Are Pixel-Wise Metrics Reliable for Sparse-View Computed Tomography Reconstruction?

Tianyu Lin[1] **Xinran Li**[1,2] **Chuntung Zhuang**[1] **Qi Chen**[1] **Yuanhao Cai**[1]
Kai Ding[3] **Alan L. Yuille**[1] **Zongwei Zhou**[1,*]

[1]Johns Hopkins University [2]Yale University [3]Johns Hopkins Medicine

Code, dataset, and models: https://github.com/MrGiovanni/CARE

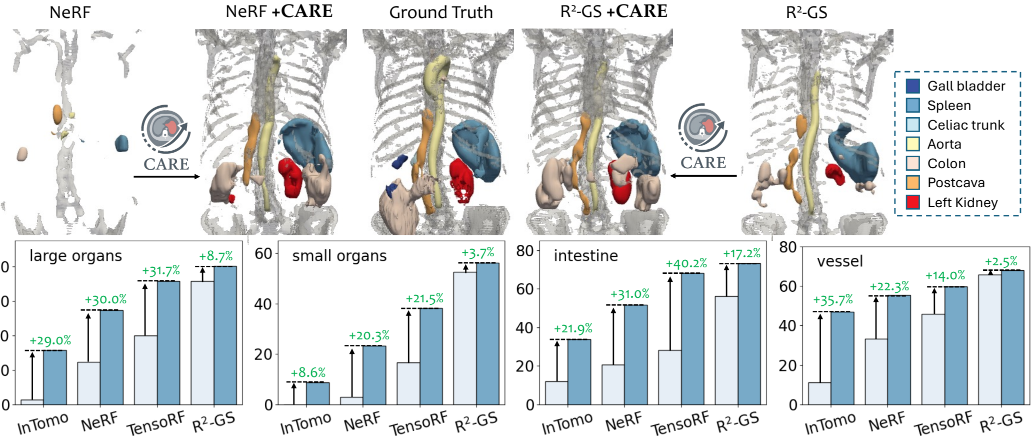

Figure 1: **CARE Improves Structural Completeness in Sparse-View CT Reconstruction.** *Top*: Qualitative comparison of reconstructed CT scans from NeRF and R^2-GS, with and without our proposed CARE on clinically important structures. *Bottom*: Quantitative gains in structural completeness across four reconstruction methods. CARE consistently improves results, confirming its model-agnostic and plug-and-play nature.

Abstract

Widely adopted evaluation metrics for sparse-view CT reconstruction—such as Structural Similarity Index Measure and Peak Signal-to-Noise Ratio—prioritize pixel-wise fidelity but often fail to capture the completeness of critical anatomical structures, particularly small or thin regions that are easily missed. To address this limitation, we propose a suite of novel anatomy-aware evaluation metrics designed to assess structural completeness across anatomical structures, including large organs, small organs, intestines, and vessels. Building on these metrics, we introduce CARE, a Completeness-Aware Reconstruction Enhancement framework that incorporates structural penalties during training to encourage anatomical preservation of significant structures. CARE is model-agnostic and can be seamlessly integrated into analytical, implicit, and generative methods. When applied to these methods, CARE substantially improves structural completeness in CT reconstructions, achieving up to **+32%** improvement for large organs, **+22%** for small organs, **+40%** for intestines, and **+36%** for vessels.

[*]Correspondence to Zongwei Zhou (ZZHOU82@JH.EDU)

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

1 Introduction

Reducing radiation dose is a major concern in computed tomography (CT), where a standard image acquisition typically requires 800 to 1500 X-ray projections [39, 17]. These high doses pose long-term health risks, especially for vulnerable patients such as children, pregnant individuals, and those undergoing frequent imaging [64, 9, 29]. One promising research direction to lower radiation is *sparse-view CT*, which aims to reconstruct high-quality images from a significantly reduced number of projection views—often fewer than 50 views [63, 50, 59]. While this approach can dramatically lower radiation dose, it introduces significant challenges: reconstructions from limited data often lose fine anatomical details [30]. These details are critical for clinical applications such as surgical planning, radiotherapy, and longitudinal disease monitoring [58, 56]. This raises a central research question: *how can we ensure the completeness of clinically important anatomical structures in sparse-view CT reconstruction?*

Over the past decade, researchers have reported steady improvements in sparse-view CT reconstruction by optimizing pixel-wise metrics—e.g., Peak Signal-to-Noise Ratio (PSNR) [24] and Structural Similarity Index Measure (SSIM) [53]. *However, higher PSNR or SSIM scores do not guarantee better clinical utility (analyzed in §3.1).* These metrics compute global averages of per-voxel intensity differences, which means they are largely insensitive to the presence or absence of small but clinically important anatomical structures. For example, missing the gall-bladder, adrenal glands, celiac artery, or a segmental vein affects only a tiny fraction of the total scan volume—often less than 0.0001%—and may change PSNR/SSIM by just the third or fourth decimal place. As a result, the reconstructed CT scans that appear high-quality under PSNR/SSIM may still miss important anatomical structures and be unsuitable for clinical applications [15]. Our internal reader study at Johns Hopkins University, involving 21 board-certified radiologists, assessed CT reconstructions produced by state-of-the-art CT reconstruction methods. The results raised significant concerns, including missing anatomical structures, implausible hallucinations, and severe artifacts, suggesting a growing disconnect between algorithmic novelty and clinical realism. This highlights a pressing need for new evaluation metrics that can directly assess whether anatomical structures are preserved in sparse-view CT reconstruction.

To assess the anatomical structures in the reconstructed CT scans, conventional practice requires radiologists to manually inspect the scans—a process that is subjective, time-consuming, and impractical to scale. Fortunately, recent advances in *medical image segmentation*, powered by strong model architectures [3, 27, 36, 37, 4] and large-scale annotated datasets [54, 32, 40, 31, 14], now enable AI models to automatically segment multiple anatomical structures with high accuracy and consistency. These advances greatly reduce the need for manual inspection and motivate a key hypothesis: automated segmentation can serve as a scalable metric to assess whether anatomical structures are preserved in reconstructed scans.

We introduce CARE, a *Completeness-Aware Reconstruction Enhancement* framework that defines new evaluation metrics—derived from medical image segmentation—and uses them to directly supervise diffusion models, enhancing the structural completeness of preexisting sparse-view CT reconstruction methods. Our contributions are:

1. **Anatomy-Aware CT Reconstruction Metric**: We leverage nnU-Net (under Apache-2.0 license) [26] as an anatomy segmentator, trained on more than 3,000 voxel-wise annotated CT scans [33], to build a suite of anatomy-aware metrics that measure the anatomical completeness of four clinically important categories of anatomical structures (detailed in Figure 2) in reconstructed CT scans.

2. **Anatomy-Aware CT Reconstruction Framework**: We go beyond evaluation by using the anatomy-aware metrics as direct supervision. During training, CARE compares the segmentation output of the reconstructed CT scans with the segmentator's prediction and back-propagates a penalty whenever any anatomical structure is missing or incomplete. This structural penalty is fully differentiable and agnostic to the underlying forward model, allowing CARE to be applied effectively to a range of preexisting reconstruction methods, including Neural Radiance Fields (NeRF), Gaussian Splatting (GS), and their variants. Instead of training a separate reconstruction model for each patient, using diffusion models to enhance the CT scans reconstructed by existing methods enables the model to learn anatomical priors from a large population and generalize across patients, thus making CARE patient-agnostic.

Extensive experiments show that applying CARE to preexisting sparse-view CT reconstruction methods produces striking gains (summarized in Figure 1; detailed in §4). Compared with nine preexisting methods, our CARE achieves up to **32%** for large organs, **+22%** for small organs, **+40%** for intestinal structures, and **+36%** for vascular structures. These gains are both statistically significant and clinically important, but worryingly, such improvements are entirely hidden when evaluated using widely adopted pixel-wise metrics like PSNR and SSIM (see Table 2). This highlights a critical pitfall in current evaluation practices and raises concerns about their continued usage in clinical reconstruction studies.

2 Related Works

Anatomy-Aware Evaluation. Conventional image quality metrics, such as PSNR [24] and SSIM [53], focus primarily on pixel-wise fidelity but fail to reliably reflect clinical significance, particularly regarding the completeness of fine anatomical structures. This limitation is especially pronounced under challenging imaging conditions, including sparse-view or low-dose acquisitions. Despite recent efforts, research explicitly targeting anatomy-aware evaluation metrics remains scarce [49].

Beyond pixel-wise fidelity, several perceptual image quality metrics have been proposed to better align with human visual judgment. The Fréchet Inception Distance (FID) [22] and Kernel Inception Distance (KID) [7] evaluate distributional similarity between sets of real and generated images using features from a pre-trained Inception network. Similarly, LPIPS [65] leverages deep features to measure perceptual dissimilarity between individual image pairs. GMSD [57] assesses quality based on gradient magnitude similarity, capturing structural distortions more effectively than PSNR or SSIM. More recently, GLIPS [2] combines global and local perceptual cues to evaluate photorealism in AI-generated images. While these metrics improve upon traditional pixel-wise measures in natural image domains, they remain agnostic to anatomical semantics and do not explicitly assess the presence, completeness, or geometric fidelity of clinically relevant structures in medical imaging.

CT Reconstruction Methods. Traditional reconstruction methods, such as Filtered Back Projection (FBP) [41] and iterative algorithms like Simultaneous Algebraic Reconstruction Technique (SART) [1] and Adaptive Steepest Descent-Projection onto Convex Sets (ASD-POCS) [47], are widely employed due to their simplicity and effectiveness. However, these methods exhibit significant quality degradation under sparse-view or low-dose scenarios [16]. Early work by Zhang [64] and Chen [12] mitigated this issue by incorporating a prior CT scan as strong regularisation, enabling acceptable image quality with substantially fewer projections. To overcome the remaining limitations, recent advances have leveraged neural rendering methods, including NeRF [38], InTomo [60], TensoRF [10], Neural Attenuation Fields (NAF) [62], and SAX-NeRF [8], to synthesize views and significantly improve 3D reconstruction fidelity. In parallel, Gaussian splatting methods such as R^2-GS [61] have also emerged, demonstrating promising results in achieving high-quality reconstructions with enhanced anatomical details. However, their capability to accurately reconstruct anatomical structures remains unverified, primarily due to the absence of suitable evaluation metrics.

Generative AI Models. Denoising Diffusion Probabilistic Models [23, 13, 34] have become strong alternatives to GANs [21] and VAEs [28] for high-fidelity image synthesis. In medicine, Stable-Diffusion–style implementations [51, 35] are already used to suppress CT noise and streak artefacts. Very recent work shows that diffusion priors can also tackle *sparse-view* CT, reporting state-of-the-art PSNR/SSIM with as few projections [50, 59, 48]. Yet these methods still optimise only pixel-wise losses and may hallucinate or omit small organs and vessels. Our CARE closes this gap by adding an anatomy-aware segmentation loss to the diffusion objective, explicitly safeguarding clinically critical structures while preserving the low-dose advantage of sparse-view acquisition.

3 CARE: Completeness-Aware Reconstruction Enhancement

3.1 Motivation: Pitfalls of Conventional Pixel–Wise Metrics

PSNR and SSIM dominate sparse-view CT reconstruction benchmarks because they are easy to compute and correlate well with global visual fidelity on natural images. In the clinical setting, however, these per-voxel averages are blind to errors that matter most. Figure 2 shows a common situation that even though having reasonable pixel-wise metrics, reconstruction could still fail in anatomical preservation. The reason is straightforward: small organs, intestines and vessels occupy

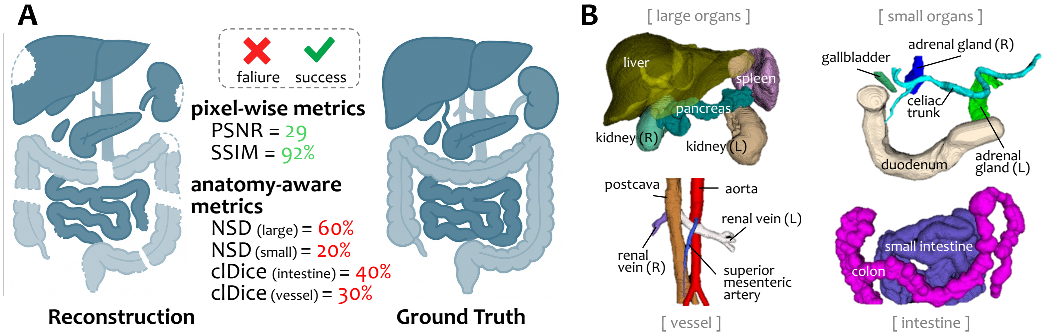

Figure 2: **Pixel-wise Metrics Overlook Structural Errors in the Focused Anatomical Structures. A.** Example pixel-wise metrics' pitfall. Pixel-wise metrics are insensitive to anatomical preservation failure. **B.** The four types of organs evaluated by CARE.

far less than 0.01 % of the volume, so their absence barely changes the third decimal place of SSIM and PSNR. Clinical evaluation, by contrast, hinges on whether these structures are present at all. Our own experiments (detailed in §4.2) confirm the gap: classical analytical methods such as FDK and SART rank mid-pack on pixel metrics but top on anatomy-aware scores, whereas several neural renderers achieve the opposite pattern. These observations motivate a shift from intensity similarity towards metrics that directly measure anatomical completeness.

3.2 Anatomy-Aware CT Reconstruction Metrics

We propose segmentation-based anatomy-aware CT reconstruction metrics to evaluate CT reconstruction quality. We utilize a frozen multi-organ segmentation model to compute metrics for four different kinds of anatomical structures. For large and small organs, we use the NSD metric due to its sensitivity to boundary accuracy and tolerance for minor shape deviations. For vessels and intestines, we apply the clDice [46] metric, given its effectiveness in capturing tubular structures and preserving topological connectivity. Finally, we report these metrics separately for each anatomical category[2].

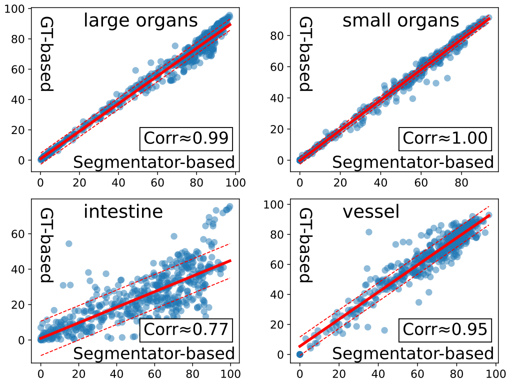

Figure 3: **Correlation Between Ground Truth (GT) Based and Segmentator Based Anatomy-Aware Metrics.** High correlation scores indicate that the anatomy segmentator is a strong substitute for human expert.

Anatomy Segmentator. A great CT reconstruction result should preserve the anatomical structures of the original CT scans, thus having similar segmentation results from a fixed segmentator between the original and reconstructed CT scans. Instead of relying on manually annotated ground-truth labels to derive our anatomy-aware metrics, we employ a frozen nnU-Net [26]—termed the *anatomy segmentator*—to capture clinically relevant fidelity. We choose nnU-Net framework for its self-configuring architecture and consistently state-of-the-art segmentation performance across diverse medical-imaging tasks. It was trained internally on the JHH dataset[3] with 3,151 CT scans annotated over three years by a team of 21 radiologists, with all annotations independently reviewed by one of three senior radiologists not involved in the initial annotation process to ensure quality [31, 55]. The anatomy segmentator will be publicly released. Leveraging a segmentator to construct our anatomy-aware metrics enables effortless deployment across large-scale CT reconstruction evaluations, eliminating the dependency

[2] These anatomical categories are divided based on clinical guidelines (see details in §A.1)

[3] A private CT segmentation dataset from Johns Hopkins Hospital (see details in §A.1.2).

on manual annotations. Figure 3 shows that the anatomy-aware metrics building on human annotated ground truth and segmentator's prediction have high correlation, further indicating the robustness of segmentator-based metrics. This further allows our anatomy-aware metrics to be applied to CT scans in the absence of segmentation labels.

Normalized Surface Dice (NSD). Both large organs and small organs can be viewed as compact, surface-dominated objects. NSD evaluates their geometric agreement by asking what fraction of the predicted surface S_P and reference surface S_G lie within a tolerance band of width τ:

$$\mathrm{NSD}_\tau = \frac{\left|\{x \in S_P : d(x, S_G) \le \tau\}\right| + \left|\{x \in S_G : d(x, S_P) \le \tau\}\right|}{|S_P| + |S_G|} \tag{1}$$

NSD is *scale-invariant* since every surface element located farther than τ incurs the same penalty. For example, a uniform 3 mm displacement with $\tau = 2$ mm therefore drives $\mathrm{NSD}_{2\mathrm{mm}}$ to 0 regardless how big the organ is.

On the contrary, the widely used Dice similarity coefficient (DSC) is scale-variant. DSC measures volumetric overlap via $\mathrm{DSC}(P, G) = \frac{2|P \cap G|}{|P|+|G|}$. As a result, the sensitivity of DSC to the same boundary shift is diluted by organ size. The expected loss satisfies $\Delta\mathrm{DSC} \approx \delta/(2R_{\mathrm{eff}})$ with $R_{\mathrm{eff}} = V_G/S_G$. For instance, with a 50 mm-radius liver, a 3 mm outer-shell error lowers DSC by merely 0.06—far too small to flag what clinicians deem a major discrepancy.

By focusing on boundary proximity rather than voxel volume, NSD retains millimetre-level sensitivity and offers a more faithful assessment of non-tubular organ segmentations than DSC.

Center-line Dice (clDice). Vessels and intestine are elongated, branching and hollow. Their clinical relevance stems from *connectivity*: a single broken segment in the superior-mesenteric artery may compromise whole-organ perfusion. clDice operates on skeletonised centre-lines and is sensitive to both continuity and topology, penalising even thin discontinuities that DSC or NSD might overlook. However, the ground-truth annotations for intestinal structures in most public datasets are often noisy or incomplete [20], which likely depresses the intestine–metric correlation by about 0.77 shown in Figure 3. Curating higher-quality bowel annotations—and re-evaluating clDice under those labels—will therefore be an important direction for future work.

Our anatomy-aware formulation thus matches the geometric character of each organ class: surface-sensitive NSD for compact organs, topology-sensitive clDice for slender, branching tubes. Together, they provide an automatic, quantitative, and clinically meaningful view through which to evaluate reconstruction algorithms.

3.3 Anatomy-Aware CT Reconstruction Framework

We propose a simple latent diffusion framework named CARE for CT enhancement, as shown in Figure 4. CARE starts by training an autoencoder and a latent diffusion model using a high-quality CT dataset, then integrating anatomy-guided supervision with additional pixel space constraints.

Preliminary. We begin by adapting a KL-regularized variational autoencoder to CT domain with the JHH dataset. Let X denote an input CT image in pixel space. Note that to consider inter-frame consistency and retain the original architecture of the autoencoder network, we stack three adjacent CT slices to be the three channels of the input image, giving $X \in \mathbb{R}^{H\times W\times 3}$. The encoder $\mathcal{E}_{\theta_E}$ maps X to a latent tensor $z = \mathcal{E}_{\theta_E}(X) \in \mathbb{R}^{h\times w\times c}$. A paired decoder $\mathcal{D}_{\theta_D}$ reconstructs the input as $\hat{X} = \mathcal{D}_{\theta_D}(z)$. The resulting autoencoder model reconstructs scans faithfully yet produces well-behaved latents that remain fixed for all later stages.

Using this frozen autoencoder, every CT image is mapped to the latent space for the diffusion process. During training, a de-noising UNet sees the noisy latent of the CT image concatenated with a deterministic, anatomy-preserving degradation of the original latent. The text embedding encodes contrast phase information of the CT scan. Its sole objective is to predict the injected noise, conditioning on the degraded latent steers the learning process toward anatomically consistent de-noising. Therefore, the model acquires an internal ability to transform low-quality latents to high-quality, structure-preserving counterparts. The autoencoder and latent diffusion model build the bedrock of our CARE framework.

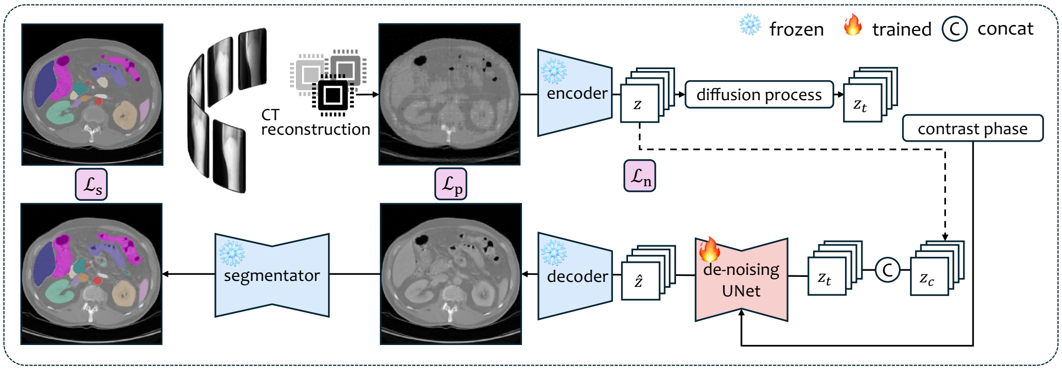

Figure 4: **CARE Framework.** Given the frozen anatomy segmentator, autoencoder, and the pretrained latent diffusion model, we then adapt the latent diffusion model to real reconstructed CT scans. CARE can be integrated into any reconstruction method to perform its enhancement capability. The overall training is supervised by three loss terms: the noise-prediction loss $\mathcal{L}_n$, pixel-space reconstruction loss $\mathcal{L}_p$, and anatomy-guidance loss $\mathcal{L}_s$.

CARE. With the pretrained latent diffusion model, we then adapt the de-noising UNet to real CT reconstructions by introducing anatomy-guided supervision.

Each CT image $X \in \mathbb{R}^{H\times W\times 3}$ is first compressed by the frozen autoencoder into a latent representation through $z = \mathcal{E}_{\theta_\mathrm{E}}(X) \in \mathbb{R}^{h\times w\times c}$. In the forward diffusion process, a noisy sequence $\{z_t\}_{t=0}^{T}$ is obtained by

$$z_t = \sqrt{\bar{\alpha}_t} z_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, \qquad \epsilon \sim \mathcal{N}(0, I), t \in \{1, \dots, T\}, \tag{2}$$

where $T = 1,000$, $z_0 \triangleq z$ and $\{\bar{\alpha}_t\}_{t=1}^{T}$ is a fixed variance-preserving schedule.

For each training data pair, we take a ground-truth CT image X_gt and its sparse-view reconstruction counterpart X_rec (e.g. FDK-50 views) and encode them as $z_\mathrm{gt} = \mathcal{E}_{\theta_\mathrm{E}}(X_\mathrm{gt})$ and $z_\mathrm{rec} = \mathcal{E}_{\theta_\mathrm{E}}(X_\mathrm{rec})$. The forward diffusion step is applied to z_gt to produce z_t, after which the input of the de-noising UNet is formed by concatenation with the actual reconstruction latent,

$$\tilde{z}_t = \mathrm{Concat}[z_t, z_\mathrm{rec}] \in \mathbb{R}^{h\times w\times 2c}. \tag{3}$$

The de-noising UNet ϵ_θ receives $\tilde{z}_t$, the diffusion timestep t, and a fixed text embedding c_phase of the contrast phase information, and predicts the noise residual $\hat{\epsilon} = \epsilon_\theta(\hat{z}_t, t, c_\mathrm{phase})$. Then the first supervision term of the de-noising UNet is the standard *Noise-prediction loss* $\mathcal{L}_\mathrm{n}$

$$\mathcal{L}_\mathrm{n} = \mathbb{E}_{t,\epsilon,X_\mathrm{gt}} \left\| \epsilon - \epsilon_\theta\big(\mathrm{Concat}[z_t, z_\mathrm{rec}], t, c_\mathrm{phase}\big) \right\|_2^2. \tag{4}$$

$\mathcal{L}_\mathrm{n}$ focuses on accurate latent de-noising; z_rec supplies anatomy information and provides prostructural cues during the de-noising process, thereby allowing the diffusion model to anchor the global CT shape and spatial layout while generating fine structural details.

To introduce anatomy-guided supervision, we augment the latent-space objective by incorporating two additional pixel space constraints. Specifically, after getting the predicted noise $\hat{\theta}$, we then reverse the diffusion process in Equation (2) to estimate the input latent via

$$\hat{z}_0 = \frac{1}{\sqrt{\bar{\alpha}_t}} \left(z_t - \sqrt{1-\bar{\alpha}_t}\hat{\epsilon} \right). \tag{5}$$

Then, $\hat{z}_0$ is decoded to the pixel space via the autoencoder. The overall fidelity in pixel space is supervised by an L1 regularization term named *Pixel-space reconstruction loss* $\mathcal{L}_\mathrm{p}$

$$\mathcal{L}_\mathrm{p} = \| \mathcal{D}_{\theta_\mathrm{D}}(\hat{z}_0) - X_\mathrm{gt} \|_1. \tag{6}$$

Most importantly, an anatomical completeness supervision is promoted with a segmentation guidance term from the anatomy segmentator $S(\cdot)$ mentioned in §3.2, named *Anatomy-guidance loss* $\mathcal{L}_\mathrm{s}$

$$\mathcal{L}_\mathrm{s} = \ell_\mathrm{seg}\big(S(\mathcal{D}_{\theta_\mathrm{D}}(\hat{z}_0)), S(X_\mathrm{gt})\big), \tag{7}$$

where ℓ_{seg} denotes cross entropy loss function.

Overall, the training objective of CARE is the weighted sum of *Noise-prediction loss* $\mathcal{L}_\mathrm{n}$ (Equation 4), *Pixel-space reconstruction loss* $\mathcal{L}_\mathrm{p}$ (Equation 6), and *Anatomy-guidance loss* $\mathcal{L}_\mathrm{s}$ (Equation 7),

$$\mathcal{L}_\mathrm{CARE} = \mathcal{L}_\mathrm{n} + \lambda_\mathrm{p}\mathcal{L}_\mathrm{p} + \lambda_\mathrm{s}\mathcal{L}_\mathrm{s}, \tag{8}$$

where $\lambda_\mathrm{p} = 1$ and $\lambda_\mathrm{s} = 0.001$ are the weights of $\mathcal{L}_\mathrm{p}$ and $\mathcal{L}_\mathrm{s}$ respectively. The de-noising UNet receives noisy latents of the ground-truth CT image and the concatenated latents of the reconstructed CT image. Meanwhile, adding image and anatomy supervision extends the single-term noise-prediction loss used in the latent diffusion model into a composite objective that jointly optimizes pixel-space reconstruction fidelity and anatomy-aware consistency.

During inference CARE starts from pure Gaussian noise z_T concatenated with a latent from an unseen reconstruction CT image, iteratively denoises to z_0, and decodes it through $\mathcal{D}_{\theta_\mathrm{D}}$ to deliver an anatomically complete CT image.

4 Experiments and Results

4.1 Pixel-Wise Metric Pitfalls

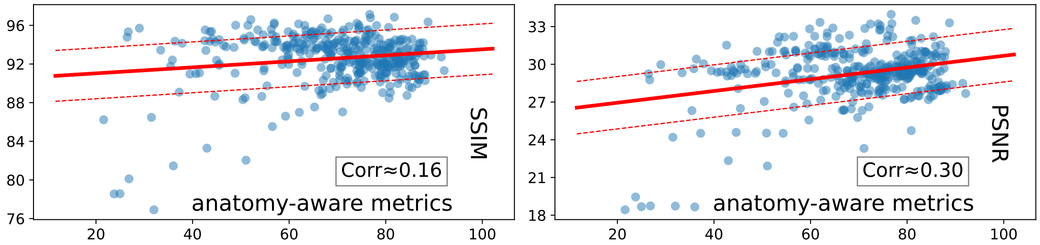

Figure 5: **Correlation Between Anatomy-Aware Metrics and Pixel-Wise Metrics.** The solid red line is a linear fit; dashed lines denote ±1σ of the residuals. Pearson correlation coefficients ($Corr_\mathrm{SSIM} \approx 0.16$, $Corr_\mathrm{PSNR} \approx 0.30$) are low, indicating that better anatomical fidelity does not necessarily yield higher SSIM or PSNR.

Figure 5 demonstrates that our anatomy-aware metrics capture reconstruction quality beyond what pixel-wise criteria can reveal. The horizontal axis shows the average anatomy-aware metrics, computed as the mean of the four groups of anatomy-aware metrics (large organs, small organs, vessel, and intestine). The vertical axes show SSIM (left) and PSNR (right). Reconstructions judged by the proposed anatomy-aware metrics to possess more faithful organ and vessel geometry show only weak association with SSIM and PSNR, confirming that superior anatomical fidelity does *not* guarantee higher pixel-similarity scores.

A further trend is the flattening of the regression lines: once a moderate anatomy-aware metric score is achieved, additional structural refinement yields progressively smaller gains in SSIM, and even less in PSNR. This saturation effect indicates that pixel-wise metrics quickly reach a ceiling, whereas the anatomy-aware metrics continue to differentiate models, underscoring the need for structure-centric assessment in clinical utility.

4.2 Benchmarking on Anatomy-Aware Metrics

We evaluated nine state-of-the-art CT reconstruction methods—including traditional methods (FDK, SART, ASD-POCS), neural rendering (InTomo, NeRF, TensoRF, SAX-NeRF), and Gaussian-splatting (R2-GS)—on our high-quality test set of 61 CT, which is by far the *largest CT reconstruction benchmark dataset*. Evaluation is conducted on both conventional pixel-wise metrics (SSIM, PSNR) and the proposed anatomy-aware metrics. Results are summarized in Table 1.

Table 1 shows that conventional pixel-wise scores can be overly optimistic: neural rendering methods such as InTomo, NeRF, and TensoRF report SSIM values above 0.83 and PSNR exceeding 24 dB, yet their anatomy-aware scores collapse—small-organ NSD approaches zero and vessel clDice rarely

Table 1: **Benchmarking CT Reconstruction Methods.** We evaluate preexisting CT reconstruction methods on high-quality CT scans using both pixel-wise metrics and our anatomy-aware metrics. We report the median and interquartile range (IQR) of these metrics. Cells are marked in blue, where a deeper color denotes a greater value. Anatomy-aware metrics reveal that poorly performing methods fail in anatomical preservation, especially in small organs, intestines, and vessels.

	pixel-wise metric		**anatomy-aware metric (ours)**			
method	SSIM	PSNR	NSD$^{\dagger}_{\text{large}}$	NSD$^{\ddagger}_{\text{small}}$	clDice$^{\triangle}_{\text{intestine}}$	clDice$^{*}_{\text{vessel}}$
InTomo [60]	82.7 (80.6,84.7)	24.3 (23.1,25.2)	2.5 (1.6,4.0)	0.0 (0.0,0.0)	7.3 (3.0,14.3)	18.0 (0.0,30.9)
NeRF [38]	86.8 (84.4,89.0)	25.7 (24.7,27.1)	22.0 (9.3,37.3)	0.2 (0.0,4.1)	17.0 (9.6,26.3)	34.5 (20.1,59.0)
TensoRF [11]	88.4 (87.0,90.7)	27.8 (26.9,28.9)	41.3 (21.2,55.8)	11.4 (0.0,28.7)	29.2 (20.4,37.7)	47.2 (34.7,58.4)
R^2-GS [61]	93.2 (91.9,93.9)	28.6 (27.0,29.5)	76.7 (55.8,86.8)	53.0 (41.8,68.8)	54.5 (44.9,68.0)	69.7 (57.6,78.5)
NAF [62]	92.3 (91.1,94.1)	29.4 (28.8,30.6)	84.3 (76.4,87.7)	60.9 (51.3,72.0)	71.3 (61.3,79.0)	76.1 (67.2,82.8)
FDK [18]	92.4 (91.1,94.2)	29.5 (28.8,30.8)	83.9 (78.6,87.6)	63.3 (52.7,70.1)	70.1 (60.1,80.8)	77.4 (69.4,82.3)
SART [1]	92.5 (91.1,94.1)	29.4 (28.8,30.6)	84.7 (76.8,88.6)	62.3 (53.2,75.3)	71.1 (59.5,79.9)	78.8 (70.4,84.4)
ASD-POCS [47]	92.5 (91.3,94.1)	29.5 (28.7,30.7)	84.9 (77.3,88.5)	61.9 (51.0,74.8)	72.0 (60.4,80.0)	75.8 (67.4,82.8)
SAX-NeRF [8]	93.7 (92.8,95.0)	29.7 (29.0,30.8)	84.8 (73.9,89.5)	65.8 (53.0,78.2)	71.5 (56.8,82.5)	78.8 (63.0,85.1)

†large organs include liver, kidney, pancreas, and spleen. The results are given in NSD (↑).

‡small organs include gallbladder, adrenal gland, celiac trunk, and duodenum. The results are given in NSD (↑).

$^{\triangle}$intestine refers to colon and small intestine. The results are given in clDice (↑).

*vessel include aorta, postcava, superior mesenteric artery, veins, and renal vein. The results are given in clDice (↑).

surpasses 35, revealing poor recovery of clinically critical structures. When evaluation pivots to anatomy fidelity, the methods ranking is effectively reshuffled: the traditional reconstruction baselines (FDK, SART, ASD-POCS) achieve the best or near-best performance across all four anatomical categories despite mediocre SSIM/PSNR, whereas high-SSIM models like R^2-GS or SAX-NeRF lag on organ and vessel integrity. The results further highlight a persistent difficulty in reconstructing fine anatomy: while large-organ NSD reaches the mid-80s for the strongest methods, small-organ surfaces remain challenging—even the top performer attains a median NSD below 70, underscoring the need for structure-aware objectives if CT reconstructions are to be clinically reliable. These findings demonstrate that pixel-wise metrics alone are insufficient and anatomy-aware evaluation provides a more discriminative and clinically meaningful assessment of reconstruction quality.

Reader Study. To validate the clinical relevance of our anatomy-aware metrics, we conducted a reader study with 21 board-certified radiologists who independently evaluated a subset of reconstructed scans from the nine benchmarked methods. Radiologists were presented with anonymized image pairs (reconstruction vs. ground truth) and asked to rate anatomical completeness across the four structure categories defined. All 21 radiologists agreed that higher pixel-wise metrics does not guarantee better anatomy preservation. The rankings derived from their assessments showed strong agreement with our anatomy-aware metrics. This confirms that our proposed metrics align closely with expert clinical judgment, whereas conventional pixel-wise scores do not reflect perceptual or diagnostic fidelity.

Discussion. Conventional pixel–wise criteria markedly over-estimate the usefulness of sparse-view reconstructions. For instance, InTomo records an SSIM of 0.83 yet recovers virtually none of the small-organ anatomy (median NSD = 0.0) in Table 1. Such disparity underscores the need for task-aligned evaluation: the proposed anatomy-aware metrics reveal structural failures that remain invisible to SSIM and PSNR.

4.3 Diffusion-Based CT Enhancement

To demonstrate the enhancement performance of CARE, we integrated it into each reconstruction method under a sparse-view reconstruction setting, i.e., CT reconstruction with 50 views of X-ray images. Table 2 reports the quantitative gains in both pixel-wise and anatomy-aware metrics. Figure 6 shows the qualitative results of reconstruction with and without CARE. CARE significantly improves reconstruction quality both in pixel-level fidelity and anatomical structures.

Under extreme sparsity in CT reconstruction, CARE-enhanced reconstructions exhibit remarkable anatomical completeness: for example, InTomo+CARE increases large-organ NSD from 2.4% to 30.9%, vessel clDice from 4.7% to 47.9%, and intestine clDice from 6.3% to 33.5%. NeRF+CARE and TensoRF+CARE show similar uplifts, confirming that our anatomy-guided loss robustly transfers structural priors even in low-data regimes.

Table 2: **CARE-enhanced Reconstruction Evaluation.** Evaluate preexisting CT reconstruction methods on high-quality CT scans with CARE using both pixel-wise metrics and our anatomy-aware metrics. Note that the results are based on 36 CT scans that CARE has never been trained on. We report the median and interquartile range (IQR) of these metrics and perform the Mann-Whitney U test for statistical analysis. Cells are marked in color only if CARE shows there is a significant difference ($p < 0.05$) with the original reconstruction, while green if the CARE enhancement results have improvement, and red otherwise. Deeper color represents greater difference.

		pixel-wise metric		**anatomy-aware metric (ours)**			
method		SSIM	PSNR	NSD$_{\text{large}}$	NSD$_{\text{small}}$	clDice$_{\text{intestine}}$	clDice$_{\text{vessel}}$
InTomo	50 views	82.7 (80.9,84.7)	24.5 (23.3,25.1)	2.4 (1.6,3.6)	0.0 (0.0,0.0)	6.3 (1.2,17.4)	4.7 (0.0,21.6)
	+CARE	76.7 (74.0,79.7)	22.0 (20.6,22.8)	30.9 (26.4,40.6)	7.1 (3.3,13.3)	33.5 (24.4,43.6)	47.9 (35.4,62.2)
NeRF	50 views	87.7 (84.5,89.4)	26.6 (24.7,27.4)	22.1 (8.2,35.8)	0.2 (0.0,4.3)	17.2 (9.0,28.0)	29.0 (19.8,42.1)
	+CARE	80.5 (78.5,83.0)	23.4 (21.9,24.4)	58.0 (49.1,63.0)	25.3 (14.2,30.1)	53.5 (47.4,59.6)	55.0 (45.3,65.6)
TensoRF	50 views	89.3 (87.3,91.1)	28.0 (27.3,29.7)	44.6 (21.8,53.5)	12.1 (1.1,28.4)	29.5 (19.2,36.4)	46.1 (36.8,58.5)
	+CARE	88.8 (87.2,90.1)	27.5 (27.2,28.2)	75.3 (65.7,79.3)	38.2 (30.2,46.4)	67.9 (62.8,75.8)	58.8 (51.0,68.4)
R^2-GS	50 views	93.3 (92.1,94.4)	28.8 (27.4,29.7)	76.3 (65.9,84.6)	59.1 (41.1,68.3)	53.9 (45.6,68.1)	64.5 (59.4,78.6)
	+CARE	89.6 (87.9,90.7)	27.3 (26.2,28.9)	82.5 (74.0,88.3)	55.7 (42.5,67.3)	75.4 (68.0,82.8)	67.6 (61.8,76.9)
NAF	50 views	92.5 (91.1,94.0)	29.4 (28.9,30.8)	83.7 (76.2,86.8)	61.0 (51.1,70.2)	71.6 (62.2,77.8)	75.4 (70.1,81.8)
	+CARE	92.8 (91.5,93.5)	29.6 (28.6,30.3)	87.9 (83.8,91.3)	68.1 (52.4,76.7)	81.8 (74.1,85.9)	71.8 (63.4,80.8)
FDK	50 views	92.5 (91.1,94.2)	29.5 (28.9,30.9)	83.8 (77.3,86.8)	62.4 (52.4,70.6)	68.2 (59.8,76.9)	76.5 (66.2,80.3)
	+CARE	92.4 (91.5,93.4)	29.7 (29.0,30.6)	87.1 (82.3,90.9)	67.0 (55.6,77.1)	81.5 (75.5,85.7)	71.7 (66.4,83.2)
SART	50 views	92.5 (91.1,94.2)	29.5 (28.9,31.0)	84.6 (76.9,88.1)	61.3 (53.7,75.4)	70.0 (61.0,77.6)	78.7 (65.6,83.6)
	+CARE	93.3 (92.7,94.3)	30.3 (29.5,30.9)	88.2 (84.2,90.8)	68.5 (56.0,75.5)	83.5 (77.8,87.5)	75.2 (62.2,83.3)
ASD-POCS	50 views	92.3 (91.6,94.2)	29.5 (28.9,30.7)	83.8 (76.5,87.2)	61.0 (51.4,73.6)	71.3 (61.8,77.4)	74.7 (67.4,81.1)
	+CARE	92.4 (91.8,93.5)	29.4 (28.4,30.1)	88.1 (82.6,90.9)	67.8 (56.2,75.4)	80.7 (74.9,84.9)	76.4 (66.6,79.9)
SAX-NeRF	50 views	93.7 (92.8,95.0)	29.6 (29.0,31.0)	82.4 (73.9,88.9)	66.5 (55.0,78.0)	68.0 (56.3,77.9)	79.3 (62.4,85.1)
	+CARE	92.5 (91.2,93.3)	29.4 (28.8,30.1)	88.7 (81.9,90.9)	67.1 (54.5,75.1)	81.5 (76.0,86.8)	72.0 (62.9,84.1)

These results highlight CARE's transformative impact: by embedding anatomical supervision directly into the diffusion framework, our method consistently elevates structural fidelity across diverse reconstruction backbones and sampling densities, without sacrificing general applicability or requiring additional scanning hardware.

Reader Study. We further evaluated the perceptual and clinical impact of CAREthrough a blinded reader study involving the same cohort of 21 radiologists. Participants assessed side-by-side reconstructions—with and without CARE—across diverse anatomical regions and reconstruction backbones. In all cases, radiologists preferred the CARE-enhanced version, citing improved visibility of small organs (e.g., gallbladder, adrenal glands), vascular continuity (e.g., celiac trunk, renal veins), and intestinal integrity. Notably, even when baseline reconstructions achieved high PSNR/SSIM, radiologists consistently identified missing or distorted anatomy that CAREsuccessfully restored. These findings underscore CARE's ability to translate quantitative gains in anatomy-aware metrics into tangible clinical improvements.

Discussion. The proposed CARE framework offers a model- and patient-agnostic avenue to inject anatomical priors into existing pipelines. Paired with CARE, diverse backbones—from classical FDK to NeRF and Gaussian-splatting—achieve substantial gains across all anatomy-aware metrics.

5 Discussion and Conclusion

Discussion. Our experiments in §4.2 and §4.3 show that CARE consistently improves the anatomical fidelity of sparse-view CT reconstructions across a wide range of baselines, including analytical, implicit, and generative methods. Further analysis in §E.3 demonstrates that CARE performs similar enhancement ability across arterial and portal venous phase CT scans, revealing CARE's potential in expanding to other CT phases. These improvements are particularly notable in small and complex structures like small organs and vessels, where segmentation-derived metrics such as NSD and clDice reveal gains of up to 40%, even when pixel-wise metrics remain nearly unchanged. Notably, CARE is trained once and applied directly to unseen cases without fine-tuning, demonstrating strong patient-agnostic generalization. This robustness arises from its design: segmentation-guided supervision

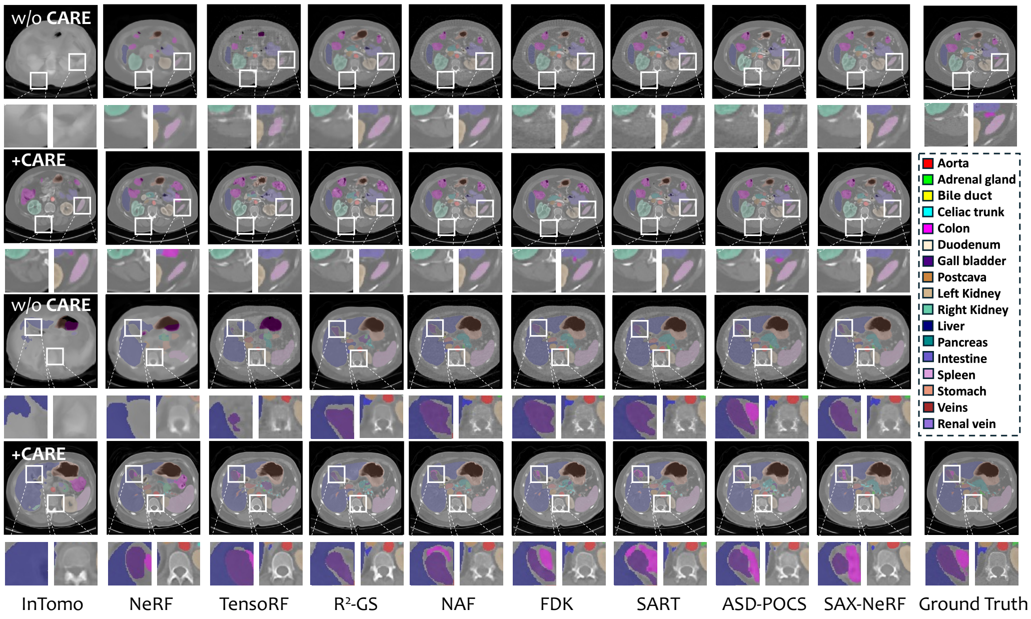

Figure 6: **Visual Comparison.** Visualizations of CT reconstruction with or without CARE on two example CT images. Please zoom in for a better view. These maps reveal that the most prominent changes occur at organ boundaries and within bone structures, where CAREmore effectively reduces artifacts and enhances contrast. See more detailed comparison in §E.2.

encodes population-level shape priors, and the diffusion model integrates them without requiring additional annotations or per-case adaptation. Its straightforward design enables seamless integration into existing reconstruction pipelines as a plug-in enhancement, without requiring modifications to or retraining of the underlying models. Moreover, CAREfeatures a modular architecture that facilitates the incorporation of future advances in segmentation models, thereby enhancing the reliability of anatomical guidance and enabling the evaluation of an expanding set of anatomical structures. Overall, CARE offers a flexible solution to enforce anatomical plausibility in CT reconstruction, particularly in settings where high-contrast structures are under-represented in pixel-domain loss functions.

Conclusion. This work reframes CT reconstruction quality through an anatomy-centric viewpoint and introduces CARE, a diffusion-based enhancement module that can be integrated into any reconstruction algorithm. A new suite of segmentation-driven metrics exposes structural deficiencies overlooked by SSIM and PSNR, and extensive experiments on nine baselines under sparse-view settings demonstrate that CARE consistently elevates anatomical fidelity, delivering up to 35% improvements in vessel clDice under 50-view settings—without additional acquisition burden. A limitation of our method is that its improvement becomes marginal when the baseline reconstruction already preserves most anatomical structures. In such high-quality cases, i.e., the reconstructions from SAX-NeRF, CARE offers limited added benefit relative to its computational cost. Our method brings algorithm goals in line with real clinical needs, making it easier to produce anatomy-aware CT reconstructions under sparse-view settings and sets a new standard for future reconstruction research.

Acknowledgments and Disclosure of Funding

This work was supported by the Lustgarten Foundation for Pancreatic Cancer Research, the Patrick J. McGovern Foundation Award, and the National Institutes of Health (NIH) under Award Number R01EB037669. We would like to thank the Johns Hopkins Research IT team in IT@JH for their support and infrastructure resources where some of these analyses were conducted; especially DISCOVERY HPC. We thank Jaimie Patterson for writing a news article about this project; thank Hamed Hooshangnejad, Heng Li, Wenxuan Li, and Guofeng Zhang for their helpful suggestions throughout the project. Paper content is covered by patents pending.

References

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

[2] Memoona Aziz, Umair Rehman, Muhammad Umair Danish, and Katarina Grolinger. Global-local image perceptual score (glips): Evaluating photorealistic quality of ai-generated images. *IEEE Transactions on Human-Machine Systems*, 2025.

[3] Pedro RAS Bassi, Wenxuan Li, Yucheng Tang, Fabian Isensee, Zifu Wang, Jieneng Chen, Yu-Cheng Chou, Yannick Kirchhoff, Maximilian Rokuss, Ziyan Huang, Jin Ye, Junjun He, Tassilo Wald, Constantin Ulrich, Michael Baumgartner, Saikat Roy, Klaus H. Maier-Hein, Paul Jaeger, Yiwen Ye, Yutong Xie, Jianpeng Zhang, Ziyang Chen, Yong Xia, Zhaohu Xing, Lei Zhu, Yousef Sadegheih, Afshin Bozorgpour, Pratibha Kumari, Reza Azad, Dorit Merhof, Pengcheng Shi, Ting Ma, Yuxin Du, Fan Bai, Tiejun Huang, Bo Zhao, Haonan Wang, Xiaomeng Li, Hanxue Gu, Haoyu Dong, Jichen Yang, Maciej A. Mazurowski, Saumya Gupta, Linshan Wu, Jiaxin Zhuang, Hao Chen, Holger Roth, Daguang Xu, Matthew B. Blaschko, Sergio Decherchi, Andrea Cavalli, Alan L. Yuille, and Zongwei Zhou. Touchstone benchmark: Are we on the right way for evaluating ai algorithms for medical segmentation? *Conference on Neural Information Processing Systems*, 2024.

[4] Pedro RAS Bassi, Xinze Zhou, Wenxuan Li, Szymon Płotka, Jieneng Chen, Qi Chen, Zheren Zhu, Jakub Prządo, Ibrahim E Hamacı, Sezgin Er, et al. Scaling artificial intelligence for multi-tumor early detection with more reports, fewer masks. *arXiv preprint arXiv:2510.14803*, 2025.

[5] Ander Biguri, Manjit Dosanjh, Steven Hancock, and Manuchehr Soleimani. Tigre: a matlab-gpu toolbox for cbct image reconstruction. *Biomedical Physics & Engineering Express*, 2(5):055010, 2016.

[6] Ander Biguri, Reuben Lindroos, Robert Bryll, Hossein Towsyfyan, Hans Deyhle, Richard Boardman, Mark Mavrogordato, Manjit Dosanjh, Steven Hancock, and Thomas Blumensath. Arbitrarily large iterative tomographic reconstruction on multiple gpus using the tigre toolbox. *arXiv preprint arXiv:1905.03748*, 2019.

[7] Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*, 2018.

[8] Yuanhao Cai, Jiahao Wang, Alan Yuille, Zongwei Zhou, and Angtian Wang. Structure-aware sparse-view x-ray 3d reconstruction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11174–11183, 2024.

[9] Kai Cao, Yingda Xia, Jiawen Yao, Xu Han, Lukas Lambert, Tingting Zhang, Wei Tang, Gang Jin, Hui Jiang, Xu Fang, et al. Large-scale pancreatic cancer detection via non-contrast ct and deep learning. *Nature Medicine*, pages 1–11, 2023.

[10] Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu, and Hao Su. Tensorf: Tensorial radiance fields. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision – ECCV 2022*, pages 333–350, Cham, 2022. Springer Nature Switzerland.

[11] Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu, and Hao Su. Tensorf: Tensorial radiance fields. In *European conference on computer vision*, pages 333–350. Springer, 2022.

[12] Guang-Hong Chen, Jie Tang, and Shuai Leng. Prior image constrained compressed sensing (piccs): a method to accurately reconstruct dynamic ct images from highly undersampled projection data sets. *Medical physics*, 35(2):660–663, 2008.

[13] Qi Chen, Xiaoxi Chen, Haorui Song, Zhiwei Xiong, Alan Yuille, Chen Wei, and Zongwei Zhou. Towards generalizable tumor synthesis. In *IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 11147–11158, 2024.

[14] Qi Chen, Xinze Zhou, Chen Liu, Hao Chen, Wenxuan Li, Zekun Jiang, Ziyan Huang, Yuxuan Zhao, Dexin Yu, Junjun He, et al. Scaling tumor segmentation: Best lessons from real and synthetic data. *arXiv preprint arXiv:2510.14831*, 2025.

[15] Li Sze Chow and Raveendran Paramesran. Review of medical image quality assessment. *Biomedical signal processing and control*, 27:145–154, 2016.

[16] Aanuoluwapo Clement David-Olawade, David B Olawade, Laura Vanderbloemen, Oluwayomi B Rotifa, Sandra Chinaza Fidelis, Eghosasere Egbon, Akwaowo Owoidighe Akpan, Sola Adeleke, Aruni Ghose, and Stergios Boussios. Ai-driven advances in low-dose imaging and enhancement—a review. *Diagnostics*, 15(6):689, 2025.

[17] Aanuoluwapo Clement David-Olawade, David B. Olawade, Laura Vanderbloemen, Oluwayomi B. Rotifa, Sandra Chinaza Fidelis, Eghosasere Egbon, Akwaowo Owoidighe Akpan, Sola Adeleke, Aruni Ghose, and Stergios Boussios. Ai-driven advances in low-dose imaging and enhancement—a review. *Diagnostics*, 15(6):689, 2025.

[18] Lee A Feldkamp, Lloyd C Davis, and James W Kress. Practical cone-beam algorithm. *Journal of the Optical Society of America A*, 1(6):612–619, 1984.

[19] EM Geraghty, JM Boone, JP McGahan, and K Jain. Normal organ volume assessment from abdominal ct. *Abdominal imaging*, 29(4):482–490, 2004.

[20] Eliot Gibson, Wenqi Li, Carole H. Sudre, Laurent Fidon, Dzhoshkun Shakir, Guotai Wang, Zach Eaton-Rosen, Richard Gray, Tom Doel, Paul Bentley, Ying Hu, Marisa Holden, Chang-Wai Chen, Dean C. Barratt, and Daniel Rueckert. Automatic multi-organ segmentation on abdominal ct with fully convolutional networks. *IEEE Transactions on Medical Imaging*, 37(9):1822–1834, 2018.

[21] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[22] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

[23] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

[24] Aljoscha Hore and Djemel Ziou. Image quality metrics: Psnr vs. ssim. In *2010 20th International Conference on Pattern Recognition*, pages 2366–2369. IEEE, 2010.

[25] Ziyan Huang, Haoyu Wang, Zhongying Deng, Jin Ye, Yanzhou Su, Hui Sun, Junjun He, Yun Gu, Lixu Gu, Shaoting Zhang, et al. Stu-net: Scalable and transferable medical image segmentation models empowered by large-scale supervised pre-training. *arXiv preprint arXiv:2304.06716*, 2023.

[26] Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2):203–211, 2021.

[27] Fabian Isensee, Tassilo Wald, Constantin Ulrich, Michael Baumgartner, Saikat Roy, Klaus Maier-Hein, and Paul F Jaeger. nnu-net revisited: A call for rigorous validation in 3d medical image segmentation. *arXiv preprint arXiv:2404.09556*, 2024.

[28] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2014.

[29] A Koch, T Gruber-Rouh, S Zangos, K Eichler, T Vogl, and L Basten. Radiation protection in ct-guided interventions: Does real-time dose visualisation lead to a reduction in radiation dose to participating radiologists? a single-centre evaluation. *Clinical Radiology*, 79(6):e785–e790, 2024.

[30] Megan Lantz, Emil Y. Sidky, Ingrid S. Reiser, Xiaochuan Pan, and Gregory Ongie. Enhancing signal detectability in learning-based ct reconstruction with a model-observer inspired loss function. *arXiv preprint arXiv:2402.10010*, 2024.

[31] Wenxuan Li, Pedro RAS Bassi, Tianyu Lin, Yu-Cheng Chou, Xinze Zhou, Yucheng Tang, Fabian Isensee, Kang Wang, Qi Chen, Xiaowei Xu, et al. Scalemai: Accelerating the development of trusted datasets and ai models. *arXiv preprint arXiv:2501.03410*, 2025.

[32] Wenxuan Li, Chongyu Qu, Xiaoxi Chen, Pedro RAS Bassi, Yijia Shi, Yuxiang Lai, Qian Yu, Huimin Xue, Yixiong Chen, Xiaorui Lin, et al. Abdomenatlas: A large-scale, detailed-annotated, & multi-center dataset for efficient transfer learning and open algorithmic benchmarking. *Medical Image Analysis*, page 103285, 2024.

[33] Wenxuan Li, Alan Yuille, and Zongwei Zhou. How well do supervised models transfer to 3d image segmentation? In *International Conference on Learning Representations*, 2024.

[34] Xinran Li, Yi Shuai, Chen Liu, Qi Chen, Qilong Wu, Pengfei Guo, Dong Yang, Can Zhao, Pedro RAS Bassi, Daguang Xu, et al. Text-driven tumor synthesis. *arXiv preprint arXiv:2412.18589*, 2024.

[35] Tianyu Lin, Zhiguang Chen, Zhonghao Yan, Weijiang Yu, and Fudan Zheng. Stable diffusion segmentation for biomedical images with single-step reverse process. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, pages 656–666, Cham, 2024. Springer Nature Switzerland.

[36] Jie Liu, Yixiao Zhang, Jie-Neng Chen, Junfei Xiao, Yongyi Lu, Bennett A Landman, Yixuan Yuan, Alan Yuille, Yucheng Tang, and Zongwei Zhou. Clip-driven universal model for organ segmentation and tumor detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21152–21164, 2023.

[37] Jie Liu, Yixiao Zhang, Kang Wang, Mehmet Can Yavuz, Xiaoxi Chen, Yixuan Yuan, Haoliang Li, Yang Yang, Alan Yuille, Yucheng Tang, et al. Universal and extensible language-vision models for organ segmentation and tumor detection from abdominal computed tomography. *Medical Image Analysis*, page 103226, 2024.

[38] 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. *Communications of the ACM*, 65(1):99–106, 2021.

[39] Samaneh Mostafapour, Marcel Greuter, Johannes H. van Snick, Adrienne H. Brouwers, Rudi A. J. O. Dierckx, Joyce van Sluis, and Adriaan A. Lammertsma. Ultra-low dose ct scanning for pet/ct. *Medical Physics*, 51(1):139–155, 2024.

[40] Chongyu Qu, Tiezheng Zhang, Hualin Qiao, Jie Liu, Yucheng Tang, Alan Yuille, and Zongwei Zhou. Abdomenatlas-8k: Annotating 8,000 abdominal ct volumes for multi-organ segmentation in three weeks. In *Conference on Neural Information Processing Systems*, volume 21, 2023.

[41] G. N. Ramachandran and A. V. Lakshminarayanan. Three-dimensional reconstruction from radiographs and electron micrographs: application of convolutions instead of fourier transforms. *Proceedings of the National Academy of Sciences*, 68(9):2236–2240, 1971.

[42] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021.

[43] Saikat Roy, Gregor Koehler, Constantin Ulrich, Michael Baumgartner, Jens Petersen, Fabian Isensee, Paul F Jaeger, and Klaus H Maier-Hein. Mednext: transformer-driven scaling of convnets for medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 405–415. Springer, 2023.

[44] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE transactions on pattern analysis and machine intelligence*, 45(4):4713–4726, 2022.

[45] Fritz Schick. Automatic segmentation and volumetric assessment of internal organs and fatty tissue: what are the benefits? *Magnetic Resonance Materials in Physics, Biology and Medicine*, 35(2):187–192, 2022.

[46] Suprosanna Shit, Johannes C Paetzold, Anjany Sekuboyina, Ivan Ezhov, Alexander Unger, Andrey Zhylka, Josien PW Pluim, Ulrich Bauer, and Bjoern H Menze. cldice-a novel topology-preserving loss function for tubular structure segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16560–16569, 2021.

[47] 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.

[48] Yiran Sun, Hana Baroudi, Tucker Netherton, Laurence Court, Osama Mawlawi, Ashok Veeraraghavan, and Guha Balakrishnan. Difr3ct: Latent diffusion for probabilistic 3d ct reconstruction from few planar x-rays. *arXiv preprint arXiv:2408.15118*, 2024.

[49] Yipeng Sun, Yixing Huang, Zeyu Yang, Linda-Sophie Schneider, Mareike Thies, Mingxuan Gu, Siyuan Mei, Siming Bayer, Frank G Zöllner, and Andreas Maier. Eagle: an edge-aware gradient localization enhanced loss for ct image reconstruction. *Journal of Medical Imaging*, 12(1):014001–014001, 2025.

[50] Pinhuang Tan, Mengxiao Geng, Jingya Lu, Liu Shi, Bin Huang, and Qiegen Liu. Msdiff: Multi-scale diffusion model for ultra-sparse view ct reconstruction. *arXiv preprint arXiv:2405.05814*, 2024.

[51] Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, Dhruv Nair, Sayak Paul, William Berman, Yiyi Xu, Steven Liu, and Thomas Wolf. Diffusers: State-of-the-art diffusion models. `https://github.com/huggingface/diffusers`, 2022.

[52] Yihe Wang, Nan Huang, Taida Li, Yujun Yan, and Xiang Zhang. Medformer: A multi-granularity patching transformer for medical time-series classification. *Advances in Neural Information Processing Systems*, 37:36314–36341, 2024.

[53] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[54] Jakob Wasserthal, Manfred Meyer, Hanns-Christian Breit, Joshy Cyriac, Shan Yang, and Martin Segeroth. Totalsegmentator: robust segmentation of 104 anatomical structures in ct images. *arXiv preprint arXiv:2208.05868*, 2022.

[55] Yingda Xia, Qihang Yu, Linda Chu, Satomi Kawamoto, Seyoun Park, Fengze Liu, Jieneng Chen, Zhuotun Zhu, Bowen Li, Zongwei Zhou, et al. The felix project: Deep networks to detect pancreatic neoplasms. *medRxiv*, 2022.

[56] Yao Xu, Jiazhou Wang, and Weigang Hu. Prior-image-based low-dose ct reconstruction for adaptive radiation therapy. *Physics in Medicine & Biology*, 69(21), 2024.

[57] Wufeng Xue, Lei Zhang, Xuanqin Mou, and Alan C Bovik. Gradient magnitude similarity deviation: A highly efficient perceptual image quality index. *IEEE transactions on image processing*, 23(2):684–695, 2013.

[58] Demin Yang, Haochen Shi, Bolun Zeng, and Xiaojun Chen. 2d/3d registration based on biplanar x-ray and ct images for surgical navigation. *Computer Methods and Programs in Biomedicine*, 257:108444, 2024.

[59] Liutao Yang, Jiahao Huang, Guang Yang, and Daoqiang Zhang. Ct-sdm: A sampling diffusion model for sparse-view ct reconstruction across all sampling rates. *arXiv preprint arXiv:2409.01571*, 2024.

[60] Guangming Zang, Ramzi Idoughi, Rui Li, Peter Wonka, and Wolfgang Heidrich. Intratomo: self-supervised learning-based tomography via sinogram synthesis and prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1960–1970, 2021.

[61] Ruyi Zha, Tao Jun Lin, Yuanhao Cai, Jiwen Cao, Yanhao Zhang, and Hongdong Li. R^2-gaussian: Rectifying radiative gaussian splatting for tomographic reconstruction. *arXiv preprint arXiv:2405.20693*, 2024.

[62] 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*, pages 442–452. Springer, 2022.

[63] Guofeng Zhang, Ruyi Zha, Hao He, Yixun Liang, Alan Yuille, Hongdong Li, and Yuanhao Cai. X-lrm: X-ray large reconstruction model for extremely sparse-view computed tomography recovery in one second. *arXiv preprint arXiv:2503.06382*, 2025.

[64] Hao Zhang, Grace J Gang, Hao Dang, and J Webster Stayman. Regularization analysis and design for prior-image-based x-ray ct reconstruction. *IEEE transactions on medical imaging*, 37(12):2675–2686, 2018.

[65] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.

Appendix

Table of Contents

A Datasets, Models and Implementation Details

A.1 Datasets

We evaluate four different kinds of anatomical structures: large organs, small organs, intestine and vessel. Large organs and small organs are divided according to clinical guidelines, showing in Table 3.

Table 3: **Grouping strategy for large and small organs.** We partition organs by typical in vivo volumes reported in radiologic volumetry studies. The 100 mL and 10 mm cutoff aligns with standard thresholds for distinguishing major organs from accessory structures in abdominal atlases.

Group	**Representative mean volume (mL)**	**References**
Large organs (> ∼100 mL)	Liver ≈1200–1700; Kidney ≈150–200; Spleen ≈215; Pancreas ≈71–83	[19]
Small organs / structures (< ∼50 mL or < 10 mm diameter)	Gallbladder ≈30–50 mL; Each adrenal gland ≈4–6 mL; Celiac trunk diameter ≈6–12 mm; Duodenum lumen 2–3 cm	[45]

A.1.1 Large-scale Reconstruction Benchmark Dataset

Table 4: **Evaluation dataset size of the CT reconstruction methods.** Among all the CT reconstruction methods we evaluate, NeRF [38] and TensoRF [10] uses 20 CT scans but no human body CT included, SAX-NeRF [8] use 15 CT scans with only 9 human body CT scans. Our reconstruction benchmark dataset is by far the *largest* in the field of CT reconstruction.

Method	**Total CT Scan**	**Human Body CT Scan**
InTomo [60]	7	4
NeRF [38]	20	0
TensoRF [10]	20	0
R2-GS [61]	15	5
NAF [62]	5	4
FDK [18]	1	0
SART [1]	1	1
ASD-POCS [47]	2	0
SAX-NeRF [8]	15	9
CARE	**61**	**61**

As shown in Table 4, our reconstruction benchmark dataset contains 61 high-quality CT scans, with 36 in arterial phase and 25 in portal venous phase, which is by far the largest CT reconstruction dataset. The real projections of these CT scans are not available. Thus, we use TIGRE [5, 6] package to generate synthetic projections following previous works [8, 61]. The detailed meta information about our reconstruction evaluation dataset is listed in Table 5.

Table 5: **Statistics of the CT reconstruction benchmark dataset.**

Dataset	**CT scan***	**Age**†	**Female***	**Male***	**In-plane spacing**‡	**White***	**Black***	**Asian***	**Other***	**Resolution**$^{\triangle}$	**Height**$^{\triangle}$	**Voxel spacing**‡
Training	25	64.2 ± 8.9	10 (40.0)	15 (60.0)	0.70 (0.68, 0.79)	6 (24.0)	4 (16.0)	5 (20.0)	6 (24.0)	512 × 512	350 ± 302	0.74 ± 0.12
Test	36	64.8 ± 9.2	16 (44.4)	20 (55.6)	0.77 (0.70, 0.98)	4 (11.1)	7 (19.4)	10 (27.8)	5 (13.9)	512 × 512	427 ± 294	1.20 ± 1.04

† Age in years (mean ± SD).

‡ Spacing in millimeters. Applies to *In-plane spacing* and *Voxel spacing*. Values reported as median (IQR) or mean ± SD as shown.

$^{\triangle}$ Image resolution and height in pixels. Resolution reported as width × height; height reported as mean ± SD.

* Counts; some columns also include percentages in parentheses, e.g., *Female*, *Male*, and race categories. Applies to *CT scan*, *Female*, *Male*, *White*, *Black*, *Asian*, *Other*.

For the anatomy-aware CT reconstruction metrics (§3.2), we report the results of all 61 CT scans. For the anatomy-aware CT reconstruction framework CARE (§3.3), we report the results of a subset of 36 CT scans (23 in arterial phase and 13 in portal venous phase), and the remaining 25 CT scans (13 in arterial phase and 12 in portal venous phase) were used for the training set. The detailed meta information about our reconstruction evaluation dataset is listed in Table 6.

A.1.2 JHH Dataset for CARE Pretraining

Table 6: **Statistics of the JHH dataset.** All the CT scans in the dataset have a resolution of 512 × 512. This dataset is collected using four major vendors: GE (39%), Siemens (38%), Phillips (12%), and Toshiba (11%).

Dataset	CT scan*	Age†	Female*	Male*	In-plane spacing‡	White*	Black*	Asian*	Other*	Height$^\triangle$	Voxel spacing‡	PDAC*	PNET*	Cyst*
Training	3151	64.9 ± 8.9	1524 (48.4)	1627 (51.6)	0.71 (0.66, 0.77)	623 (19.8)	619 (19.6)	632 (20.1)	661 (21.0)	678 ± 201	0.72 ± 0.09	1119	591	429
Test	1958	64.9 ± 8.9	1008 (51.5)	950 (48.5)	0.73 (0.68, 0.79)	409 (20.9)	394 (20.1)	389 (19.9)	397 (20.3)	690 ± 196	0.74 ± 0.09	554	327	512

† Age in years (mean ± SD).
‡ Spacing in millimeters. Applies to *In-plane spacing* and *Voxel spacing*. Values reported as median (IQR) or mean ± SD as shown.
$^\triangle$ Height in pixels (mean ± SD).
* Counts; some columns also include percentages in parentheses, e.g., *Female*, *Male*, and race categories. Applies to *CT scan*, *Female*, *Male*, *White*, *Black*, *Asian*, *Other*, and pancreatic tumor types: *PDAC*, *PNET*, *Cyst*.

The JHH dataset contains 6,212 high-quality artifact-free CT scans of 2,870 patients, where 3,107 are in the arterial phase and 3,105 are in the venous phase. The CT scans in this dataset have a spacing of 0.5 mm, providing 4 million CT slices in total. This dataset can be split into a training set of 3,151 CT scans and a testing set of 1,958 CT scans. For all 6,212 CT scans of the dataset, a total of 25 anatomical structures are annotated by a group of 21 radiologists for over three years, including the 17 structures we use as anatomy-aware CT reconstruction metrics. The 25 annotated anatomical structures are: aorta, left adrenal gland, right adrenal gland, common bile duct, celiac artery, colon, duodenum, gall bladder, postcava, left kidney, right kidney, liver, pancreas, pancreatic duct, superior mesenteric artery, intestine, spleen, stomach, veins, left renal vein, right renal vein, common bile duct, pancreatic PDAC, pancreatic cyst, and pancreatic PNET.

This dataset is used both in training the anatomy segmentator, autoencoder, and the latent diffusion model to build the anatomy-aware CT reconstruction metrics and is prepared for the CARE framework. Note that the anatomy segmentator is trained on the training set, whereas the autoencoder and latent diffusion model are trained on the entire 6,212 CT scans.

A.2 Models

We provide full access to our model checkpoints to contribute to the open-source community. We release the model checkpoints of our anatomy segmentator, autoencoder, and the CARE models in our codebase.

A.3 Implementation Details

A.3.1 Anatomy Segmentator

The proposed anatomy segmentator used in both the anatomy-aware CT reconstruction metrics and the CARE framework is implemented as a nnU-Net [26]. We extend the training plan of nnU-Net to be trained on a 48GB NVIDIA RTX 6000 GPU, with all data preprocessing, training, and inference settings set as default. This model is trained on the training set of our JHH dataset, with 3,151 expertly annotated CT scans.

A.3.2 Reconstruction Baselines

The experiments are conducted on three traditional reconstruction methods (FDK [18], SART [1], ASD-POCS [47]), five NeRF-based reconstruction methods (InTomo [60], NeRF [38], TensoRF [10], NAF [62], SAX-NeRF [8]), and a Gaussian-Spalting-based method R^2-GS [61]. We implement these methods using the CT reconstruction toolbox provided by SAX-NeRF[4] [8] and R^2-GS[5] [61]. For each algorithm, the training setting is set as default. All of these reconstruction experiments are run on an eight NVIDIA RTX 6000 GPU server, each with 48 GB of memory.

[4] https://github.com/caiyuanhao1998/SAX-NeRF
[5] https://github.com/Ruyi-Zha/r2_gaussian

A.3.3 Autoencoder Training

The autoencoder model is initialized with the checkpoints provided by Stable Diffusion v1.5 [42]. We set the weight of reconstruction loss and perceptual loss (detailed in §D.1) to be $\lambda_{rec} = \lambda_{per} = 1$. The weight of the KL regularization term is set to $\beta = 1 \times 10^{-6}$. The autoencoder model is trained on the JHH CT dataset (§A.1.2) for 150,000 iterations. We use AdamW optimizer during training.

A.3.4 Latent Diffusion Training

When training the latent diffusion model, we adopted the super resolution training scheme for the model to gain intrinsic enhancement ability. To create low-quality CT images, we followed SR3 [44] to use bicubic interpolation with anti-aliasing enabled to downsample the CT image and upsample it back to the original resolution. The downsampling factor is set to 4.

During training, the text prompt regarding different phases is fixed: the prompt of an arterial phase CT image is “An Arterial CT slice.”, and the prompt of a portal phase CT image is “A Portal-venous CT slice.”. The de-noising UNet is trained on the same JHH dataset as used in the autoencoder training, with AdamW optimizer and 50,000 training iterations.

A.3.5 CARE Training

The text prompt of CARE is identical to the latent diffusion training stage. The model is finetuned on 25 CT scans (as mentioned in §A.1.1) for 50,000 iterations with the AdamW optimizer for each given CT reconstruction method. We set the weights of the losses to $\lambda_{\mathrm{p}} = 1$ and $\lambda_{\mathrm{s}} = 0.001$.

Our diffusion model is implemented by the diffusers [51] package with a backbone of Stable Diffusion v1.5 [42]. All three training stages of CARE (§A.3.3, §A.3.4 and §A.3.5) are done on an eight RTX 8000 GPUs server, each with 48 GB of memory.

B Anatomy Segmentator

As mentioned in §A.3.3, the anatomy segmentator is trained on the training set of our JHH dataset with 3,151 CT scans. In §3.2 of the paper, we showed the high correlation (in Figure 3) of the anatomy-aware CT reconstruction metrics with the segmentator's pseudolabel and the ground truth label, to support our assertion that employing the anatomy segmentator for these metrics is justifiable.

Here, we also provide the segmentation results of our anatomy segmentator on the testing set of our JHH dataset with 1,958 CT scans. Table 7 shows that our anatomy segmentator achieves excellent performance over the large-scale testing set comparing to three state-of-the-art segmentation methods in the Touchstone 1.0 [3] benchmark: MedNeXt [43], MedFormer [52] and STU-Net-B [25].

Table 7: **Anatomy Segmentator's Performance on the Testing Set of JHH Dataset.** Segmentation performance is reported as the median and interquartile range (IQR) of Dice and IoU (%) for selected anatomical structures across the testing set with 1,958 CT scans. For each structure, the best result is **bolded** and the sencond best underlined.

Anatomical Structures	Segmentator (Ours)	MedNeXt [43]	MedFormer [52]	STU-Net-B [25]
Large Organs				
Liver	96.9 (96.4,97.3)	96.5 (95.8,96.9)	96.7 (96.2,97.1)	**97.2** (96.8,97.5)
Kidney Left	**97.7** (97.4,98.0)	97.2 (96.8,97.6)	97.1 (96.7,97.5)	96.9 (96.4,97.3)
Kidney Right	**97.7** (97.5,98.0)	97.3 (97.0,97.7)	97.0 (96.5,97.4)	96.8 (96.3,97.1)
Spleen	96.4 (95.5,97.1)	96.5 (95.7,97.2)	**96.8** (96.0,97.3)	96.0 (95.4,96.9)
Pancreas	86.5 (81.5,89.3)	**88.1** (85.3,90.5)	87.2 (84.0,89.6)	85.9 (82.1,88.3)
Small Organs				
Gall Bladder	90.9 (86.1,93.4)	90.5 (85.7,92.8)	91.3 (87.4,93.5)	**91.7** (88.2,93.8)
Adrenal Gland Left	84.6 (75.2,88.7)	**86.1** (77.3,89.4)	85.4 (76.9,88.2)	83.2 (75.0,87.0)
Adrenal Gland Right	80.1 (71.7,83.5)	**81.4** (73.9,86.5)	81.2 (72.8,85.4)	79.5 (70.7,83.2)
Celiac Artery	65.6 (56.5,73.4)	66.8 (57.2,74.3)	**67.1** (59.5,75.2)	65.9 (57.0,73.1)
Duodenum	**86.4** (82.4,89.5)	85.3 (80.5,88.1)	84.8 (80.2,87.4)	85.1 (81.2,88.3)
Intestinal Structures				
Colon	84.4 (82.6,91.1)	85.1 (83.0,91.3)	**85.2** (83.7,92.1)	84.2 (82.1,90.5)
Intestine	75.6 (69.1,81.4)	76.8 (70.4,82.3)	**76.9** (71.3,83.0)	75.0 (68.9,80.7)
Vascular Structures[†]				
Aorta	**92.1** (87.4,95.0)	91.3 (86.8,94.2)	90.9 (86.1,93.9)	91.5 (87.0,94.5)
Postcava	85.9 (79.3,87.4)	**86.0** (80.4,88.2)	85.2 (78.9,86.9)	84.7 (78.1,86.3)
Superior Mesenteric Artery	66.5 (58.2,74.5)	67.1 (59.1,75.2)	**67.3** (60.3,76.5)	66.8 (58.4,74.8)

[†] Since there is no ground truth label for renal veins (left or right) in the test set, we don't report the corresponding metrics here.

C Details and Disscussions of Anatomy-Aware Metrics

C.1 NSD vs. DSC: an Example of Sensitivity to Boundary Shifts

Consider an organ with volume V and surface area S. When the entire boundary shifts uniformly by a distance δ (e.g., organ boundary displaced by 3 mm), the volume discrepancy can be approximated as:

$$\Delta V \approx S \cdot \delta \tag{9}$$

The corresponding drop in Dice Similarity Coefficient (DSC) can be approximated by:

$$|\Delta\text{DSC}| \approx \frac{\Delta V}{2V} = \frac{S \cdot \delta}{2V} = \frac{\delta}{2R_{\text{eff}}}, \quad \text{where} \quad R_{\text{eff}} \equiv \frac{V}{S} \tag{10}$$

For a spherical organ with radius $R = 50$mm and $\delta = 3$mm, we compute the effective radius R_{eff} as:

$$R_{\text{eff}} = \frac{\frac{4}{3}\pi R^3}{4\pi R^2} = \frac{R}{3} \approx 16.67\text{mm} \tag{11}$$

This results in:

$$|\Delta\text{DSC}| \approx \frac{\delta}{2R_{\text{eff}}} = \frac{3}{2 \times 16.67} \approx 0.09 \tag{12}$$

This demonstrates that even with clinically significant displacement, DSC may only decrease a little, not reflecting the severe anatomical misalignment. In contrast, Normalized Surface Dice (NSD) is independent of organ volume and instead penalizes boundary mismatches directly. For a mismatch tolerance $\tau = 2$mm, a uniform 3 mm shift would lead NSD to fall to zero, reflecting the anatomical error more faithfully.

Overall, DSC is volume-biased and may underestimate boundary errors in large structures, while NSD maintains consistent sensitivity across anatomical scales.

C.2 High Correlation of GT-based and Segmentator-based Anatomy-Aware Metrics

Figure 3 summarizes the high correlation between GT-based and segmentator-based anatomy-aware metrics over all nine CT reconstruction methods, which indicates that the proposed anatomy segmentator is a reasonable substitute for ground truth labels to build the anatomy-aware metrics.

Here, we also provide the detailed correlation scatter plots of all nine CT reconstruction methods on GT-based and segmentator-based anatomy-aware metrics. These correlation plots consistently give the same conclusion. Note that the correlation coefficient of small organs in the InTomo method is 0.0 (in Figure 7) because all the segmentation metrics are zeros. Moreover, the inadequate correlation coefficient of the intestine highlights an existing issue with the annotation of intestinal structures: the clinical experts concentrate solely on the volumetric precision of the annotation, overlooking the tubular form and connectivity.

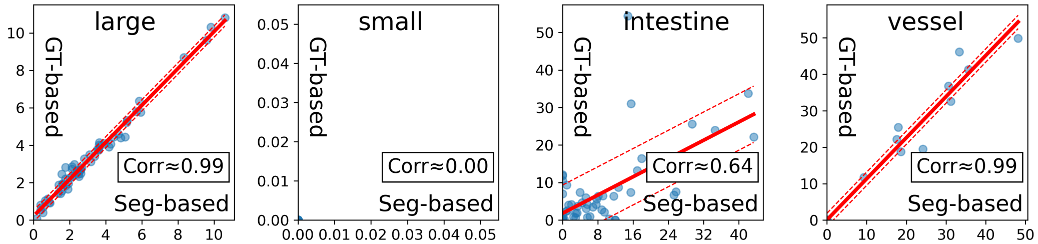

Figure 7: **Correlation Between GT-based and Segmentator-based Anatomy-Aware Metrics on InTomo [60].** The solid red line is a linear fit and the dashed lines denote ±1σ of the residuals. High correlation shows that the proposed anatomy segmentator is a strong substitute for ground truth labels when building anatomy-aware metrics.

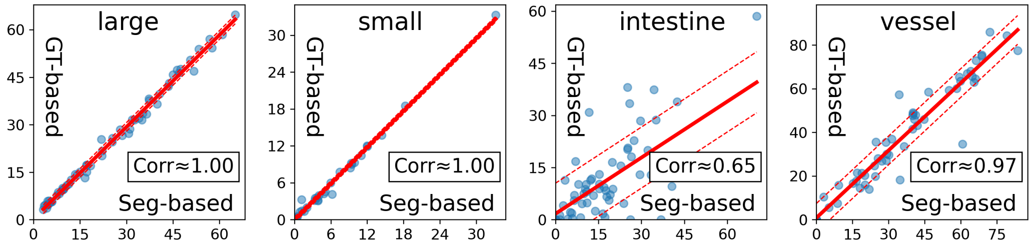

Figure 8: **Correlation Between GT-based and Segmentator-based Anatomy-Aware Metrics on NeRF [38].** The solid red line is a linear fit and the dashed lines denote ±1σ of the residuals. High correlation shows that the proposed anatomy segmentator is a strong substitute for ground truth labels when building anatomy-aware metrics.

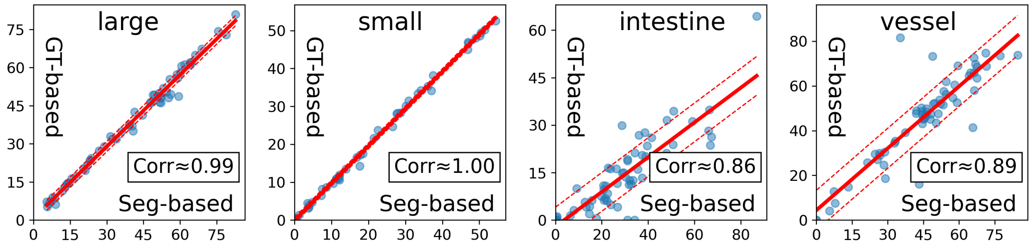

Figure 9: **Correlation Between GT-based and Segmentator-based Anatomy-Aware Metrics on TensoRF [11].** The solid red line is a linear fit and the dashed lines denote ±1σ of the residuals. High correlation shows that the proposed anatomy segmentator is a strong substitute for ground truth labels when building anatomy-aware metrics.

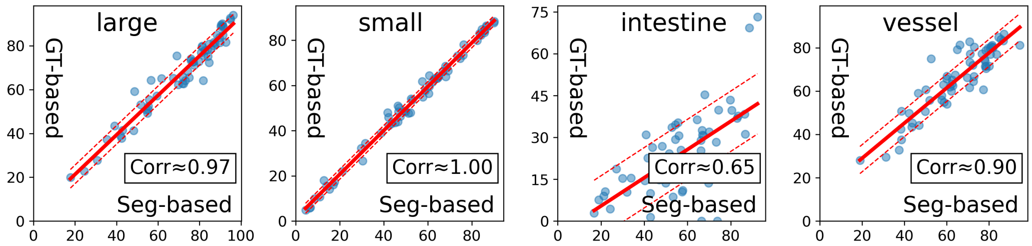

Figure 10: **Correlation Between GT-based and Segmentator-based Anatomy-Aware Metrics on $\mathbf{R^2}$-GS [61].** The solid red line is a linear fit and the dashed lines denote $\pm 1\sigma$ of the residuals. High correlation shows that the proposed anatomy segmentator is a strong substitute for ground truth labels when building anatomy-aware metrics.

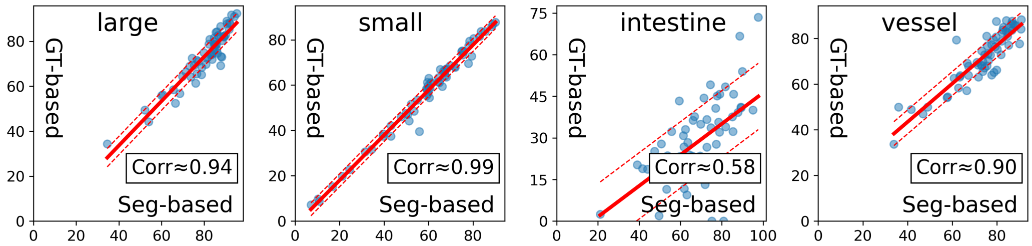

Figure 11: **Correlation Between GT-based and Segmentator-based Anatomy-Aware Metrics on NAF [62].** The solid red line is a linear fit and the dashed lines denote $\pm 1\sigma$ of the residuals. High correlation shows that the proposed anatomy segmentator is a strong substitute for ground truth labels when building anatomy-aware metrics.

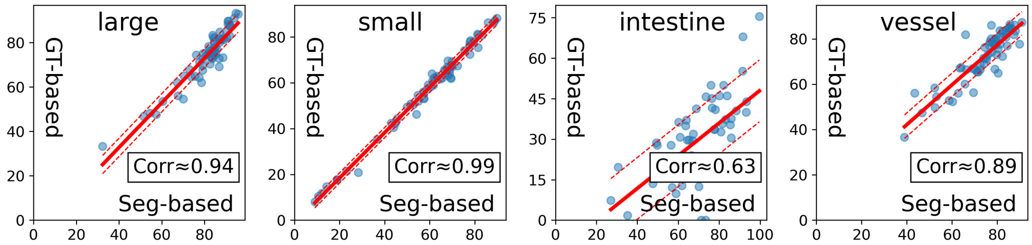

Figure 12: **Correlation Between GT-based and Segmentator-based Anatomy-Aware Metrics on FDK [18].** The solid red line is a linear fit and the dashed lines denote $\pm 1\sigma$ of the residuals. High correlation shows that the proposed anatomy segmentator is a strong substitute for ground truth labels when building anatomy-aware metrics.

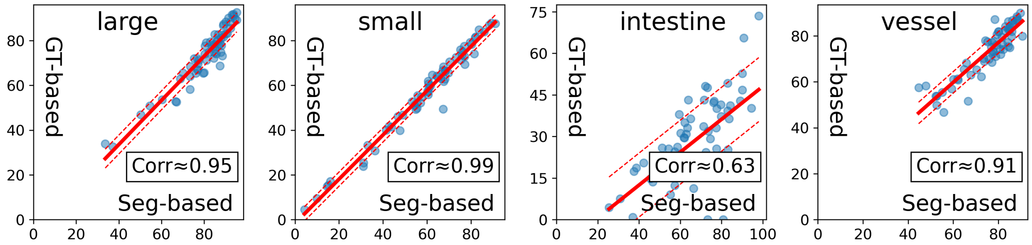

Figure 13: **Correlation Between GT-based and Segmentator-based Anatomy-Aware Metrics on SART [1].** The solid red line is a linear fit and the dashed lines denote $\pm 1\sigma$ of the residuals. High correlation shows that the proposed anatomy segmentator is a strong substitute for ground truth labels when building anatomy-aware metrics.

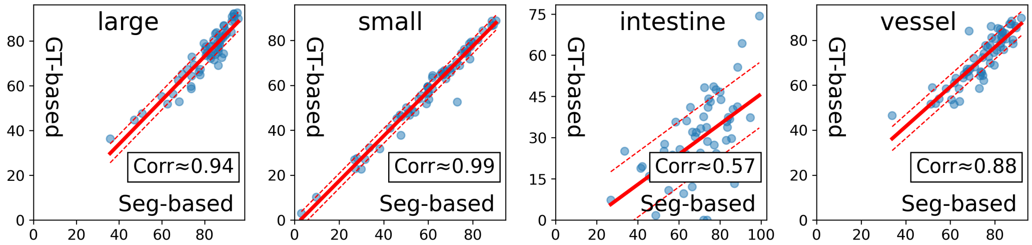

Figure 14: **Correlation Between GT-based and Segmentator-based Anatomy-Aware Metrics on ASD-POCS [47].** The solid red line is a linear fit and the dashed lines denote $\pm 1\sigma$ of the residuals. High correlation shows that the proposed anatomy segmentator is a strong substitute for ground truth labels when building anatomy-aware metrics.

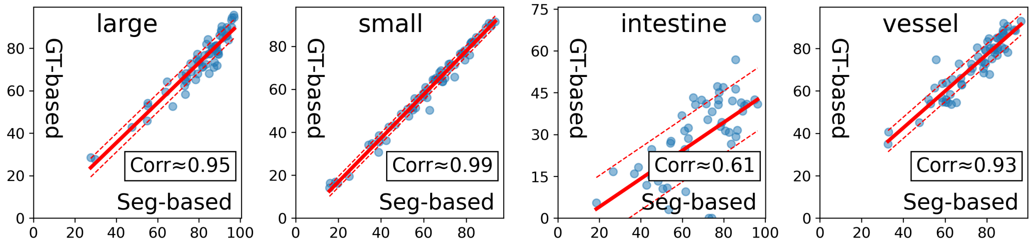

Figure 15: **Correlation Between GT-based and Segmentator-based Anatomy-Aware Metrics on SAX-NeRF [8].** The solid red line is a linear fit and the dashed lines denote $\pm 1\sigma$ of the residuals. High correlation shows that the proposed anatomy segmentator is a strong substitute for ground truth labels when building anatomy-aware metrics.

C.3 Poor Correlation between Pixel-Wise Metrics and Anatomical Preservation Performance

Figure 5 summarizes the poor correlation between pixel-wise metrics and anatomy-aware metrics over six CT reconstruction methods: FDK [18], SART [1], ASD-POCS [47], NAF [62], SAX-NeRF [8], R^2-GS [61], where NeRF [38], TensoRF [10], and InTomo [60] are eliminated due to large number of zero metric samples. This indicates that better anatomical preservation does not ensure higher pixel-wise metrics.

Here, we also provide the detailed correlation scatter plots of all nine CT reconstruction methods on all pixel-wise and anatomy-aware metrics. These correlation plots consistently give the same conclusion.

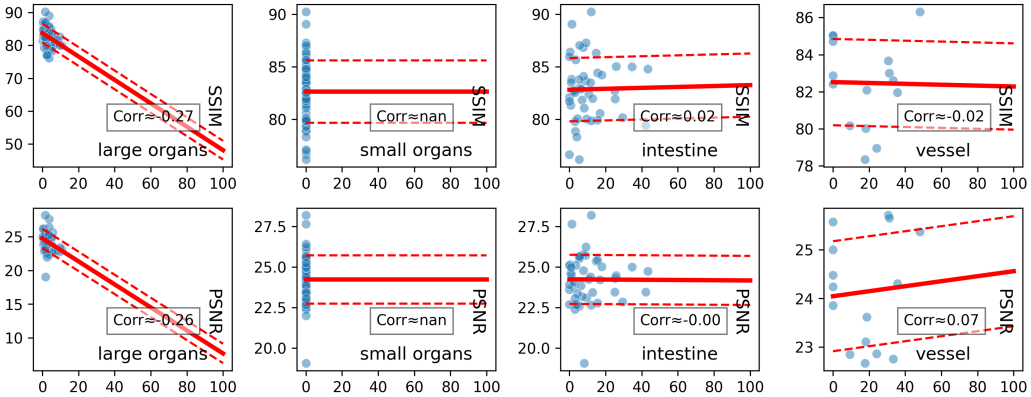

Figure 16: **Correlation Between Anatomy-Aware Metrics and Pixel-Wise Metrics of InTomo [60].** The solid red line is a linear fit and the dashed lines denote $\pm 1\sigma$ of the residuals. Low correlation indicates that better anatomical preservation does not guarantee higher pixel-wise metrics.

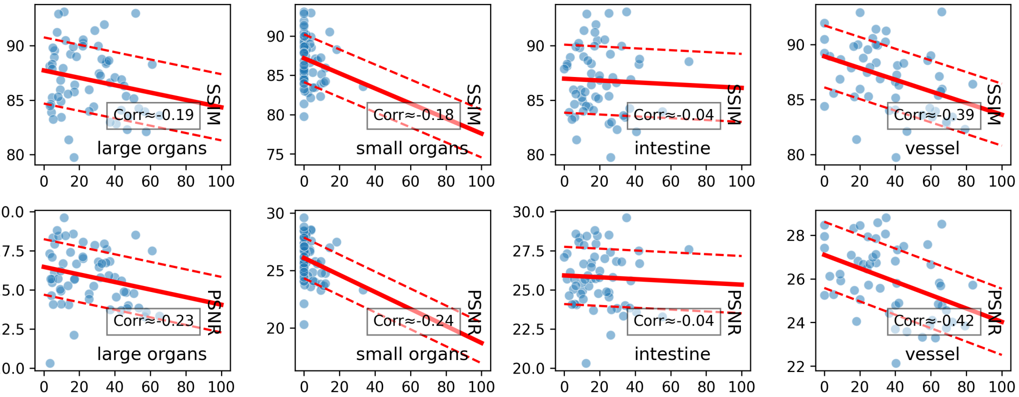

Figure 17: **Correlation Between Anatomy-Aware Metrics and Pixel-Wise Metrics of NeRF [38].** The solid red line is a linear fit and the dashed lines denote $\pm 1\sigma$ of the residuals. Low correlation indicates that better anatomical preservation does not guarantee higher pixel-wise metrics.

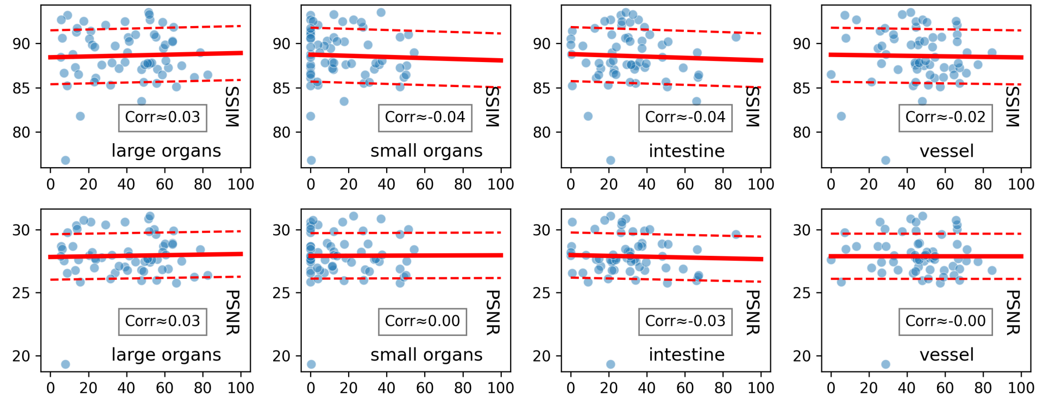

Figure 18: **Correlation Between Anatomy-Aware Metrics and Pixel-Wise Metrics of TensoRF [11].** The solid red line is a linear fit and the dashed lines denote ±1σ of the residuals. Low correlation indicates that better anatomical preservation does not guarantee higher pixel-wise metrics.

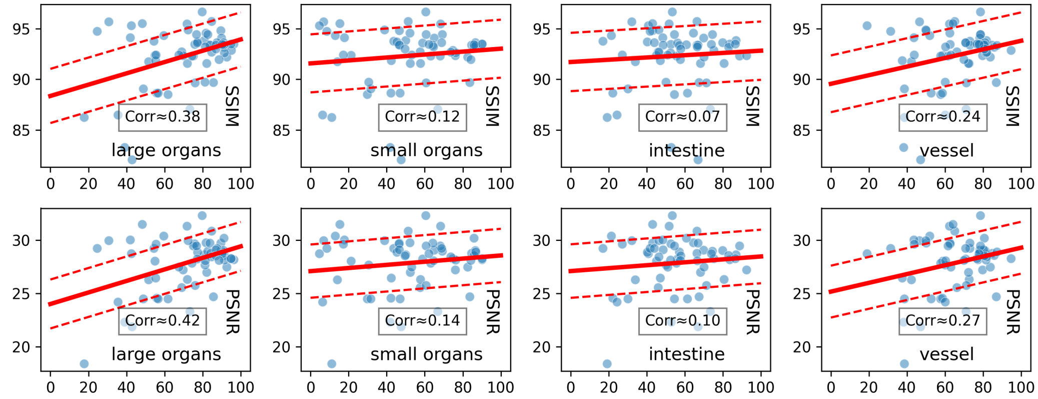

Figure 19: **Correlation Between Anatomy-Aware Metrics and Pixel-Wise Metrics of R^2-GS [61].** The solid red line is a linear fit and the dashed lines denote ±1σ of the residuals. Low correlation indicates that better anatomical preservation does not guarantee higher pixel-wise metrics.

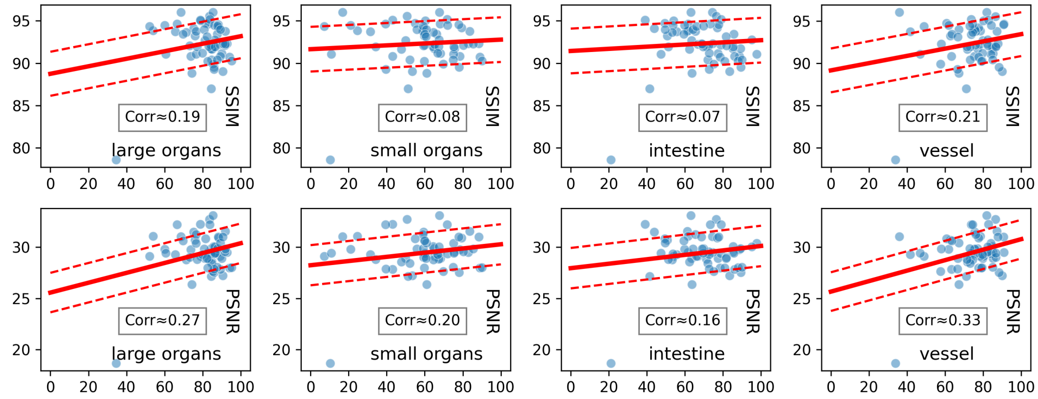

Figure 20: **Correlation Between Anatomy-Aware Metrics and Pixel-Wise Metrics of NAF [62].** The solid red line is a linear fit and the dashed lines denote ±1σ of the residuals. Low correlation indicates that better anatomical preservation does not guarantee higher pixel-wise metrics.

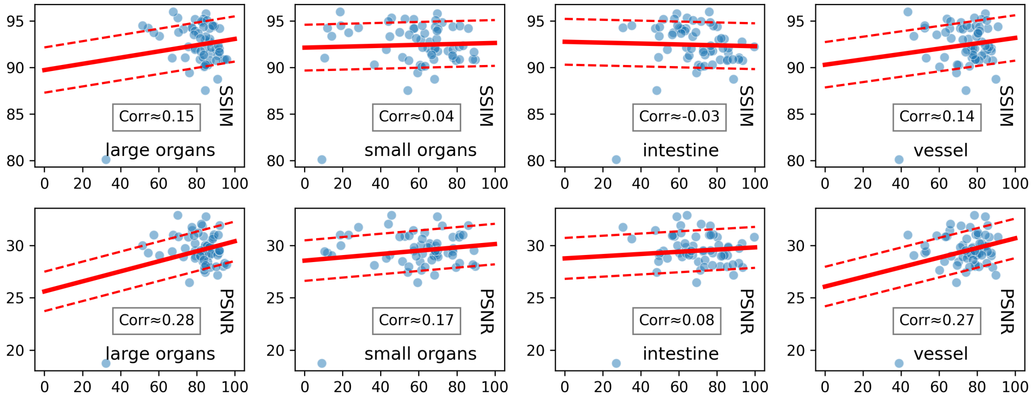

Figure 21: **Correlation Between Anatomy-Aware Metrics and Pixel-Wise Metrics of FDK [18].** The solid red line is a linear fit and the dashed lines denote $\pm 1\sigma$ of the residuals. Low correlation indicates that better anatomical preservation does not guarantee higher pixel-wise metrics.

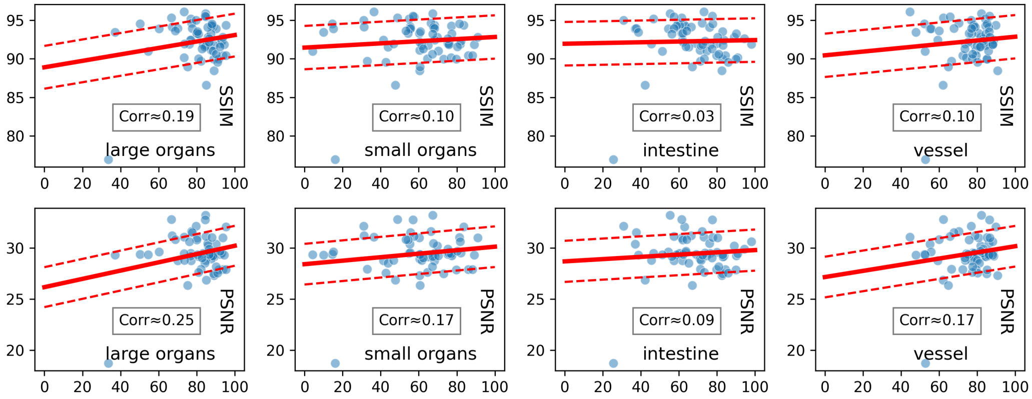

Figure 22: **Correlation Between Anatomy-Aware Metrics and Pixel-Wise Metrics of SART [1].** The solid red line is a linear fit and the dashed lines denote $\pm 1\sigma$ of the residuals. Low correlation indicates that better anatomical preservation does not guarantee higher pixel-wise metrics.

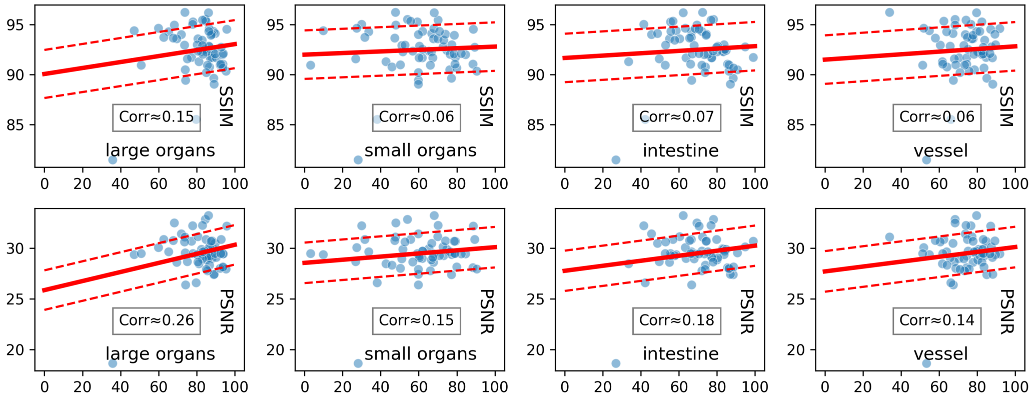

Figure 23: **Correlation Between Anatomy-Aware Metrics and Pixel-Wise Metrics of ASD-POCS [47].** The solid red line is a linear fit and the dashed lines denote $\pm 1\sigma$ of the residuals. Low correlation indicates that better anatomical preservation does not guarantee higher pixel-wise metrics.

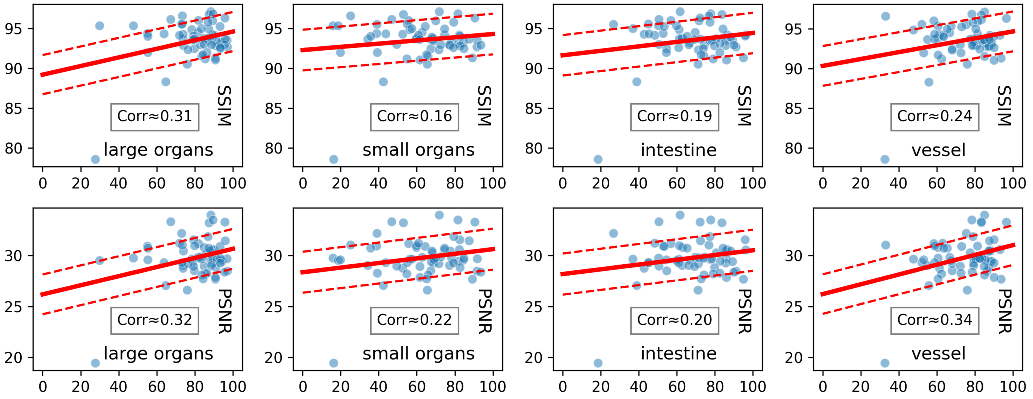

Figure 24: **Correlation Between Anatomy-Aware Metrics and Pixel-Wise Metrics of SAX-NeRF [8].** The solid red line is a linear fit and the dashed lines denote $\pm 1\sigma$ of the residuals. Low correlation indicates that better anatomical preservation does not guarantee higher pixel-wise metrics.

D Technical Details of CARE

In §3.3, we discuss the *Preliminary* stage of the CARE framework: Autoencoder and Latent Diffusion Model. In this section, we provide more technical details of the autoencoder in §D.1, and provide more technical specifications of the latent diffusion model in §D.2.

D.1 Autoencoder

In this initial stage, we finetuned a KL-regularized Variational Autoencoder (KL-VAE), using the JHH CT scans. The encoder of the VAE compresses input CT images into latent representations (denoted as z), while the decoder reconstructs the original images from these latent variables. The training objective aims to optimize both image reconstruction quality and the regularization imposed by the Kullback-Leibler divergence to ensure meaningful latent representations. These trained autoencoder serve as the foundational components for subsequent diffusion model training.

Specifically, let X denote an input CT image in pixel space. Note that to consider inter-frame consistency and retain the original architecture of the autoencoder network, we stack three adjacent CT slices to be the three channels of the input image, giving $X \in \mathbb{R}^{H\times W\times 3}$. The encoder $\mathcal{E}_{\theta_E}$ maps X to a latent representation $z = \mathcal{E}_{\theta_E}(X) \in \mathbb{R}^{h\times w\times c}$. A paired decoder $\mathcal{D}_{\theta_D}$ reconstructs the input as $\hat{X} = \mathcal{D}_{\theta_D}(z)$. Training adapts the KL-VAE pretrained on natural images (checkpoint provided by Stable Diffusion v1.5 [42]) to the target CT domain by minimising a weighted sum of three complementary criteria:

$$\begin{aligned}\mathcal{L}_{\text{AE}} = \lambda_{\text{rec}} \underbrace{\|X - \hat{X}\|_1}_{\text{reconstruction}} + \underbrace{\lambda_{\text{per}} \sum_l w_l \big\|\phi_l(X) - \phi_l(\hat{X})\big\|_2^2}_{\text{perceptual}} \\ + \beta \underbrace{\mathrm{KL}\big(q_{\theta_E}(z \mid X) \| \mathcal{N}(0, I)\big)}_{\text{KL regularisation}}\end{aligned} \tag{13}$$

where:

- **Reconstruction term $\mathcal{L}_{\text{rec}}$** directly penalises pixel-wise errors, preserving low-frequency fidelity.
- **Perceptual term $\mathcal{L}_{\text{per}}$** computes the ℓ_2 distance between intermediate feature maps $\phi_l(\cdot)$ of a fixed vision backbone (here VGG-16 pretrained on ImageNet). It sharpens textures and maintains semantic consistency that pure pixel metrics miss.
- **KL term $\mathcal{L}_{\text{KL}}$** pulls the encoder's approximate posterior $q_{\theta_E}(z \mid X)$ towards the unit Gaussian prior, preventing latent collapse and ensuring the aggregated latent distribution remains close to $\mathcal{N}(0, I)$. This normalization is crucial for the subsequent diffusion prior, whose noise schedule assumes unit-variance latents.

The weights λ_{rec}, λ_{per}, and β control the fidelity–regularity trade-off between generative capacity and roughly standard-normal latents. Unlike the adversarially-augmented auto-encoders used in Stable Diffusion, we fintuned on the three penalizations above directly, simplifying the optimization process and eliminating adversary-induced instabilities without sacrificing downstream sample quality. After fintuning the KL-VAE model on Stable Diffusion's checkpoint, frozen latents z serve as training data for the latent-space diffusion model described below.

D.2 Latent Diffusion Model

This stage pretrains a latent diffusion model with the JHH dataset to enhance low-quality CT scans while preserving anatomical structures. Each high-quality CT image $X \in \mathbb{R}^{H \times W \times 3}$ is first compressed by the frozen KL-VAE encoder $\mathcal{E}_{\theta_{\mathcal{E}}}$ into a latent representation $z = \mathcal{E}_{\theta_{\mathcal{E}}}(X) \in \mathbb{R}^{h \times w \times c}$. In the forward diffusion process a noisy sequence $\{z_t\}_{t=0}^{T}$ is obtained by

$$z_t = \sqrt{\bar{\alpha}_t} z_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon, \qquad \epsilon \sim \mathcal{N}(0, I),\ t \in \{1, \dots, T\}, \tag{14}$$

where $z_0 \triangleq z$ and $\{\bar{\alpha}_t\}_{t=1}^{T}$ is a fixed variance-preserving schedule. To expose structural cues during de-noising we degrade z with a deterministic operator $\mathcal{H}(\cdot)$ and concatenate it with z_t along the channel dimension,

$$\tilde{z}_t = \mathrm{Concat}[z_t, \mathcal{H}(z)] \ \in \mathbb{R}^{h \times w \times 2c}. \tag{15}$$

where $\mathcal{H}(\cdot)$ is a downsampling process with a factor of 4 and followed by an upsampling process back to the original resolution.

A de-noising UNet ϵ_θ receives $\tilde{z}_t$, the diffusion timestep t, and a fixed text embedding c_{phase} of the contrast phase information, and predicts the noise residual $\hat{\epsilon} = \epsilon_\theta(\tilde{z}_t, t, c_{\text{phase}})$. Then, the network parameters are optimized with the standard noise-prediction loss

$$\mathcal{L}_{\text{LDM}} = \mathbb{E}_{t,\epsilon,X} \left\| \epsilon - \epsilon_\theta\big(\mathrm{Concat}[z_t, \mathcal{H}(z)], t, c_{\text{phase}}\big) \right\|_2^2. \tag{16}$$

Because only $\mathcal{L}_{\text{LDM}}$ is used, learning focuses on accurate latent de-noising; $\mathcal{H}(z)$ supplies anatomy so the de-noising UNet acquires an intrinsic enhancement capability without extra image-space supervision.

E Experimental Results of CARE

E.1 Benchmarking on Other Segmentation Metrics

Table 8: **Benchmarking CT Reconstruction Methods on *NSD*.** We evaluate preexisting CT reconstruction methods on high-quality CT scans using both pixel-wise metrics and our anatomy-aware metrics. We report the median and interquartile range (IQR) of these metrics. Cells are marked in blue, where a deeper color denotes a greater value.

	pixel-wise metric		**anatomy-aware metric (ours)**			
method	SSIM	PSNR	$\text{NSD}_{\text{large}}$	$\text{NSD}_{\text{small}}$	$\text{NSD}_{\text{intestine}}$	$\text{NSD}_{\text{vessel}}$
InTomo [60]	82.7 (80.6,84.7)	24.3 (23.1,25.2)	2.7 (1.9,4.1)	0.0 (0.0,0.0)	2.0 (0.3,3.6)	0.0 (0.0,0.0)
NeRF [38]	86.8 (84.4,89.0)	25.7 (24.7,27.1)	21.7 (9.5,37.6)	0.1 (0.0,3.9)	5.1 (3.4,9.6)	5.8 (1.9,16.6)
TensoRF [11]	88.4 (87.0,90.7)	27.8 (26.9,28.9)	41.3 (20.3,50.5)	10.9 (0.0,28.3)	9.4 (4.8,14.8)	20.5 (10.1,40.2)
R^2-GS [61]	93.2 (91.9,93.9)	28.6 (27.0,29.5)	72.6 (57.1,80.7)	54.0 (42.3,67.9)	15.1 (10.5,19.8)	62.1 (47.9,75.6)
NAF [62]	92.3 (91.1,94.1)	29.4 (28.8,30.6)	74.7 (69.0,82.2)	61.1 (47.0,68.6)	18.2 (13.2,24.6)	71.3 (64.0,81.9)
FDK [18]	92.4 (91.1,94.2)	29.5 (28.8,30.8)	75.3 (69.2,82.0)	62.2 (50.4,69.3)	18.1 (14.1,24.6)	72.3 (62.6,81.7)
SART [1]	92.5 (91.1,94.1)	29.4 (28.8,30.6)	76.2 (67.2,82.8)	61.8 (49.3,72.5)	18.0 (12.9,24.6)	73.6 (61.3,83.5)
ASD-POCS [47]	92.5 (91.3,94.1)	29.5 (28.7,30.7)	76.3 (67.4,82.7)	63.0 (47.8,71.8)	18.6 (13.9,24.5)	72.2 (64.6,82.9)
SAX-NeRF [8]	93.7 (92.8,95.0)	29.7 (29.0,30.8)	75.9 (67.0,82.5)	64.4 (49.8,76.3)	17.5 (13.5,24.2)	70.5 (56.2,85.5)

Table 9: **Benchmarking CT Reconstruction Methods on *DSC*.** We evaluate preexisting CT reconstruction methods on high-quality CT scans using both pixel-wise metrics and our anatomy-aware metrics. We report the median and interquartile range (IQR) of these metrics. Cells are marked in blue, where a deeper color denotes a greater value.

	pixel-wise metric		**anatomy-aware metric (ours)**			
method	SSIM	PSNR	$\text{DSC}_{\text{large}}$	$\text{DSC}_{\text{small}}$	$\text{DSC}_{\text{intestine}}$	$\text{DSC}_{\text{vessel}}$
InTomo [60]	82.7 (80.6,84.7)	24.3 (23.1,25.2)	7.9 (4.6,10.6)	0.0 (0.0,0.0)	2.4 (0.1,4.9)	0.0 (0.0,0.0)
NeRF [38]	86.8 (84.4,89.0)	25.7 (24.7,27.1)	32.1 (15.7,49.8)	0.0 (0.0,2.0)	6.8 (3.1,14.4)	8.3 (1.7,20.0)
TensoRF [11]	88.4 (87.0,90.7)	27.8 (26.9,28.9)	54.5 (32.3,68.0)	7.0 (0.0,20.9)	14.0 (4.9,20.8)	22.9 (12.9,37.8)
R^2-GS [61]	93.2 (91.9,93.9)	28.6 (27.0,29.5)	79.3 (73.0,86.0)	46.9 (35.0,61.6)	25.2 (16.7,32.5)	57.1 (40.5,65.9)
NAF [62]	92.3 (91.1,94.1)	29.4 (28.8,30.6)	84.1 (76.5,87.5)	52.9 (39.6,63.9)	29.7 (21.2,40.0)	63.3 (56.6,73.0)
FDK [18]	92.4 (91.1,94.2)	29.5 (28.8,30.8)	84.2 (77.3,87.8)	53.5 (42.0,64.0)	29.6 (20.6,40.1)	63.3 (56.0,72.9)
SART [1]	92.5 (91.1,94.1)	29.4 (28.8,30.6)	83.8 (77.5,87.8)	52.3 (40.4,64.2)	28.5 (20.4,40.5)	63.6 (55.2,74.3)
ASD-POCS [47]	92.5 (91.3,94.1)	29.5 (28.7,30.7)	84.2 (77.0,88.3)	52.6 (40.3,65.6)	29.2 (21.5,40.4)	62.4 (54.4,73.0)
SAX-NeRF [8]	93.7 (92.8,95.0)	29.7 (29.0,30.8)	83.4 (75.4,86.7)	54.2 (39.1,70.6)	29.7 (20.5,38.8)	63.5 (50.6,75.3)

E.2 Qualitative Results of CARE

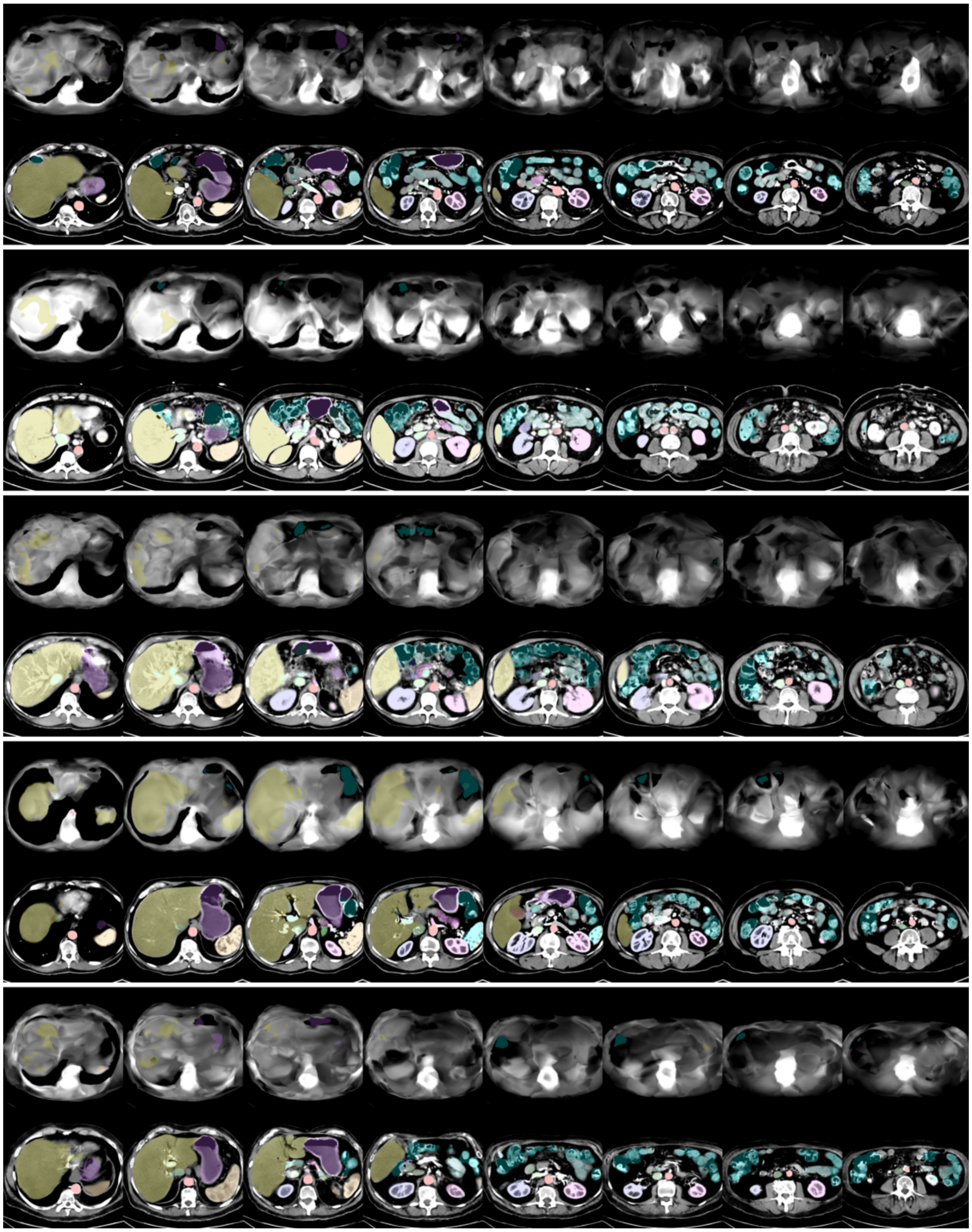

Figure 25: **Qualitative results of InTomo [60] with and without CARE.** Every two rows showcase the results of eight CT slices of a single CT scan. CARE performs great anatomical preservation during enhancement. The images are center-cropped to eliminate the background. A soft-tissue window (300/50 HU) was used to enhance abdominal organ visibility.

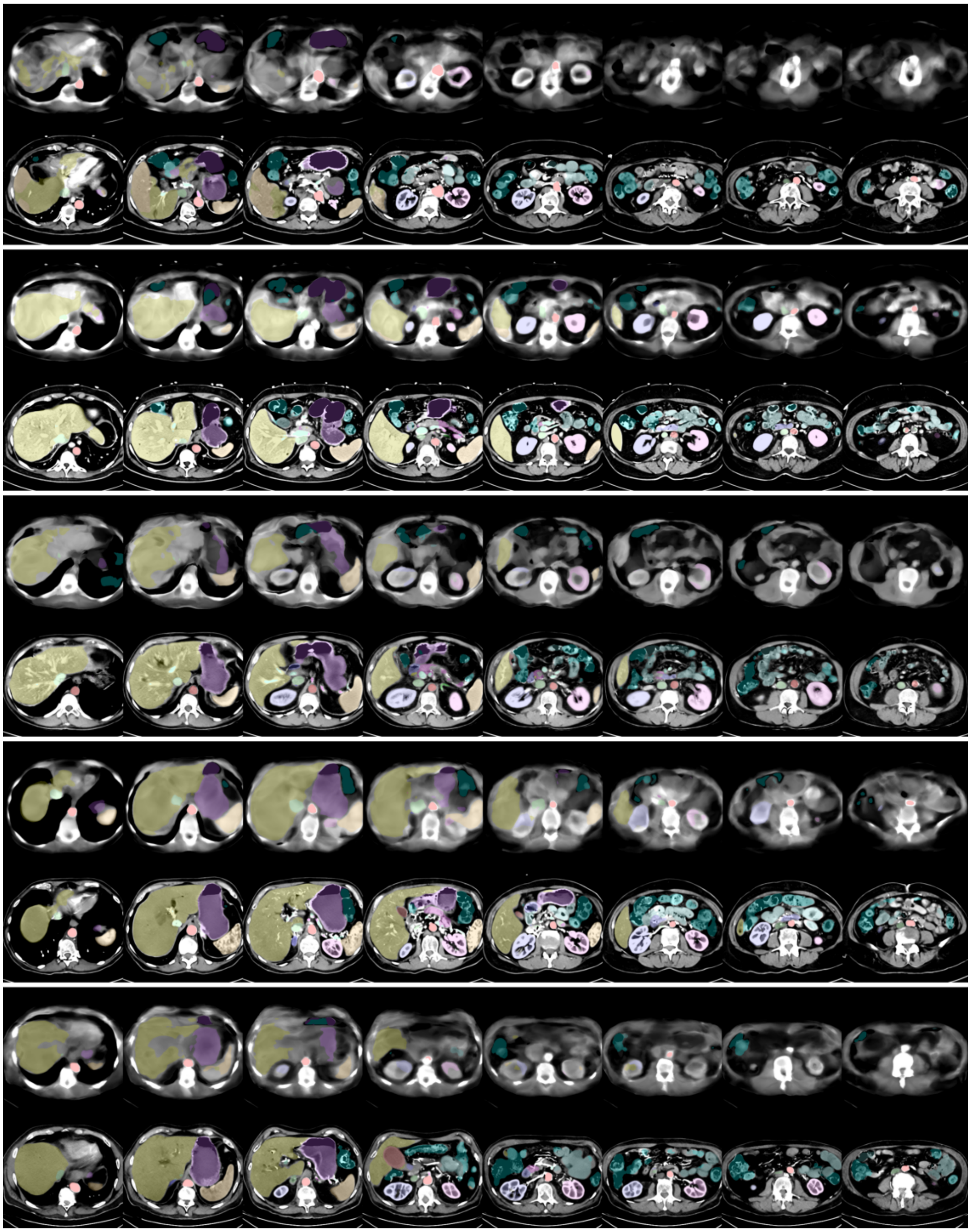

Figure 26: **Qualitative results of NeRF [38] with and without CARE.** Every two rows showcase the results of eight CT slices of a single CT scan. CARE performs great anatomical preservation during enhancement. The images are center-cropped to eliminate the background. A soft-tissue window (300/50 HU) was used to enhance abdominal organ visibility.

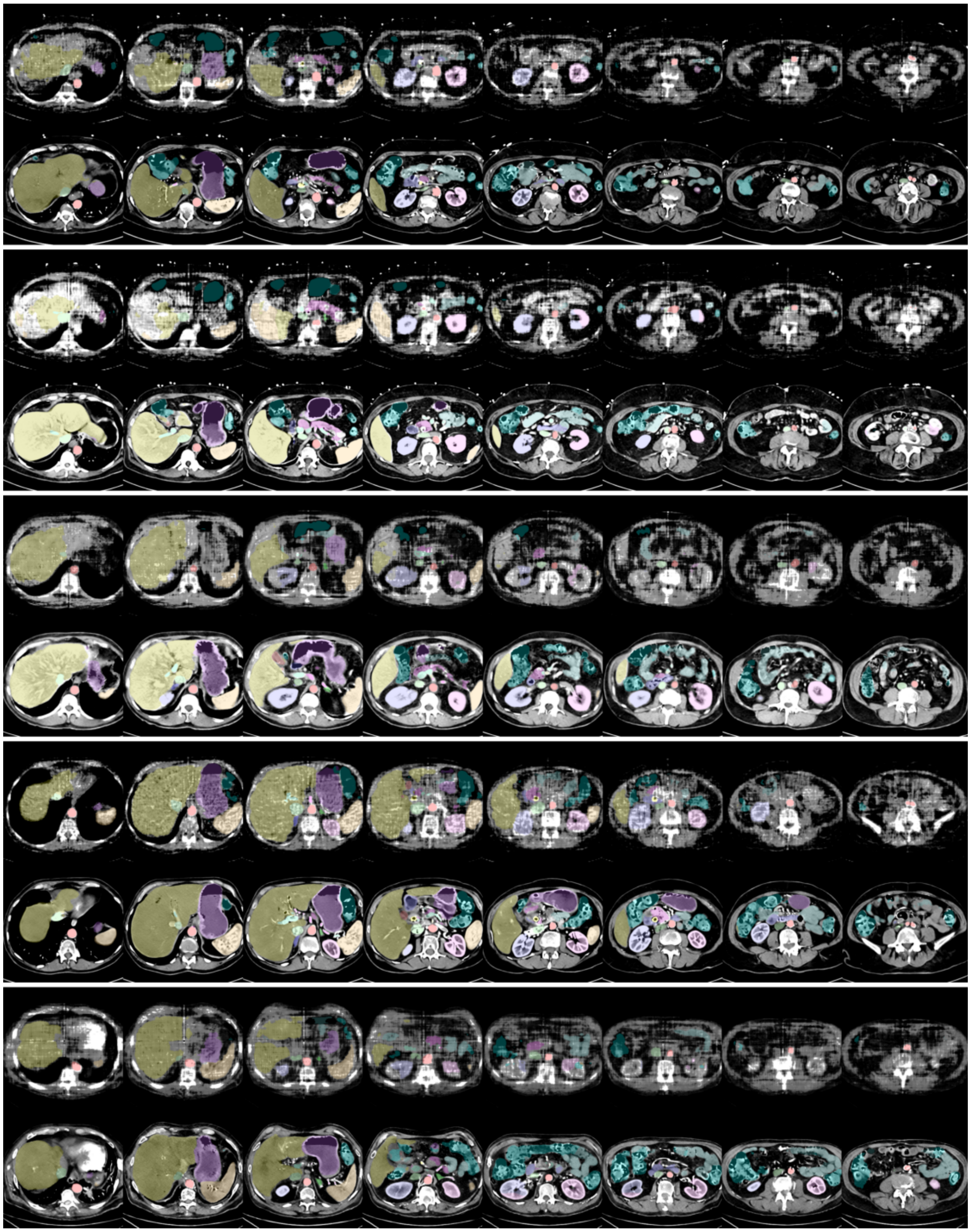

Figure 27: **Qualitative results of TensoRF [11] with and without CARE.** Every two rows showcase the results of eight CT slices of a single CT scan. CARE performs great anatomical preservation during enhancement. The images are center-cropped to eliminate the background. A soft-tissue window (300/50 HU) was used to enhance abdominal organ visibility.

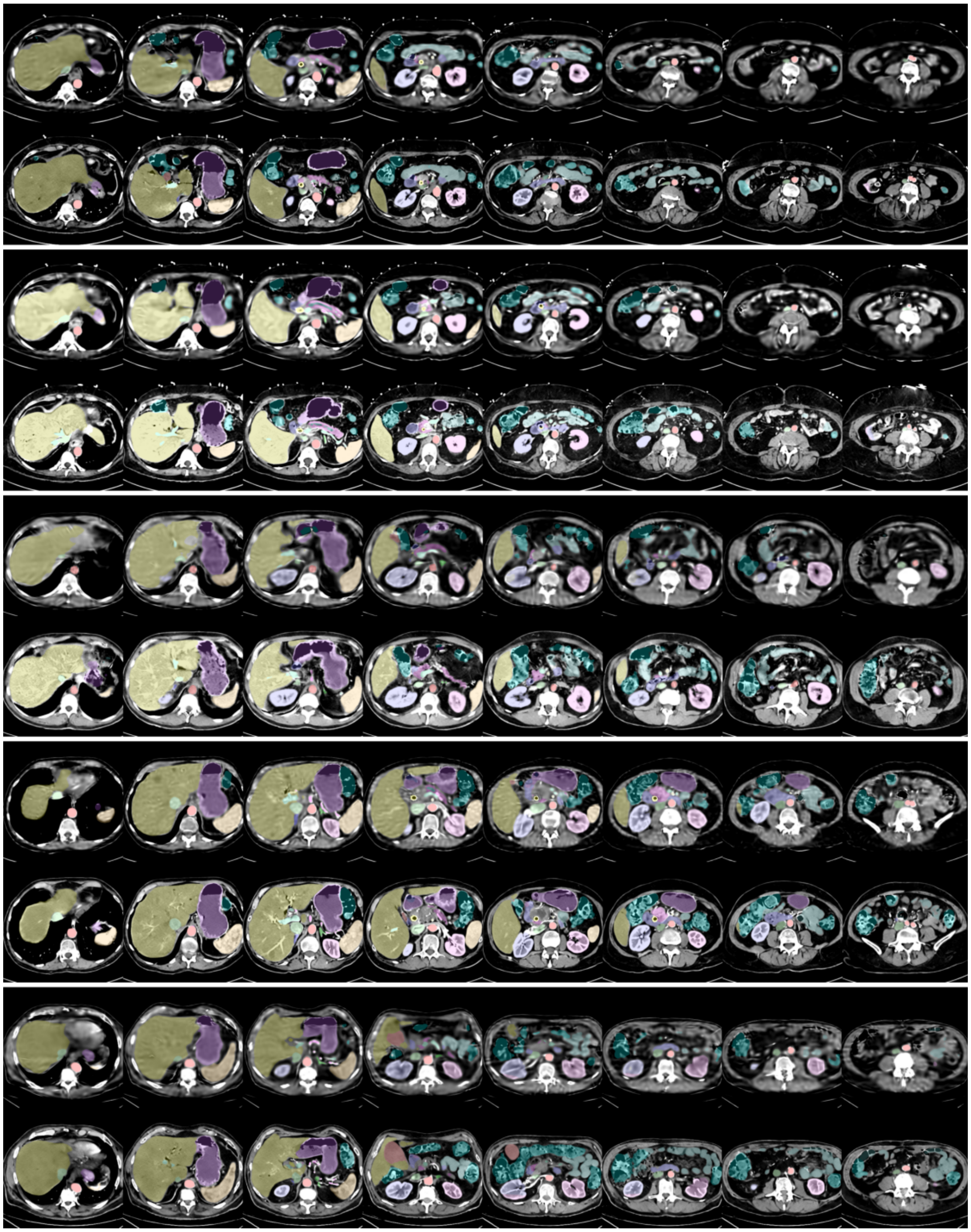

Figure 28: **Qualitative results of R^2-GS [61] with and without CARE.** Every two rows showcase the results of eight CT slices of a single CT scan. CARE performs great anatomical preservation during enhancement. The images are center-cropped to eliminate the background. A soft-tissue window (300/50 HU) was used to enhance abdominal organ visibility.

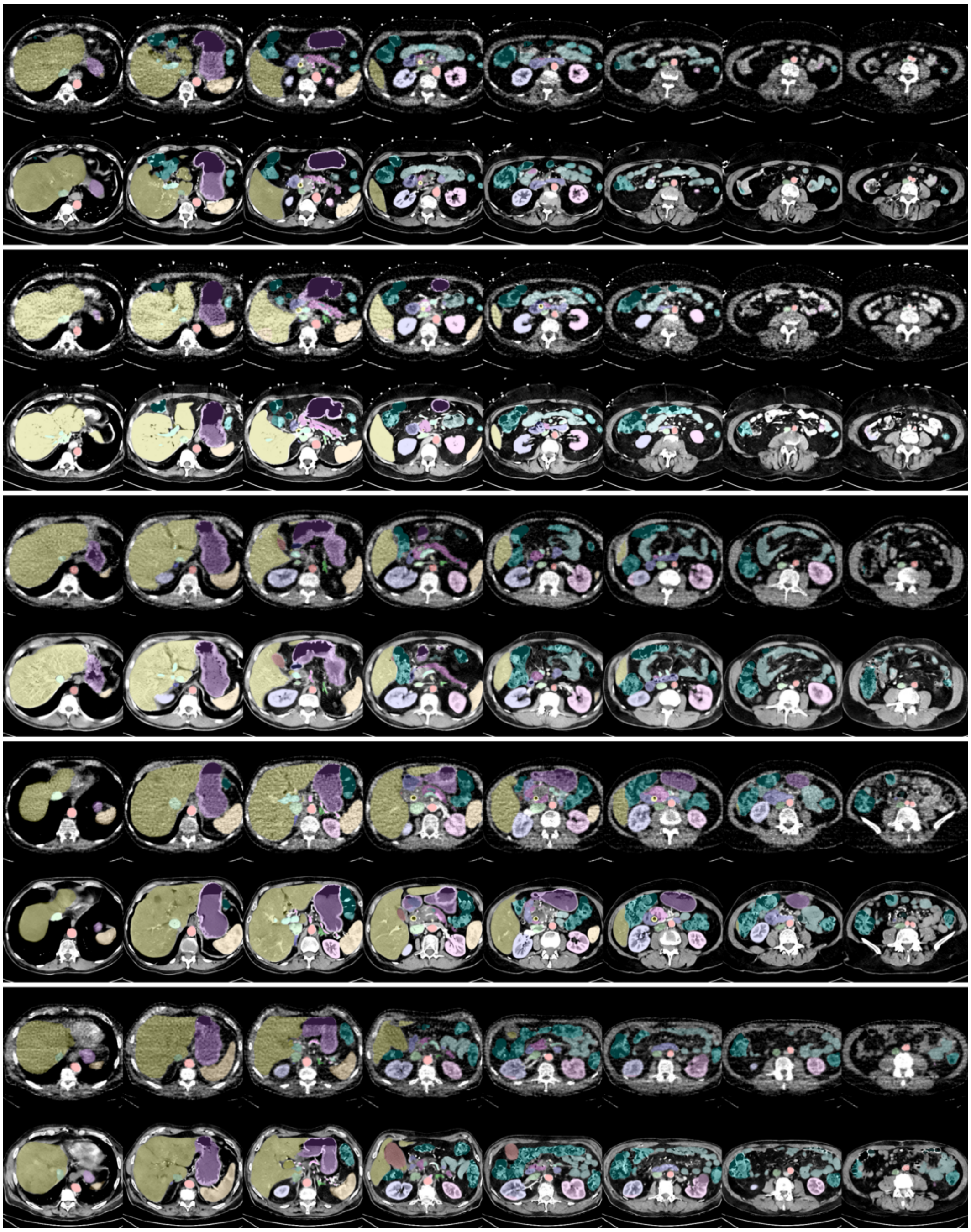

Figure 29: **Qualitative results of NAF [62] with and without CARE.** Every two rows showcase the results of eight CT slices of a single CT scan. CARE performs great anatomical preservation during enhancement. The images are center-cropped to eliminate the background. A soft-tissue window (300/50 HU) was used to enhance abdominal organ visibility.

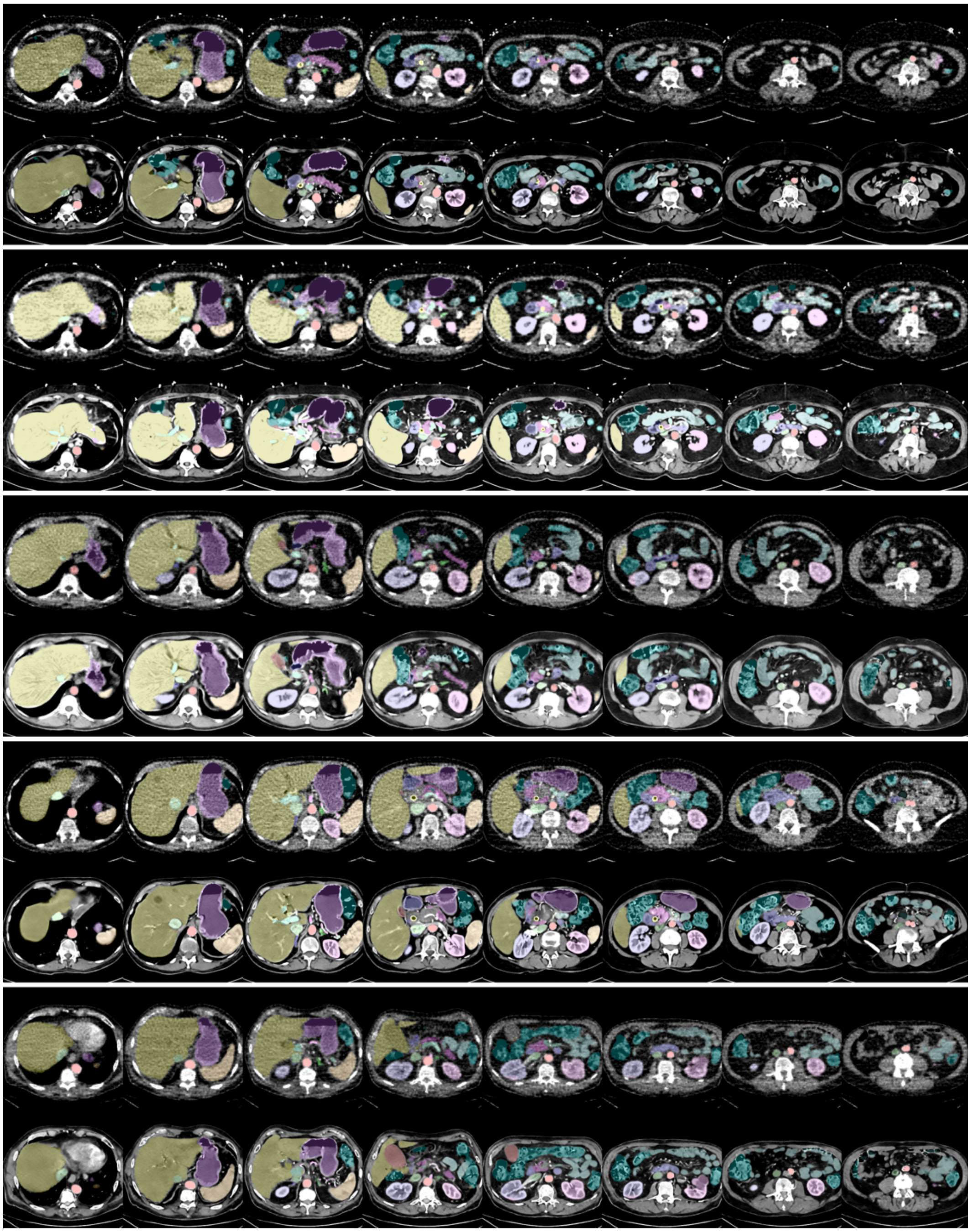

Figure 30: **Qualitative results of FDK [18] with and without CARE.** Every two rows showcase the results of eight CT slices of a single CT scan. CARE performs great anatomical preservation during enhancement. The images are center-cropped to eliminate the background. A soft-tissue window (300/50 HU) was used to enhance abdominal organ visibility.

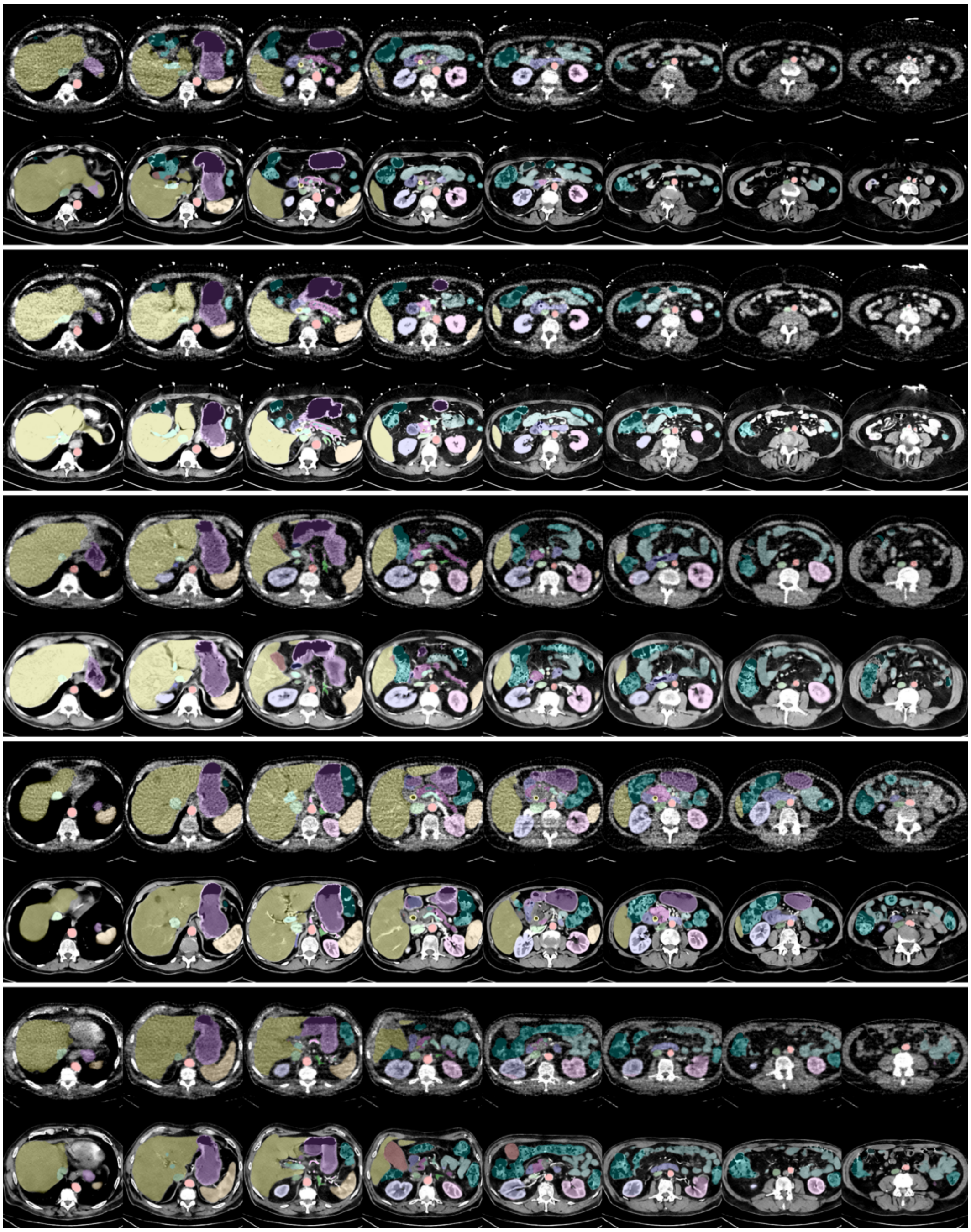

Figure 31: **Qualitative results of SART [1] with and without CARE.** Every two rows showcase the results of eight CT slices of a single CT scan. CARE performs great anatomical preservation during enhancement. The images are center-cropped to eliminate the background. A soft-tissue window (300/50 HU) was used to enhance abdominal organ visibility.

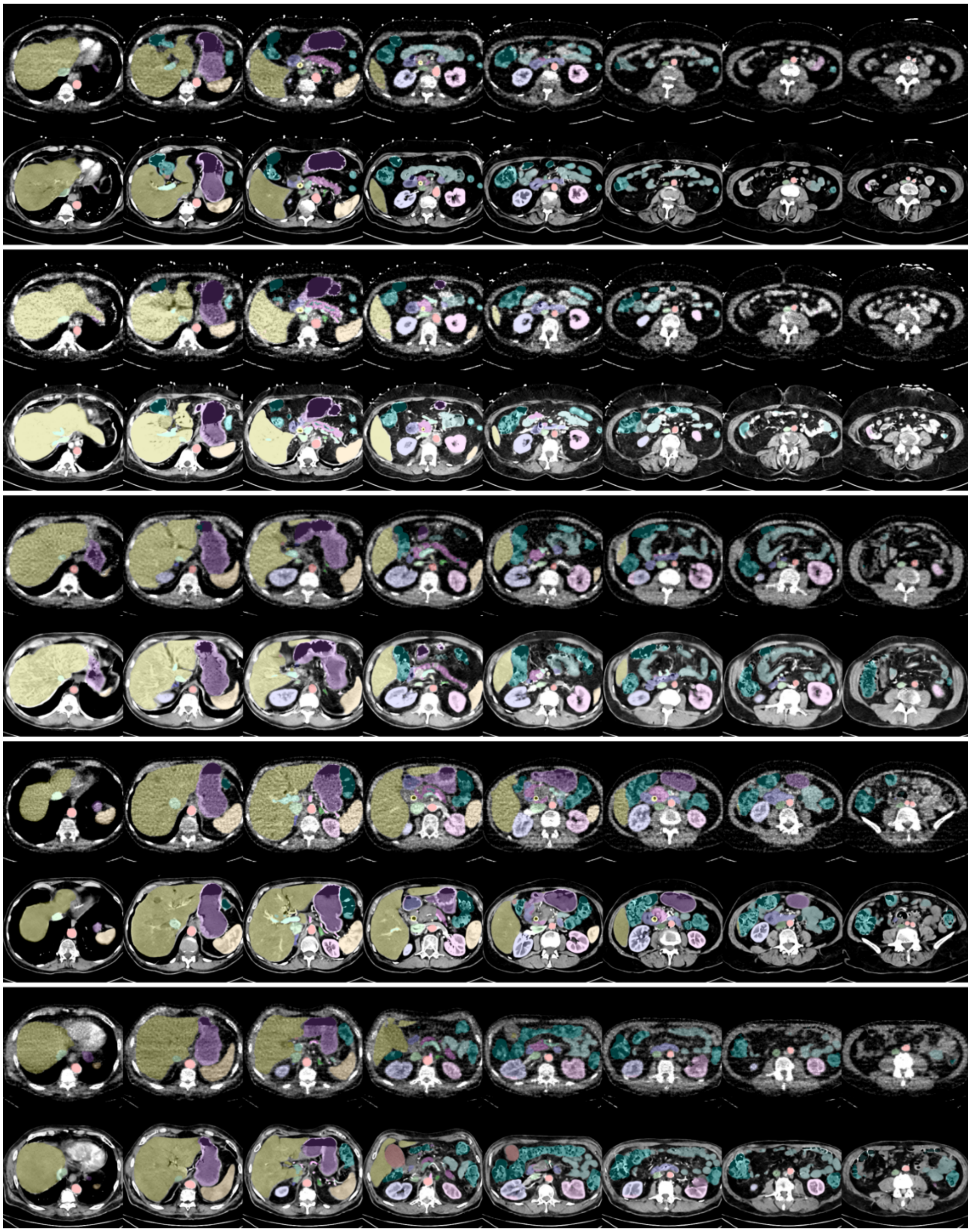

Figure 32: **Qualitative results of ASD-POCS [47] with and without CARE.** Every two rows showcase the results of eight CT slices of a single CT scan. CARE performs great anatomical preservation during enhancement. The images are center-cropped to eliminate the background. A soft-tissue window (300/50 HU) was used to enhance abdominal organ visibility.

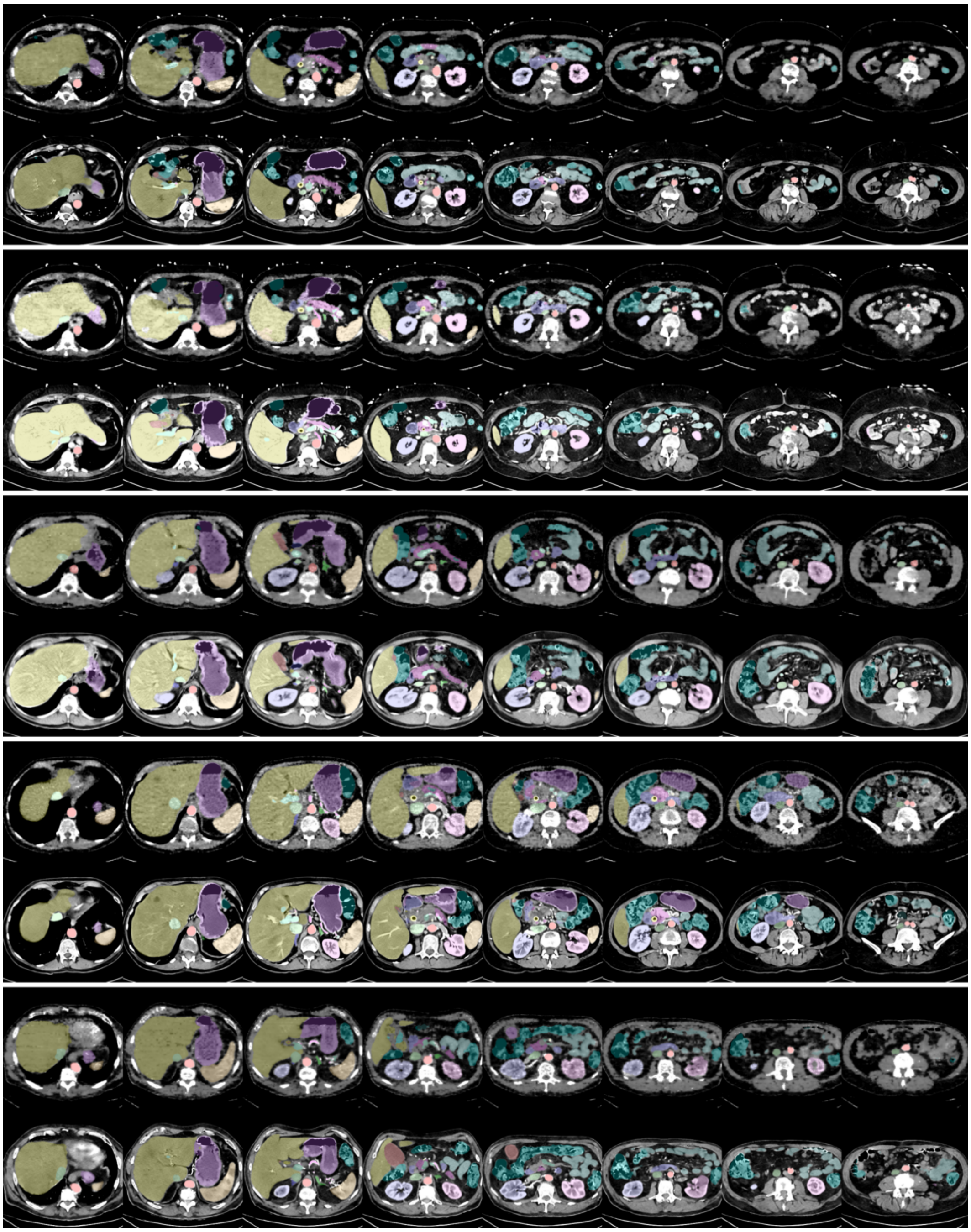

Figure 33: **Qualitative results of SAX-NeRF [8] with and without CARE.** Every two rows showcase the results of eight CT slices of a single CT scan. CARE performs great anatomical preservation during enhancement. The images are center-cropped to eliminate the background. A soft-tissue window (300/50 HU) was used to enhance abdominal organ visibility.

E.3 CARE on Arterial and Portal-venous Phase CT Scans

Table 10: **CARE-enhanced Reconstruction Evaluation on *Arterial* Phase CT Scans.** Evaluate preexisting CT reconstruction methods on high-quality CT scans with CARE using both pixel-wise metrics and our anatomy-aware metrics. Note that the results are based on 23 arterial CT scans that CARE has never been trained on. We report the median and interquartile range (IQR) of these metrics and perform the Mann-Whitney U test for statistical analysis. Cells are marked in color only if CARE brings a significant difference ($p < 0.05$), while green if the CARE enhancement results have improvement, and red otherwise. Deeper color represents greater difference.

		pixel-wise metric		**anatomy-aware metric (ours)**			
method		SSIM	PSNR	$\text{NSD}_{\text{large}}$	$\text{NSD}_{\text{small}}$	$\text{clDice}_{\text{intestine}}$	$\text{clDice}_{\text{vessel}}$
InTomo	50 views	83.4 (81.5,84.9)	24.7 (23.5,25.4)	2.4 (1.5,3.3)	0.0 (0.0,0.0)	8.1 (1.2,27.2)	9.3 (0.0,18.6)
	+CARE	76.3 (74.5,80.4)	22.2 (21.1,23.0)	28.8 (20.8,34.8)	5.5 (1.7,10.5)	30.2 (23.7,39.6)	47.4 (34.6,60.8)
NeRF	50 views	88.1 (85.3,90.1)	27.0 (24.8,27.7)	21.9 (7.7,31.8)	0.1 (0.0,2.8)	17.1 (9.7,25.9)	28.9 (24.1,44.7)
	+CARE	82.1 (79.6,83.1)	24.0 (22.5,24.7)	58.2 (46.2,61.0)	24.4 (13.8,28.4)	50.4 (39.3,56.1)	57.7 (45.4,68.8)
TensoRF	50 views	90.2 (87.8,92.3)	28.7 (27.6,30.2)	41.3 (21.6,53.3)	13.9 (2.5,27.8)	31.1 (20.9,35.8)	44.7 (33.2,57.4)
	+CARE	88.9 (87.4,90.2)	27.9 (27.3,28.7)	73.1 (61.3,77.7)	35.0 (21.2,44.6)	66.3 (60.2,73.4)	58.1 (48.7,60.8)
R^2-GS	50 views	93.5 (92.3,94.8)	29.3 (28.0,29.7)	75.8 (63.5,84.2)	60.5 (41.3,68.1)	53.4 (42.9,66.6)	64.0 (58.8,75.7)
	+CARE	89.6 (88.4,91.6)	27.8 (26.3,29.0)	80.8 (73.8,86.7)	56.2 (42.0,67.5)	72.0 (66.6,78.8)	66.8 (62.3,71.5)
NAF	50 views	93.8 (91.4,94.3)	29.9 (29.0,31.3)	83.2 (74.5,85.4)	60.9 (49.2,69.0)	71.3 (61.5,76.7)	75.4 (67.0,79.1)
	+CARE	92.8 (91.3,93.6)	29.9 (28.5,30.4)	87.0 (80.9,90.1)	63.0 (48.7,74.3)	80.8 (73.1,84.2)	68.2 (60.5,74.3)
FDK	50 views	93.7 (91.3,94.5)	29.8 (29.2,31.3)	83.0 (75.8,84.6)	61.7 (48.6,71.2)	64.4 (58.2,76.5)	77.2 (64.8,79.7)
	+CARE	92.8 (91.5,93.7)	30.3 (29.0,31.0)	85.7 (79.1,88.7)	65.8 (54.1,75.5)	81.4 (72.6,84.9)	67.8 (65.8,73.2)
SART	50 views	93.6 (91.6,94.6)	30.2 (29.4,31.4)	83.4 (75.3,85.4)	60.9 (52.1,70.0)	66.4 (58.4,75.5)	78.8 (66.8,82.6)
	+CARE	93.4 (92.7,94.6)	30.6 (29.8,31.2)	87.1 (81.6,90.1)	68.1 (54.0,75.6)	83.4 (75.8,86.1)	66.5 (55.6,75.3)
ASD-POCS	50 views	93.6 (91.8,94.5)	29.9 (29.2,31.3)	83.0 (72.9,85.8)	66.0 (49.7,69.9)	70.5 (57.6,77.0)	74.9 (67.8,79.9)
	+CARE	92.9 (91.9,93.5)	29.9 (28.8,30.6)	86.1 (81.5,89.7)	67.0 (52.3,73.8)	80.3 (72.6,84.6)	70.0 (59.7,76.6)
SAX-NeRF	50 views	93.9 (93.0,95.9)	30.5 (29.5,31.5)	82.8 (72.9,88.6)	65.0 (54.3,77.3)	66.2 (54.4,77.5)	79.3 (65.3,84.3)
	+CARE	92.7 (90.9,93.4)	29.4 (28.8,30.3)	86.2 (80.2,90.0)	61.7 (54.2,73.4)	81.8 (72.2,86.3)	68.1 (56.3,75.9)

Table 11: **CARE-enhanced Reconstruction Evaluation on *Portal Venous* Phase CT Scans.** Evaluate CT reconstruction methods on high-quality CT scans with CARE using both pixel-wise metrics and our anatomy-aware metrics. Note that the results are based on 13 portal venous CT scans that CARE has never been trained on. We report the median and interquartile range (IQR) of these metrics and perform the Mann-Whitney U test for statistical analysis. Cells are marked in color only if CARE brings significant difference ($p < 0.05$), while green if the CARE enhancement results have improvement, and red otherwise. Deeper color represents greater difference.

		pixel-wise metric		**anatomy-aware metric (ours)**			
method		SSIM	PSNR	$\text{NSD}_{\text{large}}$	$\text{NSD}_{\text{small}}$	$\text{clDice}_{\text{intestine}}$	$\text{clDice}_{\text{vessel}}$
InTomo	50 views	81.7 (79.9,83.0)	23.9 (22.8,24.6)	2.3 (1.9,3.6)	0.0 (0.0,0.0)	5.9 (2.7,15.9)	0.0 (0.0,15.6)
	+CARE	77.3 (73.3,78.1)	21.5 (20.2,22.3)	42.0 (27.9,47.6)	9.9 (5.9,15.2)	42.3 (33.5,47.1)	48.6 (38.7,63.4)
NeRF	50 views	85.6 (84.2,88.3)	25.3 (24.7,26.7)	30.7 (14.1,43.8)	0.2 (0.0,7.5)	19.1 (7.0,32.2)	29.2 (11.7,40.0)
	+CARE	79.0 (78.4,81.0)	22.2 (21.1,22.6)	57.7 (56.0,66.5)	27.3 (14.3,31.5)	56.6 (52.6,61.1)	52.9 (45.4,56.7)
TensoRF	50 views	87.3 (86.1,90.3)	27.6 (26.8,28.0)	47.9 (23.7,53.2)	8.6 (0.0,30.9)	22.8 (14.7,37.4)	51.2 (44.5,58.1)
	+CARE	88.7 (86.5,89.7)	27.3 (26.9,27.5)	78.1 (71.6,82.0)	40.6 (39.2,46.8)	71.3 (65.8,77.4)	63.0 (51.8,75.2)
R^2-GS	50 views	92.4 (89.7,93.5)	27.7 (26.3,28.7)	80.8 (67.4,84.7)	53.0 (41.8,73.6)	54.5 (50.9,73.8)	71.0 (60.1,81.0)
	+CARE	88.4 (84.8,89.9)	27.3 (24.9,27.8)	83.7 (74.4,89.7)	55.1 (45.6,60.2)	80.7 (76.1,84.9)	76.3 (56.8,86.2)
NAF	50 views	92.0 (90.4,92.8)	29.3 (28.0,29.6)	84.7 (81.1,87.6)	61.1 (52.2,74.9)	71.9 (63.6,85.6)	79.2 (73.0,82.5)
	+CARE	92.6 (92.0,92.8)	29.5 (29.0,30.0)	91.0 (86.4,91.8)	72.3 (59.8,77.6)	82.9 (78.4,90.5)	83.8 (77.9,88.4)
FDK	50 views	92.0 (90.3,93.1)	29.1 (28.3,29.5)	85.7 (84.1,86.9)	63.3 (57.5,69.3)	70.1 (66.5,83.8)	75.9 (71.2,82.4)
	+CARE	92.2 (91.3,92.8)	29.3 (29.0,29.7)	88.9 (86.1,91.4)	72.9 (62.0,79.0)	81.8 (76.3,86.5)	83.9 (74.8,88.0)
SART	50 views	91.9 (90.6,92.6)	29.2 (28.2,29.6)	87.5 (77.1,90.2)	67.2 (55.4,77.2)	71.1 (63.5,90.0)	77.8 (64.8,84.7)
	+CARE	93.3 (92.7,93.9)	30.0 (29.5,30.3)	89.8 (86.5,91.4)	70.3 (61.7,75.2)	85.7 (79.7,88.2)	83.6 (79.5,88.2)
ASD-POCS	50 views	92.1 (90.7,92.9)	29.2 (28.2,29.7)	86.4 (79.6,88.8)	60.2 (59.7,78.4)	72.0 (67.1,80.3)	74.5 (67.4,83.5)
	+CARE	92.2 (91.6,92.8)	28.8 (28.3,29.3)	88.8 (84.2,91.4)	71.0 (60.3,75.7)	81.5 (78.7,86.5)	83.6 (78.7,86.3)
SAX-NeRF	50 views	93.2 (92.3,94.1)	29.3 (28.8,29.6)	79.3 (75.9,90.0)	68.8 (56.2,79.1)	71.1 (61.0,79.8)	79.3 (58.6,87.2)
	+CARE	92.3 (91.5,93.3)	29.7 (28.9,29.9)	90.8 (86.3,91.8)	70.5 (63.7,78.9)	80.5 (77.2,90.0)	83.8 (76.5,88.4)

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction accurately reflect the paper's contributions by clearly framing the limitations of pixel-wise metrics, introducing our anatomy-aware evaluation framework and CARE enhancement pipeline, and previewing the empirical gains on multiple reconstruction backbones.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper discusses the limitations of the work performed by the authors in § 5.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: The paper does not include theoretical results. Its contributions lie in proposing a suite of anatomy-aware evaluation metrics, introducing a diffusion-based CT enhancement framework CARE, and providing thorough empirical validation, rather than in formal assumptions or mathematical proofs.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide the experimental details both in § 3.3 and in the supplementary materials.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Note that we only provide open access to code and model checkpoints; our data would be kept private. The code is attached in the supplementary materials with a link to the model checkpoint attached.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The experimental setting/details of training and testing are provided in § 3.3 and the supplementary materials.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The paper performs the Mann-Whitney U test at Table 2 to mark significance difference between our results with baselines.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The paper provides sufficient information on the computer resources in the supplementary materials.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: We preserve anonymity in all submitted materials and strictly conform the NeurIPS Code of Ethics in every respect. All patient-identifiable information was anonymized during preprocessing to ensure privacy protection, and the released dataset contains no identifiable information.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: See § 1 and § 5.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The paper has cited, credited and listed the licenses of all the papers of the code bases used in § 1 and supplementary materials.

    Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The paper introduce new code and models and they are all well documented in the supplementary materials.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper only uses pre-acquired CT scans, and does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: The paper only uses pre-acquired CT scans, and does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: The core method development in this research does not involve LLMs.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (https://neurips.cc/Conferences/2025/LLM) for what should or should not be described.