This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

SAFE-Diff: Structurally Anchored Diffusion for Anatomically Faithful CT Image Super-Resolution

Sneha[0009−0006−2176−2030], Nivedita Gupta[0009−0005−4454−8806], and Angshuman Paul[0000−0002−0935−0256]

Indian Institute of Technology Jodhpur
m24ma2010@iitj.ac.in

Abstract. Super-resolution (SR) of medical images requires a delicate balance between structural accuracy and perceptual sharpness. Diffusion models demonstrate strong texture generation but suffer from high computational costs and hallucination risks. We propose SAFE-Diff, a lightweight two-stage framework designed to balance perceptual detail enhancement with structural accuracy. In the first stage, a deterministic residual prediction net recovers the anatomical structure of the CT image. In the second stage, we employ a diffusion refiner to enhance high-resolution reconstruction. The refiner is conditioned on the output of stage one, ensuring anatomical consistency. We implement a truncated two-step diffusion trajectory via DDIM sampling. Finally, a Stationary Wavelet Transform (SWT/ISWT) fusion integrates low-frequency components from the first stage with high-frequency details from the refined image to produce the final SR image. On publicly available datasets, our approach achieves at least 46.9% improvement in LPIPS compared to various state-of-the-art methods. Our codebase is available at https://github.com/nature0012300/SAFE-Diff.git.

Keywords: CT Super-Resolution · Diffusion Models · Stationary Wavelet Transform.

1 Introduction

In CT super-resolution, while perceptual realism is desirable, preserving anatomical structure is equally critical. However, there exists an inherent trade-off between distortion and perceptual quality [3]. The balance between these two ensures that the super-resolved images are both perceptually clear for human interpretation and anatomically precise for reliable diagnostic measurement. Diffusion models [14,19,23] have emerged as powerful tools for super-resolution. Despite their perceptual strength, diffusion models may sacrifice structural fidelity, incur high computational cost, and introduce inconsistencies [17,1].

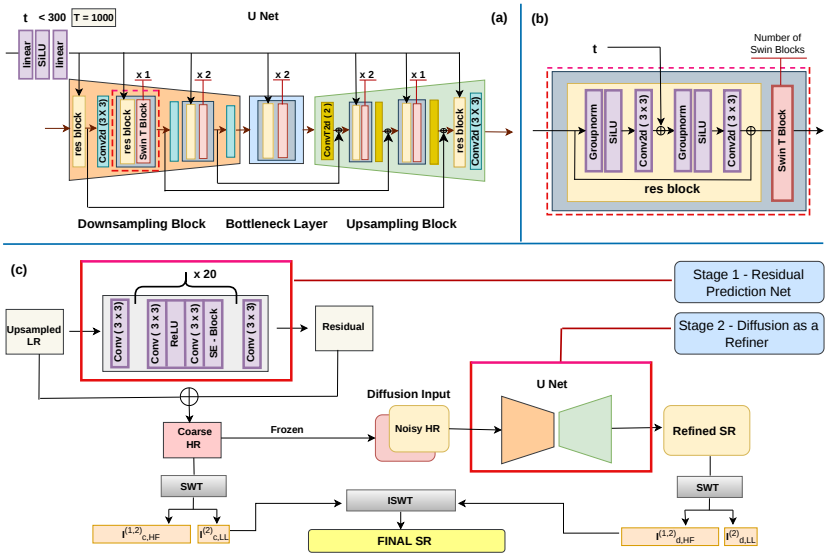

Fig. 1. Architecture of SAFE-Diff. (a) Diffusion U-Net for refinement. (b) Residual Block (res block) within the U-Net where t denotes the timestep. (c) Overall pipeline: Stage 1 predicts the residual map to produce Coarse HR (Upsampled LR + Residual). Stage 2 refines it via truncated diffusion to produce a Refined SR. Both outputs are decomposed via 2-level SWT into low-frequency ($I^{(2)}_{LL}$) and high-frequency ($I^{(1,2)}_{HF}$) sub-bands.The Final SR is reconstructed via ISWT by fusing ($I^{(2)}_{c,LL}$) from Stage 1 with the remaining sub-bands from Stage 2.

We propose SAFE-Diff, a two-stage framework designed to produce anatomically faithful CT images (Fig. 1). Our approach combines the structural stability of residual learning with the generative power of diffusion. Instead of pure pixel-space generation, we adopt a truncated diffusion refinement paradigm. By starting the *diffusion training* process at an intermediate timestep ($t_{max} = 300$) rather than from pure noise ($T = 1000$), we encourage structurally consistent refinement. During inference, sampling starts from a lower truncation point ($t_{inf} = 98$) for fast deterministic refinement. Finally, a Stationary Wavelet Transform (SWT) integrates stable low-frequency structures from Stage 1 with sharp high-frequency details from the Stage 2 refiner.

The core novelty of our work is a principled "structurally anchored" design that explicitly balances the perception-distortion trade-off. To achieve this balance, our work makes the following contributions. **First,** we propose a Stage 1 residual prediction architecture with Squeeze-and-Excitation (SE) [11] blocks. By estimating the structural residual between the ground truth and upsampled input, we achieve an anatomically grounded coarse reconstruction. **Second,** we introduce diffusion as a refiner for stage 1 output using a *truncated inference trajectory* that begins at a mid-point timestep ($t_{inf} = 98$) and completes in 2

DDIM [20] steps. **Third,** we utilize a 2-level SWT fusion strategy using `db2` [5] that preserves anatomical structures from Stage 1 while incorporating high-frequency details from the diffusion refiner. **Fourth,** we propose a lightweight architecture ∼22.5 M parameters that maintains computational efficiency while preserving anatomical fidelity and detail.

2 Methodology

2.1 Preliminaries

We adopt a Denoising Diffusion Probabilistic Models (DDPM) [10] framework. Given a low-resolution image $\mathbf{I}_{LR}$ and ground truth $\mathbf{I}_{GT}$, we define $\mathbf{x}_{up} = \text{Bicubic}(\mathbf{I}_{LR})$ as the initial upsampled baseline for our dual-stage framework. The reverse process is conditioned on the low-resolution input and follows the standard forward diffusion formulation:

$$q(\mathbf{x}_t|\mathbf{I}_{GT}) = \mathcal{N}(\mathbf{x}_t; \sqrt{\bar{\alpha}_t}\mathbf{I}_{GT}, (1-\bar{\alpha}_t)\mathbf{I}_n), \quad \bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s, \tag{1}$$

where $\mathbf{x}_t$ represents the noisy input at timestep t, $\bar{\alpha}_t$ denotes the cumulative noise schedule, $\mathbf{I}_n$ denotes the identity matrix and $\alpha_t = 1$ - β_t.

The reverse process is parameterized by a conditional U-Net trained to predict the added Gaussian noise ϵ at randomly sampled timesteps $t \in [1, t_{max}]$, conditioned on the Stage 1 output. The backward step is defined by the following:

$$\mathcal{L}_{diff} = \mathbb{E}_{\mathbf{I}_{GT},\epsilon,t}\left[\left|\epsilon - \epsilon_\theta(\underbrace{\sqrt{\bar{\alpha}_t}\mathbf{I}_{GT} + \sqrt{1-\bar{\alpha}_t}\epsilon}_{x_t}, t, \mathbf{I}_c)\right|^2\right], \tag{2}$$

where $t \sim \text{Uniform}(1, t_{max})$, $\epsilon \sim \mathcal{N}(0, \mathbf{I})$, ϵ_θ is the predicted noise and $\mathbf{I}_c$ denotes the conditioning input (Stage 1 output).

To facilitate frequency-domain fusion, we utilize a 2-level SWT with the 'db2' filter, a well-established baseline for decomposition of image into sub-bands. Unlike the standard Discrete Wavelet Transform (DWT), which downsamples images and can introduce blocky artifacts due to its shift-variant nature [18]. The shift-invariant SWT preserves spatial dimensions. The 2-level SWT decomposes the image into 8 sub-bands, LL (low-frequency) and LH, HL, HH (high-frequency horizontal, vertical and diagonal) at each level, from coarse structures (Level 2) to fine details (Level 1).

$$f^2_{\text{SWT}}(I) = \left\{I^{(2)}_{LL},\ I^{(2)}_{LH},\ I^{(2)}_{HL},\ I^{(2)}_{HH},\ I^{(1)}_{LL},\ I^{(1)}_{LH},\ I^{(1)}_{HL},\ I^{(1)}_{HH}\right\}. \tag{3}$$

2.2 Overview of Model Architecture

As shown in Fig. 1, our model consists of two main stages followed by a wavelet fusion step to achieve both anatomical fidelity and perceptual sharpness.

Stage 1: Residual Prediction Net This module recovers anatomical structures to estimate the missing high-frequency details from a low-resolution ($\mathbf{I}_{LR}$) input. The network comprises of 20 blocks, where each block integrates a Squeeze-and-Excitation (SE) module for channel-wise attention. SE modules emphasize anatomically relevant features in the predicted residual map $\mathbf{r}_m$. Given a bicubic upsampled input $\mathbf{x}_{up}$, the model predicts a residual map $\mathbf{r}_m$ to produce a 'Coarse HR' image ($\mathbf{I}_c = \mathbf{x}_{up} + \mathbf{r}_m$).

Loss Function: We train this stage using L_1 loss between ground truth residual ($\mathbf{I}_{GT}$ - $\mathbf{x}_{up}$) and predicted residual ($\mathbf{r}_m$).

Stage 2: Truncated Diffusion Refinement This module acts as a perceptual refiner to restore fine textures missing in the Coarse HR ($\mathbf{I}_c$). We employ a denoising U-Net composed of residual and Swin Transformer blocks [15], enabling efficient modeling of long-range dependencies. We initialize refinement from an intermediate timestep $t < t_{max}$. We add noise to the ground truth HR ($\mathbf{I}_{GT}$) to create the starting noisy state $\mathbf{x}_t$ to obtain Refined SR ($\mathbf{I}_d$) as:

$$\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{I}_{GT} + \sqrt{1-\bar{\alpha}_t}\epsilon \tag{4}$$

The U-Net learns to predict the noise ϵ_θ conditioned on the $\mathbf{I}_c$. The input to the Diffusion Refinement Module is noisy state $\mathbf{x}_t$ concatenated channel-wise with $\mathbf{I}_c$. During inference, truncated DDIM sampling is used.

Loss Functions: Stage 2 is optimized using a combination of diffusion, LPIPS ($\mathcal{L}_{lpips}$ with weight λ_1), and reconstruction ($\mathcal{L}_{char}$ with weight λ_2) losses:

$$\mathcal{L}_{stage2} = (1-\lambda_1-\lambda_2)\mathcal{L}_{diff}(\epsilon, \epsilon_\theta) + \lambda_1\mathcal{L}_{lpips}(\mathbf{I}_d, \mathbf{I}_{GT}) + \lambda_2\mathcal{L}_{char}(\mathbf{I}_d, \mathbf{I}_{GT}). \tag{5}$$

As Standard Mean Squared Error (MSE) often results in blurred anatomical boundaries and Mean Absolute Error (MAE) is more robust but lacks differentiability at the origin. We employ the **Charbonnier Loss** [4], a smooth, differentiable approximation of the L_1 norm defined as

$$\mathcal{L}_{\text{char}} = \sqrt{\|\mathbf{I}_d - \mathbf{I}_{GT}\|^2 + \delta^2}, \tag{6}$$

where $\delta = 10^{-3}$. The LPIPS loss encourages perceptual similarity between the synthesized output and high-resolution CT image.

Inference via Truncated DDIM Sampling During inference, we employ a Denoising Diffusion Implicit Model (DDIM) sampler [20].

Initialization: The LR input $\mathbf{I}_{LR}$ is upsampled to the target resolution using bicubic interpolation, followed by the Stage 1 to obtain the coarse image $\mathbf{I}_c$. Instead of sampling from pure noise, diffusion is initialized by injecting controlled noise into $\mathbf{I}_c$ at a truncated timestep $t_{\text{inf}} = 98$.

$$\mathbf{x}'_{t_{inf}} = \sqrt{\bar{\alpha}_{t_{inf}}}\mathbf{I}_c + \sqrt{1-\bar{\alpha}_{t_{inf}}}\epsilon, \quad \epsilon \sim \mathcal{N}(0, \mathbf{I}) \tag{7}$$

Deterministic Denoising: Starting from $\mathbf{x}'_{t_{inf}}$ to obtain $\mathbf{I}_d$ (x'_{t_o}) , we perform 2 inference steps. At each step i, the model predicts the noise component ϵ_θ

conditioned on $\mathbf{I}_c$. The update follows the deterministic DDIM rule ($\eta = 0$):

$$\mathbf{x}'_{t_{i-1}} = \sqrt{\bar{\alpha}_{t_{i-1}}} \underbrace{\left(\frac{\mathbf{x}'_{t_i} - \sqrt{1-\bar{\alpha}_{t_i}}\epsilon_\theta}{\sqrt{\bar{\alpha}_{t_i}}} \right)}_{\text{Predicted } \mathbf{I}_0} + \sqrt{1-\bar{\alpha}_{t_{i-1}}}\epsilon_\theta \tag{8}$$

Frequency Fusion via SWT/ISWT We conclude the process with a SWT fusion. The final output I_f is reconstructed via Inverse SWT (ISWT). This preserves stable anatomy in the low-frequency band while recovering fine textures in the high-frequency bands. This yields a perceptually sharp yet anatomically faithful reconstruction.

3 Experiments and Results

Dataset: We evaluate SAFE-Diff on 2D slices extracted from the official 3D training CT volumes of the LiTS [2] and KiTS [7,8] datasets. The volumes are clipped to HU range [-1024, 1024] and normalized to 8-bit range [0, 255]. To strictly prevent data leakage, all splits are patient-disjoint. This partitioning yielded 43,889/ 4,476/ 10,301 (train/validation/test) from 130 LiTS volumes, and 53,718/ 4,197/ 13,338 (train/validation/test) from 210 KiTS volumes. For $\times 4$ super-resolution, the model takes 128×128 low-resolution (LR) slices as input and predicts 512×512 high-resolution (HR) slices as output. The LR slices are obtained by bicubic downsampling of the original 512×512 HR slices.

Implementation Details: The framework is implemented in PyTorch on an NVIDIA A6000 GPU with FP16 mixed precision. Stage 1 uses Adam (lr $= 1 \times 10^{-4}$) with L_1 loss. Stage 2 employs a U-Net augmented with Swin Transformer blocks (8×8 shifted-window attention), a four-level hierarchy with channels scaling from 64 to 512, attention heads (0, 4, 8, 16), and 128-dimensional sinusoidal time embeddings projected via MLP into residual blocks. The diffusion process uses a cosine noise schedule with 1000 timesteps, training restricted to $t \in [0, 300]$, with Stage 1 frozen. Stage 2 is optimized with AdamW (lr $= 1.2 \times 10^{-4}$, weight decay $= 0.01$) using a composite loss of diffusion MSE, LPIPS ($\lambda_1 = 0.35$), and Charbonnier ($\lambda_2 = 0.40$), with a 5-epoch linear warmup followed by cosine annealing. All hyperparameters, including the diffusion truncation levels used during training ($t_{max} = 300$) and inference ($t_{inf} = 98$), were selected via Bayesian search to optimize validation performance. Table 1 summarizes the computational efficiency of all diffusion-based methods.

Table 1. Model efficiency comparison among diffusion-based methods under $\times 4$ upsampling. Inference time is measured per 512×512 CT slice. The best results are highlighted in bold.

Method	ResShift	Diwa	SRDiff	SinSR	SAFE-Diff
Parameters (M)	116.69	91.89	**10.84**	116.69	22.65
Time (ms)↓	537	12618	2572	**149**	236

Table 2. Quantitative comparison under ×4 super-resolution on LiTS and KiTS. We report mean ± standard deviation over the test images.

Data	Method	LPIPS↓	PSNR↑	SSIM↑	FID↓
LiTS	BiCubic	0.258 ± 0.040	32.70 ± 1.94	0.90 ± 0.031	78.02
	VDSR	0.124 ± 0.059	**40.79 ± 2.56**	**0.95 ± 0.028**	22.39
	ResShift	0.049 ± 0.019	37.38 ± 1.48	0.87 ± 0.030	**3.61**
	Diwa	0.069 ± 0.033	38.89 ± 1.99	0.93 ± 0.030	13.75
	SRDiff	0.068 ± 0.021	39.85 ± 2.77	0.94 ± 0.037	7.75
	SinSR	0.050 ± 0.017	36.83 ± 1.39	0.87 ± 0.029	14.50
	SAFE-Diff (Ours)	**0.026 ± 0.012**	40.36 ± 2.75	0.94 ± 0.035	4.80
KiTS	BiCubic	0.225 ± 0.042	32.88 ± 1.94	0.92 ± 0.022	70.23
	VDSR	0.094 ± 0.046	**42.00 ± 2.02**	**0.97 ± 0.016**	17.68
	ResShift	0.041 ± 0.015	38.21 ± 1.35	0.88 ± 0.031	3.93
	Diwa	0.067 ± 0.074	39.38 ± 2.12	0.94 ± 0.050	12.81
	SRDiff	0.059 ± 0.020	41.31 ± 2.17	0.96 ± 0.020	6.17
	SinSR	0.041 ± 0.014	37.82 ± 1.34	0.89 ± 0.030	9.59
	SAFE-Diff (Ours)	**0.023 ± 0.022**	41.71 ± 2.11	0.96 ± 0.020	**3.71**

Evaluation Metrics: We evaluate super-resolution performance using PSNR [6] and SSIM [22] to assess reconstruction fidelity, LPIPS [24] and FID [9] to measure perceptual quality and distributional similarity. Together, these metrics provide a balanced evaluation of both pixel-level accuracy and visual realism.

3.1 Quantitative and Qualitative Results

We quantitatively compare SAFE-Diff with Bicubic, VDSR [13], SRDiff [14], ResShift [23], SinSR [21] and Diwa [16]. To ensure a fair comparison, all baseline models were trained using the same preprocessing pipeline and data splits. As shown in Table 2, SAFE-Diff achieves the best LPIPS score (0.026) while maintaining competitive PSNR and SSIM values. While VDSR attains the highest PSNR (40.79), it exhibits poorer perceptual quality (LPIPS 0.124, FID 22.39), indicating over-smoothed reconstructions. In contrast, SAFE-Diff achieves a comparable PSNR (40.36 dB) while reducing LPIPS by 79% and FID by 17.59 relative to VDSR. Among diffusion-based methods, SinSR shows weaker reconstruction fidelity, while Diwa provides moderate performance across metrics. Compared to diffusion-based baselines like SRDiff and ResShift, SAFE-Diff improves both PSNR and LPIPS simultaneously and balances the trade-off [3]. Similar trends are observed on the KiTS dataset Table 2, confirming the robustness of the proposed framework.

Visual Comparison: As illustrated by visual results in Fig. 2, a clear divergence is observed between pixel-wise fidelity and perceptual clarity. While baseline models like VDSR preserve global anatomical structure, they produce blurry edges. In contrast, perception-driven models often look sharp but may introduce structural deviations. Our approach bridges this gap by producing high-frequency sharpness without introducing the structural artifacts.

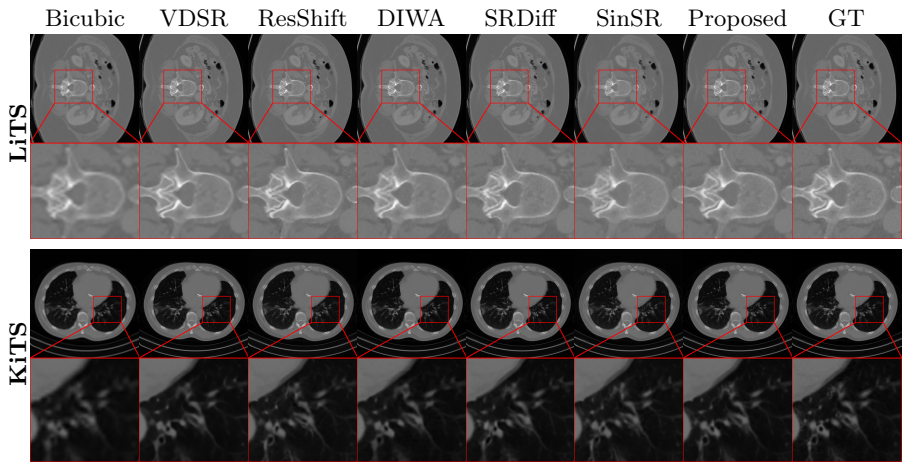

Fig. 2. Qualitative comparison of ×4 super-resolution on LiTS and KiTS datasets. From left to right: Bicubic, VDSR, ResShift, DIWA, SRDiff, SinSR, Proposed (Ours), and Ground Truth (GT).

Ablation Studies: We evaluate the contribution of each module in Table 3. Stage 1 alone achieves a PSNR (41.20), surpassing the best distortion-based baseline VDSR (40.78 dB). This highlights the strong structural reconstruction capability of the proposed module. The slight reduction in PSNR after diffusion refinement reflects a controlled trade-off for perceptual enhancement. Although, stage 1 and stage 2 individually improve distortion and perceptual metrics, their integration with SWT fusion yields the most balanced structural fidelity and perceptual realism. It yields a marginal improvement in PSNR compared to not using the SWT fusion, while maintaining a nearly identical LPIPS. However, a slight degradation in FID indicates that performance gains are not uniform across all metrics.

Table 3. Ablation study of the proposed architecture under ×4 upsampling on LiTS and KiTS. Checkmarks (✓) denote inclusion of architectural components. Best results per dataset are highlighted in bold.

Dataset	Stage 1	Stage 2	SWT Fusion	LPIPS↓	PSNR↑	SSIM↑	FID↓
LiTS	✓	–	–	0.120 ± 0.058	$\mathbf{41.20 \pm 2.68}$	$\mathbf{0.95 \pm 0.030}$	20.69
	–	✓	–	0.034 ± 0.012	38.81 ± 2.83	0.94 ± 0.035	5.73
	–	✓	✓	0.052 ± 0.015	37.05 ± 2.43	0.93 ± 0.034	9.80
	✓	✓	–	$\mathbf{0.025 \pm 0.011}$	40.20 ± 2.74	0.94 ± 0.036	**3.18**
	✓	✓	✓	0.026 ± 0.012	40.36 ± 2.75	0.94 ± 0.035	4.80
KiTS	✓	–	–	0.091 ± 0.046	$\mathbf{42.60 \pm 2.13}$	$\mathbf{0.97 \pm 0.016}$	17.49
	–	✓	–	0.033 ± 0.017	39.27 ± 2.19	0.95 ± 0.021	5.69
	–	✓	✓	0.050 ± 0.015	37.29 ± 2.05	0.95 ± 0.021	8.85
	✓	✓	–	$\mathbf{0.022 \pm 0.021}$	41.55 ± 2.10	0.96 ± 0.020	**2.64**
	✓	✓	✓	0.023 ± 0.022	41.71 ± 2.11	0.96 ± 0.020	3.71

Table 4. Cross-dataset evaluation under ×4 upsampling for both training directions. Best values are highlighted in bold.

Method	LPIPS↓	PSNR↑	SSIM↑	FID↓
Train: KiTS → Test: LiTS				
BiCubic	0.258 ± 0.040	32.70 ± 1.94	0.90 ± 0.031	78.02
VDSR	0.129 ± 0.056	$\mathbf{38.90 \pm 1.89}$	$\mathbf{0.95 \pm 0.028}$	23.47
ResShift	0.060 ± 0.023	36.37 ± 1.26	0.87 ± 0.032	5.66
Diwa	0.094 ± 0.040	37.42 ± 1.71	0.93 ± 0.025	21.07
SRDiff	0.079 ± 0.025	38.31 ± 2.09	0.94 ± 0.033	9.32
SinSR	0.061 ± 0.030	35.95 ± 1.19	0.87 ± 0.028	18.62
SAFE-Diff (Ours)	$\mathbf{0.038 \pm 0.012}$	38.78 ± 2.00	0.94 ± 0.033	**5.37**
Train: LiTS → Test: KiTS				
BiCubic	0.225 ± 0.042	32.88 ± 1.94	0.92 ± 0.022	70.23
VDSR	0.099 ± 0.045	$\mathbf{39.87 \pm 1.93}$	$\mathbf{0.96 \pm 0.017}$	18.21
ResShift	0.053 ± 0.019	36.75 ± 1.37	0.88 ± 0.030	7.96
Diwa	0.071 ± 0.073	38.20 ± 1.78	0.94 ± 0.050	14.13
SRDiff	0.069 ± 0.027	39.35 ± 1.91	0.96 ± 0.020	7.90
SinSR	0.057 ± 0.019	36.23 ± 1.59	0.88 ± 0.031	19.39
SAFE-Diff (Ours)	$\mathbf{0.051 \pm 0.024}$	39.42 ± 1.87	0.95 ± 0.022	**3.29**

Cross-Dataset Generalization Experiment: We evaluate cross-dataset generalization across LiTS and KiTS (Table 4). The model is trained on one dataset and is directly evaluated on the other. In both transfer directions, SAFE-Diff maintains superior perceptual fidelity while remaining competitive in distortion-based metrics. These results show that the model generalizes across datasets without retraining.

Downstream Segmentation Validation: To further assess anatomical fidelity, we conducted a downstream organ segmentation experiment on the LiTS and KiTS datasets using nnU-Net v2 [12]. The segmentation model was trained exclusively on original high-resolution images. To perform the downstream task, we evaluated the trained model on both ground truth images and the images super-resolved by SAFE-Diff. Performance was measured using the Dice Similarity Coefficient (DSC) on 2D slices as shown in Table 5. The visual results

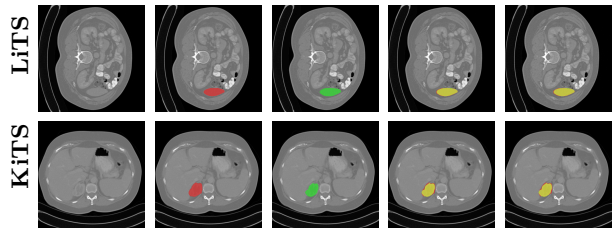

Fig. 3. Qualitative downstream segmentation results using nnU-Net v2 on LiTS and KiTS. Columns show: (1) SAFE-Diff super-resolved image, (2) predicted mask (red), (3) ground-truth mask (green), (4) overlay on the SR image, and (5) overlay on the original image, where yellow indicates overlap. The strong spatial agreement demonstrates preservation of anatomically relevant structures.

Table 5. Downstream organ segmentation results using nnUNet-v2 (mean ± standard deviation of Dice score).

Input Type	LiTS	KiTS
Ground Truth Images	$\mathbf{0.970 \pm 0.123}$	$\mathbf{0.969 \pm 0.132}$
Super-resolved (SAFE-Diff)	0.969 ± 0.132	0.963 ± 0.156

of organ segmentation (Fig. 3) show that super-resolved images remain nearly identical to those obtained on real images.

4 Conclusion

We present SAFE-Diff, a two-stage framework utilizing diffusion as a dedicated refiner for CT super-resolution. It focuses on enhancing fine details while mitigating hallucination risks and preserving anatomical integrity. The proposed SWT-based frequency fusion ensures that stable low-frequency structures are retained while high-frequency details are recovered. The resulting super-resolved images closely match the ground truth and remain anatomically faithful, as validated by SOTA metrics and near-identical Dice scores in downstream segmentation tasks. Future work includes arbitrary-scale and 3D volumetric SR, modality-specific wavelets for CT/MRI, and extension to other imaging modalities.

Acknowledgments. This project has been supported by Srijan - Centre for Generative AI by Meta and IndiaAI.

Disclosure of Interests. The authors declare that they have no conflict of interest.

References

1. Bhadra, S., Kelkar, V.A., Brooks, F.J., Anastasio, M.A.: On hallucinations in tomographic image reconstruction. IEEE transactions on medical imaging **40**(11), 3249–3260 (2021)
2. Bilic, P., Christ, P., Li, H.B., Vorontsov, E., Ben-Cohen, A., Kaissis, G., Szeskin, A., Jacobs, C., Mamani, G.E.H., Chartrand, G., et al.: The liver tumor segmentation benchmark (lits). Medical image analysis **84**, 102680 (2023)
3. Blau, Y., Michaeli, T.: The perception-distortion tradeoff. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6228–6237 (2018)
4. Charbonnier, P., Blanc-Féraud, L., Aubert, G., Barlaud, M.: Deterministic edge-preserving regularization in computed imaging. IEEE Transactions on image processing **6**(2), 298–311 (1997)
5. Daubechies, I.: Orthonormal bases of compactly supported wavelets. Communications on pure and applied mathematics **41**(7), 909–996 (1988)
6. Gonzalez, R.C., Woods, R.E.: Digital Image Processing. Prentice Hall, 3rd edn. (2008)
7. Heller, N., Isensee, F., Maier-Hein, K.H., Hou, X., Xie, C., Li, F., Nan, Y., Mu, G., Lin, Z., Han, M., et al.: The state of the art in kidney and kidney tumor segmentation in contrast-enhanced ct imaging: Results of the kits19 challenge. Medical Image Analysis p. 101821 (2020)

8. Heller, N., Sathianathen, N., Kalapara, A., Walczak, E., Moore, K., Kaluzniak, H., Rosenberg, J., Blake, P., Rengel, Z., Oestreich, M., et al.: The kits19 challenge data: 300 kidney tumor cases with clinical context, ct semantic segmentations, and surgical outcomes. arXiv preprint arXiv:1904.00445 (2019)
9. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium (2018)
10. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
11. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7132–7141 (2018)
12. Isensee, F., Wald, T., Ulrich, C., Baumgartner, M., Roy, S., Maier-Hein, K., Jaeger, P.F.: nnu-net revisited: A call for rigorous validation in 3d medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 488–498. Springer (2024)
13. Kim, J., Lee, J.K., Lee, K.M.: Accurate image super-resolution using very deep convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1646–1654 (2016)
14. Li, H., Yang, Y., Chang, M., Chen, S., Feng, H., Xu, Z., Li, Q., Chen, Y.: Srdiff: Single image super-resolution with diffusion probabilistic models. Neurocomputing **479**, 47–59 (2022)
15. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10012–10022 (2021)
16. Moser, B.B., Frolov, S., Raue, F., Palacio, S., Dengel, A.: Waving goodbye to low-res: A diffusion-wavelet approach for image super-resolution. In: 2024 International Joint Conference on Neural Networks (IJCNN). pp. 1–8. IEEE (2024)
17. Pfaff, L., Wagner, F., Vysotskaya, N., Thies, M., Maul, N., Mei, S., Wuerfl, T., Maier, A.: No-new-denoiser: A critical analysis of diffusion models for medical image denoising. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 568–578. Springer (2024)
18. Rockinger, O.: Image sequence fusion using a shift-invariant wavelet transform. In: Proceedings of international conference on image processing. vol. 3, pp. 288–291. IEEE (1997)
19. Saharia, C., Ho, J., Chan, W., Salimans, T., Fleet, D.J., Norouzi, M.: Image super-resolution via iterative refinement. IEEE transactions on pattern analysis and machine intelligence **45**(4), 4713–4726 (2022)
20. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)
21. Wang, Y., Yang, W., Chen, X., Wang, Y., Guo, L., Chau, L.P., Liu, Z., Qiao, Y., Kot, A.C., Wen, B.: Sinsr: diffusion-based image super-resolution in a single step. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 25796–25805 (2024)
22. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing **13**(4), 600–612 (2004)
23. Yue, Z., Wang, J., Loy, C.C.: Resshift: Efficient diffusion model for image super-resolution by residual shifting. Advances in Neural Information Processing Systems **36**, 13294–13307 (2023)

24. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018)