K-Material SDFs for Neural Attenuation Fields

Daksh K. Shah
Computer Science and Engineering
University of California, Santa Cruz
dakshah@ucsc.edu

Abstract

Computed Tomography (CT) carries significant ionizing radiation risks, driving the need for sparse-view reconstruction. While current implicit neural representations can recover volumetric attenuation fields and surface geometry from sparse projections, they are limited by manually tuned attenuation bounds and rigid two-material constraints. This thesis proposes a unified, scalable architecture for automated, multi-material surface reconstruction. We replace independent material networks with a shared latent backbone and introduce a fully differentiable K-material sequential soft selector to model an arbitrary number of overlapping tissues. To eliminate manual tuning, we automate attenuation bounding using a Gaussian Mixture Model (GMM). Additionally, we implement a scheduled auxiliary floater loss to mitigate geometric hallucinations common in hash-encoded networks under extreme sparsity. Evaluated on clinical Cone-Beam CT (CBCT) datasets, our model achieves superior 3D volumetric fidelity in complex, multi-tissue regions and demonstrates enhanced robustness under moderate sparsity. However, performance drops on high-frequency density shifts highlight remaining trade-offs between shared latent efficiency and decoupled network capacity.

1 Introduction

1.1 Motivation

The guiding principle in medical imaging, according to the CDC, is to keep radiation exposure as low as reasonably achievable while still preserving diagnostic quality. This often means using alternative imaging modalities such as MRI or ultrasound where appropriate. However, Computed Tomography (CT) remains clinically irreplaceable for a wide class of diagnostic tasks such as bone fracture assessment, tumor staging, and surgical planning, where its combination of speed, spatial resolution, and contrast sensitivity has no practical substitute. For these applications, the question is not whether to use CT, but how to reduce the radiation burden it carries.

Sparse-view CT can reduce the number of required projections by over tenfold, but classical analytical reconstruction techniques such as FDK and FBP break down and lead to significant artifacts under these conditions. Though volumetric neural methods such as NAF and derivative works can recover sharp attenuation volumes, they lack the ability to easily extract 3D surface meshes without manually defined thresholds for post-hoc surface extraction algorithms such as Marching Cubes

Compounding this, real anatomy is not homogenous, comprised of bone, soft tissue, and air, each with distinct attenuation coefficients and densities. A single implicit field cannot cleanly separate them without explicit structural priors, and the baseline two-material NeAS [16] requires manual tuning bounds for each scene, which is impractical in clinical situations.

In this work, we aim to close these gaps by extending neural attenuation surface reconstruction to handle automatic multi-material decomposition and robust surface geometry recovery under clinical sparse-view conditions.

1.2 The Promise of Neural Fields

Neural Radiance Fields (NeRF) [8] demonstrated that a continuous function representing a scene, parameterized by a multi-layer perceptron (MLP), can synthesize novel views with remarkable fidelity from sparse observation. This is a capability that traditional explicit representations, such as voxel grids and meshes, are unable to match without dense input. Because implicit neural representations encode a scene as network weights rather than discrete samples, they are resolution-independent and can capture fine-geometric details without the cubic memory cost in volumetric grids. Critically, the Beer-Lambert law governing X-ray attenuation shares the same line-integral structure as NeRF's volume rendering, making the transfer to CT natural. Neural Attenuation Fields (NAF) [15] demonstrated that swapping color for attenuation coefficients within the same framework as NeRF produces high-quality novel X-ray views from sparse projections.

The introduction of neural signed distance functions extended these representations beyond appearance to geometry itself. Rather than requiring post-hoc surface extraction via Marching Cubes with manually chosen thresholds, NeuS [14] demonstrated that a neural SDF's zero-level set defines the surface by construction, making geometry a direct output of the optimization. NeAS [16] simultaenously recovers attenuation and surface geometry from sparse X-ray projections. However, its reliance on manually tunred, scene-specific attenuation bounds and non-differentiable two-material selector limits its applicability to real clinical data, the gap that this work addresses.

1.3 Challenges

Reconstructing volumes from sparse-view CBCT presents significant challenges across both classical and neural methods. Classical analytical techniques, such as FDK, assume dense angular sampling and produce severe streak artifacts and noise under sparse conditions. Conversely, while implicit neural representations (INRs) mitigate these streaks, they are highly susceptible to geometric hallucinations, or "floaters," where spurious density accumulates in empty space. Additionally, extracting distinct multi-material surfaces from these neural volumes currently relies on impractical, scene-specific manual thresholding.

1.4 Contributions

In this thesis, we address the limitations of existing implicit neural representations for sparse-view CT by extending the NeAS architecture to handle complex, real-world data and multi-material boundaries. Our core contributions are as follows:

- **Unified Multi-Material Architecture:** We propose a shared latent space representation that replaces the independent per-material attenuation MLPs of the baseline with a single shared backbone and K lightweight output heads. This reduces parameter count, promotes feature sharing across material boundaries, and serves as the foundation for scaling to arbitrary K.
- **Differentiable K-Material Priority Selector:** We introduce a soft, fully differentiable sequential occupancy filter that scales to an arbitrary number of nested material surfaces. Each material's contribution at a given point is governed by a sigmoid-based priority membership weight, enabling smooth gradient propagation through all K SDF fields and eliminating the hard piecewise selector of the two-material baseline.
- **Scheduled Floater Loss:** We formulate a scheduled auxiliary loss that explicitly supervises zero-attenuation rays during the first 20% of training. By decoupling air rays from the primary intensity batch and penalizing spurious density early in optimization, this loss directly mitigates the hallucinated geometry artifacts introduced by hash encoding under real-world sparse-view conditions.
- **Real-World Sparse-View Evaluation:** Unlike the original NeAS framework, which was validated on synthetic phantoms with known ground-truth geometry, we evaluate our proposed pipeline on clinical CBCT data across four anatomical regions (Abdomen, Chest, Foot, Jaw) and four extreme sparsity configurations (50 down to 5 views). We report improvements in both 2D projection-domain and 3D volumetric PSNR and SSIM on three of four scenes, demonstrating the practical utility of the proposed extensions on data without ground-truth meshes.

2 Background & Preliminaries

2.1 Principles of Computed Tomography

Computed Tomography (CT) reconstructs a three-dimensional representation of an objects internal structure from a set of two-dimensional X-ray projections acquired at varying angles using a rotating X-ray source and detectors. The physical basis for CT imaging is the attenuation, or signal energy loss as X-ray photons pass through a medium. This attenuation decays according to the Beer-Lambert law:

$$I = I_0 \exp\left(-\int_{\ell} \mu(x)\, dt\right)$$

where I_0 is the initial intensity, I is the measured intensity at the detector, and $\mu(x)$ is the material's linear attenuation coefficient along ray path ℓ. In practice, following the normalization convention of NAF [15] and NeAS [16], pixel intensities are scaled to [0,1] and the scene is fit within a unit sphere, so μ is treated as a dimensionless quantity relative to this normalized scale rather than a physical unit. Different tissues attenuate X-rays to different degrees, where dense structures such as bone exhibit a high μ and softer tissues exhibit lower values. The goal of CT reconstruction is to recover the volumetric attenuation field $\mu(x)$ from a set of line integral measurements.

Cone-Beam CT (CBCT) is a variant of CT where a wide cone-shaped X-ray beam and 2D flat-panel detector are rotated around the subject, acquiring a full set of projections in a single rotation. Compared to conventional fan-beam CT, CBCT offers a lower radiation dose and compact acquisition geometry, making it better suited for clinical settings. However, CBCT reconstruction quality is highly sensitive to the number of acquired projections. Sparse-view CBCT is a variant of CBCT where only a small subset of views is acquired to further reduce radiation exposure. However, this introduces significant noise and streak artifacts into the reconstructions produced by classical/analytic methods. Mitigating these artifacts while preserving diagnostic quality under sparse-view conditions is the primary problem this work addresses.

2.2 Sparse-View CT and Classical Reconstruction

The most widely used analytic reconstruction for CT are Filtered Back Projection (FBP) [12] and its cone-beam extension, the Feldkamp-Davis-Kress (FDK) algorithm [3]. Both methods utilize a ramp filter to the acquired projections and back-project the filtered data into the reconstructed volume. Their primary advantage is their computational efficiency, however they are prone to streak artifacts and noise because of the assumption of a dense, uniformly sampled set of projection angles. Under sparse-view conditions, the angular undersampling will violate this assumption and amplify these artifacts, particularly in low-contrast soft tissue.

Iterative algorithms aim to mitigate these artifacts, including Conjugate Gradient Least Squares (CGLS), Simultaneous Algebraic Reconstruction Technique (SART) [1], and Adaptive Steepest Descent Projection Onto Convex Sets (ASD-POCS) [13] model the CT acquisition as a system of linear equations, solving for the attenuation volume by minimizing a data-fidelity term with optional regularization priors such as Total Variation (TV). While these approaches substantially suppress streak artifacts compared to analytic methods, they require hundreds to thousands of forward and back-projection iterations over the full volume, making them computationally prohibitive for clinical use. This limitation motivates the adoption of neural reconstruction approaches, which achieve competitive reconstruction quality at a fraction of the inference cost.

2.3 Implicit Scene Representations

Traditional 3D representations such as voxel grids, point clouds, and triangle meshes describe geometry explicitly by storing discrete samples of a scene. While more intuitive, these representations are resolution-constrained: voxel grids scale cubically with more detail and meshes require known topology. Implicit neural representations (INRs) offer an alternative by encoding a scene as a continuous function approximated by a neural network, mapping a spatial coordinate $x \in \mathbb{R}^3$ directly to physical quantities such as color, density, or attenuation. Because these representations are continuous and resolution-independent, INRs can capture finer geometric details without the memory overhead of explicit geometry.

2.4 Neural Radiance Fields

Neural Radiance Fields (NeRF) [8] demonstrated that a scene can be represented as a neural network mapping a 3D position x and viewing direction $\mathbf{d}$ to a volume density σ and emitted color $\mathbf{c}$. Novel views are synthesized by integrating these quantities along camera rays via differentiable volume rendering. Müller et al. [9] later introduced multi-resolution hash encoding in Instant-NGP as a significantly faster alternative to NeRF's original positional encoding, achieving one to two orders of magnitude speedup while maintaining reconstruction quality. Both encoding strategies are used within NeAS and are discussed in detail in Section 3.1.1.

2.5 Attenuation Fields

Recent works such as Neural Attenuation Fields (NAF) [15] extend the architecture used in NeRF to the realm of X-rays. Utilizing an MLP with hash encoding to learn attenuation coefficients along sampled rays, NAF constructs novel views via Beer-Lambert volume rendering rather than the color/density rendering used in RGB NeRF. This adaptation is natural given that X-ray attenuation follows the same line-integral formulation that volumetric rendering approximates.

However, NAF and similar methods treat the scene purely as a continuous attenuation volume, with no explicit notion of surface geometry. Extracting surfaces requires post-hoc application of Marching Cubes with a manually chosen density threshold, which produces noisy, artifact-prone meshes. This is the core limitation that NeAS directly addresses.

2.6 Signed Distance Fields (SDFs)

A Signed Distance Field (SDF) is a scalar field $f : \mathbb{R}^3 \to \mathbb{R}$ that maps any 3D coordinate to its signed distance to the nearest surface. By convention, $f(x) > 0$ outside the surface, $f(x) = 0$ on the surface boundary, and $f(x) < 0$ inside the object. This zero-level set $\{x \mid f(x) = 0\}$ provides a continuous definition of the surface geometry. The concept of parameterizing an SDF as a neural network, allowing a model to learn the surface of an object directly from data, was first introduced by Park et al. [10] in DeepSDF.

While DeepSDF demonstrated that SDFs can be learned, it does not guarantee the unit gradient property that geometrically valid SDFs require. To address this, IGR [4] utilized the Eikonal equation $\|\nabla f(x)\| = 1$ as an Eikonal regularization loss formulated as a soft penalty on deviations from unit gradient norm, which has become standard in modern neural SDF methods.

Unlike the soft density fields used in NeRF — where manual threshold selection and marching cubes post-processing is needed to extract surfaces — NeuS [14] demonstrated how neural SDFs could be used directly within a volume rendering framework by converting signed distance values into rendering weights, ensuring that the learned zero-level set corresponds precisely to the reconstructed surface. This eliminated the need for manual thresholding, making surface extraction a natural byproduct of the rendering optimization rather than a post-processing step. NeAS [16] extends this application of Neural SDFs to X-ray/CT attenuation.

2.7 Surface Reconstruction from Volume Data

A core challenge in volumetric reconstruction is extracting clean surface meshes from scalar fields. The most widely used algorithm for this is Marching Cubes [7], which tessellates a surface by evaluating the scalar volume at the corners of a voxel grid and connecting edges where the field crosses a user-defined threshold. While effective when working with known thresholds, this is a significant limitation in CT. The thresholds for materials in CT are often scene-dependent, and under sparse-view conditions the recovered attenuation volume is noisy enough that small threshold changes can substantially alter the extracted geometry.

Poisson Surface Reconstruction [5] offers an alternative by fitting a watertight surface to an oriented point cloud by solving a Poisson equation. However, this approach requires a dense point cloud with normals as input, which is not directly obtainable from CT attenuation volumes and impractical for our purposes.

NAF-style implicit attenuation fields, on the other hand, learn a continuous volumetric attenuation function that is optimzed to reproduce novel X-ray projections. Because the training objective

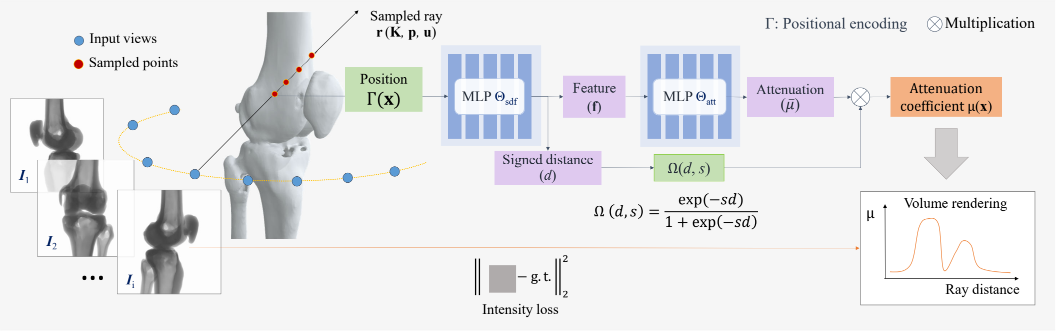


Figure 1: Overview of the 1-M NeAS baseline architecture

constrains the Lambert-Beer line integral rather than geometry, the network will not necessarily learn surface boundaries. Extracting surfaces therefore still requires post-hoc Marching Cubes with a manually chosen threshold, inheriting the same problems. This limitation directly motivates NeuS [14] and NeAS [16], which replace the soft density field with a neural SDF whose zero-level set is the surface by construction, eliminating the need for manual thresholding and making geometry the output of the reconstruction.

3 Methodology

3.1 Baseline Implementation (NeAS)

Our baseline architecture builds upon the Neural Attenuation Surface (NeAS) framework, an implicit neural representation approach designed to simultaneously capture surface geometry and attenuation coefficient fields. The model represents the 3D scene using two decoupled multilayer perceptrons (MLPs), as illustrated in Figure 1. The two-material extension of this architecture is shown in Figure 2.

For a 3D point x sampled along a ray r, the input is first mapped to a higher-dimensional space using a positional encoding function $\Gamma(x)$. The Θ_{sdf} network takes this encoded position and outputs a signed distance $d \in \mathbb{R}$ and an intermediate K-dimensional feature vector $f \in \mathbb{R}^K$.

To bridge the geometric representation with the volumetric rendering process, a Surface Boundary Function (SBF) is applied to the signed distance d. The SBF is defined using a sigmoid approximation to facilitate smooth gradient propagation during training:

$$\Omega(d, s) = \frac{\exp(-sd)}{1 + \exp(-sd)}$$

where s is a learnable parameter that governs the steepness of the surface boundaries.

The intermediate feature vector f is subsequently passed to the attenuation network, Θ_{att}, which predicts a raw attenuation parameter $\bar{\mu}$. To ensure that attenuation is strictly positive and bounded, the final layer of Θ_{att} utilizes the activation function $\alpha\sigma(x) + \beta$, where α and β define the specific attenuation range for the material. The final attenuation coefficient $\mu(x)$ at point x is then computed by modifying the raw attenuation with the SBF:

$$\mu(x) = \Omega(d, s)\bar{\mu}$$

Volume rendering is performed by discretizing the Beer-Lambert law introduced in Section 2 via the quadrature rule, approximating the continuous line integral as a sum over N stratified samples along each ray, where $\delta_j = t_{j+1} - t_j$ is the distance between adjacent sampled points.

The network is optimized using a combined loss function that minimizes the Mean Squared Error (MSE) of the pixel intensity:

$$\mathcal{L}_{int} = \sum_{r \in \mathcal{R}} \|\hat{I}(r) - I(r)\|_2^2$$

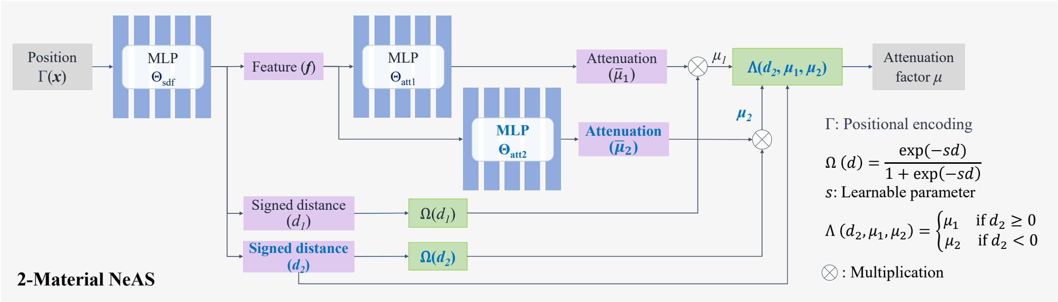

Position $\Gamma(\boldsymbol{x})$
MLP Θ_{sdf}
Feature (f)
MLP Θ_{att1}
MLP Θ_{att2}
Attenuation ($\bar{\mu}_1$)
Attenuation ($\bar{\mu}_2$)
μ_1
μ_2
$\Lambda(d_2, \mu_1, \mu_2)$
Attenuation factor μ
Signed distance (d_1)
Signed distance (d_2)
$\Omega(d_1)$
$\Omega(d_2)$
2-Material NeAS
Γ: Positional encoding
$\Omega(d) = \frac{\exp(-sd)}{1+\exp(-sd)}$
s: Learnable parameter
$\Lambda(d_2, \mu_1, \mu_2) = \begin{cases} \mu_1 & \text{if } d_2 \geq 0 \\ \mu_2 & \text{if } d_2 < 0 \end{cases}$
⊗ : Multiplication

Figure 2: Overview of the 2-M NeAS baseline architecture

$\mathcal{R}$ is a sampled batch of the rays, alongside an Eikonal regularization term to enforce valid SDF gradients:

$$\mathcal{L}_{reg} = \frac{1}{mn} \sum_{k,j} (\|\mathbf{n}_{k,j}\| - 1)^2$$

where the batch size is m, sampling size is n, and gradient of the SDF at the sampled point is **n**. Together, NeAS formulates their total loss as:

$$\mathcal{L} = \mathcal{L}_{int} + \lambda \mathcal{L}_{reg}$$

3.1.1 Encoding Methods

NeAS supports two positional encoding strategies:

Frequency encoding [8] maps input coordinates to a sinusoidal Fourier feature space via:

$$\gamma(p) = \big(\sin(2^0 \pi p),\, \cos(2^0 \pi p),\, \ldots,\, \sin(2^{L-1} \pi p),\, \cos(2^{L-1} \pi p)\big)$$

counteracting the spectral bias of MLPs to recover fine geometric detail.

Multi-resolution hash encoding [9] instead stores trainable feature vectors in a hash-table-backed multi-resolution voxel grid, trilinearly interpolating and concatenating features across L resolution levels before decoding with a compact MLP. Hash encoding achieves comparable reconstruction quality at approximately 3.5× the training speed of frequency encoding, though it is more susceptible to noise artifacts from hash collisions under sparse-view conditions. For this reason, we evaluate exclusively with hash encoding throughout this work to reduce compute, while our floater regularization (Section 3.2.3) directly addresses its susceptibility to noisy geometry on real-world data.

3.1.2 Coarse-to-Fine Optimization

Initially introduced in BARF [6] for joint pose and radiance field optimization, coarse-to-fine frequency annealing addresses the tendency of neural networks to overfit high-frequency features before the global scene structure has converged. The core idea is to mask out high-frequency encoding components early in training and progressively introduce them as optimization proceeds, ensuring the network first learns coarse geometry before refining fine detail.

NeAS adopts this strategy by integrating a weight mask into the encoding process [16]. The k-th frequency component is reweighted by:

$$w_k(\tau) = \begin{cases} 0 & \tau < k \\ \frac{1-\cos((\tau-k)\pi)}{2} & 0 \leq \tau - k < 1 \\ 1 & \tau - k \geq 1 \end{cases}$$

where τ increases linearly with the number of training iterations. At the start of training, only the lowest-frequency components are active; higher frequencies are smoothly introduced as τ grows. For hash encoding, the same mask is applied across resolution levels rather than frequency bands, suppressing fine-resolution voxel features early in optimization. This schedule stabilizes pose refinement and prevents the attenuation and SDF networks from locking onto noise-driven high-frequency patterns before the coarser scene structure has been established.

3.1.3 Determining Attenuation Bounds

The final layer of each attenuation head uses the activation $\alpha \cdot \sigma(x) + \beta$, where σ is the sigmoid function, α controls the range of predicted attenuation values, and β sets the lower bound. Together, they constrain each material's predicted attenuation to the interval $[\beta,\ \alpha + \beta]$, ensuring physically plausible and non-overlapping material ranges.

The baseline approach determines α and β by first training a 1M-NeAS model and then performing a histogram-based analysis of the resulting attenuation volume. Once the 1M model converges, the full voxel grid is sampled to produce a predicted attenuation volume, and voxels below a small background threshold (0.01) are discarded as air. A histogram of the remaining non-background values is computed, and local maxima are identified via peak detection, requiring a minimum prominence of 1% of the non-zero voxel count to suppress noise-driven spurious peaks.

If a single dominant peak is detected, the scene is treated as a single-material configuration. The parameters are fixed to $\alpha = 10$ and $\beta = 0.1$, providing a deliberately wide activation range that encompasses the observed attenuation values without over-constraining the network.

If two or more peaks are detected, a bimodal (two-material) distribution is assumed, corresponding to distinct tissue classes such as soft tissue and bone. The valley between the two largest peaks is identified as the inter-material threshold t. Material-specific parameters are then assigned as:

$$\beta_1 = 0.01, \quad \alpha_1 = t - \beta_1, \tag{1}$$

$$\beta_2 = \beta_1 + \alpha_1, \quad \alpha_2 = v_{\max} - \beta_2, \tag{2}$$

where $v_{\max}$ denotes the maximum observed attenuation value, ensuring that Material 1 spans $[\beta_1,\ \beta_1 + \alpha_1]$ and Material 2 spans $[\beta_2,\ \beta_2 + \alpha_2]$ without overlap. The parameters are then scaled uniformly so that $\alpha_2 + \beta_2 = 1$, normalising the two-material activation range to a canonical unit interval and improving consistency across anatomical scenes.

3.1.4 Challenges in the Baseline Architecture

While NeAS provides a strong foundation for integrating surface extraction into attenuation fields, adapting this framework for real-world CT data introduces significant challenges.

Manual Attenuation Bounds: A major limitation of the baseline NeAS architecture is the reliance on manually defining the activation parameters α and β. These parameters strictly determine the ranges of attenuation values the network can predict. When applied to complex real-world datasets containing diverse, unknown, or overlapping material densities, manually determining and tuning these thresholds becomes highly impractical. This architectural constraint necessitates a shift toward a more dynamic, self-supervised material grouping approach.

Hallucinated Geometry/Floaters: When tested on real-world datasets, especially under extreme sparse-view conditions. We observed that the baseline architecture frequently struggles with geometric hallucinations, commonly referred to as “floaters”. In configurations utilizing hash-encoding, these artifacts manifest as rough surfaces and numerous nonexistent points in empty space. These anomalies are likely due to hash collisions or inherent noise in real-world measurements that the under-constrained network fails to penalize. Because the baseline loss function only optimizes for final ray intensity, it does not adequately suppress localized regions of zero-attenuation that inappropriately accumulate density, requiring the introduction of a dedicated mask-based loss.

3.2 Proposed Changes

3.2.1 Shared Latent Space

The original NeAS architecture employs independent multilayer perceptrons ($\Theta_{att1}, \Theta_{att2}$) for each target material. While this approach isolates the attenuation bounds for a two-material system, it scales poorly to K materials, significantly increasing the parameter count and preventing the network from learning shared geometric features across material boundaries.

To address this, we propose a unified architecture utilizing a shared latent space. Instead of completely decoupled MLPs, our model merges the attenuation networks into a single shared backbone. The SDF network (Θ_{sdf}) outputs a global feature vector $\mathbf{f}$ that encapsulates the spatial context of the sampled point. This feature vector is passed to the shared attenuation backbone, which then branches into K separate output heads. Each head consists of a single hidden layer responsible for predicting the raw attenuation parameter $\overline{\mu}_i$ for its corresponding material layer Φ_i. This shared representation improves computational efficiency, accelerates convergence, and allows the model to utilize multi-material structures inherent in human anatomy.

3.2.2 K-Material Soft Selector

For scenes containing multiple anatomical structures, such as bone and muscle, a given spatial point x may lie within the overlapping bounding surfaces of several materials. The baseline 2-material NeAS addresses this ambiguity using a piecewise selector function Λ. However, this "hard" selection logic is non-differentiable at the boundary and does not naturally generalise beyond two materials. To resolve this, we introduce a soft, fully differentiable sequential occupancy filter, illustrated in Figure 3, that scales to an arbitrary number of materials K.

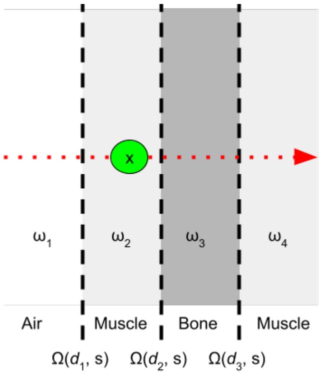


Figure 3: Soft Selector Visualization

We define an ordered set of materials $\Phi_1, \dots, \Phi_K$, where the index i encodes structural priority. Each material is associated with its own SDF d_i and a raw attenuation parameter $\overline{\mu}_i$. The Surface Boundary Function (SBF) $\Omega(d_i, s)$ is utilized to determine the occupancy of each material, outputting 1 when x is inside the surface ($d_i < 0$) and 0 when outside ($d_i > 0$).

To calculate the contribution of each material at point x, we compute a priority membership weight $w_i(x)$ as follows:

$$w_i(x) = \Omega(d_i, s) \prod_{j=i+1}^{K} (1 - \Omega(d_j, s))$$

This formulation functions as a sequential visibility mask. The product term $\prod(1 - \Omega(d_j, s))$ ensures that a material Φ_i only contributes to the final signal if the point is not already "occluded" or claimed by a higher-priority material Φ_j where $j > i$. For the highest-priority material Φ_K, the weight simplifies to $w_K(x) = \Omega(d_K, s)$.

The total attenuation coefficient $\mu(x)$ is determined by the weighted sum of all raw material attenuations:

$$\mu(x) = \sum_{i=1}^{K} \overline{\mu}_i \cdot w_i(x)$$

This soft selector offers several advantages over the original piecewise framework:

- **Differentiability**: By replacing the hard switch with a sigmoid-based weighted sum, we maintain a fully differentiable rendering process. This facilitates smoother gradient propagation during backpropagation, enhancing learning efficiency for both the attenuation fields Θ_{att} and the SDFs Θ_{sdf}.
- **Complex Layering**: The sequential logic allows the model to represent complex nested or overlapping structures (e.g., skin, muscle, and bone) without imposing rigid geometric constraints on how surfaces relate.

- **Boundary Refinement**: The learnable parameter s governs the steepness of the transitions. This enables a coarse-to-fine strategy where boundaries are initially "soft" to assist global optimization before sharpening to extract accurate 3D geometry.

Unlike the hard selector used in the original 2M-NeAS, this priority-based membership imposes no limit on the number of materials and ensures that distinct attenuation ranges $(\beta, \alpha + \beta)$ are maintained for each structure to ensure accurate surface representation.

3.2.3 Mask-Based Floater Regularization

In the baseline NeAS pipeline, training rays are sampled uniformly from all pixels in each projection, meaning the 512 rays per projection batch may contain any mixture of zero-attenuation (air) and non-zero attenuation rays. In contrast, our proposed pipeline explicitly decouples these two ray populations: 512 rays are sampled exclusively from non-zero attenuation pixels to prioritize learning the primary attenuation signal, with an additional independent batch of 128 zero-attenuation rays used solely for floater regularization. This separation ensures that the primary intensity loss $\mathcal{L}_{int}$ is not diluted by air rays, while still providing explicit supervision over empty space to suppress spurious geometry. To formalize this, we introduce a scheduled auxiliary regularization, formulated as:

$$\mathcal{L}_{aux} = \sum_{r \in \mathcal{R}_{air}} \|\hat{I}_{air}(r)\|_2^2$$

We apply $\mathcal{L}_{aux}$ for the first 20% of training epochs, where $\hat{I}_{air}(r)$ is the predicted intensity along rays known to pass entirely through air and therefore render to zero attenuation. The auxiliary rays are sampled as an additional batch, independent of the primary $\mathcal{L}_{int}$ sampling, to avoid diluting the non-zero attenuation training signal. By only applying the regularization early on, we prevent unconstrained floaters from forming early, and $\mathcal{L}_{aux}$ is highest in these early epochs where floaters are most prevalent, and decreases as the regularization suppresses them. Keeping the two ray sets separate and time-limiting the regularization to early epochs prevents the network from over-fitting to zero-attenuation rays, which could otherwise suppress valid low-attenuation features in the primary training signal.

3.2.4 Optimization and Training Pipeline

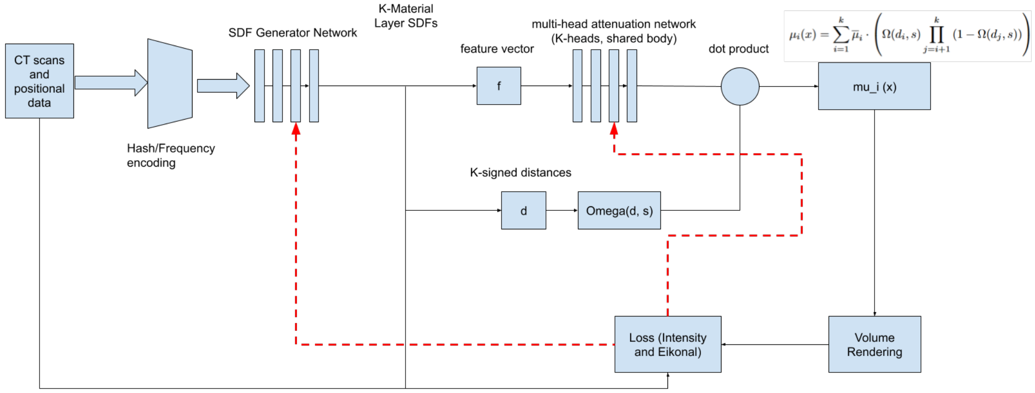

CT scans and positional data
Hash/Frequency encoding
SDF Generator Network
K-Material Layer SDFs
feature vector
f
multi-head attenuation network (K-heads, shared body)
dot product
$\mu_i(x) = \sum_{i=1}^{k} \overline{\mu}_i \cdot \left(\Omega(d_i, s) \prod_{j=i+1}^{k} (1 - \Omega(d_j, s)) \right)$
mu_i (x)
K-signed distances
d
Omega(d, s)
Loss (Intensity and Eikonal)
Volume Rendering

Figure 4: An overview of the proposed pipeline

The proposed training pipeline, illustrated in Figure 4, follows the same overall structure as NeAS. Input positional data is encoded via hash or frequency encoding before being passed to the SDF Generator Network, which outputs K signed distances $d_1, \ldots, d_K$ and a shared feature vector $\mathbf{f}$. The feature vector is forwarded to the unified multi-head attenuation network, whose K output heads each consist of a single hidden layer, producing raw attenuation parameters $\overline{\mu}_1, \ldots, \overline{\mu}_K$. The K-material soft selector then combines these with the corresponding boundary values $\Omega(d_i, s)$ to produce the final attenuation coefficient $\mu(x)$, which is integrated along each ray via Beer-Lambert volume rendering using stratified sampling.

The network is optimized end-to-end using the Adam optimizer with a step-decay learning rate schedule, minimizing the combined loss $\mathcal{L} = \mathcal{L}_{int} + \mathcal{L}_{aux} + \lambda\mathcal{L}_{reg}$. The SDF network, shared attenuation backbone, and the boundary sharpness parameter s are all co-optimized within a single optimizer, allowing the surface boundary steepness to adapt alongside the geometry and attenuation fields throughout training. To stabilize early optimization, a coarse-to-fine frequency annealing schedule is applied: such that the bandwidth parameter τ grows linearly from $\tau_0 = 2.0$ to the maximum encoding frequency L over the first half of training iterations, after which it is held fixed. This prevents the network from prematurely fitting high-frequency noise before the coarser scene structure has converged. Gradients from both the intensity and Eikonal terms propagate back through the renderer into all K SDF fields and the shared attenuation backbone jointly. The Eikonal regularization is applied independently to each of the K SDF fields and averaged, ensuring that the unit-gradient constraint is enforced uniformly across all material surfaces regardless of K.

3.2.5 Determining the Range of Attenuation Values

A Gaussian Mixture Model (GMM) [2] is an unsupervised probabilistic model that represents a distribution as a weighted sum of K Gaussian components, well-suited to modelling multi-modal scalar distributions where distinct modes correspond to separable material classes. The optimal K is selected via the Bayesian Information Criterion (BIC) [11], which balances goodness-of-fit against model complexity to avoid overfitting.

Rather than manually estimating attenuation bounds as in the original NeAS, we adopt a data-driven approach to determine the α and β parameters for each material configuration. For each anatomical scene, we first train a 1M-NeAS model with a deliberately wide attenuation range (high α), allowing the network to freely capture the full distribution of attenuation coefficients present in the scene.

Once trained, the learned volumetric attenuation field is sampled over the full voxel grid, and voxels below a small background threshold are discarded as air. A Gaussian Mixture Model (GMM) is then fit to the resulting non-background attenuation distribution via Expectation-Maximization (EM) for $K = 1, \ldots, 4$ components. The optimal number of materials is selected by minimizing the Bayesian Information Criterion (BIC), which penalizes model complexity to avoid overfitting to noise in the distribution.

Material boundaries are placed at the density valley between adjacent Gaussian peaks, with the leftmost boundary fixed at zero and the rightmost set to 15% above the 99.5th percentile of the distribution. For each interval $[a_i, b_i]$, the material parameters are assigned as $\beta_i = a_i$ and $\alpha_i = b_i - a_i$ to ensure material boundaries do not overlap. The resulting values are used for configurations for 1M to 4M experiments for that anatomic region.

4 Experimental Setup

All models were trained on NVIDIA A4000 GPUs with 16GB of VRAM. Training runs for 1000 epochs (50,000 iterations) per configuration, with model checkpoints saved every 50 epochs. Training progress and metrics are logged via Weights & Biases. Per epoch, 512 non-zero attenuation rays are sampled per projection for the primary intensity loss, with an additional 128 zero-attenuation rays sampled exclusively during the first 200 epochs (20% of training) for the floater regularization loss.

4.1 Dataset

First introduced in NAF [15], we evaluate both pipelines on clinical Cone-Beam CT (CBCT) data covering four anatomical regions: Abdomen, Chest, Foot, and Jaw. The dataset is provided in TIGRE format, with scene-dependent ray sampling depths determined by the physical extent of each anatomical region: 576 points per ray for Abdomen, 320 for Foot and Jaw, and 192 for Chest. Unlike the phantom-based evaluation in the original NeAS paper, which provided ground-truth meshes from helical CT for surface reconstruction assessment, this dataset consists of real clinical CBCT scans without corresponding ground-truth geometry. Consequently, evaluation is limited to image quality metrics on held-out projections. Each scene provides 50 projections for training and 50 for validation, where the validation projections are used exclusively for quantitative evaluation and are withheld during training. The four scenes vary considerably in material complexity and the ratio of air to tissue, providing a diverse benchmark for evaluating the proposed method.

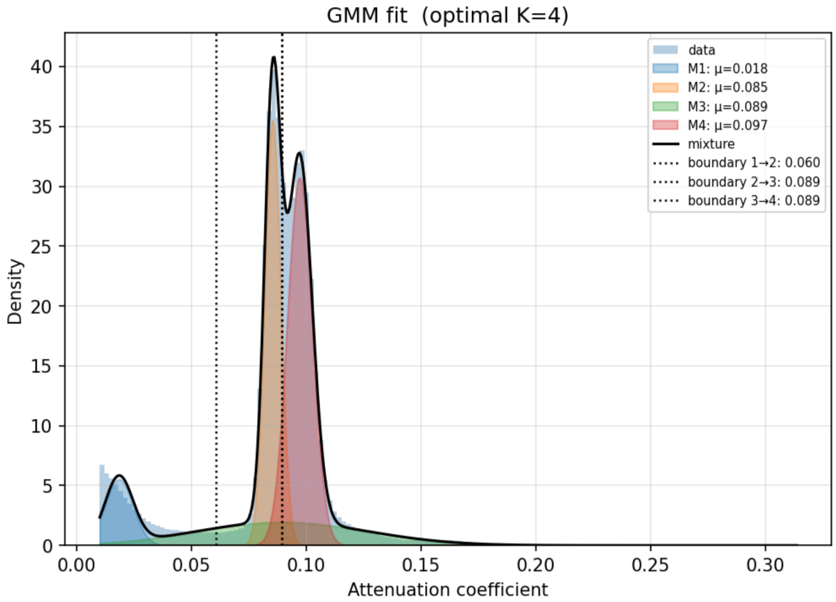


Figure 5: Optimal $K = 4$ GMM fit to the attenuation distribution of a converged modified 1M-NeAS abdomen volume. Vertical dotted lines indicate inter-material boundaries placed at density valleys to establish non-overlapping $[\beta_i, \beta_i + \alpha_i]$ activation intervals. X-Axis is the Normalized Attenuation Coefficient μ

4.2 Sparsity Configurations

To evaluate the robustness of both pipelines under limited acquisition conditions, we test across four input sparsity levels: 50, 20, 10, and 5 views. The 50 configuration serves as the standard baseline. All subsets are sampled uniformly from the 50 training projections to ensure consistent angular coverage across sparsity levels.

4.3 Evaluation Metrics

We evaluate performance using both 2D projection-domain and 3D volume-domain metrics. For projection-domain evaluation, we report Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM) computed on the 50 held-out validation projections. To assess the quality of the reconstructed CT volume, we additionally report volumetric PSNR and SSIM computed directly on the 3D output, providing a more comprehensive measure of reconstruction fidelity beyond held-out view synthesis.

5 Results & Analysis

5.1 Baseline vs. Proposed Model

Tables 1–4 present quantitative comparisons between NeAS and our proposed pipeline across 2D and 3D PSNR and SSIM metrics, with the best score per metric bolded. Qualitative projection comparisons are shown in Tables 6–9, where differences in sharpness and artifact suppression between the two pipelines are most visible.

Volume density comparisons are shown in Figures 6–13, where each image shows ground truth slices in the top half and predicted slices in the bottom half.

A primary limitation of the baseline NeAS architecture is its restriction to two materials. As shown in Tables 1 and 2, our proposed pipeline successfully extends reconstruction to arbitrary material counts. Notably, in the Abdomen scene, reconstruction fidelity scales positively with material complexity, peaking at the 4M configuration with a 3D PSNR of 33.248. This indicates that our differentiable

K-material soft selector effectively isolates distinct anatomical tissues. Conversely, the proposed model underperformed the baseline across all metrics in the Jaw scene (Table 4), an anatomical edge case that is analyzed further in Section 6.

Table 1: Abdomen: Quantitative comparison across material configurations.

	2D PSNR ↑		2D SSIM ↑		3D PSNR ↑		3D SSIM ↑	
Config	NeAS	Ours	NeAS	Ours	NeAS	Ours	NeAS	Ours
1M	46.825	47.086	0.992	**0.994**	31.416	32.656	0.816	0.843
2M	46.285	47.060	0.992	0.993	31.225	32.424	0.823	0.836
3M	—	46.848	—	0.993	—	32.724	—	0.843
4M	—	**47.303**	—	**0.994**	—	**33.248**	—	**0.861**

Table 2: Chest: Quantitative comparison across material configurations.

	2D PSNR ↑		2D SSIM ↑		3D PSNR ↑		3D SSIM ↑	
Config	NeAS	Ours	NeAS	Ours	NeAS	Ours	NeAS	Ours
1M	45.689	45.490	0.991	0.992	31.698	31.746	0.919	0.916
2M	**46.358**	45.872	0.992	0.992	32.053	**32.177**	0.917	0.921
3M	—	45.474	—	0.992	—	31.878	—	0.916
4M	—	45.968	—	**0.993**	—	32.176	—	**0.922**

Table 3: Foot: Quantitative comparison across material configurations.

	2D PSNR ↑		2D SSIM ↑		3D PSNR ↑		3D SSIM ↑	
Config	NeAS	Ours	NeAS	Ours	NeAS	Ours	NeAS	Ours
1M	42.440	42.743	0.982	0.983	31.231	31.500	**0.895**	0.894
2M	42.250	42.725	0.982	0.981	31.086	31.307	0.889	0.891
3M	—	42.449	—	0.983	—	31.590	—	0.890
4M	—	**43.124**	—	**0.984**	—	**31.718**	—	0.892

Table 4: Jaw: Quantitative comparison across material configurations.

	2D PSNR ↑		2D SSIM ↑		3D PSNR ↑		3D SSIM ↑	
Config	NeAS	Ours	NeAS	Ours	NeAS	Ours	NeAS	Ours
1M	**39.524**	34.319	**0.964**	0.954	**33.846**	31.731	0.880	0.801
2M	37.212	34.043	0.959	0.951	32.986	31.607	**0.813**	0.792
3M	—	34.232	—	0.954	—	31.714	—	0.801
4M	—	34.343	—	0.953	—	31.701	—	0.798

Table 5: Chest Comparison across 5, 10, 20, and 50 views, on 2-material.

	2D PSNR ↑		2D SSIM ↑		3D PSNR ↑		3D SSIM ↑	
# Views	NeAS	Ours	NeAS	Ours	NeAS	Ours	NeAS	Ours
5	31.746	31.286	0.915	0.916	21.517	21.730	0.550	0.547
10	35.145	35.922	0.940	0.947	23.310	24.435	0.638	0.671
20	41.163	41.153	0.976	0.974	27.088	27.376	0.781	0.785
50	46.358	45.872	0.992	0.992	32.053	32.177	0.917	0.921

5.2 Performance on Sparser Views

While both models exhibit expected performance degradation as angular sampling decreases, the proposed architecture demonstrates slightly higher robustness in mid-sparsity regimes. Specifically, at the 10-view threshold, the proposed model outperforms the baseline in both 2D PSNR (35.922 vs. 35.145) and 3D PSNR (24.435 vs. 23.310). At extreme sparsity (5 views), both models suffer significant degradation, which is visually reflected by a loss of fine structural detail and increased blurring, as shown in the qualitative comparisons in Table 10.

6 Discussion & Limitations

6.1 Successes

The proposed pipeline successfully addresses several critical bottlenecks in applying implicit neural representations to clinical CT data. Primarily, the architecture scales robustly to arbitrary material counts without the need for manual hyperparameter tuning. By replacing the decoupled attenuation MLPs for each material as in the baseline with a unified MLP with a shared latent space and differentiable K-material soft selector, the network is forced to learn coherent geometric features across tissue boundaries. This structural prior is validated by the quantitative results in the Abdomen and Chest scenes, where increasing the material complexity to 4M resulted in the highest overall 3D PSNR and SSIM scores.

6.2 Current Limitations

Despite the successes, the proposed architecture fails with high-frequency details, as shown by the Jaw scene as it underperformed relative to the baseline in all metrics. The Jaw dataset is characterized by complex, dense, and fine boundaries such as teeth and jawbones mixed with air and soft tissue. It is possible this is caused by the shared attenuation backbone, which may excel at learning smoother material transitions but remains bottlenecked by sharper ones.

Furthermore, while the Gaussian Mixture Model (GMM) approach successfully automates the determination of attenuation bounds, it introduces a strict sequential dependency. Because the GMM requires sampling a learned volumetric attenuation field, a complete 1M model must be trained to convergence before the α and β parameters can be calculated for any multi-material configuration. This doubles the effective training time and creates a computational bottleneck. Additionally, the multi-material model becomes entirely dependent on the quality of this 1M prior; if the initial 1M reconstruction contains significant noise or hallucinated geometry that are common under extreme sparse-view conditions, then the GMM may fit to these artifacts rather than true anatomical densities and propagate suboptimal bounds into the K-material optimization phase.

7 Conclusion & Future Work

In this thesis, we introduced a unified, scalable approach to multi-material surface reconstruction in sparse-view CT. By extending the Neural Attenuation Surface (NeAS) framework with a shared latent backbone and a differentiable K-material sequential soft selector, we demonstrated that implicit neural representations can be effectively scaled to arbitrary material counts without requiring rigid, hard-coded boundaries. Our quantitative and qualitative results on clinical datasets, particularly in the Abdomen and Chest scenes, validate that this unified architecture successfully isolates distinct tissue structures and maintains robust 3D volumetric fidelity.

However, the model's underperformance on the Jaw dataset indicates that while a shared latent space excels at modeling continuous soft tissue transitions, it struggles to resolve the extreme, high-frequency density shifts characteristic of complex bone-to-air boundaries. Future work must explore architectural adaptations, such as adaptive frequency routing or hybrid network capacities, to better capture these fine geometric details without sacrificing the efficiency of a shared backbone.

Additionally, the reliance on a Gaussian Mixture Model to automate attenuation bounds requires further study. While the GMM successfully eliminates manual parameter tuning, its sequential dependency on a fully converged 1M prior creates a computational bottleneck and risks propagating noise artifacts into the final multi-material optimization. To bypass this limitation and further stabilize reconstructions under extreme sparsity (e.g., fewer than 10 views), future research should investigate the integration of explicit 3D structural priors and targeted importance sampling. By embedding volumetric priors directly into the training loop, it may be possible to accelerate implicit attenuation fields and dynamically guide material decomposition without the overhead of the current two-step GMM approach.

References

[1] A.H. Andersen and A.C. Kak. Simultaneous algebraic reconstruction technique (sart): A superior implementation of the art algorithm. *Ultrasonic Imaging*, 6(1):81–94, 1984.

[2] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977.

[3] L. A. Feldkamp, L. C. Davis, and J. W. Kress. Practical cone-beam algorithm. *J. Opt. Soc. Am. A*, 1(6):612–619, Jun 1984.

[4] Amos Gropp, Lior Yariv, Niv Haim, Matan Atzmon, and Yaron Lipman. Implicit geometric regularization for learning shapes. In *Proceedings of Machine Learning and Systems 2020*, pages 3569–3579. 2020.

[5] Michael Kazhdan, Matthew Bolitho, and Hugues Hoppe. Poisson surface reconstruction. In *Proceedings of the Fourth Eurographics Symposium on Geometry Processing*, SGP '06, page 61–70, Goslar, DEU, 2006. Eurographics Association.

[6] Chen-Hsuan Lin, Wei-Chiu Ma, Antonio Torralba, and Simon Lucey. Barf: Bundle-adjusting neural radiance fields. In *IEEE International Conference on Computer Vision (ICCV)*, 2021.

[7] William E. Lorensen and Harvey E. Cline. Marching cubes: A high resolution 3d surface construction algorithm. In *Proceedings of the 14th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '87, page 163–169, New York, NY, USA, 1987. Association for Computing Machinery.

[8] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020.

[9] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics*, 41(4):1–15, July 2022.

[10] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[11] Gideon Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978.

[12] L. A. Shepp and B. F. Logan. The fourier reconstruction of a head section. *IEEE Transactions on Nuclear Science*, 21(3):21–43, 1974.

[13] Emil Y Sidky and Xiaochuan Pan. Image reconstruction in circular cone-beam computed tomography by constrained, total-variation minimization. *Physics in Medicine & Biology*, 53(17):4777, 2008.

[14] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *arXiv preprint arXiv:2106.10689*, 2021.

[15] Ruyi Zha, Yanhao Zhang, and Hongdong Li. Naf: Neural attenuation fields for sparse-view cbct reconstruction. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 442–452. Springer, 2022.

[16] Chengrui Zhu, Ryoichi Ishikawa, Masataka Kagesawa, Tomohisa Yuzawa, Toru Watsuji, and Takeshi Oishi. Neas: 3d reconstruction from x-ray images using neural attenuation surface, 2025.

8 Additional Figures

Table 6: Abdomen: Qualitative projection comparison across 4 views (2-Material)

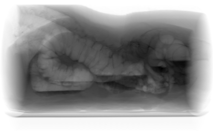
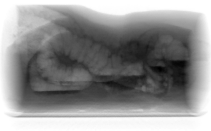
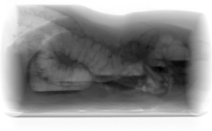
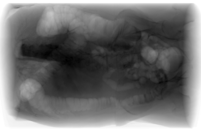
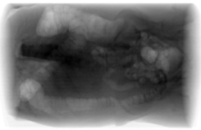
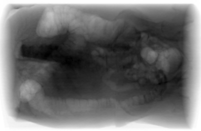
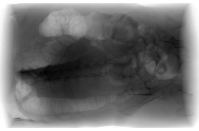
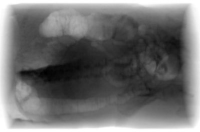
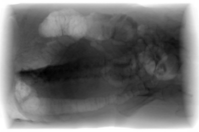
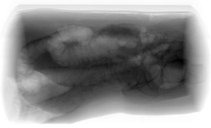
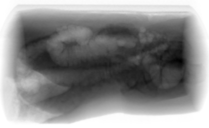
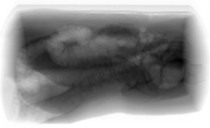
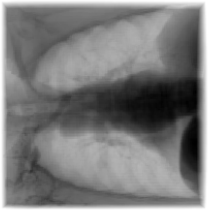
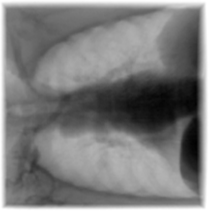
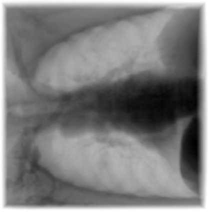
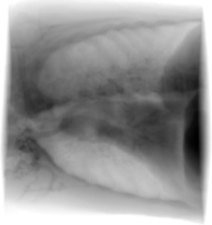
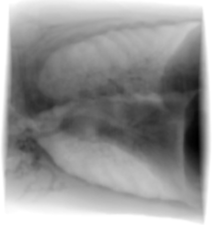
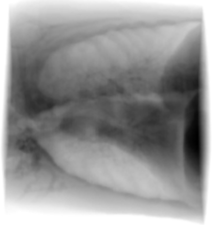
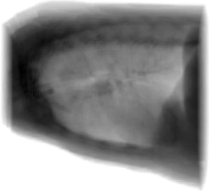
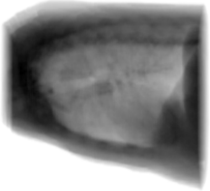
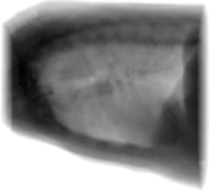
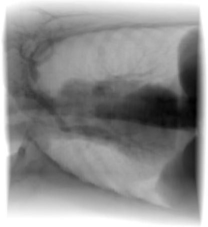
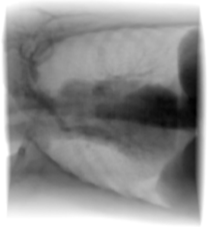
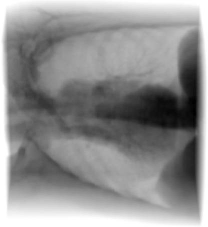
View	Ground Truth	NeAS	Proposed
View 1			
View 15			
View 30			
View 45			

Table 7: Chest: Qualitative projection comparison across 4 views (2-Material)

View	Ground Truth	NeAS	Proposed
View 1			
View 15			
View 30			
View 45			

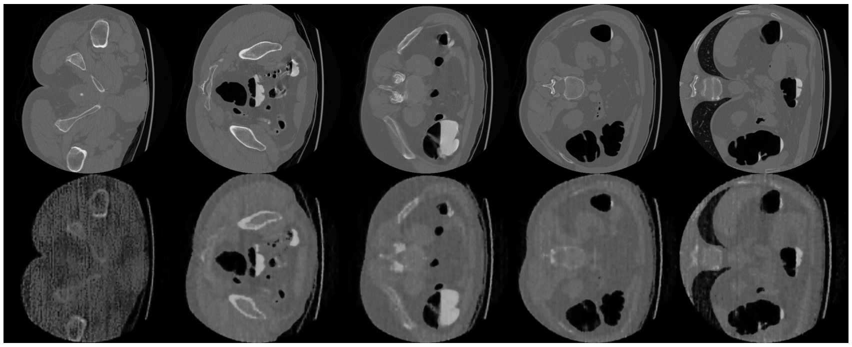
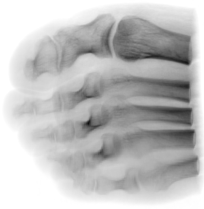
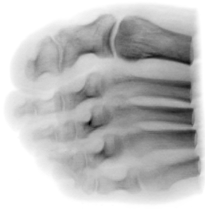
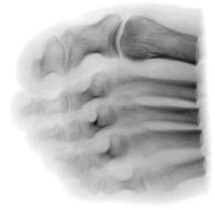
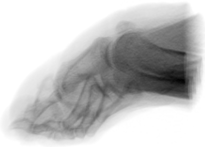
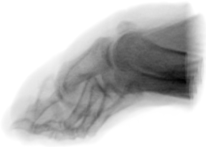
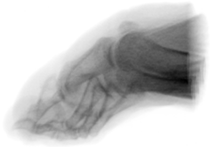
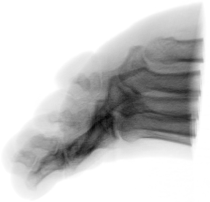
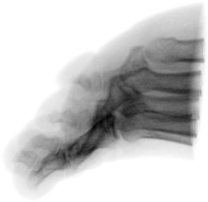
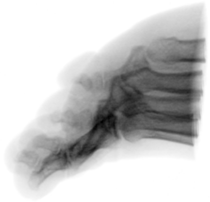
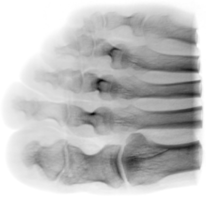
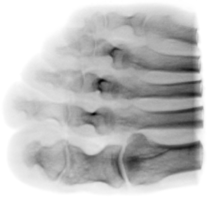
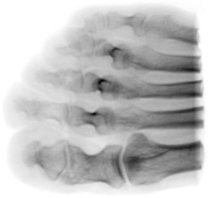


Figure 6: Abdomen: 2-M density reconstruction (NeAS). Top: Ground Truth, Bottom: Predicted.

Table 8: Foot: Qualitative projection comparison across 4 views (2-Material)

View	Ground Truth	NeAS	Proposed
View 1			
View 15			
View 30			
View 45			

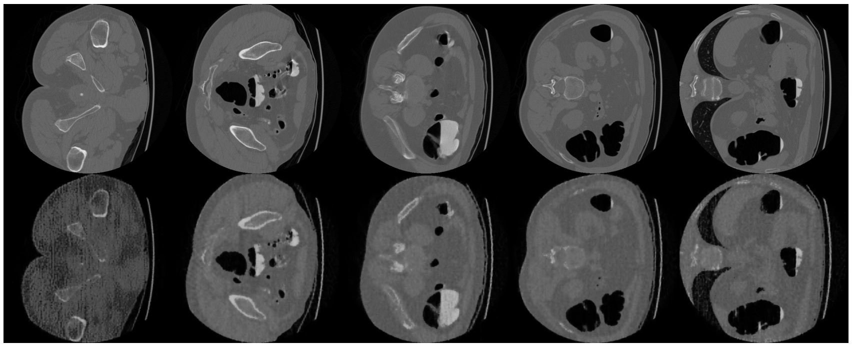
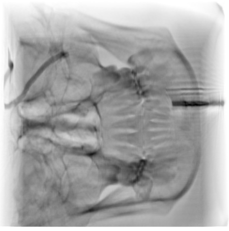
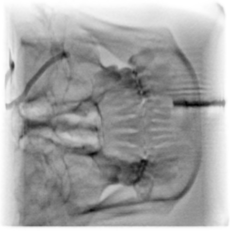
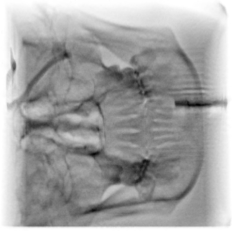
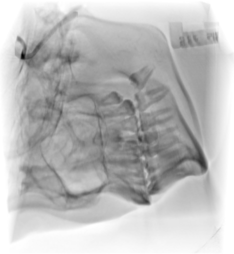
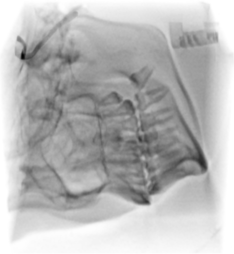
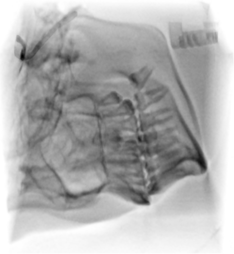
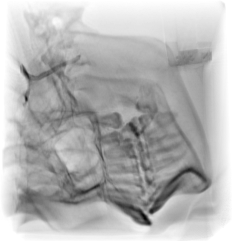
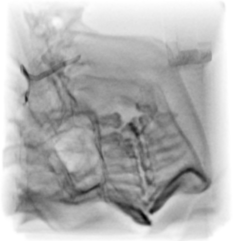
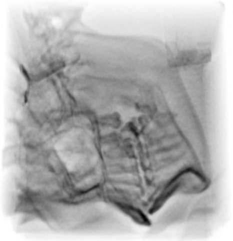
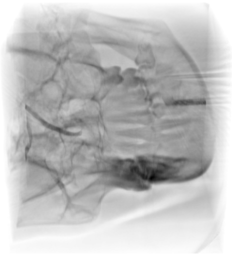
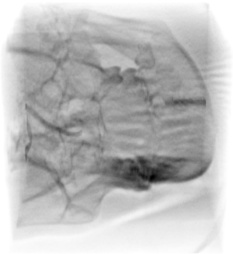
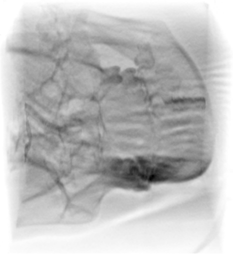


Figure 7: Abdomen: 2-M density reconstruction (Ours). Top: Ground Truth, Bottom: Predicted.

Table 9: Jaw: Qualitative projection comparison across 4 views (2-Material)

View	Ground Truth	NeAS	Proposed
View 1			
View 15			
View 30			
View 45			

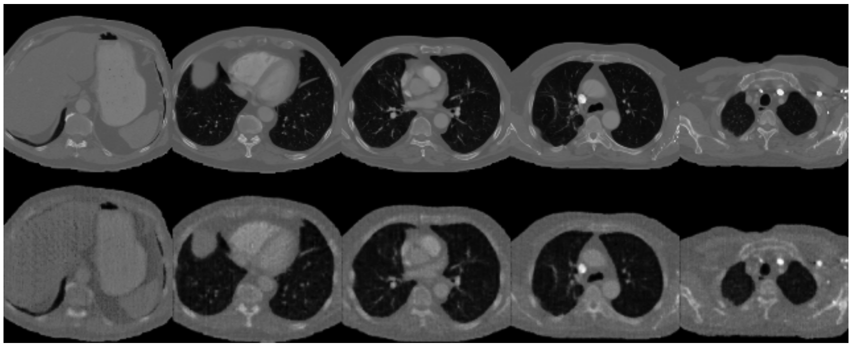
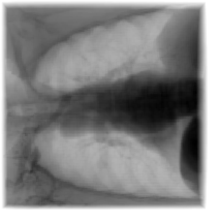
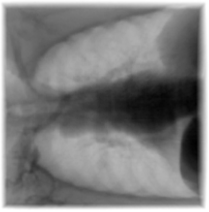
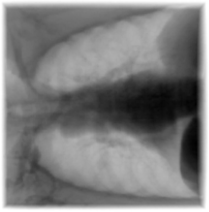
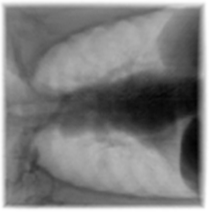
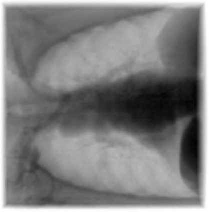
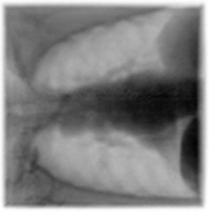
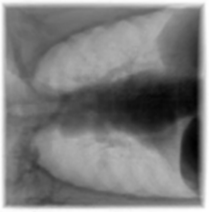
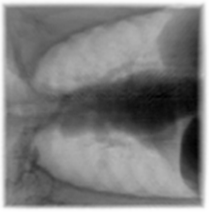
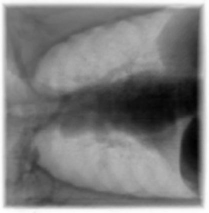


Figure 8: Chest: 2-M density reconstruction (NeAS). Top: Ground Truth, Bottom: Predicted.

Table 10: Chest: Qualitative projection comparison across sparsity configurations (View 1)

Views	NeAS	Proposed
GT		
50		
20		
10		
5		

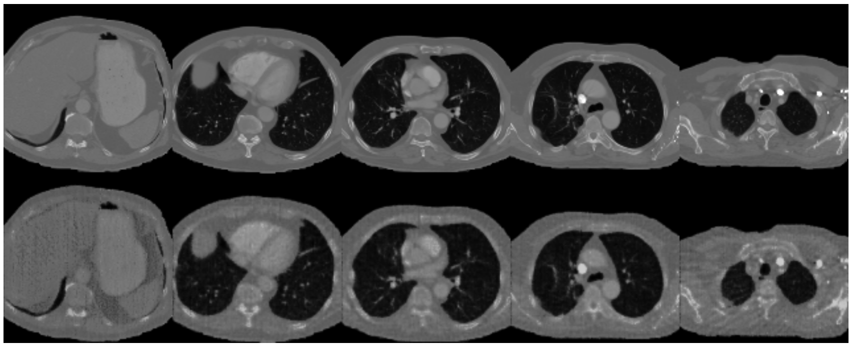


Figure 9: Chest: 2-M density reconstruction (Ours). Top: Ground Truth, Bottom: Predicted.

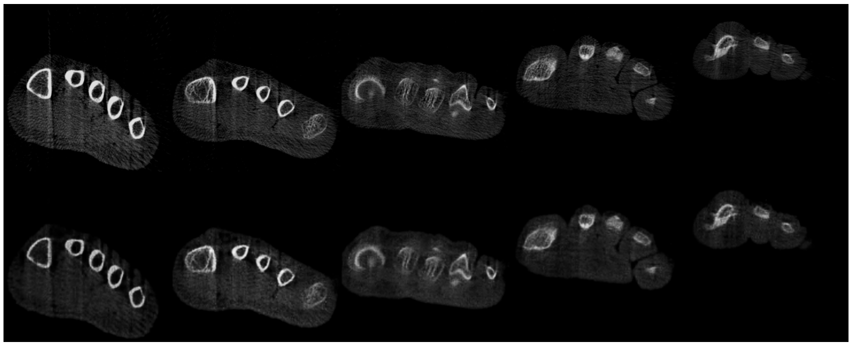


Figure 10: Foot: 2-M density reconstruction (NeAS). Top: Ground Truth, Bottom: Predicted.

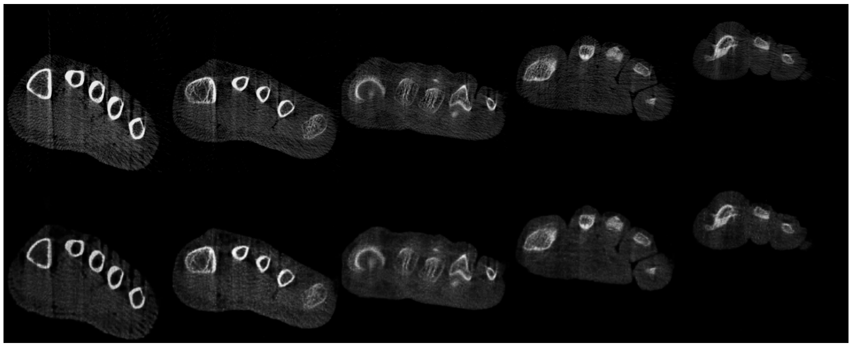


Figure 11: Foot: 2-M density reconstruction (Ours). Top: Ground Truth, Bottom: Predicted.

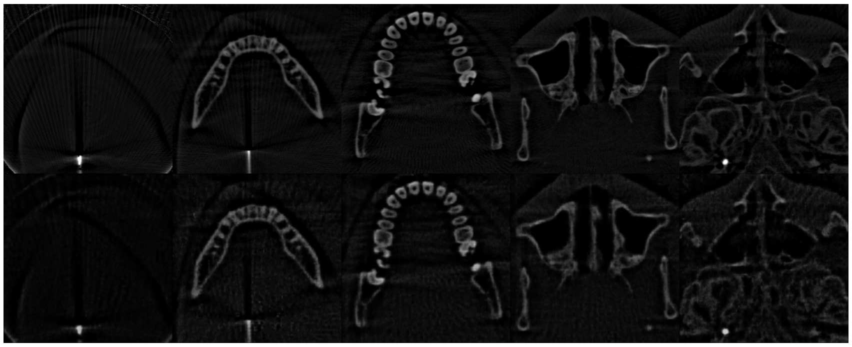


Figure 12: Jaw: 2-M density reconstruction (NeAS). Top: Ground Truth, Bottom: Predicted.

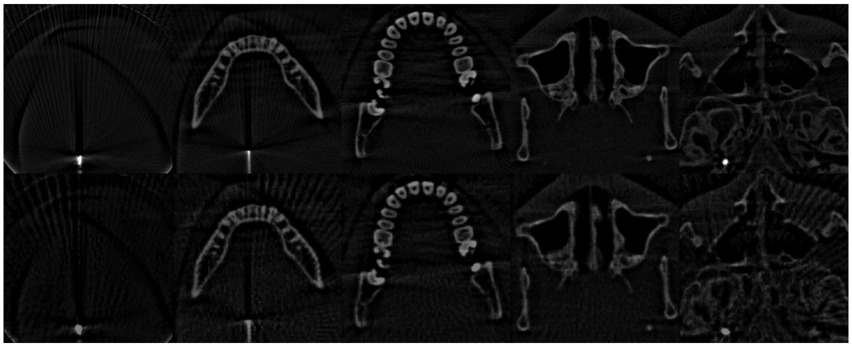


Figure 13: Jaw: 2-M density reconstruction (Ours). Top: Ground Truth, Bottom: Predicted.