MSP-SR: Multi-Stage Probabilistic Generative Super Resolution with Scarce High-Resolution Data

Ruike Zhu[†1] **Matthew Weston**[1] **Hanwen Zhang**[1] **Arindam Banerjee**[1]

[1]Siebel School of Computing and Data Science , University of Illinois at Urbana-Champaign , Urbana, Illinois, USA

Abstract

Several application domains, especially in science and medicine, benefit tremendously from acquiring high-resolution images of objects and phenomena of interest. Recognizing this need, generative models for super-resolution (SR) have emerged as a promising approach for such data generation. However, when training data are scarce due to high acquisition costs, such models struggle and often fail to capture the true data distribution due to insufficient data and domain knowledge. While transfer learning, domain adaptation, or few shot learning of such generative models can be a reasonable approach, most existing large scale generative models have been (pre)trained on natural images and it is unclear if such models can be seamlessly transferred to say medical images.
In this paper, we propose Multi-Stage Probabilistic Super Resolution (MSP-SR), a cascaded few-shot learning framework for super-resolution through multi-stage transfer learning. At a high level, MSP-SR first transfers a generative model from out-of-domain to in-domain, e.g., from natural to medical images, and then from in-domain to the target application. We present the details based on conditional diffusion models and validate MSP-SR on multiple Magnetic Resonance Imaging (MRI) datasets, demonstrating that MSP-SR persistently and usually significantly outperforms direct fine-tuning (DFT) approaches as well as SR baselines. Further, MSP-SR empirically provides more accurate characterization of uncertainty in SR compared to DFT.

[†]Corresponding author: ruikez2@illinois.edu

1 INTRODUCTION

Several application domains, especially in science and medicine, e.g., MRI (magnetic resonance imaging) or CT (Computed Tomography) [Greenspan, 2009, Umehara et al., 2018, Zhang and An, 2017], benefit tremendously from acquiring high-resolution images of objects and phenomena of interest. Recognizing this need, generative models for super-resolution (SR), especially diffusion models [Saharia et al., 2022b, Liu et al., 2023, Shang et al., 2024], have emerged as a promising approach for such data generation.

Unlike conventional CNN- [Dong et al., 2015], transformer- [Lu et al., 2022], and GAN-based [Wang et al., 2021b,a, Zhang et al., 2021a] methods, diffusion models excel at handling the fundamentally ill-posed nature of super-resolution problems [Kabanikhin, 2008]. Their generative formulation provides inherent uncertainty awareness and enhanced capacity to model complex data distributions [Wang et al., 2020], making them particularly suited for medical imaging.

However, practical implementation faces significant hurdles in scientific domains. Unlike natural images, medical high-resolution data acquisition demands specialized equipment with substantial time and financial investments, e.g., MRI scanners require advanced hardware and prolonged scan durations, creating patient discomfort and institutional burdens. This scarcity creates a critical paradox: while medical images contain exceptionally complex anatomical structures requiring millimeter-level reconstruction accuracy, data-starved models often develop harmful over-reliance on limited priors. Conventional direct fine-tuning (DFT) approaches in such few-shot scenarios typically produce inaccurately confident predictions, a perilous outcome where hallucinated anatomical details could lead to clinical misdiagnosis.

Towards resolving this challenge, we propose **M**ulti-**S**tage **P**robabilistic **S**uper-**R**esolution (MSP-SR), a cascaded few-shot medical image super-resolution framework based on generative models. To mitigate the constraints imposed

by scarce medical data, we develop a multi-stage learning framework that enables the model to pre-train and extract visual features from abundant natural image domains, subsequently transferring and adapting these learned representations to the medical imaging context. Utilizing SR3 (Super-Resolution via Repeated Refinement) [Saharia et al., 2022b] as our foundational architecture and employing ControlNet [Zhang et al., 2023] to facilitate nuanced transfer learning, we introduce additional training constraints through innovative loss penalties. These mechanism ensure more accurate data generation by maintaining appropriate uncertainty levels, particularly for ambiguous regions where limited training data provides insufficient reconstruction guidance.

Our framework implements a three-stage training methodology that leverages diverse datasets to progressively enhance model specialization. The initial Out-of-Domain (OOD) pre-training stage utilizes SR3 [Saharia et al., 2022b] with the low-resolution COCO dataset [Lin et al., 2014], extracting generalized visual features from diverse natural images. The subsequent In-Domain (ID) stage adapts the model to medical imaging characteristics using low-resolution T2-weighted scans from the IXI dataset [IXI, 2023], which contains MR images from healthy subjects. The final Target-Domain (TD) stage fine-tunes the model on specific low-resolution-high-resolution (LR-HR) pairs from T2-weighted FastMRI [Zbontar et al., 2018] and BrainTumor [Chaki and Wozniak, 2023] datasets, along with T1-weighted OASIS [Marcus et al., 2007] scans. This dataset progression enables precise brain MRI super-resolution while maintaining cross-modality generalization through controlled domain transfer.

Ablation studies validated our framework's effectiveness by examining variants without ControlNet, Out-of-Domain pre-training, In-Domain fine-tuning, and using target data training alone. Quantitative analysis demonstrated our full pipeline's superior performance in detail preservation and training stability during the TD stage. The cascaded transfer learning framework facilitates effective knowledge transfer both across domains (Out-of-Domai to In-Domain) and within domains (In-Domain to Target-Domain), providing a more accurate characterization of uncertainty and improving feature extraction in few-shot scenarios.

In summary, this paper's contributions include:

1. **Novel Multi-stage Learning Framework.** We propose a novel cascaded learning framework that achieves high generation accuracy compared with other SR models under few-shot conditions.

2. **Cross-domain Knowledge Transfer.** Our framework effectively transfers knowledge from natural image domains to medical imaging, enabling robust performance despite limited medical training data.

3. **Uncertainty-aware Generation.** Our framework reduces dependency on limited training data, providing a precise characterization of uncertainty and ensuring more accurate data generation.

2 RELATED WORK

Data-Efficient Super-Resolution. Data-efficient super-resolution has been extensively discussed in medical imaging literature. Greenspan [2009] presents an overview of early methods that required multiple low-resolution images to resolve high-resolution outputs, contrasting with modern learned approaches [Li et al., 2021b] including our own. Early techniques employed iterative back-propagation for consistency checking, while newer diffusion techniques [Song et al., 2023a] allow learning-based SR methods to avoid explicit down-sampling modeling.

Li et al. [2021b] discusses recent deep learning-based super-resolution techniques for medical imaging, where acquiring high-resolution images remains a major bottleneck. Approaches include recursive neural networks to limit parameters [Kim et al., 2016, Tai et al., 2017], GANs for training on small datasets [Wang et al., 2024a, Mansoor et al., 2018, Ensemble] despite training stability concerns, and smaller deep models like U-nets [Park et al., 2018] with additional regularization. Notably, [Ensemble] draws inspiration from [Zhu et al., 2017] to improve consistency and reduce hallucination.

Diffusion-Based Super-Resolution. The advent of diffusion models has created significant advances in super-resolution techniques. Palette [Saharia et al., 2022a] and SR3 [Saharia et al., 2022b] pioneered the application of diffusion models to image restoration tasks. Training-free methods like DPS [Chung et al., 2022] extend traditional diffusion models to solve non-linear inverse problems, enabling restoration across diverse real-world corruptions without requiring training. Meanwhile, learning-based methods such as I^2SB [Liu et al., 2023] and SinSR [Wang et al., 2024b] achieve improved performance through model training, where I^2SB develops a score-based framework for direct distribution mapping, and SinSR enables efficient single-step inference through learned diffusion.

Medical Image Quality Assessment. While traditional image quality metrics like PSNR and SSIM provide quantitative measures for super-resolution performance, they may not fully capture anatomical accuracy or clinical utility in medical imaging applications. Recent research has highlighted specialized evaluation methodologies: Zhang et al. Zhang et al. [2021b] proposed task-driven assessment using numerical observers for diagnostic tasks, Kelkar et al. Kelkar et al. [2022] introduced medical image-specific statistical divergence metrics to detect anatomical hallucinations, and Li et al. Li et al. [2021a] developed meth-

ods to analyze covariance structures for identifying unintended smoothing of anatomical textures. These specialized approaches complement traditional metrics by providing deeper insights into clinically relevant feature preservation. While our current work utilizes established metrics for comparison with prior methods, incorporating these specialized medical imaging evaluation protocols represents an important direction for future research to ensure anatomical fidelity in super-resolution enhancement.

3 MULTI-STAGE PROBABILISTIC FRAMEWORK

The proposed training framework implements a three-stage cascading approach to address few-shot learning challenges. The generative model undergoes progressive refinement through out-of-domain (OOD) pre-training, in-domain (ID) fine-tuning, and target-domain (TD) adaptation. This hierarchical strategy enables diffusion models to perform effectively with limited training data by facilitating knowledge transfer across domains. The complete training pipeline is detailed in Algorithm 1, with each stage's methodology and rationale examined in the following sections.

Problem Formulation. Following [Song et al., 2021], we formulate super-resolution as a linear inverse problem to recover unknown signals y from observed measurements x. Given scarce HR target data $\boldsymbol{x}_{\text{gt}} = \{x_i\}_{i=1}^{M}$, the corresponding LR target data $\boldsymbol{y}_{\text{gt}} = \{y_i\}_{i=1}^{M}$ is formulated as $\boldsymbol{y}_{\text{gt}} = \boldsymbol{A}\boldsymbol{x}_{\text{gt}} + \boldsymbol{\eta}$, where $\boldsymbol{A}$ denotes the linear downsampling matrix and $\boldsymbol{\eta}$ represents noise [Song et al., 2021].

To enable coarse-to-fine information flow, we incorporate two additional large-scale datasets: out-of-domain (OOD) data $\boldsymbol{x}_{\text{ood}} = \{x_i\}_{i=1}^{N_1}$ and in-domain (ID) data $\boldsymbol{x}_{\text{id}} = \{x_i\}_{i=1}^{N_2}$, where $N_1 > N_2 \gg M$. A bicubic degradation matrix $\tilde{\boldsymbol{A}}$ generates their LR counterparts: $\tilde{\boldsymbol{y}}_{\text{ood}} = \tilde{\boldsymbol{A}}\boldsymbol{x}_{\text{ood}}$ and $\tilde{\boldsymbol{y}}_{\text{id}} = \tilde{\boldsymbol{A}}\boldsymbol{x}_{\text{id}}$.

3.1 TRAINING STAGES

Low-resolution Out-of-Domain Model Pre-Training. This stage constructs a model for $p(\boldsymbol{x}_{\text{ood}}|\tilde{\boldsymbol{y}}_{\text{ood}})$ using abundant OOD data. We utilize COCO [Lin et al., 2014] data as $\boldsymbol{y}_{\text{ood}}$ with its LR counterparts $\tilde{\boldsymbol{y}}_{\text{ood}}$ to extract coarse-grained features. SR3 [Saharia et al., 2022b] serves as the backbone model, ensuring framework generality across applications.

Low-resolution In-Domain ControlNet Pre-Training. The ID stage leverages IXI [IXI, 2023] brain MRI datasets for model adaptation. We generate LR counterparts $\tilde{\boldsymbol{y}}_{\text{id}} = \tilde{\boldsymbol{A}}\boldsymbol{x}_{\text{id}}$ using downsampling matrix $\tilde{\boldsymbol{A}}$ for the low-resolution image in IXI data. Rather than simple fine-tuning, we integrate ControlNet [Zhang et al., 2023] by connecting the pre-trained diffusion model's U-Net to zero convolutional layers (Fig. 1), enabling simultaneous in-domain knowledge acquisition $p(\boldsymbol{x}_{\text{id}}|\tilde{\boldsymbol{y}}_{\text{id}})$ and OOD information preservation.

High-resolution Target-Domain ControlNet Fine-Tuning. To further fine-tune the system, the final stage aligns ControlNet [Zhang et al., 2023] with distribution $p(\boldsymbol{x}_{\text{gt}}|\boldsymbol{y}_{\text{gt}})$ using HR data $\boldsymbol{x}_{\text{gt}}$ (FastMRI [Zbontar et al., 2018], BrainTumor [Chaki and Wozniak, 2023], OASIS [Marcus et al., 2007]) and corresponding LR data $\boldsymbol{y}_{\text{gt}} = \boldsymbol{A}\boldsymbol{x}_{\text{gt}}$. Theoretically, $\boldsymbol{A}$ represents the true degradation process in medical imaging, which differs from the bicubic downsampling matrix $\tilde{\boldsymbol{A}}$ used in previous stages and is unknown to us. In practice, we approximate this true degradation by using bicubic downsampling to generate the target domain LR data $\boldsymbol{y}_{\text{gt}}$. Our experimental results demonstrate that this bicubic approximation achieves satisfactory performance in modeling the complex medical imaging degradation process.

ControlNet Integration We adopt ControlNet [Zhang et al., 2023] in Out-of-Domain and In-Domain stages to enable efficient domain adaptation while preserving pre-trained knowledge. ControlNet creates a trainable copy of the encoding layers from the pre-trained U-Net, connected through zero-initialized convolutional layers. During training, the original U-Net weights remain frozen, while the ControlNet branch learns domain-specific features. The outputs from both branches are combined via element-wise addition, allowing the model to maintain general super-resolution capabilities from the OOD stage while acquiring medical imaging knowledge in other two stages. This architecture ensures stable training and prevents catastrophic forgetting of previously learned features.

3.2 CONDITIONAL GENERATIVE MODEL (CGM)

Gaussian Diffusion Process. The backbone architecture is illustrated using the target-domain (TD) fine-tuning stage as an example. For an input image pair $\{\boldsymbol{x} : \boldsymbol{x}_{\text{gt}}, \boldsymbol{y} : \boldsymbol{y}_{\text{gt}}\}$, the model generates output $\{\boldsymbol{x}_0 : \boldsymbol{x}_{\text{gt}}\}$ through the reverse diffusion process. The framework follows the conditional diffusion process and optimizes a neural denoising model that receives source image $\boldsymbol{y}$ and noisy target image $\boldsymbol{x}_t$ as inputs to produce the denoised image $\boldsymbol{x}_0$.

Following DDPM [Ho et al., 2020], the unconditional diffusion process progressively adds Gaussian noise to the clean input $\boldsymbol{x}_0$ over T iterations.

$$p(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0) = \prod_{t=1}^{T} p(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}), \tag{1}$$

$$p(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) = \mathcal{N}(\boldsymbol{x}_t|\sqrt{1-\beta_t}\,\boldsymbol{x}_{t-1}, \beta_t\mathbf{I}). \tag{2}$$

where the parameter $\beta_{1:T}$ $(0 < \beta_t < 1)$ determines the variance of the added noise. Let $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{t'=1}^{t} \alpha_{t'}$. The relationship between the noisy image $\boldsymbol{x}_t$

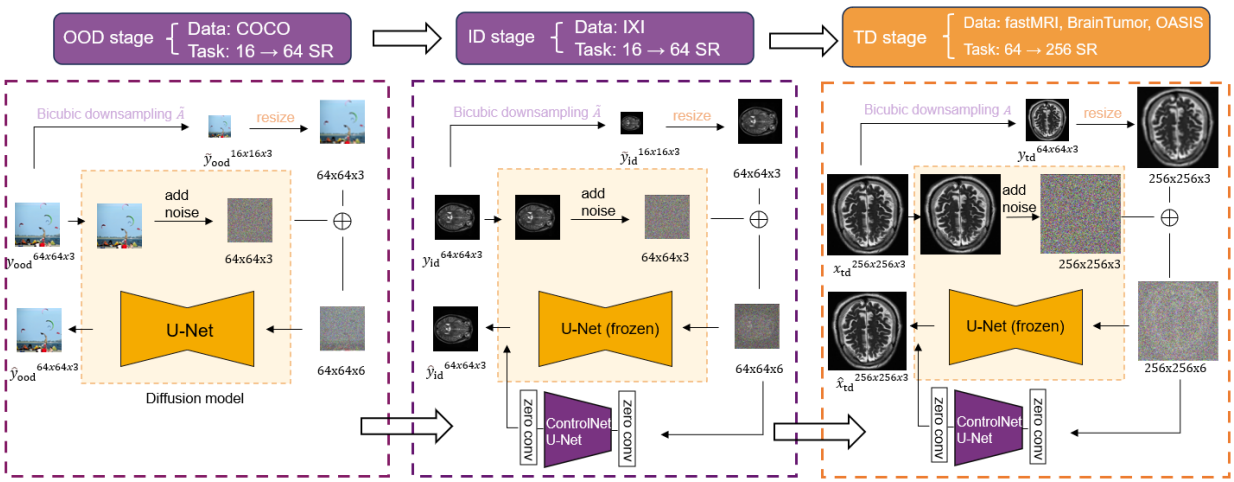

Figure 1: Overview of the MSP-SR framework: a three-stage approach incorporating out-of-domain pre-training on COCO (16→64), ControlNet-assisted in-domain fine-tuning on IXI (16→64), and target-domain adaptation on medical datasets (64→256). Each stage uses bicubic downsampling with progressive resolution and domain transfer from natural to medical images.

and the original image $\boldsymbol{x}_0$ can then be expressed as:

$$\boldsymbol{x}_t = \sqrt{\bar{\alpha}_t}\,\boldsymbol{x}_0 + \sqrt{1-\bar{\alpha}_t}\,\boldsymbol{\epsilon}, \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})\,. \tag{3}$$

Recovering $\boldsymbol{x}_0$ from a Gaussian noise input $\boldsymbol{x}_T$ enables the generation of new samples. Although $p(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$ approximates a Gaussian distribution when the noise variance β_t is sufficiently small, directly estimating $p(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$ remains intractable. Instead, when conditioned on $\boldsymbol{x}_0$, the inverse conditional probability $p(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0)$ becomes tractable as follows:

$$p(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0) = \mathcal{N}(\boldsymbol{x}_{t-1}; \mu(\boldsymbol{x}_t, \boldsymbol{x}_0), \tilde{\beta}_t \mathbf{I}), \tag{4}$$

where

$$\mu(\boldsymbol{x}_t, \boldsymbol{x}_0) = \frac{1}{\sqrt{\alpha_t}}\left(\boldsymbol{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_t\right), \tag{5}$$
$$\tilde{\beta}_t = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\cdot\beta_t.$$

Here, $\mu(\boldsymbol{x}_t, \boldsymbol{x}_0)$ represents the mean of the Gaussian distribution derived from the noisy image $\boldsymbol{x}_t$ and clean image $\boldsymbol{x}_0$, where $\boldsymbol{\epsilon}_t$ is the forward process noise that enables effective denoising through appropriate scaling with α_t and $\bar{\alpha}_t$.

Building upon this foundation, our framework extends to a conditional diffusion model that conditions on the low-resolution input $\boldsymbol{y}$. This conditioning provides crucial prior information during the early stages of the diffusion denoising process, guiding the generation towards semantically consistent outputs. The objective here is to learn a diffusion model q_θ to approximate the inverse conditional probability as follows [Saharia et al., 2022b]:

$$q(\boldsymbol{x}_T) = \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{6}$$

$$q_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{y}) = \mathcal{N}(\boldsymbol{x}_{t-1}; \mu_\theta(\boldsymbol{x}_t, t, \boldsymbol{y}), \tilde{\beta}_t \mathbf{I}), \tag{7}$$

$$q_\theta(\boldsymbol{x}_{0:T}|\boldsymbol{y}) = q(\boldsymbol{x}_T)\prod_{t=1}^{T} q_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{y}). \tag{8}$$

Ideally, we want $\mu_\theta(\boldsymbol{x}_t, t, \boldsymbol{y})$ to output the conditional equivalent of μ in Equation 5. While Equation 5 describes the unconditional case, our conditional framework requires the mean to be guided by the conditioning input $\boldsymbol{y}$. Since $\boldsymbol{x}_t$ is known during inference, we only need to predict the noise term $\boldsymbol{\epsilon}_t$ conditioned on $\boldsymbol{y}$. Therefore, $\mu_\theta(\boldsymbol{x}_t, t, \boldsymbol{y})$ is parameterized via the conditional noise predictor $\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t, \boldsymbol{y})$ as

$$\mu_\theta(\boldsymbol{x}_t, t, \boldsymbol{y}) = \frac{1}{\sqrt{\alpha_t}}\left(\boldsymbol{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t, \boldsymbol{y})\right). \tag{9}$$

3.3 MODEL CONSISTENCY

Drawing inspiration from the consistency model Song et al. [2023b], we enhance input-output correspondence by introducing a consistency loss l_{CON} alongside the standard reconstruction loss l_{GT} used in diffusion models to train $\boldsymbol{\epsilon}_\theta$. While consistency models ensure temporal consistency across different timesteps in the diffusion trajectory, our consistency loss enforces correspondence between the input low-resolution image and the generated high-resolution output through a degradation process.

Following [Ho et al., 2020], the reconstruction loss l_{GT} optimizes the conditional model μ_θ rather than $\boldsymbol{\epsilon}_\theta$ (based on

Equation 9), minimizing a variant of the ELBO with true image x_0 and conditioning y as inputs.

$$l_{\text{GT}} = \mathbb{E}_{t,\boldsymbol{x}_0,\boldsymbol{\epsilon}}\left[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\sqrt{\bar{\alpha}_t}\boldsymbol{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_t, t, \boldsymbol{y})\|^2\right]. \tag{10}$$

For understanding the consistency loss l_{CON}, assume we are at t step now. The denoised clean image of noisy image $\boldsymbol{x}_t$ at t step will be denoted as $\hat{\boldsymbol{x}}_{0,t}$, and can be crudely obtained from Equation 3 as

$$\hat{\boldsymbol{x}}_{0,t} = \frac{\boldsymbol{x_t} - \sqrt{1-\bar{\alpha}}\hat{\boldsymbol{\epsilon}}}{\sqrt{\bar{\alpha}}}, \tag{11}$$

where $\hat{\boldsymbol{\epsilon}} = \epsilon_\theta(\boldsymbol{x_t}, t, \boldsymbol{y})$ is the predicted noise at t step, and $\boldsymbol{x}_t$ is the noisy image of target image $\boldsymbol{x}_{\text{gt}}$.

Then, if we construct $\tilde{\boldsymbol{A}}\hat{\boldsymbol{x}}_{0,t}$, i.e., downsample $\hat{\boldsymbol{x}}_{0,t}$ with the bicubic downsample operator $\tilde{\boldsymbol{A}}$, then we expect that to be close to the LR image $\boldsymbol{y}$, i.e., $\tilde{\boldsymbol{A}}\hat{\boldsymbol{x}}_{0,t} \approx \boldsymbol{y}$. Then, taking expectation over all t, we get the consistency loss and the total combined loss of diffusion model as:

$$l_{\text{CON}} = \mathbb{E}_t[\|\tilde{\boldsymbol{A}}\hat{\boldsymbol{x}}_{0,t} - \boldsymbol{y}\|^2], \tag{12}$$

$$L = \gamma l_{\text{GT}} + (1-\gamma) l_{\text{CON}}. \tag{13}$$

where γ is an adjustable hyperparameter manually set to 0.5 in experiments for convenience, the optimal value can be obtained by parameter search for further study.

Algorithm 1 Training a Denoising Model μ_θ

Input: Datasets: $D_{\text{ood}}(\boldsymbol{y}_{\text{ood}}, \tilde{\boldsymbol{y}}_{\text{ood}})$,
  $D_{\text{id}}(\boldsymbol{y}_{\text{id}}, \tilde{\boldsymbol{y}}_{\text{id}})$,
  $D_{\text{td}}(\boldsymbol{x}_{\text{td}}, \boldsymbol{y}_{\text{td}})$
Output: Trained model μ_θ
for d **in** Datasets **do**
  $\boldsymbol{x} = \boldsymbol{y}_{\text{ood}}, \boldsymbol{y} = \tilde{\boldsymbol{y}}_{\text{ood}}$ **if** $d \in D_{\text{ood}}$
  $\boldsymbol{x} = \boldsymbol{y}_{\text{id}}, \boldsymbol{y} = \tilde{\boldsymbol{y}}_{\text{id}}$ **if** $d \in D_{\text{id}}$
  $\boldsymbol{x} = \boldsymbol{x}_{\text{td}}, \boldsymbol{y} = \boldsymbol{y}_{\text{td}}$ **if** $d \in D_{\text{td}}$
  while not converged **do**
    $t \sim \text{Uniform}(1, \ldots, T)$
    $(\boldsymbol{x_0}, \boldsymbol{y}) \sim p(\boldsymbol{x}, \boldsymbol{y})$
    $\bar{\alpha} \sim p(\bar{\alpha})$
    $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
    $\boldsymbol{x_t} = \sqrt{\bar{\alpha}}\boldsymbol{x_0} + \sqrt{1-\bar{\alpha}}\boldsymbol{\epsilon}$
    $\hat{\boldsymbol{\epsilon}} = \boldsymbol{\epsilon}_\theta(\boldsymbol{x_t}, t, \boldsymbol{y})$
    $l_{\text{GT}} = \|\hat{\boldsymbol{\epsilon}} - \boldsymbol{\epsilon}\|^2$
    $\hat{\boldsymbol{x}}_{0,t} = \frac{\boldsymbol{x_t} - \sqrt{1-\bar{\alpha}}\hat{\epsilon}}{\sqrt{\bar{\alpha}}}$
    $l_{\text{CON}} = \|\tilde{\boldsymbol{A}}\hat{\boldsymbol{x}}_{0,t} - \boldsymbol{y}\|^2$
    Take gradient descent w.r.t. θ: $\nabla_\theta[\gamma l_{\text{GT}} + (1-\gamma) l_{\text{CON}}]$ using Adam optimizer
  end while
end for

Algorithm 2 Sampling μ_θ in T steps

repeat
  $\boldsymbol{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
  for $t = T, \ldots, 1$ **do**
    $\boldsymbol{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ if $t > 1$, else $\boldsymbol{z} = \mathbf{0}$
    $\boldsymbol{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(\boldsymbol{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\mu_\theta(\boldsymbol{y}, \boldsymbol{x}_t, \gamma_t)\right) + \sqrt{1-\alpha_t}\boldsymbol{z}$
  end for
until converged

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Datasets for Training and Evaluation. We evaluated our framework on three target datasets: FastMRI [Zbontar et al., 2018], BrainTumor [Chaki and Wozniak, 2023], and OASIS [Marcus et al., 2007]. For training, we used COCO [Lin et al., 2014] as the out-of-domain dataset and IXI [IXI, 2023] T2 brain MRI scans as the in-domain dataset. We evaluated performance using PSNR, SSIM [Wang et al., 2004], and LPIPS [Zhang et al., 2018] metrics on the test sets. Dataset specifications are detailed in Table 1.

Table 1: Dataset Specifications for Multi-stage Training in 4× Super-resolution.

Name	Res.	Size*	Type	Stage
COCO	$16 \to 64$	98k/100/-	General	OOD
IXI	$16 \to 64$	60k/60/-	T2	ID
FastMRI	$64 \to 256$	300/10/60	T2	TD
BrainTumor	$64 \to 256$	300/10/60	T1/T2	TD
OASIS	$64 \to 256$	300/10/60	T1	TD

Note: * Sample sizes indicate train/validation/test split counts. T1/T2 denotes T1-weighted or T2-weighted magnetic resonance imaging (MRI) scans.

Implementation Details. We performed super-resolution (SR) at $2\times$, $4\times$, and $8\times$ scales with noise $\boldsymbol{\eta}$ set to zero. All datasets were cropped to consistent input dimensions and degraded using bicubic downsampling. Single-channel grayscale MRI datasets were expanded to three channels through duplication. Datasets were normalized to $[0, 1]$ and augmented with flip and rotation transforms. The SR3 model with ControlNet conditioning was trained in three stages: (1) initial training on out-of-domain (OOD) COCO dataset for 1M steps, (2) fine-tuning on in-domain (ID) IXI dataset for 1M steps, and (3) final fine-tuning on target domain (TD) datasets for 20K steps. All experiments used Adam optimizer with learning rate $1e^{-4}$.

Table 2: Inference times and parameter counts for each evaluated method.

Method	Parameter Count	Inference Time
DPS	552.81M	117.3 s/sample
I2sb	552.80M	57.0 s/sample
sinSR	118.59M	1.22 s/sample
MSP-SR	136.28M	50.0 s/sample

Note: Inference times measured on NVIDIA A100 GPU.

4.2 PERFORMANCE

Performance Evaluation Against Existing Methods. To demonstrate the effectiveness of MSP-SR framework, we evaluated against state-of-the-art super-resolution methods across three medical imaging datasets: FastMRI, BrainTumor, and OASIS, as shown in Table 3. The comparison includes both training-free approaches (DPS Chung et al. [2022]) and learning-based methods (I2SB Liu et al. [2023] and SinSR Wang et al. [2024b]). For the learning-based baselines, we utilized their provided pre-trained weights followed by direct fine-tuning (DFT) on target datasets using approximately 300 training samples per dataset. Our approach consistently outperformed existing methods across all datasets and evaluation metrics. For the FastMRI dataset, which contains homogeneous T2-weighted axial slices, MSP-SR achieved a PSNR of 28.71dB and SSIM of 0.846, surpassing all baseline methods. The framework demonstrated even more substantial improvements on the more heterogeneous BrainTumor (containing both T1- and T2-weighted images) and OASIS (comprising both axial and coronal brain MRI scans) datasets, achieving PSNR values of 27.34dB and 29.03dB, respectively.

In addition to quality metrics, we also report parameter counts and inference times for those models in Table 2. As shown in the table, SinSR's significantly faster inference time is due to its use of latent space models, while the other methods (including MSP-SR) operate in the ambient dimension, which naturally requires more computational resources. This fundamental architectural difference explains the observed variance in processing speeds across the evaluated approaches. Our MSP-SR achieves a practical balance between quality and efficiency, with superior results at a reasonable computational cost of 50 seconds per sample.

Ablation Studies Analysis. To thoroughly validate the MSP-SR framework's effectiveness and generalizability, we conducted four complementary sets of ablation experiments: (1) Domain Transfer Ablation to evaluate each training stage's contribution in cross-domain scenarios, (2) Cross-dataset Generalization to verify our approach's dataset-agnostic nature, (3) Multi-Model Validation to demonstrate applicability beyond specific models, and (4) Consistency Loss Analysis to assess our consistency regularization technique's impact.

Table 4: Quantitative comparison over different frameworks on FastMRI dataset, where the bolded values represent the best value in each evaluation metric. The results demonstrate that the MSP-SR framework achieves the majority of the best results across different SR scales.

Scale	**Training Components**				**Metrics**		
	OOD	ID	TD	CN*	PSNR↑	SSIM↑	LPIPS↓
$128 \rightarrow 256$	✓	✓	✓	✓	**34.09**	**0.922**	0.0762
	✓		✓	✓	33.03	0.913	**0.0757**
	✓	✓	✓		32.11	0.892	–
		✓	✓		28.45	0.771	0.101
			✓		5.051	0.209	0.513
$64 \rightarrow 256$	✓	✓	✓	✓	**28.71**	0.846	**0.1454**
	✓		✓	✓	28.69	**0.847**	0.1455
	✓	✓	✓		27.90	0.814	–
		✓	✓		27.04	0.764	0.147
			✓		24.31	0.686	0.173
$32 \rightarrow 256$	✓	✓	✓	✓	**22.98**	**0.734**	**0.216**
	✓		✓	✓	21.00	0.654	0.219
	✓	✓	✓		22.43	0.669	–
		✓	✓		19.00	0.558	0.271
			✓		17.38	0.477	0.284

Note: * CN indicates ControlNet fine-tuning applied with ID/TD stages.

We first examined domain transfer capabilities through systematic ablation on four key components: In-Domain (ID) fine-tuning, Out-of-Domain (OOD) pre-training, ControlNet integration, and baseline Target-Domain training. Table 4 demonstrates our multi-stage approach's consistent superiority across all super-resolution scales. OOD pre-training improved PSNR by 19.8% for 128→256 SR tasks, while multi-stage fine-tuning enhanced PSNR by 9.4% in challenging 32→256 (8×) SR scenarios, underscoring our framework's effectiveness in leveraging domain knowledge. Visual comparisons in Fig. 2 corroborate these findings, demonstrating MSP-SR's superior preservation of intricate brain features with enhanced structural details and sharper edge reconstructions.

To verify generalization across different test datasets, we evaluated our approach on BrainTumor [Chaki and Wozniak, 2023] and OASIS [Marcus et al., 2007] datasets for 4× super-resolution. Results in Table 5 show meaningful contributions from each training stage, with BrainTumor dataset showing pronounced improvements: Out-of-Domain pre-training improved PSNR by 3.13%, In-Domain fine-tuning enhanced PSNR by 5.68%, and ControlNet integration further increased PSNR by 1.75%.

Beyond dataset generalization, we investigated whether our

Table 3: Performance Comparison of MSP-SR Against State-of-the-Art Methods Across Multiple Datasets

	FastMRI			BrainTumor			OASIS		
Method	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
DPS	23.00	0.723	0.3678	20.56	0.620	0.3553	21.94	0.622	0.3456
SinSR	26.98	0.843	0.2745	22.79	0.7201	0.2680	21.95	0.711	0.2445
I2SB	16.01	0.123	0.5758	14.95	0.1446	0.6082	15.68	0.1299	0.6011
MSP-SR	**28.71**	**0.846**	**0.1450**	**27.34**	**0.811**	**0.1626**	**29.03**	**0.872**	**0.1319**

Note: Best results are highlighted in bold. LPIPS uses VGG backbone.

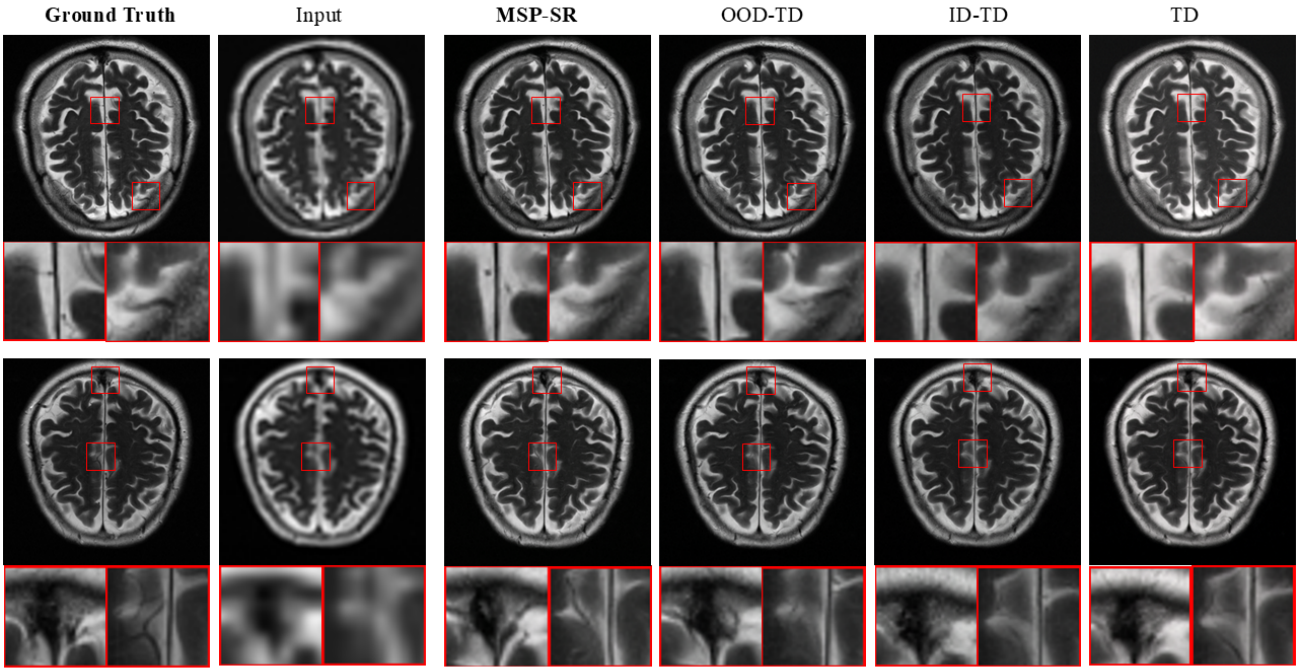

Figure 2: Visualized samples for different frameworks on 4× scale SR task. We use the COCO dataset for the Out-of-Domain stage, IXI for the In-Domain stage, and FastMRI for the Target-Domain stage. TD refers to train on FastMRI from scratch. Note that our MSP-SR (OOD-ID-TD) framework generates samples with clearer and more accurate structural details.

training strategy benefits different models. We conducted experiments using SinSR [Wang et al., 2024b] as the backbone to demonstrate broad applicability across architectures. As shown in Table 6, our training strategy consistently improved SinSR performance with gains of 1.04 dB in PSNR, confirming effective generalization across different model architectures.

Finally, we analyzed the consistency loss component's specific contribution. Its incorporation improved PSNR to 29.15dB and SSIM to 0.859 in 4× SR tasks, with visual comparisons in Fig. 5 demonstrating qualitative improvements in image detail and overall quality.

Uncertainty Analysis. To assess our multi-stage training framework's uncertainty characterization, we conducted comprehensive uncertainty quantification experi-

Table 5: Quantitative comparison on other MRI datasets in 4× scale SR, where the bolded values represent the best value in each evaluation metric. The results demonstrate that the MSP-SR framework achieves the best results across different datasets.

Training component				BrainTumor		OASIS	
OOD	ID	TD*	CN*	PSNR↑	SSIM↓	PSNR↑	SSIM↓
✓	✓	✓	✓	**27.34**	**0.811**	**29.03**	**0.872**
✓		✓	✓	25.87	0.783	28.75	0.852
✓	✓	✓		26.87	0.783	29.02	0.857
	✓	✓		26.51	0.800	28.52	0.760

Note: * CN indicates ControlNet fine-tuning applied with ID/TD stages. TD refers to the TD fine-tuning stage where we set the notumor brain and OASIS as the target dataset here.

Table 6: Quantitative comparison for SinSR over different frameworks in 4× scale SR, where the bolded values represent the best value in each evaluation metric.

Training component			FastMRI		
OOD	ID	TD	PSNR↑	SSIM↓	LPIPS↑
✓	✓	✓	**28.02**	**0.8550**	**0.1748**
✓		✓	26.98	0.8430	0.2745
	✓	✓	25.14	0.7866	0.2111
		✓	22.54	0.6451	0.2177

Table 7: Negative Log Likelihood (NLL) Comparison Across Training Configurations and SR Scales

Method	Negative Log Likelihood↓		
	2× SR	4× SR	8× SR
Training Stages (Fig. 3)			
MSP-SR Framework (Fig.4a)	-	17.177	-
OOD + TD Stages (Fig.4b)	-	18.681	-
Only TD Stages (Fig.4c)	-	26.509	-
SR Scale (Fig. 4)			
MSP-SR Framework (Fig.5a)	12.642	-	-
MSP-SR Framework (Fig.5b)	-	17.177	-
MSP-SR Framework (Fig.5c)	-	-	17.147

Note: "-" indicates non-applicable due to scale-specific training.

ments across different training configurations and super-resolution scales. Specifically, we compared uncertainty calibration across three training paradigms (MSP-SR (OOD+ID+TD), OOD+TD, and TD-only) and evaluated scale-dependent uncertainty patterns from 2× to 8× super-resolution using negative log-likelihood metrics and probabilistic distribution analysis.

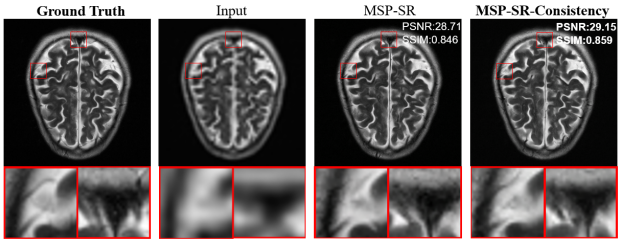

Figure 5: Visualized sample to verify the influence of consistency loss in 4× scale SR task. The consistency loss significantly improves reconstruction quality and detail preservation.

Table 7 presents quantitative evaluation using the negative log-likelihood (NLL) of pixel intensity distributions. For each low-resolution input, we generated 10 high-resolution predictions and constructed pixel-wise probability distributions using 256 intensity bins between 0 and 1. Mathematically, for a pixel location (i, j) with true intensity value $y_{i,j}$ and estimated probability distribution $\hat{p}_{i,j}$, the pixel-wise NLL is:

$$\text{NLL}_{i,j} = -\log(\hat{p}_{i,j}(y_{i,j}))$$

Here, $\hat{p}_{i,j}$ is derived from our generative model's samples without prior knowledge of $y_{i,j}$. The final NLL averages over all pixel locations:

$$\text{NLL} = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} \text{NLL}_{i,j}$$

Our MSP-SR framework achieved lower NLL (17.177) compared to both Out-of-Domain pre-training with target-domain fine-tuning and target-domain-only training, demonstrating more accurate uncertainty characterization across scales.

We visualized pixel-wise uncertainty through probability distribution heatmaps (Fig. 3, 4) for a 30-pixel segment

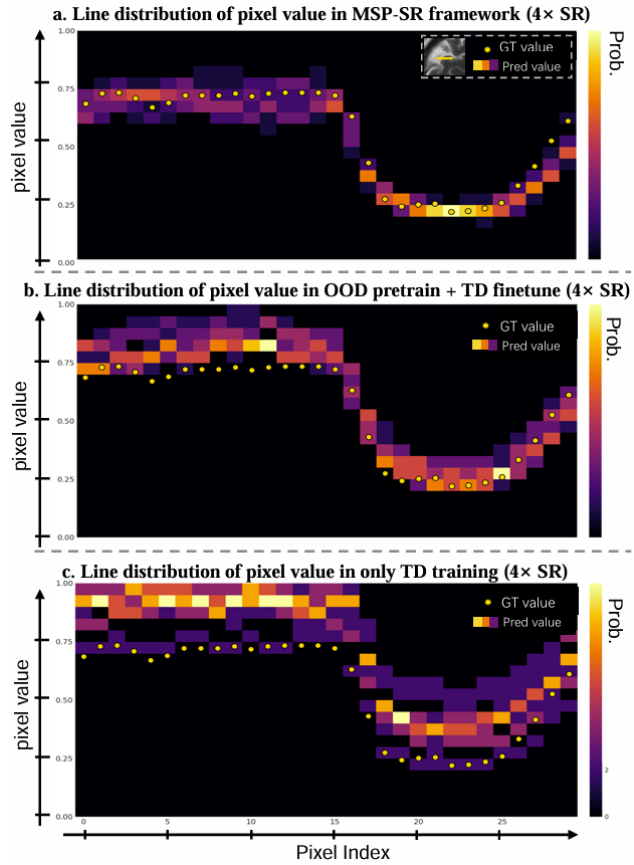

Figure 3: Uncertainty Analysis Across Training Configurations. Pixel value distributions shown for: (a) MSP-SR framework, (b) OOD pre-training with TD fine-tuning, and (c) TD-only training, demonstrating a more accurate uncertainty characterization through progressive domain transfer.

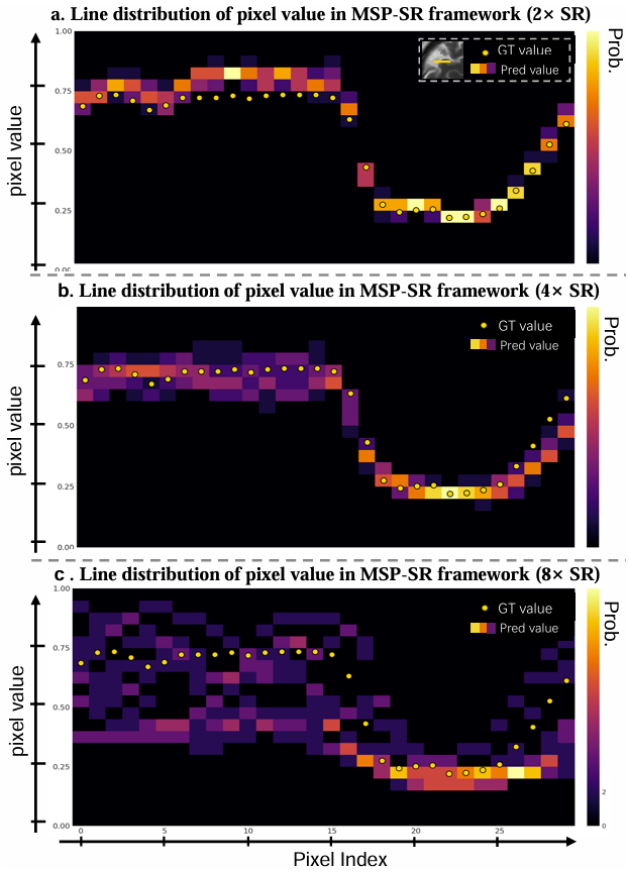

Figure 4: Scale-dependent Uncertainty Analysis in Super-resolution. Probabilistic distributions at 2×, 4×, and 8× scales illustrate increasing reconstruction uncertainty at higher resolutions.

marked by a yellow line in the MRI image, with color intensity representing probability density and yellow markers indicating ground truth values.

Comparative analysis of uncertainty characteristics (Fig. 3) reveals our MSP-SR framework's superior calibration properties. Our approach demonstrates well-calibrated predictive uncertainty where predicted distributions closely follow ground truth distributions across pixel intensities, indicating accurate uncertainty estimation without systematic bias. The OOD+TD approach shows improved alignment compared to baseline but still exhibits distribution misalignment in high-intensity regions. In contrast, TD-only training exhibits severe uncertainty pathologies with distributions concentrated around incorrect values, demonstrating uncertainty collapse with false confidence in erroneous predictions.

Analysis of increasing super-resolution scales (2× to 8×) reveals progressively wider pixel value distributions (Fig. 4), indicating heightened reconstruction uncertainty at higher scales. At 2× SR, predicted distributions maintain tight alignment with ground truth values, while 4× SR shows moderate broadening in mid-intensity ranges corresponding to gray matter regions. At 8× SR, distributions become significantly dispersed in intermediate intensity values where tissue boundaries reside, while maintaining better confidence in extreme intensities. This scale-dependent uncertainty pattern is confirmed by variance maps (Fig. 6), showing elevated uncertainty in complex brain regions such as cortical folds.

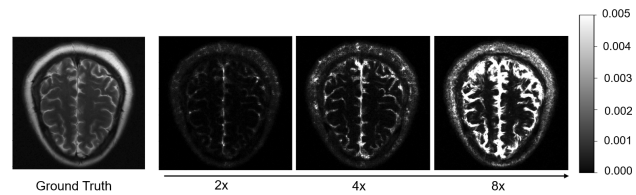

Figure 6: Variance maps generated from multiple inference outputs of same diffusion model applied to 2×,4×, and 8× super-resolution tasks. Standard deviation is normalized by value to indicate the scale of variance relative to actual intensities. The variance is high in areas where the brain's features are more distinct, such as the folds and ridges of the cortex.

5 CONCLUSION

In this paper, we presented MSP-SR, a diffusion-based cascaded framework for few-shot medical image super-resolution that leverages both in-domain and out-of-domain low-resolution data to build models with enhanced accuracy and uncertainty characterization. Experimental results across diverse MRI datasets demonstrate MSP-SR's superior fidelity compared to existing methods, while its staged fine-tuning preserves uncertainty modeling capabilities and avoids misaligned predictions common in direct fine-tuning (DFT) approaches. Future work will investigate alternative pre-training strategies to reduce hallucination and address domain-specific biases through enhanced cross-domain pre-training techniques, while validating MSP-SR's adaptability across diverse models and imaging domains.

Acknowledgements. The work was supported by National Science Foundation (NSF) through awards IIS 21-31335, OAC 21-30835, DBI 20-21898, as well as a C3.ai research award. This work used the Delta system at the National Center for Supercomputing Applications (NCSA) through allocations from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program.

References

Ixi dataset – brain development. http://brain-development.org/ixi-dataset/, 2023. Accessed: 2023-10-15.

Jyotismita Chaki and Marcin Wozniak. Brain tumor mri dataset, 2023. URL https://dx.doi.org/10.21227/1jny-g144.

Hyungjin Chung, Jeongsol Kim, Michael T Mccann, Marc L Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. *arXiv preprint arXiv:2209.14687*, 2022.

Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, 38(2):295–307, 2015.

Cycle Learning Ensemble. Ct super-resolution gan constrained by the identical, residual, and cycle learning ensemble.

Hayit Greenspan. Super-resolution in medical imaging. *The computer journal*, 52(1):43–63, 2009.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

S. I. Kabanikhin. Definitions and examples of inverse and ill-posed problems. *Journal of Inverse and Ill-posed Problems*, 16(4):317–357, 2008. doi: doi:10.1515/JIIP.2008.019. URL https://doi.org/10.1515/JIIP.2008.019.

V. A. Kelkar, D. S. Gotsis, F. J. Brooks, P. KC, K. J. Myers, R. Zeng, and M. A. Anastasio. Assessing the ability of generative adversarial networks to learn canonical medical image statistics. *IEEE Transactions on Medical Imaging*, 2022.

Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Deeply-recursive convolutional network for image super-resolution, 2016. URL https://arxiv.org/abs/1511.04491.

K. Li, W. Zhou, H. Li, and M. A. Anastasio. Assessing the impact of deep neural network-based image denoising on binary signal detection tasks. *IEEE Transactions on Medical Imaging*, 40(9):2295–2305, 2021a.

Yufei Li, Bruno Sixou, and Francois Peyrin. A review of the deep learning methods for medical images super resolution problems. *Irbm*, 42(2):120–133, 2021b.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

Guan-Horng Liu, Arash Vahdat, De-An Huang, Evangelos A Theodorou, Weili Nie, and Anima Anandkumar. I2sb: Image-to-image schr\" odinger bridge. *arXiv preprint arXiv:2302.05872*, 2023.

Zhisheng Lu, Juncheng Li, Hong Liu, Chaoyan Huang, Linlin Zhang, and Tieyong Zeng. Transformer for single image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 457–466, 2022.

Awais Mansoor, Teerit Vongkovit, and Marius George Linguraru. Adversarial approach to diagnostic quality volumetric image enhancement. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 353–356. IEEE, 2018.

Daniel S Marcus, Tracy H Wang, Jamie Parker, John G Csernansky, John C Morris, and Randy L Buckner. Open access series of imaging studies (oasis): cross-sectional mri data in young, middle aged, nondemented, and demented older adults. *Journal of cognitive neuroscience*, 19(9):1498–1507, 2007.

Junyoung Park, Donghwi Hwang, Kyeong Yun Kim, Seung Kwan Kang, Yu Kyeong Kim, and Jae Sung Lee. Computed tomography super-resolution using deep convolutional neural network. *Physics in Medicine & Biology*, 63(14):145011, 2018.

Chitwan Saharia, William Chan, Huiwen Chang, Chris A. Lee, Jonathan Ho, Tim Salimans, David J. Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models, 2022a. URL https://arxiv.org/abs/2111.05826.

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, 2022b.

Shuyao Shang, Zhengyang Shan, Guangxing Liu, LunQian Wang, XingHua Wang, Zekai Zhang, and Jinglin Zhang. Resdiff: Combining cnn and diffusion model for image super-resolution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 8975–8983, 2024.

Jiaming Song, Arash Vahdat, Morteza Mardani, and Jan Kautz. Pseudoinverse-guided diffusion models for inverse problems. In *International Conference on Learning Representations*, 2023a. URL https://openreview.net/forum?id=9_gsMA8MRKQ.

Yang Song, Liyue Shen, Lei Xing, and Stefano Ermon. Solving inverse problems in medical imaging with score-based generative models. *arXiv preprint arXiv:2111.08005*, 2021.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. 2023b.

Ying Tai, Jian Yang, and Xiaoming Liu. Image super-resolution via deep recursive residual network. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2790–2798, 2017. doi: 10.1109/CVPR.2017.298.

Kensuke Umehara, Junko Ota, and Takayuki Ishida. Application of super-resolution convolutional neural network for enhancing image resolution in chest ct. *Journal of digital imaging*, 31:441–450, 2018.

Jianyi Wang, Zongsheng Yue, Shangchen Zhou, Kelvin CK Chan, and Chen Change Loy. Exploiting diffusion prior for real-world image super-resolution. *International Journal of Computer Vision*, pages 1–21, 2024a.

Longguang Wang, Yingqian Wang, Xiaoyu Dong, Qingyu Xu, Jungang Yang, Wei An, and Yulan Guo. Unsupervised degradation representation learning for blind super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10581–10590, 2021a.

Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1905–1914, 2021b.

Yufei Wang, Wenhan Yang, Xinyuan Chen, Yaohui Wang, Lanqing Guo, Lap-Pui Chau, Ziwei Liu, Yu Qiao, Alex C Kot, and Bihan Wen. Sinsr: diffusion-based image super-resolution in a single step. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 25796–25805, 2024b.

Zhihao Wang, Jian Chen, and Steven CH Hoi. Deep learning for image super-resolution: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 43(10): 3365–3387, 2020.

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.

Jure Zbontar, Florian Knoll, Anuroop Sriram, Tullie Murrell, Zhengnan Huang, Matthew J Muckley, Aaron Defazio, Ruben Stern, Patricia Johnson, Mary Bruno, et al. fastmri: An open dataset and benchmarks for accelerated mri. *arXiv preprint arXiv:1811.08839*, 2018.

Kai Zhang, Jingyun Liang, Luc Van Gool, and Radu Timofte. Designing a practical degradation model for deep blind image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4791–4800, 2021a.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023. URL https://arxiv.org/abs/2302.05543.

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 (CVPR)*, 2018.

X. Zhang, V. A. Kelkar, J. Granstedt, H. Li, and M. A. Anastasio. Impact of deep learning-based image super-resolution on binary signal detection. *Journal of Medical Imaging*, 8(6):065501, 2021b.

YiNan Zhang and MingQiang An. Deep learning-and transfer learning-based super resolution reconstruction from single medical image. *Journal of healthcare engineering*, 2017(1):5859727, 2017.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.

Supplementary Material

Ruike Zhu[†1] **Matthew Weston**[1] **Hanwen Zhang**[1] **Arindam Banerjee**[1]

[1]Siebel School of Computing and Data Science , University of Illinois at Urbana-Champaign , Urbana, Illinois, USA

In the supplementary material, we present additional qualitative results for the MSP-SR framework. We also include detailed information on the model architecture and comprehensive descriptions of the datasets used.

A DATASET DESCRIPTION

OOD Pre-Training Stage. The initial stage utilizes the COCO dataset Lin et al. [2014], a comprehensive collection of over 330,000 images, including 200,000 labeled images across 80 object categories. We selected 100,000 images for OOD pre-training, with each image cropped to a square format based on its smaller dimension.

ID Fine-tuning Stage. The second stage employs the IXI dataset IXI [2023], which contains brain MRI scans from London hospitals. While the dataset includes T1, T2, PD-weighted, diffusion-weighted, and MR angiography images, we specifically utilize T2-weighted longitudinal and transversal brain scans for fine-tuning.

TD Fine-tuning Stage. The final stage incorporates three target-domain datasets: FastMRI Zbontar et al. [2018], BrainTumor Chaki and Wozniak [2023], and OASIS Marcus et al. [2007]. FastMRI, developed by Facebook AI Research and NYU Langone Health, provides k-space data and high-resolution reconstructed images, from which we use T2-weighted longitudinal scans. The BrainTumor dataset contains multi-modal MRI scans (T1, T2, FLAIR) of brain tumor patients with tumor type annotations; we utilize its T1/T2-weighted longitudinal scans. From the OASIS-2 longitudinal Alzheimer's study dataset, we select MRI scans from cognitively stable subjects as our target data.

B TRAINING AND EVALUATION DETAILS

Our training process consists of three stages: initial training on the out-of-domain (COCO) dataset, followed by fine-tuning on the in-domain (IXI) dataset and target-domain datasets (FastMRI, BrainTumor, OASIS). Input resolutions are set to 64×64 for COCO and IXI datasets, and 256×256 for target-domain datasets. All experiments are conducted on an NVIDIA A100-SXM4-40GB GPU with CUDA 12.2. The OOD pre-training and ID fine-tuning stages each require 1 million iterations, consuming approximately 24 hours per stage on a single A100 GPU, with the best-performing epoch selected for subsequent analysis. Target-domain fine-tuning continues for 60k-80k iterations until convergence. To maintain consistency with the baseline architecture, we adopt training hyperparameters similar to those in SR3 Saharia et al. [2022b], as detailed in Table 8.

LPIPS We evaluate the perceptual quality of generated super-resolution images using Learned Perceptual Image Patch Similarity (LPIPS) Zhang et al. [2018], a neural network-based metric that captures perceptually meaningful image differences more effectively than traditional pixel-based measures. Our implementation utilizes the VGG backbone for LPIPS computation, with evaluations conducted across 60 test samples from each target dataset.

	Batchsize	Iteration	LR	Dropout	Resolution	Opt.
COCO	4	1000000	1e-4	0.2	$16 \rightarrow 64$	Adam
IXI		1000000				
FastMRI		70000			$64 \rightarrow 256$	
BrainTumor		70000				
OASIS		70000				

Table 8: Training Configuration for Various Datasets for 4× scale SR.

C ADDITIONAL EXPERIMENT RESULTS

C.1 ADDITIONAL VISUAL RESULTS FOR OASIS AND BRAINTUMOR

Table 4 presents performance comparisons between MSP-SR and its ablated variants on the FastMRI dataset. Supplementary visual results for these experiments are provided in Figures 8, 9, and 10. Additionally, Figure 11 demonstrates visual results from OASIS Marcus et al. [2007] and BrainTumor Chaki and Wozniak [2023] datasets, complementing the quantitative analysis in Table 5.

C.2 UNCERTAINTY QUANTIFICATION ANALYSIS

Figure 7 presents the complete sample used for uncertainty analysis in Figures 4 and 3. The visualization consists of two components: a ground truth MRI image with a highlighted region of interest (left), and sets of inference samples from three training configurations (right). These configurations include: (1) OOD pre-training + ID fine-tuning + TD fine-tuning, (2) OOD pre-training + TD fine-tuning, and (3) TD-only training, each represented by five inference samples.

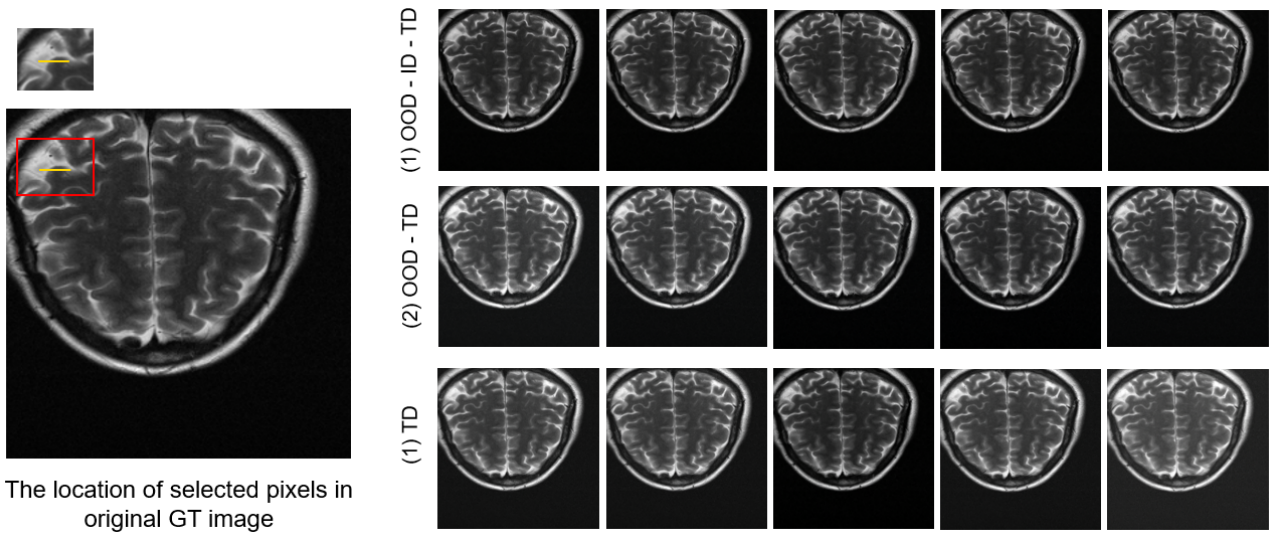

Figure 7: Visualization of Multiple 4× Inference Results Across Different Training Configurations.

D MODEL ARCHITECTURE

In this section, we present our detailed model architecture. We use the same encoder-decoder architecture for denoising Unet in all three stages. We build our model on top of SR3Saharia et al. [2022b] and ControlNetZhang et al. [2023]. Our model contains 3 components: the encoder, the decoder, and the ControlNet.

Encoder For input noise level t, the encoder uses positional encoding and 2 MLP layers to produce embedding for the noise level. The encoder contains an input convolution layer followed by five down-sampling blocks, each containing two

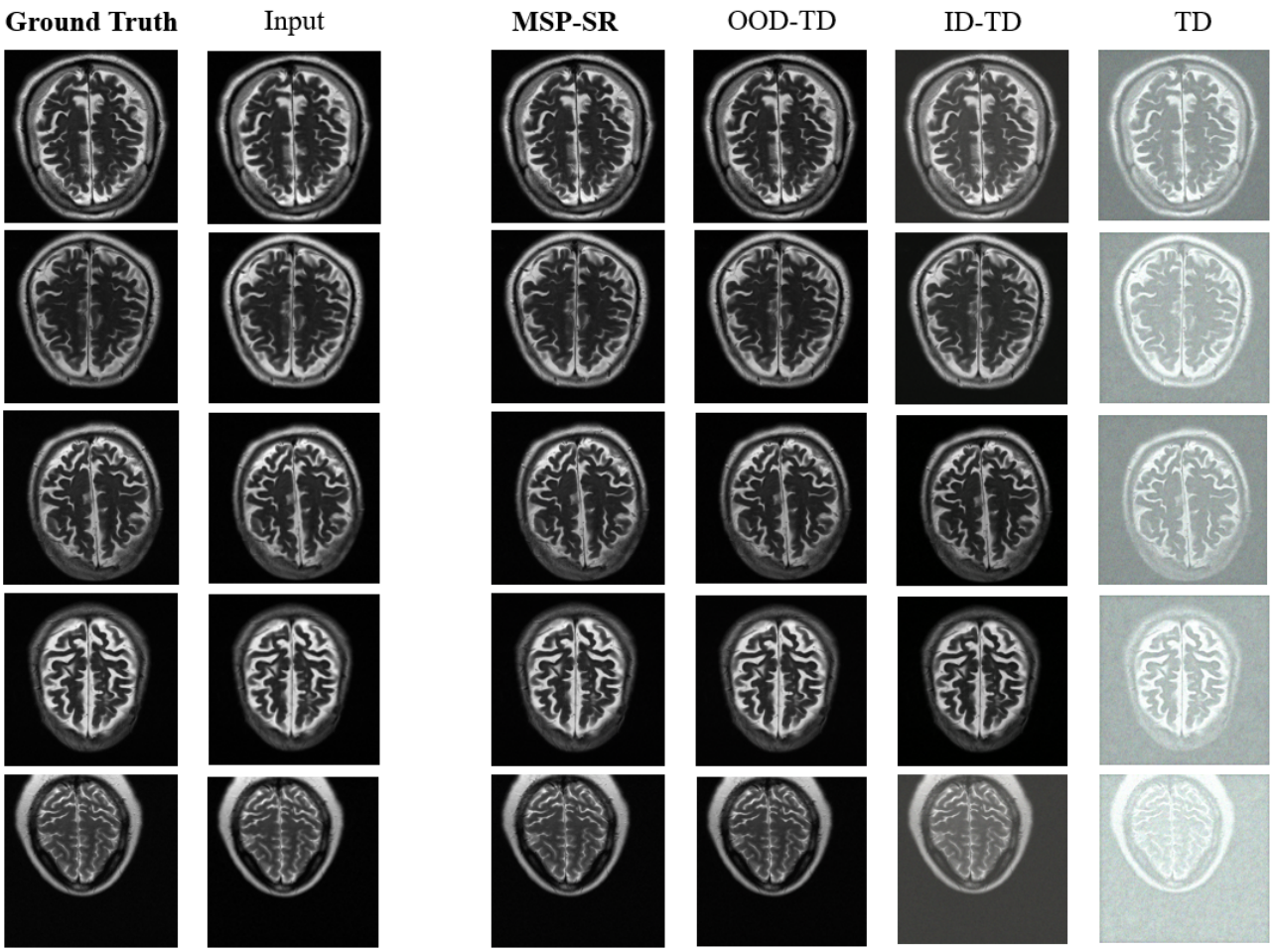

Figure 8: More visualized samples for different frameworks on 2× scale SR task.

res-net blocks. One of the down-sampling blocks contains self-attention. After down-sampling, the encoder has two middle res-net blocks and one middle attention block. We show detailed model architecture in Tab. 9

Decoder The decoder has a similar structure as the encoder. It includes five upsampling blocks, each containing three res-net blocks, one of which has self-attention following each resnet block. The decoder also applies an output convolution layer. We show the detailed model architecture in Tab. 10

ControlNet Following ControlNetZhang et al. [2023], we use an additional branch of the network for better finetuning. For the ControlNet branch, we use exactly the same architecture as the Encoder, except for the additional zero convolution blocks. The initial weight of the ControlNet branch is copied from the Encoder, and the zero conv layers are initialized to output all zeros. The outputs in the ControlNet branch after each zero-convolutions are added to each input of the decoder's upsampling block. We show the detailed model architecture in Tab. 11.

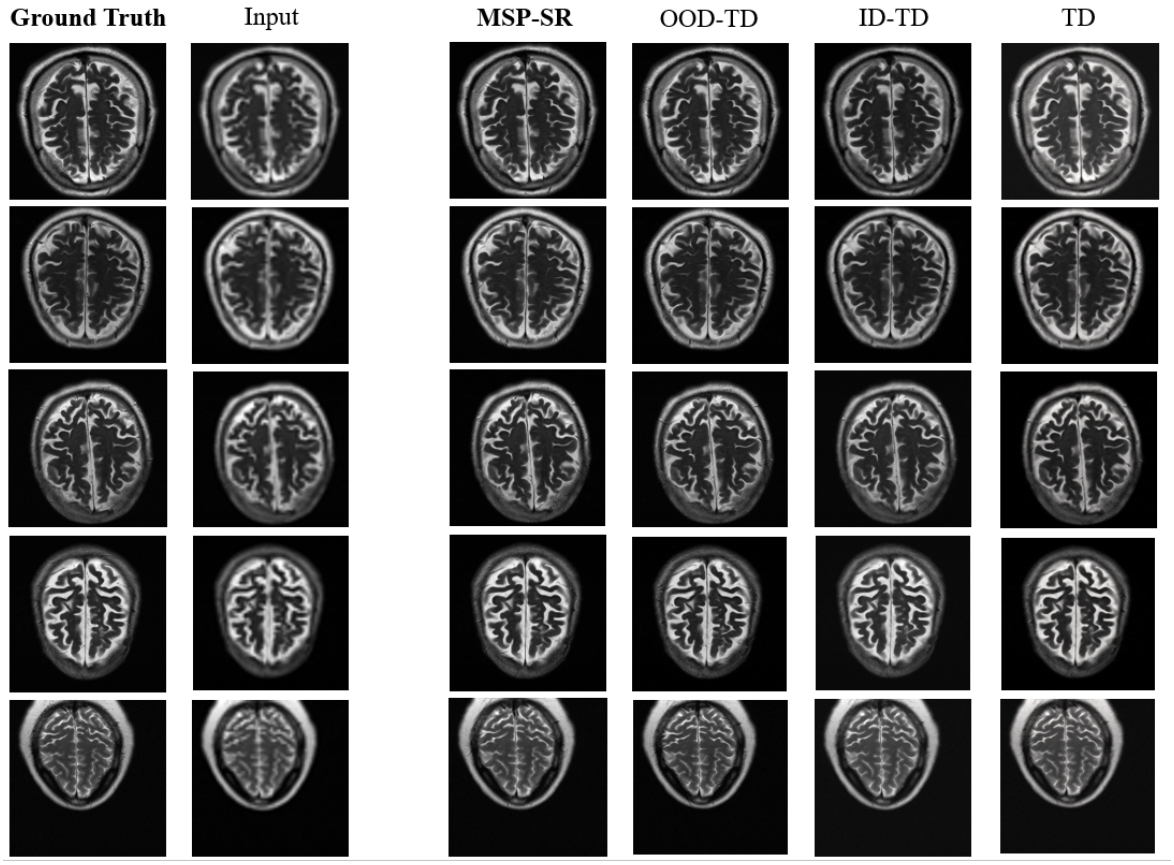

Figure 9: More visualized samples for different frameworks on 4× scale SR task.

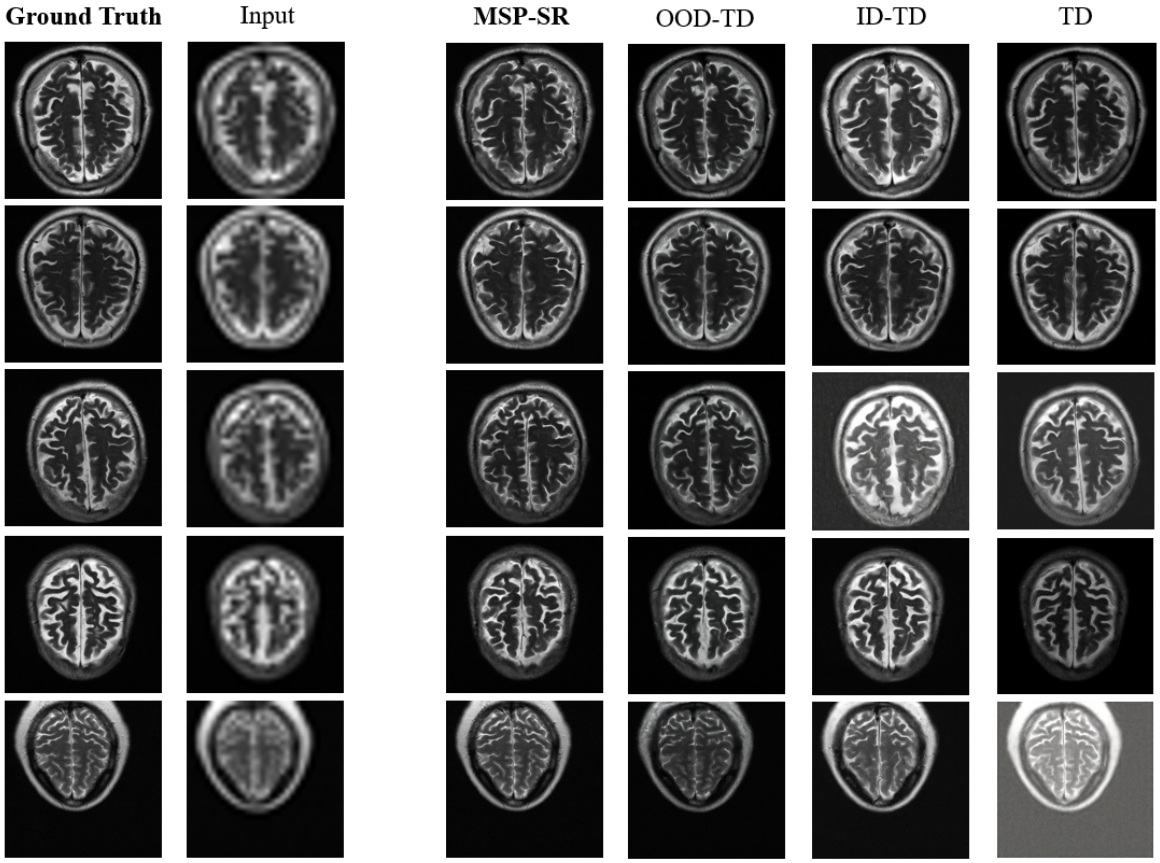

Figure 10: More visualized samples for different frameworks on 8× scale SR task.

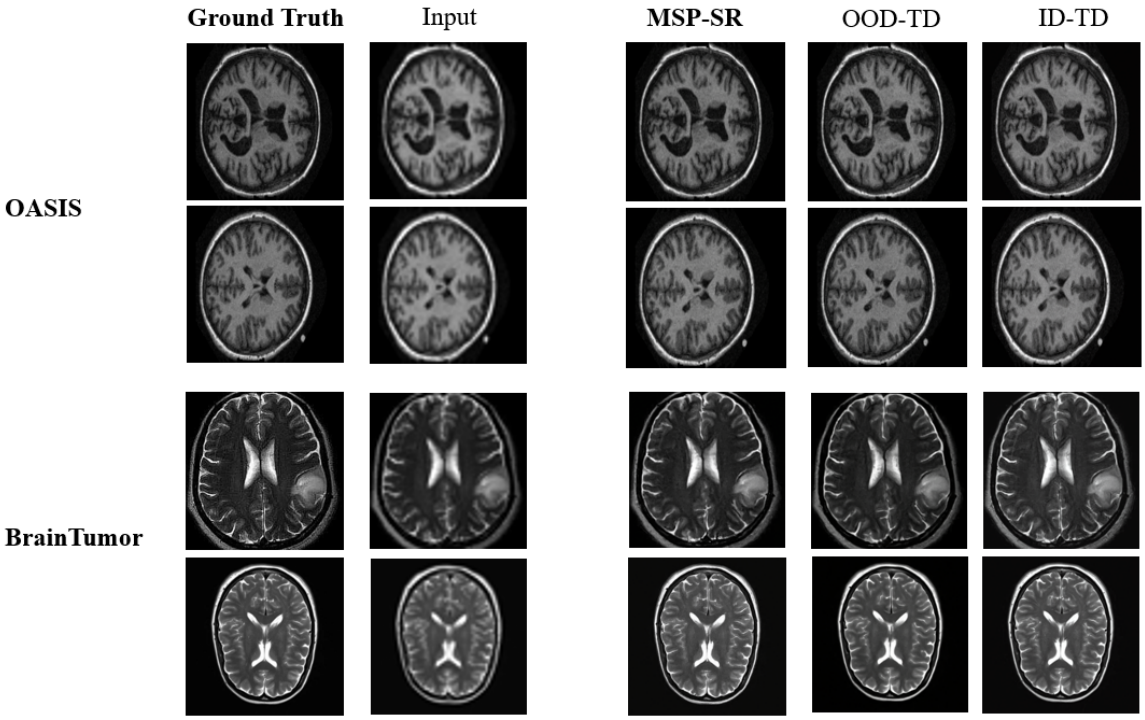

Figure 11: Visualized samples for experiments on OASIS and BrainTumor datasets.

1

Table 9: Architecture for Encoder

layers		parameters
noise_level_t	PositionalEncoding	t = t * exp(-log(1e4) * arange(64)); encoding = (sin(t), cos(t))
	mlp	in_ch:64, out_ch: 256
	Swish activation	x*signoid(x)
	mlp	in_ch:256, out_ch: 64
input	Conv2d	in_ch:6, out_ch: 64, kernel: 3x3, stride: 1, pad: 1
downsample_block_1	ResnetBlock	in_ch:64, out_ch: 64
	ResnetBlock	in_ch:64, out_ch: 64
	Downsample(Conv2d)	in_ch:64, out_ch: 64,kernel:3x3, stride:2, padding=((0,1,0,1),val=0)
downsample_block_2	ResnetBlock	in_ch:64, out_ch: 128
	ResnetBlock	in_ch:128, out_ch: 128
	Downsample(Conv2d)	in_ch:128, out_ch: 128,kernel:3x3, stride:2, padding=((0,1,0,1),val=0)
downsample_block_3	ResnetBlock	in_ch:128, out_ch: 256
	ResnetBlock	in_ch:256, out_ch: 256
	Downsample(Conv2d)	in_ch:256, out_ch: 256,kernel:3x3, stride:2, padding=((0,1,0,1),val=0)
downsample_block_4	ResnetBlock	in_ch:256, out_ch: 512
	SelfAtt	in_ch:512, out_ch: 512
	ResnetBlock	in_ch:512, out_ch: 512
	SelfAtt	in_ch:512, out_ch: 512
	Downsample(Conv2d)	in_ch:512, out_ch: 512,kernel:3x3, stride:2, padding=((0,1,0,1),val=0)
downsample_block_5	ResnetBlock	in_ch:512, out_ch: 512
	ResnetBlock	in_ch:512, out_ch: 512
middle	ResnetBlock	in_ch:512, out_ch: 512
	AttnBlock	in_ch:512
	ResnetBlock	in_ch:512, out_ch: 512

Table 10: Architecture for Decoder

layers		parameters
upsample_block_1	ResnetBlock	in_ch:1024, out_ch: 512
	ResnetBlock	in_ch:512, out_ch: 512
	ResnetBlock	in_ch:512, out_ch: 512
	Upsample(nearest_interpolate)	scale_factor=2.0
	Conv2d	in_ch:512, out_ch: 512,kernel=3x3,stride=1,padding=1)
upsample_block_2	ResnetBlock	in_ch:1024, out_ch: 512
	SelfAtt	in_ch:512, out_ch: 512
	ResnetBlock	in_ch:512, out_ch: 512
	SelfAtt	in_ch:512, out_ch: 512
	ResnetBlock	in_ch:512, out_ch: 512
	SelfAtt	in_ch:512, out_ch: 512
	Upsample(nearest_interpolate)	scale_factor=2.0
	Conv2d	in_ch:512, out_ch: 512,kernel=3x3,stride=1,padding=1)
upsample_block_3	ResnetBlock	in_ch:512, out_ch: 256
	ResnetBlock	in_ch:256, out_ch: 256
	ResnetBlock	in_ch:256, out_ch: 256
	Upsample(nearest_interpolate)	scale_factor=2.0
	Conv2d	in_ch:256, out_ch: 256,kernel=3x3,stride=1,padding=1)
upsample_block_4	ResnetBlock	in_ch:128, out_ch: 128
	ResnetBlock	in_ch:256, out_ch: 128
	ResnetBlock	in_ch:128, out_ch: 128
	Upsample(nearest_interpolate)	scale_factor=2.0
	Conv2d	in_ch:128, out_ch: 128,kernel=3x3,stride=1,padding=1)
upsample_block_5	ResnetBlock	in_ch:128, out_ch: 64
	ResnetBlock	in_ch:64, out_ch: 64
	ResnetBlock	in_ch:64, out_ch: 64
final conv	Normalize	GroupNorm,num_groups=32, num_channels=512
	Conv2d	in_ch:64, out_ch: 6, kernel: 3x3, stride: 1, pad: 1

Table 11: Architecture for ControlNet

layers		parameters
noise_level_t	PositionalEncoding	t = t * exp(-log(1e4) * arange(64)); encoding = (sin(t), cos(t))
	mlp	in_ch:64, out_ch: 256
	Swish activation	x*signoid(x)
	mlp	in_ch:256, out_ch: 64
input	Conv2d	in_ch:6, out_ch: 64, kernel: 3x3, stride: 1, pad: 1
zero_conv	Conv2d	in_ch:6, out_ch: 64, kernel: 3x3, stride: 1, pad: 1
downsample_block_1	ResnetBlock	in_ch:64, out_ch: 64
	ResnetBlock	in_ch:64, out_ch: 64
	Downsample(Conv2d)	in_ch:64, out_ch: 64,kernel:3x3, stride:2, padding=((0,1,0,1),val=0)
zero_conv	Conv2d	in_ch:6, out_ch: 64, kernel: 3x3, stride: 1, pad: 1
downsample_block_2	ResnetBlock	in_ch:64, out_ch: 128
	ResnetBlock	in_ch:128, out_ch: 128
	Downsample(Conv2d)	in_ch:128, out_ch: 128,kernel:3x3, stride:2, padding=((0,1,0,1),val=0)
zero_conv	Conv2d	in_ch:6, out_ch: 64, kernel: 3x3, stride: 1, pad: 1
downsample_block_3	ResnetBlock	in_ch:128, out_ch: 256
	ResnetBlock	in_ch:256, out_ch: 256
	Downsample(Conv2d)	in_ch:256, out_ch: 256,kernel:3x3, stride:2, padding=((0,1,0,1),val=0)
zero_conv	Conv2d	in_ch:6, out_ch: 64, kernel: 3x3, stride: 1, pad: 1
downsample_block_4	ResnetBlock	in_ch:256, out_ch: 512
	SelfAtt	in_ch:512, out_ch: 512
	ResnetBlock	in_ch:512, out_ch: 512
	SelfAtt	in_ch:512, out_ch: 512
	Downsample(Conv2d)	in_ch:512, out_ch: 512,kernel:3x3, stride:2, padding=((0,1,0,1),val=0)
zero_conv	Conv2d	in_ch:6, out_ch: 64, kernel: 3x3, stride: 1, pad: 1
downsample_block_5	ResnetBlock	in_ch:512, out_ch: 512
	ResnetBlock	in_ch:512, out_ch: 512
middle	ResnetBlock	in_ch:512, out_ch: 512
	AttnBlock	in_ch:512
	ResnetBlock	in_ch:512, out_ch: 512