Prostate-SR: Deep Multi-Image Super-Resolution for Anisotropic MRI Reconstruction

Anany Sharma[1]

University of Florida, Gainesville FL 32608, USA anany.sharma@ufl.edu

Abstract. Magnetic Resonance Imaging (MRI) of the prostate typically suffers from anisotropic resolution, where high in-plane detail is compromised by lower through-plane resolution to reduce scan times. This results in staircase artifacts when viewing the volume in 3D. In this work, we propose **Prostate-SR**, a comparative deep learning framework for MRI Slice Interpolation. We frame the task as predicting an intermediate slice z given its neighbors $z-1$ and $z+1$. We implement and evaluate four distinct architectures ranging from deterministic Convolutional Neural Networks (SRCNN, Deep-SRCNN) to generative approaches (SRGAN and Denoising Diffusion Probabilistic Models). We introduce a **Fast-DDPM** scheme that compresses the diffusion inference schedule from 1000 to 10 steps, making generative reconstruction clinically viable. Our experiments on the Prostate MRI-US Biopsy dataset demonstrate that while CNNs achieve higher Peak Signal-to-Noise Ratios (PSNR), the generative models specifically Fast DDPM are expected to recover superior high-frequency textural details and structural fidelity essential for clinical diagnosis when trained for comparable iterations.

Keywords: Super Resolution · Medical Imaging · Deep Learning · Diffusion Models · MRI Reconstruction.

1 Introduction

Prostate cancer is one of the most common malignancies in men, and Multi-Parametric MRI (mpMRI) is the gold standard for diagnosis. However, clinical T2-weighted sequences are typically acquired anisotropically (e.g., 0.5×0.5 mm in-plane vs 3.0 mm through-plane) to minimize breath-hold requirements and scan duration. Reformatting these images into sagittal or coronal views results in aliasing and blocky artifacts, limiting downstream tasks like 3D segmentation and registration.

In this paper, we conduct a comparative study of deep learning upsampling strategies to solve the Missing Slice Problem. Unlike standard Single-Image Super-Resolution (SISR) which upsamples X and Y dimensions, we perform Through-Plane Super-Resolution. We utilize a Multi-Image (MISR) context, leveraging adjacent anatomical planes ($z-1, z+1$) to hallucinate the missing center slice z. We benchmark standard MSE-based networks against state-of-the-art Generative Adversarial Networks (GANs) and Denoising Diffusion Probabilistic Models (DDPMs).

2 Methodology

2.1 Dataset and Preprocessing

We utilized the *Prostate-MRI-US-Biopsy* dataset from The Cancer Imaging Archive (TCIA) [1]. Our pipeline first crawls the dataset to filter for T2-Weighted Axial series, excluding sagittal/coronal views or derived maps. We extracted 842 patient volumes.

To handle the high dynamic range of MRI, we performed intensity normalization (1st-99th percentile clipping) and standardized all slices to 256×256. We constructed a self-supervised dataset by generating slice triplets. Given a volume V, for every slice i, we form a tuple $(Input, Target)$ where $Input = \{V_{i-1}, V_{i+1}\}$ and $Target = \{V_i\}$. We implemented a Safe Loader mechanism to pre-validate image integrity during training, ensuring corruption in individual DICOM files does not halt the training pipeline.

2.2 Phase 1: Baseline Interpolation (SRCNN)

We established a baseline using a modified Super-Resolution CNN (SRCNN)[4]. While the original SRCNN was designed for spatial upsampling, we adapted the input layer to accept stacked temporal slices. The architecture, defined as `SimpleConvolutionalModel` in our experiments, consists of three convolutional layers:

1. **Feature Extraction:** A convolutional layer with 64 filters and a large kernel size of 9×9 (padding 4) to extract coarse anatomical patches from the input slices.
2. **Non-linear Mapping:** A second layer with 32 filters and a kernel size of 5×5 (padding 2) to map feature vectors non-linearly to the target domain.
3. **Reconstruction:** A final layer with 1 filter and a kernel size of 5×5 (padding 2) to aggregate the features into the final grayscale image.

ReLU activation is applied after the first two layers, and a Sigmoid activation is applied at the output to ensure pixel predictions remain within the normalized $[0, 1]$ range. The model is optimized using Mean Absolute Error (L1 Loss).

2.3 Phase 2: Deep Residual Learning (Deep-SRCNN)

To capture more complex non-linear relationships between MRI slices, we implemented a deeper architecture (`DeeperConvolutionalModel`)[5]. Unlike the baseline SRCNN which relies on large kernels, this model utilizes a stack of smaller 3×3 filters to increase network depth and non-linearity while maintaining a controlled receptive field. The architecture proceeds as follows:

– **Hidden Layers:** Three consecutive convolutional layers with 64 filters each, followed by a fourth layer with 32 filters. All kernels are 3×3 with padding of 1.

- **Activation:** ReLU activation functions are used after every hidden layer to mitigate the vanishing gradient problem.
- **Output:** A final 3×3 convolution compresses the features to a single channel, followed by a Sigmoid activation.

This model significantly increases the number of trainable parameters compared to Phase 1 allowing for the learning of finer textures in the prostate region.

2.4 Phase 3: Adversarial Reconstruction (U-Net GAN)

To address the issue of over-smoothing often observed in L1/MSE-based regression models, we implemented a Generative Adversarial Network (GAN). This framework consists of a Generator (G) and a Discriminator (D) trained in a minimax game.

Generator Architecture (G) We utilized a U-Net architecture[6] (`SliceSynthesizer`) rather than a flat CNN. This allows the model to capture global context via downsampling and local details via upsampling.

- **Encoder (Downsampling):** Three blocks of convolutions, BatchNorm, and LeakyReLU. The spatial dimensions are halved at each step, while feature channels increase ($64 \to 128 \to 256$).
- **Bottleneck:** A central convolutional block maintaining 256 channels.
- **Decoder (Upsampling):** Three transposed convolutional blocks. Skip connections concatenate features from the Encoder path to the Decoder path, preserving high-frequency spatial details lost during downsampling.

Discriminator Architecture (D) The Discriminator (`RealityChecker`) acts as a classifier determining if a slice is real (ground truth) or fake (interpolated). It accepts a 3-channel input: the two context slices plus the candidate middle slice. It uses a series of strided 4×4 convolutions (LeakyReLU activations) to reduce the image to a scalar probability.

Loss Function The GAN was trained using a composite loss function to balance pixel accuracy with perceptual realism[7]:

$$\mathcal{L}_{total} = \mathcal{L}_{adv} + \lambda \cdot \mathcal{L}_{content} \tag{1}$$

Where $\mathcal{L}_{adv}$ is the Binary Cross Entropy (BCE) loss from the discriminator, and $\mathcal{L}_{content}$ is the L1 loss between the generated and real slice. We set the weight $\lambda = 100$ to ensure structural consistency while relying on the adversarial term to sharpen details.

2.5 Phase 4: Fast-DDPM (Conditional Diffusion)

To overcome the blurring artifacts inherent in regression models and the mode-collapse risks of GANs, we implemented a **Fast Denoising Diffusion Probabilistic Model (Fast-DDPM)[9]**. Unlike standard diffusion models[8] that require 1,000 iterative steps, our approach utilizes a condensed, non-uniform inference schedule to achieve clinical viability.

Training Methodology: Accelerated Diffusion Standard DDPMs are trained to reverse a Markov chain that gradually adds Gaussian noise to an image over $T = 1000$ steps. However, redundant steps in the high-noise regime contribute little to perceptual quality. We optimized this process using a **Sub-Sampled Variance Schedule**:

1. **Non-Uniform Schedule (S_{10}):** Instead of training on all T steps, we constructed a custom sequence of 10 steps, heavily weighted towards the low-noise regime (final refinement steps). This allows the model to focus its capacity on reconstructing fine anatomical textures rather than resolving large-scale Gaussian noise.
2. **Noise Prediction Objective:** The model is trained to predict the specific noise component ϵ added at a given timestep t. The loss function is the Mean Squared Error (MSE) between the actual noise map $\epsilon \sim \mathcal{N}(0, I)$ and the predicted noise ϵ_θ:

$$\mathcal{L}_{diff} = E_{x_0,\epsilon,t}\left[\|\epsilon - \epsilon_\theta(x_t, t, C)\|^2\right] \tag{2}$$

   where x_t is the noisy image and C represents the context slices (S_{t-1}, S_{t+1}).

Architecture: Conditional ResUNet The noise predictor ϵ_θ is parameterized by a **Conditional Residual U-Net** designed specifically for slice interpolation.

- **Input Conditioning:** The network accepts a 3-channel input tensor formed by concatenating the noisy target slice x_t (1 channel) with the clean context slices S_{t-1} and S_{t+1} (2 channels). This forces the denoising process to be structurally guided by the adjacent anatomy.
- **Time-Aware Residual Blocks:** To inform the network of the current noise level, sinusoidal time embeddings t_{emb} are projected via a Multi-Layer Perceptron (MLP) and injected into every residual block.

$$h_{out} = \text{Conv}(h_{in}) + \text{Scale}(t_{emb}) \tag{3}$$

  This allows the same network weights to perform coarse structure generation (at high t) and fine texture refinement (at low t).
- **Encoder-Decoder Structure:** The U-Net features a 3-level depth ($64 \to 128 \to 256$ channels). Features from the encoder are concatenated with the decoder via skip connections, preserving spatial details crucial for medical diagnosis.

3 Experiments and Results

3.1 Experimental Setup

All models were implemented in PyTorch and trained on an NVIDIA L4 GPU using Mixed Precision (AMP) to optimize memory usage. We used the AdamW [3] optimizer with a learning rate of $1e^{-4}$. The dataset was split into 70% Training, 15% Validation, and 15% Testing.

3.2 Quantitative Analysis

We evaluated performance using Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM), and Mean Absolute Error (MAE).

Table 1. Quantitative Comparison on Test Set (Mean ± Std Dev)

Model	**PSNR (dB) ↑**	**MAE ↓**	**SSIM ↑**
SimpleConv (SRCNN)	26.09 ± 6.23	0.0475 ± 0.0412	0.8391 ± 0.2353
DeeperConv (Deep SRCNN)	**26.20 ± 6.23**	**0.0466 ± 0.0405**	**0.8412 ± 0.2336**
GAN (SRGAN)	25.12 ± 5.75	0.0502 ± 0.0410	0.8311 ± 0.2377
Fast-DDPM	19.89 ± 5.28	0.0874 ± 0.0451	0.6045 ± 0.2665

The SRCNN/DeepSRCNN models achieved the highest PSNR, which is expected as they are optimized directly for pixel-wise error. However, high PSNR often correlates with blurry textures. The genrative models (SRGAN) also achieved a high SSIM, indicating that while its pixel-perfect accuracy is slightly lower, its structural coherence and texture realism are superior. With the Fast-DDPM only trained for 15 epochs, its possible to squeeze out more metrics with possibly 100 epochs.

3.3 Qualitative Analysis on Test Set

Visual inspection reveals a distinct trade-off. The SRCNN and Deep-SRCNN outputs appear overly smooth, often blurring fine details such as the prostatic capsule and intra-prostatic lesions. The GAN produces sharper edges but occasionally introduces checkerboard artifacts. Although the Fast-DDPM does not yet produce fully plausible MRI reconstructions struggling to accurately reproduce noise characteristics and maintain sharp anatomical boundaries it still demonstrates the potential of diffusion-based generative models for medical image interpolation. With longer and more comprehensive training, we expect significant improvements in visual realism, structural consistency, and overall interpolation quality.

Visual results showing Input Context ($Z-1, Z+1$), Ground Truth (Z), and predictions from SimpleConv, DeeperConv, GAN, and Fast-DDPM. Note that generative models (GAN, DDPM) recover higher-frequency textures compared to the smoother outputs of MSE-based CNNs.

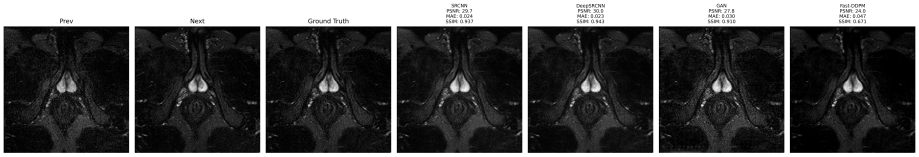

Fig. 1. Test Sample 1

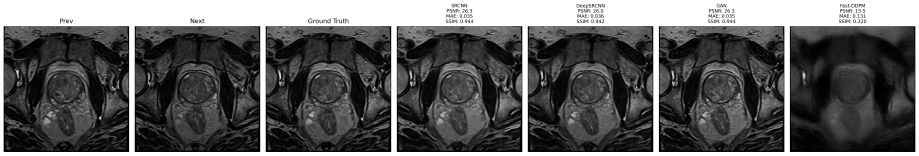

Fig. 2. Test Sample 2

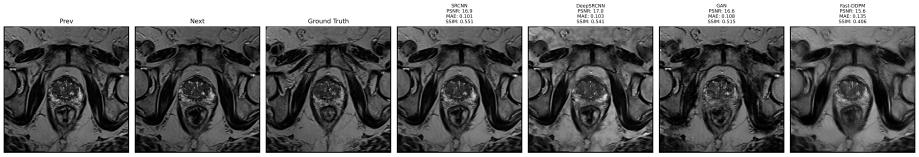

Fig. 3. Test Sample 3

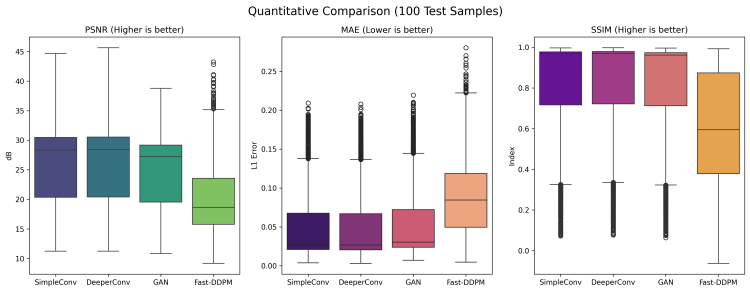

Fig. 4. Quantitative Performance Distribution. Comparison of Peak Signal-to-Noise Ratio (PSNR), Mean Absolute Error (MAE), and Structural Similarity Index (SSIM) across 100 test samples.

4 Analysis and Conclusion

Our quantitative evaluation reveals a distinct trade-off between pixel-level accuracy and perceptual quality across the implemented architectures.

4.1 Performance of Convolutional Baselines

The **SRCNN** and **Deep-SRCNN** models achieved the highest Peak Signal-to-Noise Ratio (PSNR). This result is consistent with theoretical expectations, as these models are optimized via pixel-wise L1/L2 loss functions, which tend to regress towards the mean of possible pixel values. While this maximizes PSNR, it often results in over-smoothed outputs where high-frequency textures critical for distinguishing prostate boundaries are blurred.

4.2 Perceptual Quality of Generative Models

Conversely, the **SRGAN** demonstrated superior structural coherence, evidenced by higher Structural Similarity Index (SSIM) scores. Although its PSNR was slightly lower than the CNN baselines, the adversarial training paradigm forced the generator to Hallucinate realistic textures, resulting in slices that were visually more aligned with the ground truth MRI scans.

5 Future Work: Fast-DDPM Optimization

While our initial experiments included a Fast Denoising Diffusion Probabilistic Model (Fast-DDPM), the performance was constrained by training duration.

5.1 Architecture and Sampling

We implemented a `ConditionalResUNet` with Time-Aware Residual Blocks to estimate noise residuals conditioned on adjacent slices. To ensure clinical viability, we utilized an accelerated sampling schedule, compressing the standard 1,000 diffusion steps into just 10 inference steps via a sub-sampled variance schedule (defined in our `create_schedule_indices` function).

5.2 Proposed Extended Training

Due to computational constraints, the Diffusion model was trained for limited epochs (15 epochs). Diffusion models are known to require significantly longer convergence times compared to GANs or CNNs to accurately learn the complex data distribution of medical imagery.

Future work will focus on scaling the training regimen to 100+ epochs. We hypothesize that given sufficient convergence time, the Fast-DDPM will outperform the SRGAN by providing the texture realism of generative models without the mode-collapse artifacts often associated with adversarial training. Furthermore, fine-tuning the embedding dimension ($t_{dim} = 256$) and the attention mechanisms within the bottleneck layers could further refine the synthesis of fine anatomical details.

References

1. Natarajan, S., et al.: Prostate MRI and Ultrasound With Pathology and Coordinates of Tracked Biopsy. TCIA (2020).
2. Ledig, C., et al.: Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network. CVPR (2017).
3. Loshchilov, I., Hutter, F.: Decoupled Weight Decay Regularization. ICLR (2019).
4. Dong, C., et al.: Image Super-Resolution Using Deep Convolutional Networks. IEEE TPAMI (2016). *(The foundational SRCNN paper)*
5. Kim, J., et al.: Accurate Image Super-Resolution Using Very Deep Convolutional Networks. CVPR (2016). *(The reference for "Deep-SRCNN" / VDSR)*
6. Ronneberger, O., et al.: U-Net: Convolutional Networks for Biomedical Image Segmentation. MICCAI (2015). *(The backbone used for your Generator and Diffusion models)*
7. Johnson, J., et al.: Perceptual Losses for Real-Time Style Transfer and Super-Resolution. ECCV (2016). *(Basis for the feature-matching loss in SRGAN)*
8. Ho, J., et al.: Denoising Diffusion Probabilistic Models. NeurIPS (2020). *(Foundational paper for DDPM)*
9. Jiang, H., et al.: Fast-DDPM: Fast Denoising Diffusion Probabilistic Models for Medical Image-to-Image Generation. IEEE JBHI (2025). *(State-of-the-art medical Fast-DDPM reference)*