*

It was just submitted to Journal of Dental Research for publication consideration. Manuscript is attached above. AAOF support was included in acknowledgements section of the manuscript.

Presented

Please list titles, author or co-authors of these presentation/s, year and locations:*

2026 AAO Table Clinic - Improving Spatial Resolution of Low Dosage Cone Beam Computed Tomography (CBCT) Scans through Neural Networks
2026 AAO William Profit Resident Scholar Competition - Improving Spatial Resolution of Low Dosage Cone Beam Computed Tomography (CBCT) Scans through Neural Networks

Was AAOF support acknowledged?

If so, please describe:

It was acknowledged for support in all acknowledgement sections.

Internal Review

Reviewer comments

It appears that larger sample sizes are required to further realize the proposed aims. The study resulted in two presentations. A manuscript is currently in review by Journal of Dental Research.

Reviewer Status*

Approved

File Attachment Summary

Applicant File Uploads

- JDR-26-0639_Proof.pdf

Journal of
Dental Research

Enhancing Spatial Resolution of Low Dose CBCT Scans through Neural Networks

Journal:	*Journal of Dental Research*
Manuscript ID	JDR-26-0639
Manuscript Type:	Research Reports
Date Submitted by the Author:	01-May-2026
Complete List of Authors:	Nelson, John; The University of North Carolina at Chapel Hill Adams School of Dentistry, Orthodontics Moyer, Ilan; Massachusetts Institute of Technology, Mechanical Engineering Atkinson, Sophia; The University of North Carolina at Chapel Hill Adams School of Dentistry, Biomedical Sciences Wu, Di; The University of North Carolina at Chapel Hill Adams School of Dentistry, Biomedical Sciences Mitchell, Kelly; The University of North Carolina at Chapel Hill Adams School of Dentistry, Orthodontics Tyndall, Don; The University of North Carolina at Chapel Hill Adams School of Dentistry, Department of Diagnostic Sciences Jacox, Laura; University of North Carolina at Chapel Hill School of Dentistry, Biomedical Sciences; The University of North Carolina at Chapel Hill Adams School of Dentistry, Orthodontics
Keywords:	Computed tomography, Digital imaging/radiology, Engineering, Imaging, Orthodontic(s), Oral & Maxillofacial Surgery
Abstract:	Introduction: Cone beam computed tomography (CBCT) provides three-dimensional diagnostic imaging at higher radiation doses than conventional radiographs. Reducing dose lowers spatial resolution, creating a tradeoff between patient safety and diagnostic quality. Neural network super-resolution suggests a possible solution, with three fundamental challenges that we address. First, training data is typically derived from synthetic downsampling rather than real acquisition characteristics. Second, there is no quantitative definition of success relative to inherent scan-to-scan variation. Third, there is no principled framework for selecting loss functions that can discriminate genuine quality differences from scan-to-scan variation. Methods: Real paired CBCT scans were acquired at three dose settings (Low, Medium, High) on a Carestream 9600 scanner using anthropomorphic phantom skulls. Baseline variation was established from 20 consecutive same-setting scans, defining a quantitative goalpost for enhancement quality. A discriminative power framework ranked candidate loss functions by their ability to separate between-setting quality differences from baseline variation, and the top-ranking components were combined into a data-driven hybrid loss. A 3D U-Net with residual learning was trained on 643 voxel patches and validated on

	held-out phantom volumes and nine human subjects. Results: All three models significantly reduced the quality gap between low resolution and standard resolution scans, measured as the percentage of error closed toward baseline variation. The Low-to-High model closed 67% of this gap ($p < 0.001$, $r = +0.94$), the Medium-to-High (Phantom) model 44% ($p < 0.001$, $r = +0.99$), and the Medium-to-High (Human) model 25% ($p < 0.001$, $r = +0.87$) with human data used only in validation. Conclusions: This study demonstrates real paired CBCT acquisitions capture actual dose-reduction characteristics for network training. A baseline variation measure provides an objective goalpost for enhancement quality. A discriminative power framework enables data-driven loss function selection. These results suggest potential for meaningful dose reduction in CBCT imaging.

Title: Enhancing Spatial Resolution of Low Dose CBCT Scans through Neural Networks

Abstract

Introduction: Cone beam computed tomography (CBCT) provides three-dimensional diagnostic imaging at higher radiation doses than conventional radiographs. Reducing dose lowers spatial resolution, creating a tradeoff between patient safety and diagnostic quality. Neural network resolution enhancement suggests a possible solution, but three challenges remain. First, training data is typically derived from synthetic downsampling rather than real acquisition characteristics, limiting upscaling performance. Second, there is no quantitative definition of success relative to inherent scan-to-scan background variation. Third, there is no principled framework for selecting appropriate error metrics that can discriminate genuine quality differences from scan-to-scan variation.

Methods: Real paired CBCT scans were acquired at three dose settings (Low, Medium, High) on a Carestream 9600 scanner using anthropomorphic phantom skulls. Baseline scanner variation was adopted as a quantitative goalpost for enhancement quality, and was established from 20 consecutive same-setting scans. A discriminative power framework ranked candidate loss functions by their ability to separate quality differences from baseline variation. A 3D U-Net with residual learning was trained on 64^3 voxel patches and validated on unseen phantom volumes and nine human subjects, leading to three independently trained models.

Results: All three models significantly reduced the quality gap between low resolution and standard resolution scans, measured as the percentage of error closed toward baseline variation. The Low-to-High model closed 67% of this gap ($p < 0.001$, $r = +0.94$), the Medium-to-High (Phantom) model 44% ($p < 0.001$, $r = +0.99$), and the Medium-to-High (Human) model 25% ($p < 0.001$, $r = +0.87$) with human data used only in validation.

Conclusions: This study demonstrates real paired CBCT acquisitions capture actual dose-reduction characteristics for network training. A baseline variation measure provides an objective goalpost for enhancement quality. A discriminative power framework enables data-driven loss function selection. These results suggest potential for meaningful dose reduction in CBCT imaging.

Introduction

Cone beam computed tomography (CBCT) provides three-dimensional (3D) volumetric imaging that is increasingly used across dentistry, with larger field-of-view scans used in orthodontics and oral surgery for evaluating dental impactions, skeletal asymmetries, and surgical planning (Hounsfield 1995; Gupta and Ali 2013; LightForce Product Launch 2024 2024 Jan 30; CBCT | Invisalign Provider). However, effective CBCT dosages (68–1073 μSv depending on protocol) remain substantially greater than conventional two-dimensional dental radiographs (1–30 μSv) (Shin et al. 2014; Ludlow et al. 2015; Radiation doses in dental radiology 2017 Aug 7). The effective dose can be reduced by adjusting field of view, voxel size, current, voltage, and exposure time, but spatial resolution, defined as the ability to discern fine anatomical boundaries and detail, is proportional to these dosage parameters. Even when voxel dimensions remain constant, lower dose settings produce images with more noise and reduced contrast between adjacent structures, effectively lowering spatial resolution. Thus low dose, low spatial resolution (LSR) scans expose patients to less radiation but contain more noise than standard dose, high spatial resolution (HSR) scans acquired at the same voxel size,

limiting diagnostic power. Pediatric patients carry up to a 3-fold increase relative radiation risk compared to adults (National Research Council (U.S.) 2006; Harrison et al. 2023), making the "As Low As Reasonably Achievable" (ALARA) principle particularly important for orthodontic imaging where repeated scans over multi-year observation and treatment courses are common (Abdelkarim 2019; X-Rays Radiographs). Given the benefits of 3D imaging contrasted with its potential risks, there is an opportunity to explore whether LSR scans might be computationally enhanced to approach HSR diagnostic quality while minimizing radiation exposure.

The simplest computational approach to enhancing LSR images is interpolation, which estimates missing values from neighboring voxels. Interpolation can physically increase the size of an image but only produces smoother images without recovering detail. Neural networks can learn complex relationships from data and offer a promising path forward. CBCT volumetric reconstruction uses filtered back projection, which distributes signal information from any single physical point across adjacent voxels (Schofield et al. 2020). This means that neighboring voxels may contain relevant information about the original signal that is not perceptible to the human eye but that a computational learning approach might recover. However, most existing resolution enhancement approaches create training pairs by artificially downsampling high quality images, preventing the model from learning the complex physical interactions associated with quality reduction in real CBCT scans (Umehara et al. 2018; Yin et al. 2023). Generative adversarial networks (GANs) have been proposed as a convincing alternative, but they hallucinate fine detail that may not correspond to real anatomy, an unacceptable risk in medical imaging (Ledig et al. 2017). Super-resolution training on real paired CBCT acquisitions at different dose settings holds promise to exploit real-world physics-driven differences to improve upscaling performance but remains unexplored.

Even with appropriate training data, there is no established quantitative definition of "good enough" for enhanced medical images. Reader studies, where clinicians compare original and enhanced images, carry a familiarity bias. There is evidence that readers rate objectively superior images lower when enhancement changes the appearance they are accustomed to (Geyer et al. 2015; Laurent et al. 2019). Standard computational metrics such as peak signal-to-noise ratio (PSNR) and mean squared error (MSE) avoid this perceptual bias but provide no threshold for when improvement is meaningful versus incidental. Any measured difference between two successive scans includes inherent scan-to-scan variation. An objective, reproducible goalpost is needed, defined in terms of this scan-to-scan variation rather than subjective assessment.

During training, a neural network iteratively measures the error between its output and a known target, adjusting its weights to improve future predictions. This error measure, called the loss function, determines what the network learns to prioritize. MSE, a common choice, is insufficient on its own because multiple distinct outputs can produce identical MSE values, causing the network to converge on blurred averages (Kim et al. 2019). Composite loss functions offer improved performance as standard options either do not transfer to medical imaging (i.e. perceptual losses derived from natural photographs (Johnson et al. 2016; Azizi et al. 2021)) or are confounded by the intensity scale changes between dose settings (i.e. Structural Similarity Index, SSIM (Wang et al.

2004)). More fundamentally, without a reference for how much variation is inherent to the scanner versus representative of real quality differences, there is no principled way to determine which loss components are meaningful. Loss functions are typically selected empirically or adopted from related tasks without verifying their discriminative capacity in the target imaging context (Ghosh et al. 2017; Yuan et al. 2019).

This study addresses these three challenges. First, real paired CBCT scans acquired at three dose settings on anthropomorphic phantom skulls provide training data that captures real acquisition characteristics rather than synthetic degradation. Second, baseline variation established from repeated same-setting scans defines a quantitative goalpost where success is measured as closing the gap between the between-setting error and the baseline variation of the imaging system. Third, a discriminative power framework uses this baseline variation to rank candidate loss functions by their ability to separate genuine quality differences from scan-to-scan variation, enabling data driven loss selection rather than empirical tuning. The network is trained on phantom data and validated on both phantom and human subjects excluded from training to assess generalizability, with relevance for growing orthodontic patients who undergo repeated imaging over multi-year observation and treatment courses.

Methods

Dataset

Four anthropomorphic phantom skulls (three adult, one pediatric) were used for training data. Two featured axially reconfigurable components, producing eight unique skull configurations. Each configuration was scanned at three randomly varied positions, and at each position, scans were acquired at three dose settings on a Carestream 9600 CBCT scanner (Carestream Dental, Atlanta, GA), referred to as Low (276 mGy-cm² dose area product, DAP), Medium (796 mGy-cm²), and High (1592 mGy-cm²) (**Figure 1A, 1B**). All settings shared 0.30 mm isotropic voxel size but differed in tube voltage, current, and scan duration. This produced 24 phantom-position configurations, each containing LSR/HSR scan pairs. Because phantoms remain stationary between acquisitions, differences between dose settings are predominantly dose-related rather than motion-related, providing the closest achievable approximation to identical positioning.

Nine non-growing human subjects (21-27 years old, 3 female, 6 male) about to undergo orthodontic treatment were scanned at Medium and High settings. Subjects were selected with minimal intra-oral metal to reduce scatter artifacts not represented in phantom training (inclusion/exclusion criteria in **Table S1**). No human data was used during training as this dataset served exclusively for validation. The Institutional Review Board (IRB) approved this study (IRB # 25-[redacted]), and written informed consent was obtained from all participants. For both phantom models, the dataset was split into 80% training and 20% testing. The human model used all phantom data for training and was tested against the unseen human scans. (**Figure 1C, 1D**).

Baseline Variation

To establish a quantitative goalpost for enhancement quality, twenty consecutive scans of a single phantom were acquired in a fixed position at both High and Low settings (**Figure 1A**). Pairwise error metrics computed across these same-setting scans produced

a distribution of reference scan error, the amount of variation present even between repeated scans at the same dose setting. This reference scan error, referred throughout as baseline variation, represents the best achievable agreement between any two scans and defines the measurable target for enhancement. An enhanced low dose scan is successful to the extent that its error relative to the High dose reference approaches this baseline variation. Gap closure, the percentage of the initial-to-baseline error gap closed by the model, provides a single interpretable measure of progress toward this goalpost.

Pre-processing

LSR and HSR volumes required spatial alignment due to gantry and reconstruction binning errors. Sub-voxel rigid body registration was performed using six degrees of freedom (three translational, three rotational; no scaling or warping). Alignment used phase correlation for initial estimation (Foley et al. 2016), iterative gradient refinement (Livyatan et al. 2003), and tricubic interpolation for final resampling. The High dose volume served as the fixed reference for between-setting pairs. The first acquired volume served as the reference for baseline variation pairs.

Registration quality was assessed using normalized cross-correlation (NCC), chosen over SSIM for its intensity invariance across dose settings. Root mean squared error (RMSE) and PSNR provided complementary alignment measures. Each volume was independently min-max normalized to [0,1] because different scan settings produce fundamentally different intensity distributions.

Network Architecture

The primary architecture was a 3D U-Net implemented in PyTorch (Ronneberger et al. 2015; PyTorch). A residual learning strategy was employed, where the network predicted only the correction needed to improve the input. The enhanced output, referred to as upscaled LSR (ULSR), was formed by adding this correction to the original LSR input (**Figure 1E**) (He et al. 2015). This approach is well suited to super resolution because most of the input is already correct; only edges, fine structures, and noisy regions require adjustment.

Training used 64^3 voxel patches rather than whole volumes, which reduces memory requirements, increases dataset size, and ensures the network learns the general relationship between noisy and clean tissue rather than memorizing specific anatomies. All volumes were downsampled 2x in each spatial dimension (0.5x scaling, 8x data reduction) to fit within available memory. Detailed description of memory requirements in Supplemental Methods. Augmentation was limited to interpolation-free geometric transformations (axis flips and 90-degree rotations) to avoid introducing blurring that would undermine the enhancement objective. Together, patch-based training and conservative augmentation force the network to learn generalizable image quality relationships rather than anatomy-specific patterns. Full model architecture, training hyperparameters, and loss weights are detailed in Appendix Table 2**.**

Loss Function Selection

Using the baseline variation, discriminative power (DP) was computed for each candidate loss component. DP is the ratio of between-setting error to baseline variation (**Figure 1F**). A DP near one indicates the metric cannot distinguish real quality differences

from scan-to-scan variation, while higher values indicate reliable detection of genuine quality differences. Candidate losses spanning pixel error, edge, and structural domains were ranked by DP, and the top-ranking components were combined into a weighted hybrid loss (**Figure 1G**). Loss component weights were determined by DP analysis rather than manual tuning.

Evaluation

Standard image quality metrics (PSNR, RMSE, mean absolute error (MAE), NCC) were computed at both patch and volume levels in normalized [0,1] space. The hybrid loss components used during training were also computed on validation data, with the weighted total model loss serving as the primary outcome measure. Statistical significance was assessed using non-parametric tests. The Wilcoxon signed-rank test was used for paired comparisons (the same patch or volume measured before and after network processing), and the Mann-Whitney U test for independent group comparisons (e.g., phantom vs. human). Significance was set at $p < 0.05$, with matched-pairs rank-biserial effect size (r). Pseudocode is included in the Appendix Supplemental methods.

Results

Registration and Initial Error

Sub-voxel registration was required for all scan pairs. Registration significantly improved alignment across all groups, with consistent improvements in NCC, RMSE, and PSNR in every case. Human pairs required substantially larger corrections than phantom pairs, as expected given patient repositioning between acquisitions. Full registration quality metrics and initial error breakdowns are provided in Appendix Tables 4–10.

Discriminative Power and Loss Selection

MSE had the highest discriminative power, followed by gradient MAE and local contrast (**Table 1**). Loss components were selected based on DP ranking and phantom-to-human consistency. Components that showed significant divergence between phantom and human distributions ($p < 0.05$) were excluded to ensure generalizable loss weights (Appendix Table 11).

Model Output and Performance

Three models were trained using the 3D U-Net with residual learning (**Figure 1D**). All three showed significant reduction in error from initial values (**Table 2**).

At the patch level, median gap closure was 67% for the Low-to-High model ($p < 0.001$, $r = +0.93$), 44% for the Medium-to-High (Phantom) model ($p < 0.001$, $r = +0.99$), and 25% for the Medium-to-High (Human) model ($p < 0.001$, $r = +0.87$). At the volume level, the Low-to-High model achieved 59% gap closure, the Medium-to-High (Phantom) model 40%, and the Medium-to-High (Human) model 22% (all $p < 0.05$, all $r = +1.00$). These reductions are visualized in **Figure 2**, where output distributions shift toward the baseline variation relative to initial values across all models.

Standard interpretable metrics confirmed these improvements. Both phantom models showed statistically significant improvements in PSNR, RMSE, and NCC at the

patch level (all $p < 0.001$), with large effect sizes. The Medium-to-High (Human) model showed significant structural improvement (NCC, $p < 0.001$) but did not reach significance for pixel-level metrics, consistent with the smaller quality gap and greater anatomical complexity. Volume-level results confirmed significant improvement for the Low-to-High model across all standard metrics (all $p < 0.05$, $r = +1.00$). Full metric breakdowns are provided in Appendix Tables 12-15. **Figure 3** shows representative error heatmaps demonstrating visible reduction in voxel-level discrepancy, with corrections concentrated in anatomically relevant tissue.

Discussion

This study demonstrates that a deep learning model trained exclusively on phantom CBCT scans can significantly reduce the quality gap between low dose and standard dose acquisitions, closing 67% of the error gap on phantom validation and 25% on human validation. These results were enabled by three methodological contributions that address fundamental challenges in medical image enhancement.

To our knowledge, this is the first study to use real, same-modality paired CBCT scans acquired at different dose settings for super-resolution training, rather than relying on synthetic downsampling or cross-modality reference standards such as micro-CT (Rytky et al. 2024; Chen et al. 2026). Even large-scale synthetic degradation efforts, such as Thummerer et al. applying ESRGAN to nearly 3,000 retrospectively downsampled CBCT scans for radiotherapy planning, rely on artificially constructed training pairs rather than prospectively acquired dose-paired data (Fok et al. 2024; Peng et al. 2025; Thummerer et al. 2025). Our paired approach captures the real-world physics of dose reduction rather than a synthetic approximation. **Critically, the network was trained on 64^3 voxel patches from only four physical phantoms, yet a model trained exclusively on this phantom data achieved 25% gap closure on nine human subjects never seen before in training data.** This transfer suggests the network learned generalizable image quality relationships rather than memorizing specific anatomies, a direct consequence of patch-based training that prevents the network from ever seeing a complete skull.

The reference scan error established from 20 repeated same-setting scans provides an objective, reproducible definition of enhancement quality. Gap closure measured relative to baseline variation is more informative than raw metrics alone, because it describes performance relative to the best achievable agreement between any two scans. An error gap closure of 100% means the enhanced scan's error is no greater than the variation between two reference-dose scans, making it computationally indistinguishable from a native high dose acquisition. This framework is reproducible, and any laboratory can establish scanner-specific baseline variation and use it to benchmark enhancement methods on a common scale.

The discriminative power framework provided a principled, data-driven method for selecting and weighting loss function components, in contrast to the common practice of selecting losses empirically or adopting them from related tasks without verifying their discriminative capacity in the target context. This framework also predicted a seemingly

counterintuitive result. The Low-to-High model achieved greater improvement (67% gap closure) than the Medium-to-High model (44%) despite starting from a noisier input. This is most likely explained by the larger quality gap between Low and High settings being more distinguishable from scan-to-scan variation, giving the network a stronger signal to learn from.

The Medium-to-High (Human) model's 25% error gap closure with no human training data represents a floor, not a ceiling. Three factors likely explain the discrepancy between phantom (44%) and human (25%) performance. These include greater residual registration error from patient repositioning, more complex anatomy including soft tissue heterogeneity, and transfer learning from a model trained exclusively on phantom data (Yoon et al. 2024). NCC was the only metric reaching volume-level significance for humans, consistent with its intensity-invariant robustness. Even this modest improvement suggests larger improvement is achievable with human training data, following a pre-train on phantoms / fine-tune on humans approach.

Several limitations should be acknowledged. Regarding training data, only four physical phantoms were available. Axial reconfiguration and repositioning expanded this to 24 unique phantom-position pairs, and patch extraction yielded hundreds of training samples per volume, but these remain correlated observations from a small number of underlying anatomies. No scans included metal artifacts from braces, bonded retainers, implants, or surgical plates. All data was acquired on a single Carestream 9600 scanner and generalization to other CBCT systems is currently unknown.

Regarding methodology, rigid body registration requires tricubic interpolation and introduces slight smoothing. Even under controlled phantom conditions, registration corrections of approximately 0.3 mm (one voxel width) were required to correct systematic sub-voxel gantry offset (Kim et al. 2023). Perfect voxel-to-voxel correspondence is therefore unattainable in this paired-scan design and the interpolation smoothing introduces a ceiling on sharpness gains. This is an accepted tradeoff of real paired acquisitions, which ensures the training signal reflects genuine scanner behavior rather than simulated degradation. Additionally, training at half resolution may introduce incidental denoising through spatial averaging during downsampling which would warrant investigation at native resolution with sufficient computation resources.

Regarding evaluation, all comparisons used computational metrics rather than clinical diagnostic accuracy assessments, which would require reader studies with orthodontists or radiologists evaluating specific diagnostic tasks. The computational goalpost established by the baseline variation does not replace reader studies for clinical deployment but provides a principled target to train toward before future subjective evaluation.

The Low-to-High model's 67% error gap closure suggests potential for a nearly 6-fold dose reduction (276 vs. 1592 mGy-cm² DAP) with substantial quality recovery. Notably, the Low-to-High model's output error visually approximates the initial error of the Medium-to-High model in Figure 2, suggesting the network may have recovered image quality equivalent to acquiring at the medium dose level, though a formal equivalence test comparison between the models was not performed. Future work should incorporate

human training data, expand to multiple scanner datasets, evaluate metal artifact performance, and conduct reader studies to assess relationship between quantitative gap closure metrics and diagnostic accuracy improvements. The discriminative power framework may prove broadly applicable to other imaging modalities where acquisition variation limits conventional training losses.

Conclusion

This study presents three contributions to deep learning enhancement of low dose CBCT scans. First, real paired acquisitions at different dose settings provide training data that captures real acquisition characteristics. The network learned generalizable features that transferred from phantom to human anatomy without any human training data. Second, baseline variation from repeated same-setting scans provides an objective, reproducible goalpost for enhancement quality, enabling measurement of progress toward a computationally indistinguishable result. Third, a discriminative power framework enables data-driven selection and weighting of loss functions, replacing empirical tuning with a principled measure of each component's ability to separate genuine quality differences from scan-to-scan variation. Together, these contributions closed 67% of the quality gap at a 83% dose reduction on phantom data, with results suggesting potential for meaningful dose reduction in dental CBCT imaging.

Acknowledgements

We thank all human participants for engaging with this study. We appreciate the support and assistance from members of the Jacox lab.

Funding Acknowledgements

This work was supported by the Southern Association of Orthodontists Research Award and the American Association of Orthodontists Foundation Resident Research Award ([redacted]). The project was supported by the NIDCR with a K08 grant (K08DE030235) ([redacted].) The content is solely the responsibility of the authors and does not represent the official views of the NIH.

Tables

Table 1. Loss Function Selection

Table 2. Model Performance Summary

Figures Legends

Figure 1: Study overview. **A**: Study overview showing the two-phase workflow. Phase 1 contains data acquisition and pre-processing on a Carestream 9600 CBCT scanner (Carestream Dental, Atlanta, GA), including anthropomorphic phantom and human scan collection at three dose settings, registration, and baseline variation establishment from repeated same-setting scans. Phase 2 contains neural network training using a 3D residual U-Net architecture with 64^3 voxel patch-based learning. **B**: Acquisition settings for the Carestream 9600 CBCT scanner. DAP = dose area product. All settings use 0.30 mm isotropic voxels. High represents the standard clinical acquisition setting and serves

as the reference standard. **C**: Sample summary. 8 phantoms at 3 positions (24 scan pairs), 9 human subjects for validation. L = Low, M = Medium, H = High. **D**: Three models trained with identical architecture and loss recipe. Dose reduction = percentage DAP decrease from target to input setting. The human model was trained on all phantom data and validated on unseen human subjects. **E**: Residual learning. The network predicts a correction f(LSR) added to the input. **F**: Discriminative power quantifies how well each metric separates between-setting error from baseline variation. **G**: The hybrid loss combines selected components with data-driven weights (λ) determined by discriminative power analysis (Table 1).

Figure 2: Weighted total loss distributions comparing initial error (LSR vs HSR), model output (ULSR vs HSR), and the same-setting baseline variation. **A**: Patch-level distributions (whiskers at p5/p95). **B**: Volume-level distributions (whiskers at min/max). Significance brackets show Wilcoxon signed-rank test results (*p < 0.05, p < 0.01,* **p < 0.001) with matched-pairs rank-biserial correlation (r).

Figure 3: Representative slice comparisons of LSR input, ULSR network output, and HSR reference standard. **A.** Sagittal view. Top row shows full sagittal slices with the region of interest outlined (red box) (A1-A3). Middle row shows the magnified ROI of an incisor (A4-6). Bottom row shows absolute error maps (|LSR−HSR| and |ULSR−HSR|) and a signed error improvement map (green = error reduced, red = error increased) (A7-9). Error maps are computed in normalized intensity space and rescaled to the HSR gray value range for comparison. **B.** Axial view. Same image layout as A from the axial plane, showing a cross section through the maxillary region (B1-9). **C.** Edge detail recovery. Sobel gradient-based edge error maps compare initial edge error (LSR vs HSR) and output edge error (ULSR vs HSR) (C1-2), with HSR reference (C3). Cyan and red dotted boxes highlight the region of interest, with magnified insets below (C4-6).

References

Abdelkarim A. 2019. Cone-Beam Computed Tomography in Orthodontics. Dent J. 7(3):89. https://doi.org/10.3390/dj7030089

Azizi S et al. 2021. Big Self-Supervised Models Advance Medical Image Classification. [accessed 2026 Mar 17]. http://arxiv.org/abs/2101.05224. https://doi.org/10.48550/arXiv.2101.05224

CBCT | Invisalign Provider. [accessed 2024 Sept 13]. https://www.invisalign.com/align-digital-platform/clincheck/cbct-integration

Chen P et al. 2026. Deep learning super-resolution for dental CBCT using micro-CT reference and edge loss function. J Dent. 164:106209. https://doi.org/10.1016/j.jdent.2025.106209

Fok WYR et al. 2024. Deep learning in computed tomography super resolution using multi-modality data training. Med Phys. 51(4):2846–2860. https://doi.org/10.1002/mp.16825

Foley D et al. 2016. Phase correlation applied to the 3D registration of CT and CBCT image volumes. Phys Medica Eur J Med Phys. 32(4):618–624. https://doi.org/10.1016/j.ejmp.2016.02.009

Geyer LL et al. 2015. State of the Art: Iterative CT Reconstruction Techniques. Radiology. 276(2):339–357. https://doi.org/10.1148/radiol.2015132766

Ghosh A, Kumar H, Sastry PS. 2017. Robust Loss Functions under Label Noise for Deep Neural Networks. [accessed 2026 Mar 7]. http://arxiv.org/abs/1712.09482. https://doi.org/10.48550/arXiv.1712.09482

Gupta J, Ali SP. 2013. Cone beam computed tomography in oral implants. Natl J Maxillofac Surg. 4(1):2–6. https://doi.org/10.4103/0975-5950.117811

Harrison JD et al. 2023. Effective doses and risks from medical diagnostic x-ray examinations for male and female patients from childhood to old age. J Radiol Prot Off J Soc Radiol Prot. 43(1). https://doi.org/10.1088/1361-6498/acbda7

He K, Zhang X, Ren S, Sun J. 2015. Deep Residual Learning for Image Recognition. [accessed 2024 Sept 15]. http://arxiv.org/abs/1512.03385

Hounsfield GN. 1995. Computerized transverse axial scanning (tomography): Part I. Description of system. 1973. Br J Radiol. 68(815):H166-172

Johnson J, Alahi A, Fei-Fei L. 2016. Perceptual Losses for Real-Time Style Transfer and Super-Resolution. [accessed 2026 Mar 17]. http://arxiv.org/abs/1603.08155. https://doi.org/10.48550/arXiv.1603.08155

Kim B, Han M, Shim H, Baek J. 2019. A performance comparison of convolutional neural network-based image denoising methods: The effect of loss functions on low-dose CT images. Med Phys. 46(9):3906–3923. https://doi.org/10.1002/mp.13713

Kim Y-J et al. 2023. Novel Procedure for Automatic Registration between Cone-Beam Computed Tomography and Intraoral Scan Data Supported with 3D Segmentation. Bioengineering. 10(11):1326. https://doi.org/10.3390/bioengineering10111326

Laurent G et al. 2019. Full model-based iterative reconstruction (MBIR) in abdominal CT increases objective image quality, but decreases subjective acceptance. Eur Radiol. 29(8):4016–4025. https://doi.org/10.1007/s00330-018-5988-8

Ledig C et al. 2017. Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network. [accessed 2026 Mar 17]. http://arxiv.org/abs/1609.04802. https://doi.org/10.48550/arXiv.1609.04802

LightForce Product Launch 2024. 2024. LightForce; [accessed 2024 Sept 13]. https://lf.co/sp/resources/blog/2024productlaunch

Livyatan H, Yaniv Z, Joskowicz L. 2003. Gradient-based 2-D/3-D rigid registration of fluoroscopic X-ray to CT. IEEE Trans Med Imaging. 22(11):1395–1406. https://doi.org/10.1109/TMI.2003.819288

Ludlow JB et al. 2015. Effective dose of dental CBCT-a meta analysis of published data and additional data for nine CBCT units. Dento Maxillo Facial Radiol. 44(1):20140197. https://doi.org/10.1259/dmfr.20140197

National Research Council (U.S.), editor. 2006. Health risks from exposure to low levels of ionizing radiation: BEIR VII Phase 2. National Academies Press.

Peng L et al. 2025. TOWARDS REALISTIC DATA GENERATION FOR REAL- WORLD SUPER-RESOLUTION.

PyTorch. PyTorch; [accessed 2024 Sept 15]. https://pytorch.org/

Radiation doses in dental radiology. 2017. [accessed 2024 Feb 8]. https://www.iaea.org/resources/rpop/health-professionals/dentistry/radiation-doses

Ronneberger O, Fischer P, Brox T. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. In: Navab N, Hornegger J, Wells WM, Frangi AF, editors. Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. Springer International Publishing; p 234–241. https://doi.org/10.1007/978-3-319-24574-4_28

Rytky SJO et al. 2024. Clinical Super-Resolution Computed Tomography of Bone Microstructure: Application in Musculoskeletal and Dental Imaging. Ann Biomed Eng. 52(5):1255–1269. https://doi.org/10.1007/s10439-024-03450-y

Schofield R et al. 2020. Image reconstruction: Part 1 – understanding filtered back projection, noise and image acquisition. J Cardiovasc Comput Tomogr. 14(3):219–225. https://doi.org/10.1016/j.jcct.2019.04.008

Shin HS et al. 2014. Effective doses from panoramic radiography and CBCT (cone beam CT) using dose area product (DAP) in dentistry. Dentomaxillofacial Radiol. 43(5):20130439. https://doi.org/10.1259/dmfr.20130439

Thummerer A et al. 2025. Deep learning based super-resolution for CBCT dose reduction in radiotherapy. Med Phys. 52(3):1629–1642. https://doi.org/10.1002/mp.17557

Umehara K, Ota J, Ishida T. 2018. Application of Super-Resolution Convolutional Neural Network for Enhancing Image Resolution in Chest CT. J Digit Imaging. 31(4):441–450. https://doi.org/10.1007/s10278-017-0033-z

Wang Z, Bovik AC, Sheikh HR, Simoncelli EP. 2004. Image quality assessment: from error visibility to structural similarity. IEEE Trans Image Process. 13(4):600–612. https://doi.org/10.1109/TIP.2003.819861

X-Rays Radiographs. [accessed 2024 Feb 8]. https://www.ada.org/resources/research/science-and-research-institute/oral-health-topics/x-rays-radiographs

Yin J, Xu S-H, Du Y-B, Jia R-S. 2023. Super resolution reconstruction of CT images based on multi-scale attention mechanism. Multimed Tools Appl. 82(15):22651–22667. https://doi.org/10.1007/s11042-023-14436-8

Yoon JS et al. 2024. Domain Generalization for Medical Image Analysis: A Review. Proc IEEE. 112(10):1583–1609. https://doi.org/10.1109/JPROC.2024.3507831

Yuan T et al. 2019. Signal-to-Noise Ratio: A Robust Distance Metric for Deep Metric Learning. [accessed 2026 Mar 7]. http://arxiv.org/abs/1904.02616.

Table 1. Loss Function Selection

Rank	Metric	Domain	DP	P-H p-value	Status	Weight	Rationale
1	MSE	Pixel error	4.7x	0.1401	Consistent	α = 0.70	Full-spectrum pixel guidance
2	Laplacian MAE	Texture	2.6x	0.0454*	Divergent	—	—
3	Gradient MAE	Edge	2.6x	0.1889	Consistent	β = 0.15	Edge preservation
4	Local Contrast	Edge	2.1x	0.1889	Consistent	γ = 0.15	Multi-scale sharpness
5	MAE	Pixel error	2.0x	—	—	—	—
6	NCC	Structural	1.7x	—	—	—	—
7	SSIM 3D	Structural	1.2x	—	—	—	—

Table 1: Loss function selection guided by discriminative power (DP), phantom-to-human consistency, and domain coverage. **DP**: ratio of between-setting error to baseline variation (1F). Higher values indicate better separation. **P-H p-value**: Mann-Whitney U test comparing phantom vs. human distributions. 'Consistent' ($p \geq 0.05$) indicates transferability across anatomy types. **Weight**: coefficient in the hybrid loss (1G).

Table 2. Model Performance Summary

A. Standard Imaging Metrics of Model Output

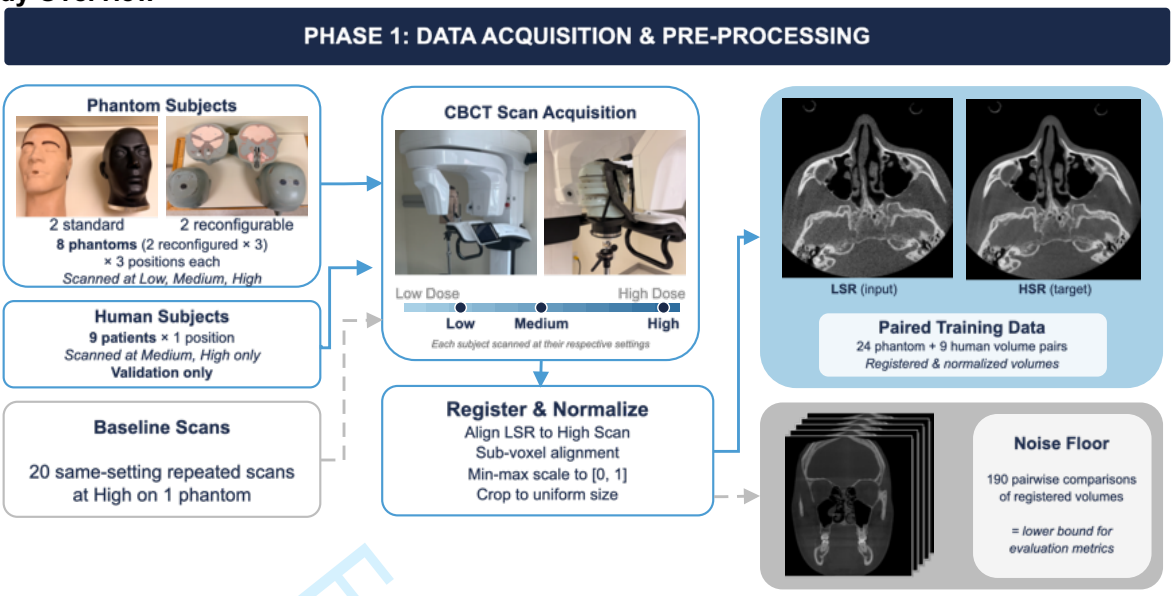
Model	Level	Metric	Error Reduction (%)	Wilcoxon p	Effect (r)
Medium → High (Phantom)	Patch (N=222)	PSNR (dB)	+5.7%	< 0.001*	+0.5041
		RMSE	+2.9%	< 0.001*	+0.6493
		NCC	+11.0%	< 0.001*	+1.000
	Volume (N=5)	PSNR (dB)	-3.7%	0.094	N/S
		RMSE	-0.5%	0.062	N/S
		NCC	+17.1%	0.031*	+1.000
Medium → High (Human)	Patch (N=436)	PSNR (dB)	+1.9%	1.000	N/S
		RMSE	+0.9%	1.000	N/S
		NCC	+8.4%	< 0.001*	+1.0000
	Volume (N=9)	PSNR (dB)	-13.9%	0.787	N/S
		RMSE	-5.7%	0.752	N/S
		NCC	+8.2%	0.002*	+1.000
Low → High (Phantom)	Patch (N=257	PSNR (dB)	+43.3%	< 0.001*	+0.4685
		RMSE	+24.7%	< 0.001*	+0.6146
		NCC	+27.5%	< 0.001*	+1.000
	Volume (N=5)	PSNR (dB)	+16.6%	0.031*	+1.000
		RMSE	+12.9%	0.031*	+1.000
		NCC	+35.3%	0.031*	+1.000

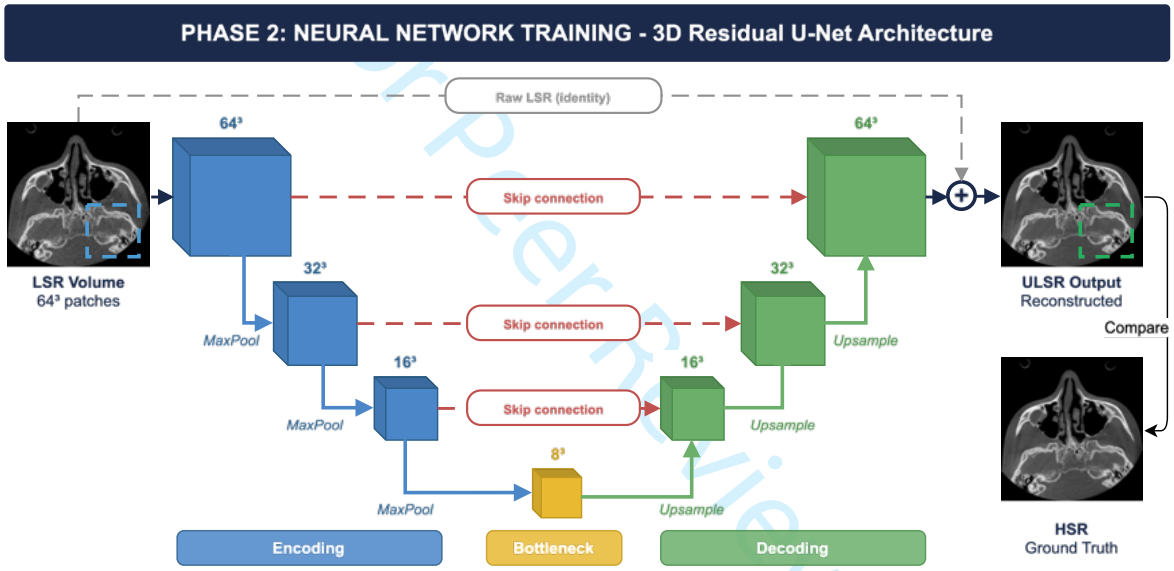
B. Model Output Error Performance

Model	Level (N = validation)	Gap Closure (%)	Wilcoxon p	Effect (r)
Medium → High (Phantom)	Patch (N=222)	44	< 0.001*	+0.9914
	Volume (N=5)	40	0.031*	+1.000
Medium → High (Human)	Patch (N=436)	25	< 0.001*	+0.8658
	Volume (N=9)	22	0.002*	+1.000
Low → High (Phantom)	Patch (N=257)	67	< 0.001*	+0.9375
	Volume (N=5)	59	0.031*	+1.000

Table 2: Model performance at both patch and volume levels. **A**: Standard interpretable metrics commonly used in imaging. **Error Reduction (%)**: percentage reduction in error from initial (pre-network) to output (post-network). For RMSE, error is the metric value. For NCC, error = 1 − NCC. For PSNR, error is the linear-scale equivalent. Positive values indicate error reduction. **Wilcoxon p**: one-sided paired signed-rank test. **Effect (r)**: matched-pairs rank-biserial correlation. N/S = not significant. **B**: Weighted training loss performance. **Gap Closure (%)**: percentage of the initial-to-baseline error gap closed by the model. Patch level metrics are computed per foreground patch extracted from each validation volume; volume-level values average all patches within each subject. Full breakdowns in Tables S12–S14.

A: Study Overview





B. Acquisition Settings

Name	Voltage (kVp)	Current (mA)	Voxel Size (mm)	DAP (mGy-cm²)
Low	91	2.0	0.30	276
Medium	120	2.5	0.30	796
High	120	5.0	0.30	1592

C. Sample Summary

Dataset	Phantoms	Positions	Scan Pairs	Scan Settings
Baseline	1	1	19	H
Training	8	3	24	L, M, H
Humans	—	—	9	M, H

D: Trained Models

Model	Dose Reduction	Train	Validate	Rationale
Medium → High (Phantom)	50%	19 phantom vols	5 phantom vols	Test easier upscaling
Medium → High (Human)	50%	24 phantom vols	9 human subjects	Test transfer learning
Low → High (Phantom)	83%	19 phantom vols	5 phantom vols	Test difficult upscaling

E: Residual Learning

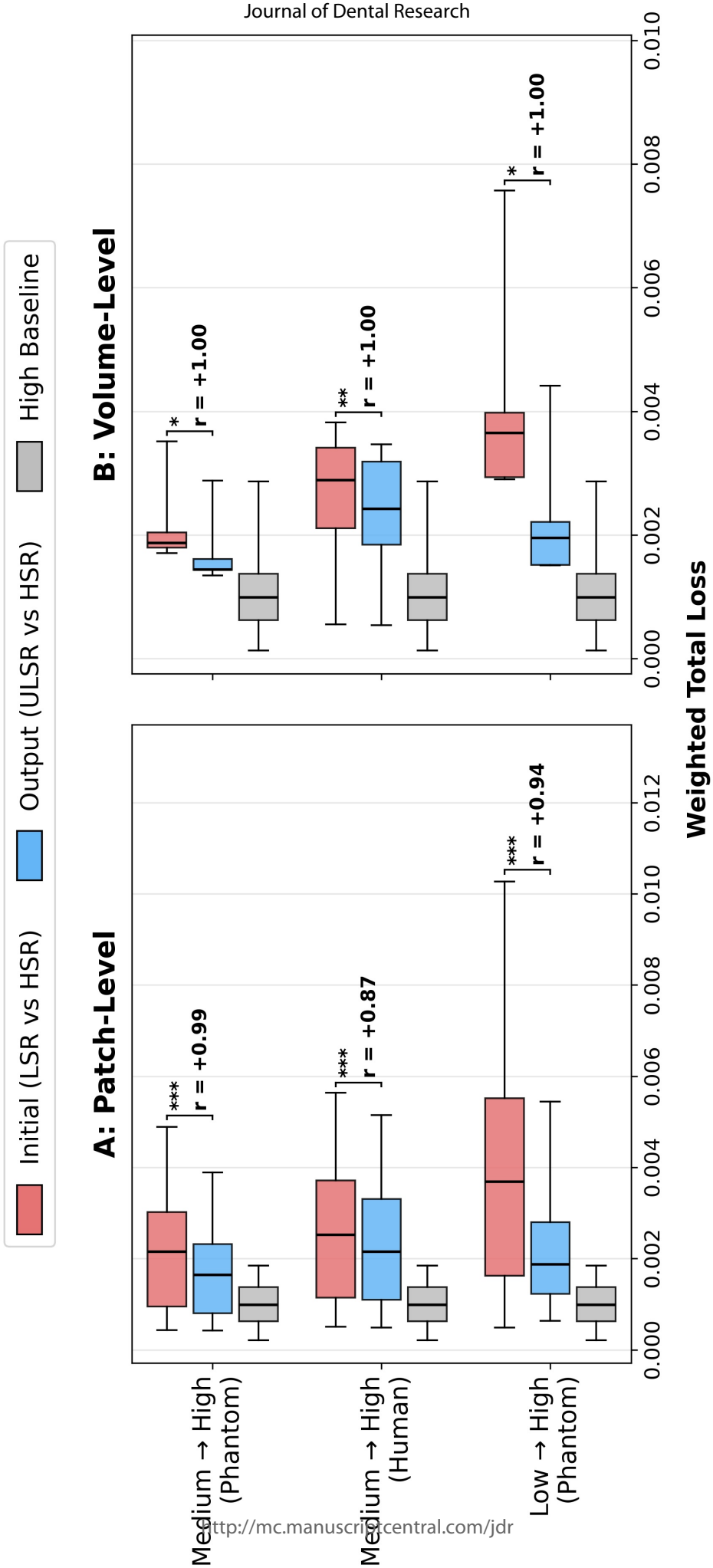
$$ULSR = LSR + f(LSR)$$

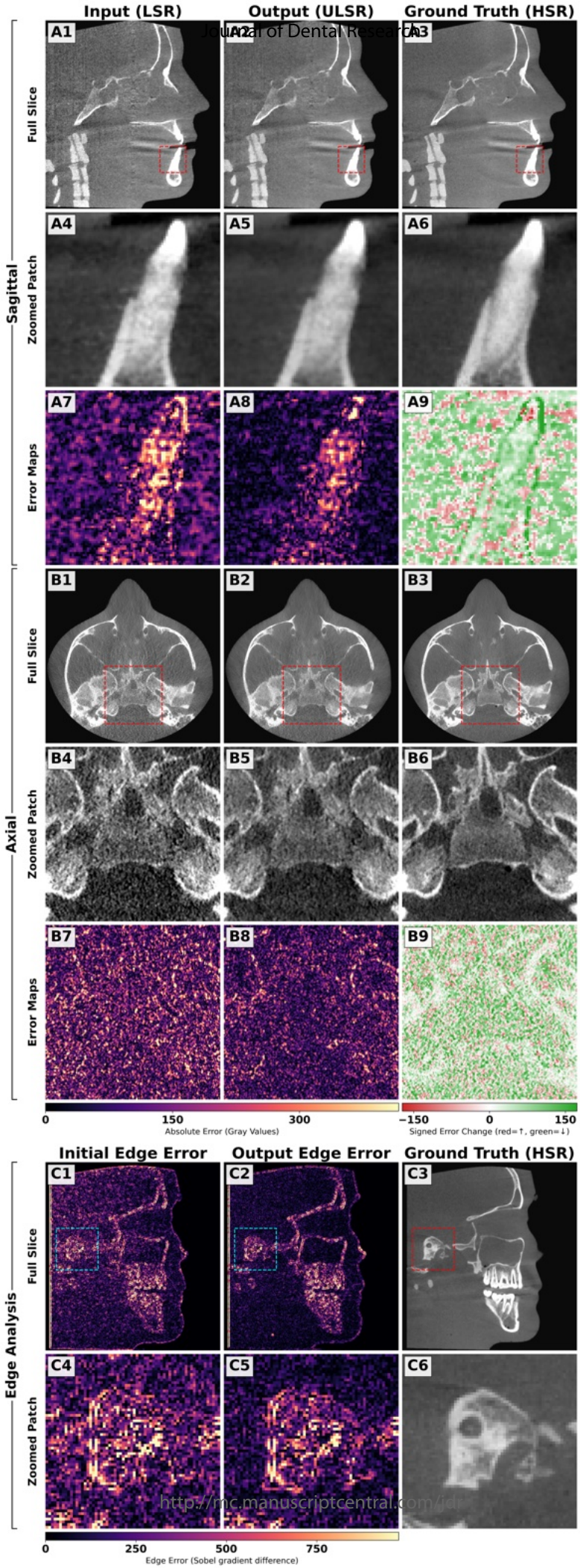
F: Discriminative Power

$$DP = \frac{between - setting\ error}{baseline\ variation}$$

G: Hybrid Loss

$$L_{total} = \sum_{i=1}^{N} \lambda_i L_i$$


Initial (LSR vs HSR)
Output (ULSR vs HSR)
High Baseline
A: Patch-Level
B: Volume-Level
Weighted Total Loss
0.000
0.002
0.004
0.006
0.008
0.010
0.012
***
r = +0.99
***
r = +0.87
***
r = +0.94
*
r = +1.00
**
r = +1.00
*
r = +1.00
Medium → High (Phantom)
Medium → High (Human)
Low → High (Phantom)



Appendix

Supplemental Methods

Memory & Compute requirements:
All training was performed at 0.5x spatial resolution (downscale factor of 2 in each dimension), reducing each volume by a factor of 8. This decision was driven by our Graphics Processing Unit (GPU) memory constraints. At native resolution (577 x 535 x 535 voxels per volume) produced a peak GPU memory for a 128^3 patch, including volume pair, intermediate activations, small batch sizes, gradients, optimizer states, that results in out-of-memory failures during training. Furthermore, to maintain the same receptive field, the computational cost of 3D convolution scales at $O(s^3)$ with patch side length, yielding an 8x increase in both memory and Floating-point Operations Per Second at 128^3 patch size. Additionally, native-resolution training may benefit from also increasing model capacity (greater network depth or feature channels) to capture the final spatial detail at full resolution, which would further increase memory and compute requirements. Follow up training at full resolution with hardware requirements is needed.

Pseudocode

1. Volume Registration

REGISTER(source, template):

Resize both volumes to a standard voxel grid
Initialize a 6-parameter rigid transform (3 translations, 3 rotations)
Optimize via gradient descent to maximize intensity-based similarity
Apply final transform using cubic interpolation

Return registered volume

2. Normalization & Patch Extraction

NORMALIZE(lsr, hsr):

Scale both volumes to [0, 1] using min-max normalization
Normalization is applied per volume pair before patch extraction
This ensures consistent intensity ranges across all scan pairs

Return normalized lsr, hsr

EXTRACT PATCHES(volume, patch_size, overlap):

Compute a sliding window grid across the 3D volume
Stride = patch_size × (1 − overlap)

For each grid position:
Extract a cubic patch of size patch_size³
Discard patches that fall entirely in background/air

```
    Return list of foreground patches

3. Model Architecture — 3D U-Net with Residual Learning
    ENCODER (repeated for each depth level):
      Apply two 3D convolutions, each followed by normalization and activation
      Save output as a skip connection
      Downsample via pooling
      Feature channels double at each level

    BOTTLENECK:
      Apply two 3D convolutions with normalization and activation

    DECODER (repeated for each depth level, in reverse):
      Upsample the feature map
      Concatenate with the corresponding encoder skip connection
      Apply two 3D convolutions with normalization and activation
      Feature channels halve at each level

    RESIDUAL LEARNING:
      prediction = input + residual
      The network learns only the correction between LSR and HSR,
        rather than reconstructing the full volume from scratch

4. Evaluation
    EVALUATE(predicted, ground_truth):

      All metrics computed in normalized [0, 1] space:
        MSE  — mean squared error (voxel-wise)
        MAE  — mean absolute error (voxel-wise)
        PSNR — peak signal-to-noise ratio
        SSIM — structural similarity index (3D)
        NCC  — normalized cross-correlation
        GMAE — gradient mean absolute error (edge preservation)

    COMPARE TO BASELINE:

      Baseline metrics are computed from repeated identical scans
        of the same phantom, representing the physical noise floor

      If the model's error falls at or below the baseline,
        it has reached the noise ceiling of the imaging hardware

      Statistical significance is assessed per metric

5. Median Error Gap Closure
    GAP CLOSURE(init_errors, output_errors, baseline_errors):

      median_init    = median error before enhancement
```

median_output = median error after enhancement
median_baseline = median error from repeated identical scans

gap = median_init − median_baseline

closure % = (median_init − median_output) / gap × 100

0% — no improvement
100% — fully reached the baseline noise floor

Supplemental Tables

Appendix Table 1. Human Subject Inclusion and Exclusion Criteria

	Criteria
Inclusion	Adults aged 18–65 in stable physical health (ASA I–II)
	Able to comprehend, consent to, and follow study requirements
Exclusion	Developmental/cognitive disability precluding informed consent
	Greater than two intra-oral implants (scattering artifacts)
	Previous history of orthognathic surgery
	Current or potential pregnancy (self-reported)
	Fixed metal appliances, bonded retainer wire, or >1 oral implant

Table 1. Inclusion and exclusion criteria for human subjects. All 9 enrolled subjects met inclusion criteria.

Appendix Table 2. Model and Training Configuration

Category	Parameter	Value
Architecture	Model	3D U-Net
	Depth	3 levels (32 → 64 → 128 → 256 bottleneck)
	Activation	ReLU
	Residual learning	ULSR = LSR + f(LSR)
Training	Optimizer	AdamW (lr = 2×10^{-4})
	LR schedule	Cosine warm restarts (T_0 = 40, T_mult = 2)
	Epochs	130
	Batch size	8
	Patch size	64^3 voxels
	Min foreground	10% per patch
	Precision	bfloat16 (CUDA AMP)
Loss	MSE (λ = 0.70)	—
	Gradient MAE (λ = 0.15)	—
	Local Contrast (λ = 0.15)	—
Augmentation	Transforms	Axis flips, 90° rotations
	Not used	Continuous rotation, intensity augmentation
Data	Normalization	Independent min-max to [0, 1]
	Voxel size	0.30 mm isotropic
	Hardware/Software	NVIDIA RTX 4090 (24 GB), PyTorch 2

Table 2: Configuration shared by all three models. Each loss component is normalized by its initial magnitude before weighting.

Appendix Table 3. Human Subject Demographics

Subject	Age (years)	Sex
01	22	F
02	27	M
03	21	M
04	22	M
05	21	M
06	22	M
07	22	M
08	23	F
09	23	F

Table 3: Demographics of 9 enrolled human subjects. Age: 22.6 ± 1.8 years (range 21–27). Sex: 6 male, 3 female.

Appendix Table 4. Pre-Network Scan Quality After Registration

Scan Group	Type	N	Correction (mm / deg)	NCC (mean ± SD)	RMSE (mean ± SD)	PSNR dB (mean ± SD)
High repeated	Baseline	19	0.39 / 0.02	0.9930 ± 0.0010	0.0122 ± 0.0008	38.32 ± 0.62
Medium → High (Phantom)	Training	24	0.52 / 0.03	0.9909 ± 0.0022	0.0184 ± 0.0057	35.04 ± 2.28
Low → High (Phantom)	Training	24	0.57 / 0.05	0.9852 ± 0.0035	0.0275 ± 0.0101	31.63 ± 2.54
Medium → High (Human)	Validation	9	1.51 / 1.11	0.9707 ± 0.0076	0.0217 ± 0.0074	34.13 ± 4.40

Table 4: Consolidated registration quality and initial error. **Correction**: mean translation (mm) / rotation (deg) applied during alignment. Baseline rows show baseline variation (same scan setting, repeated acquisitions). Training and Validation rows show between-setting error after registration. See Tables 3d–3e and 3h–3j for detailed per-group breakdowns.

Appendix Table 5. Statistical Comparison: Phantom vs. Human Registration

Metric	Phantom Median (N=24)	Human Median (N=9)	Ratio	p-value
NCC	0.9908	0.9722	0.9813	< 0.001*
RMSE	0.0166	0.0233	1.41x	0.1513
PSNR (dB)	35.61	32.64	0.9168	0.1513
Translation (mm)	0.4837	1.4763	3.1x	0.0013*
Rotation (deg)	0.0271	0.9456	34.9x	< 0.001*

Table 5: Mann-Whitney U test comparing registration quality between phantom (N=24) and human (N=9) Medium scans. "Ratio" shows the human median relative to the phantom median. Human scans required larger alignment corrections than phantom scans, which is expected given patient repositioning between acquisitions.

Appendix Table 6. Registration Effectiveness & Improvement

Group	N	Pre-Reg NCC	Post-Reg NCC	NCC Δ%	Pre-Reg RMSE	Post-Reg RMSE	RMSE Δ%	Pre-Reg PSNR (dB)	Post-Reg PSNR (dB)	PSNR Δ (dB)	Correction (mm)	Correction (deg)
Baseline High	19	0.9843 +/- 0.0098	0.9930 +/- 0.0010	+0.88 %	0.0174 +/- 0.0052	0.0122 +/- 0.0008	-30.2 %	35.53 +/- 2.53	38.32 +/- 0.62	+2.78	0.3851 +/- 0.2219	0.0204 +/- 0.0134
Train: Medium to High	24	0.9790 +/- 0.0128	0.9909 +/- 0.0022	+1.22 %	0.0256 +/- 0.0067	0.0184 +/- 0.0057	-28.2 %	32.14 +/- 2.32	35.04 +/- 2.28	+2.91	0.5168 +/- 0.2745	0.0316 +/- 0.0310
Train: Low to High	24	0.9712 +/- 0.0113	0.9852 +/- 0.0035	+1.44 %	0.0338 +/- 0.0105	0.0275 +/- 0.0101	-18.5 %	29.74 +/- 2.23	31.63 +/- 2.54	+1.88	0.5651 +/- 0.2554	0.0531 +/- 0.0500
Human: Medium to High	9	0.9126 +/- 0.0423	0.9707 +/- 0.0076	+6.36 %	0.0375 +/- 0.0171	0.0217 +/- 0.0074	-42.2 %	29.83 +/- 5.66	34.13 +/- 4.66	+4.30	1.5115 +/- 0.8471	1.1092 +/- 0.6727

Table 6: Registration effectiveness and improvement across all scan groups. "Pre-Reg" columns show metrics before alignment. "Post-Reg" shows metrics after rigid-body registration. **Δ% columns**: NCC Δ% = percent increase in NCC. RMSE Δ% = percent reduction in RMSE. PSNR Δ = absolute gain in dB. Human scans benefit most from registration due to larger repositioning corrections.

Appendix Table 7. Same-Setting Pairwise Baseline Metrics

Setting	N	NCC	MSE	RMSE	PSNR (dB)
High	19	0.9930 +/- 0.0010	0.000149 +/- 0.000021	0.0122 +/- 0.0008	38.32 +/- 0.62

Table 7: Pairwise error metrics for same-setting baseline scans. Each value is computed by comparing consecutive same-setting scans of the same phantom. These values represent the scanner's inherent baseline variation, the minimum error expected even between identical scans. The network's goal is to reduce between-setting error to this level.

Appendix Table 8. Between-Setting Error Metrics (All Groups)

Setting Pair	Split	N	NCC	MSE	RMSE	PSNR (dB)
Medium to High	Train	19	0.9905 +/- 0.0023	0.000379 +/- 0.000281	0.0187 +/- 0.0056	34.88 +/- 2.17
	Validation	5	0.9925 +/- 0.0010	0.000338 +/- 0.000303	0.0173 +/- 0.0062	35.67 +/- 2.87
	All	24	0.9909 +/- 0.0022	0.000371 +/- 0.000280	0.0184 +/- 0.0057	35.04 +/- 2.28
Low to High	Train	19	0.9850 +/- 0.0036	0.000825 +/- 0.000707	0.0272 +/- 0.0091	31.64 +/- 2.32
	Validation	5	0.9860 +/- 0.0031	0.000989 +/- 0.001121	0.0286 +/- 0.0131	31.56 +/- 3.58
	All	24	0.9852 +/- 0.0035	0.000859 +/- 0.000784	0.0275 +/- 0.0101	31.63 +/- 2.54
Medium to High (Human)	Validation	9	0.9707 +/- 0.0076	0.000525 +/- 0.000288	0.0217 +/- 0.0074	34.13 +/- 4.66

Table 8: Error between low-resolution input scans and their corresponding High dose reference, split by training and validation subsets. The human row shows pre-network error for 9 human subjects (validation only, no human data used during training). Train/validation splits were fixed by phantom identity to prevent data leakage.

Appendix Table 9. Statistical Comparison: Medium vs. Low Between-Setting Error

Metric	Medium Median (N=24)	Low Median (N=24)	U	p-value	Significant?
MSE	0.000275	0.000606	77	< 0.001*	Yes
NCC	0.9908	0.9843	517	< 0.001*	Yes
PSNR (dB)	35.61	32.17	499	< 0.001*	Yes

Table 9: Mann-Whitney U test comparing between-setting error between Medium and Low scan pairs (both relative to High). As expected, low scans showed significantly higher error, reflecting the larger resolution gap between Low and High.

Appendix Table 10. Statistical Comparison: Phantom vs. Human Initial Error (Medium to High)

Metric	Phantom Median (N=24)	Human Median (N=9)	Ratio	U	p-value	Significant ?
NCC	0.9908	0.9722	0.9813	216	< 0.001*	Yes
MSE	0.000275	0.000544	1.98x	72	0.1513	No
RMSE	0.0166	0.0233	1.41x	72	0.1513	No
PSNR (dB)	35.61	32.64	0.9168	144	0.1513	No

Table 10: Mann-Whitney U test comparing initial between-setting error between phantom (N=24) and human (N=9) Medium to High scan pairs. "Ratio" shows the human median relative to the phantom median. Differences reflect anatomical variability between subjects and the additional challenge of patient repositioning.

Appendix Table 11. Phantom vs. Human Metric Consistency (Medium to High, foreground patches only)

Loss Component	Phantom Median (N=24)	Human Median (N=9)	Ratio	U	p-value	Status
MSE	0.0001658	0.0004125	2.49x	71	0.1401	Consistent
Local Contrast	0.01075	0.01383	1.29x	75	0.1889	Consistent

Table 11: Mann-Whitney U test comparing each training loss metric between phantom (N=24) and human (N=9) Medium to High scan pairs. Only foreground patches (foreground ratio >= 0.1) were included. Per-subject means were computed for each metric, then compared across anatomy types using a two-sided Mann-Whitney U test (alpha = 0.05). "Consistent" (p >= 0.05) indicates no significant difference between phantom and human distributions, suggesting the metric transfers well. "Divergent" (p < 0.05) indicates the metric's distribution differs significantly between phantom and human anatomy.

Appendix Table 12. Training Summary & Patch-Level Improvement

For each model, the weighted total loss was computed per patch by summing each active loss component multiplied by its training weight. The resulting per-patch distributions were compared against the High same-setting baseline variation.

Model	Init Median	Out Median	BL Median	Validation N	Wilcoxon p	Effect (r)	Gap Closure %
Medium → High (Phantom)	0.002158	0.001653	0.0009991	222	< 0.001*	+0.9914	44%
Medium → High (Human)	0.002533	0.002157	0.0009991	436	< 0.001*	+0.8658	25%
Low → High (Phantom)	0.003696	0.001883	0.0009991	257	< 0.001*	+0.9375	67%

Table 12: Training summary with patch-level paired improvement analysis. **Init/Out/BL Median**: Median per-patch weighted total loss (initial, output, baseline). **Wilcoxon p**: one-sided paired signed-rank test. **Effect (r)**: matched-pairs rank-biserial correlation. **Gap Closure %**: (median_init − median_output) / (median_init − median_baseline) × 100.

Appendix Table 13. Volume-Level Performance & Improvement

Validation losses computed as volume-wide means. Each loss component is averaged across all foreground voxels in each validation volume.

Model	N	Init Total [p5, p50, p95]	Out Total [p5, p50, p95]	Reduction	BL Total [p5, p50, p95]	Wilcoxon p	Effect (r)	Gap Closure %
Medium → High (Phantom)	5	[0.001733, **0.001874**, 0.003228]	[0.001367, **0.001444**, 0.002634]	23%	[0.0007794, **0.0007931**, 0.0008502]	0.031*	+1.000	40%
Medium → High (Human)	9	[0.0008226, **0.002896**, 0.003819]	[0.0007599, **0.002431**, 0.003434]	16%	[0.0007794, **0.0007931**, 0.0008502]	0.002*	+1.000	22%
Low → High (Phantom)	5	[0.002912, **0.003657**, 0.006859]	[0.001516, **0.001956**, 0.003980]	47%	[0.0007794, **0.0007931**, 0.0008502]	0.031*	+1.000	59%

Table 13: Volume-level weighted total validation loss with paired improvement analysis. Background patches (foreground_ratio < 0.1) excluded before aggregation. **Reduction**: Percentage decrease in median. **Wilcoxon p**: one-sided paired signed-rank test (H_1: init > output). **Effect (r)**: matched-pairs rank-biserial correlation. **Gap Closure %**: (median_init − median_output) / (median_init − median_baseline) × 100.

Appendix Table 14. Standard Interpretable Metrics

Standard metrics at both patch and volume levels. Each metric compares model output (ULSR) to high-resolution target (HSR), the initial input (LSR vs HSR), and the same-setting baseline (High).

Model	Level	Metric	Initial Median	Output Median	BL Median	Wilcoxon p	Effect (r)
Medium → High (Phantom)	Patch	PSNR (dB)	37.79	38.04	40.55	< 0.001*	+0.5041
		RMSE	0.01290	0.01253	0.009390	< 0.001*	+0.6493
		MAE	0.009310	0.009031	0.006497	0.450	N/S
		NCC	0.9938	0.9945	0.9945	< 0.001*	+1.000
	Volume	PSNR (dB)	38.96	38.80	40.55	0.094	N/S
		RMSE	0.01176	0.01182	0.009390	0.062	N/S
		MAE	0.007614	0.008453	0.006497	0.781	N/S
		NCC	0.9801	0.9835	0.9945	0.031*	+1.000
Medium → High (Human)	Patch	PSNR (dB)	35.09	35.17	40.55	1.000	N/S
		RMSE	0.01760	0.01744	0.009390	1.000	N/S
		MAE	0.009837	0.01060	0.006497	1.000	N/S
		NCC	0.9765	0.9784	0.9945	< 0.001*	+1.0000
	Volume	PSNR (dB)	34.77	34.21	40.55	0.787	N/S
		RMSE	0.01932	0.02041	0.009390	0.752	N/S
		MAE	0.01124	0.01177	0.006497	0.994	N/S
		NCC	0.9671	0.9698	0.9945	0.002*	+1.000
Low → High (Phantom)	Patch	PSNR (dB)	32.86	35.32	40.55	< 0.001*	+0.4685
		RMSE	0.02276	0.01714	0.009390	< 0.001*	+0.6146
		MAE	0.01798	0.01457	0.006497	< 0.001*	+0.2619
		NCC	0.9889	0.9920	0.9945	< 0.001*	+1.000
	Volume	PSNR (dB)	34.56	35.35	40.55	0.031*	+1.000
		RMSE	0.01976	0.01722	0.009390	0.031*	+1.000
		MAE	0.01444	0.01402	0.006497	0.031*	+1.000
		NCC	0.9667	0.9784	0.9945	0.031*	+1.000

Table 14: Standard interpretable metrics at patch and volume levels. **SSIM**: Structural Similarity (higher = better). **PSNR**: Peak Signal-to-Noise Ratio in dB (higher = better). **RMSE**: Root Mean Square Error (lower = better). **MAE**: Mean Absolute Error (lower = better). **NCC**: Normalized Cross-Correlation (higher = better). Volume-level values are per-patch means averaged within each validation volume. **Wilcoxon p**: one-sided paired signed-rank test in the direction of improvement.

Appendix Table 15. Phantom vs Human Initial Difficulty (Medium → High)
Pre-network error comparison between phantom and human subjects at both patch and volume levels. Higher initial error indicates a harder upscaling task. Mann-Whitney U test (two-sided, independent groups).

Level	Metric	Phantom	Human	Mann-Whitney p	Effect (r)
Patch	Weighted Loss	0.002158	0.002534	0.005*	-0.1338
	MSE	0.0001665	0.0003098	< 0.001*	-0.2536
	Gradient MAE	0.002037	0.002432	0.022*	-0.1079
	Local Contrast	0.01144	0.01288	0.010*	-0.1228
	PSNR (dB)	37.79	35.09	< 0.001*	-0.2536
	RMSE	0.01290	0.01760	< 0.001*	-0.2536
	MAE	0.009310	0.009837	0.246	N/S
	NCC	0.9938	0.9765	< 0.001*	-0.5281
Volume	Weighted Loss	0.001874	0.002896	0.364	N/S
	MSE	0.0001483	0.0004125	0.518	N/S
	Gradient MAE	0.001663	0.002895	0.438	N/S
	Local Contrast	0.01025	0.01383	0.298	N/S
	PSNR (dB)	38.96	34.77	0.518	N/S
	RMSE	0.01176	0.01932	0.518	N/S
	MAE	0.007614	0.01124	0.606	N/S
	NCC	0.9801	0.9671	< 0.001*	-1.000

Table 15: Pre-network error for Medium → High, comparing phantom (N=222 patches, 5 volumes) vs human (N=436 patches, 9 volumes) subjects. Volume-level values are per-patch means averaged within each volume. Mann-Whitney U test (independent groups, two-sided). Negative effect = harder for humans.