An Efficient Approach to Super-Resolution with Fine-Tuning Diffusion Models

Shreshth Saini
saini.2@utexas.edu

Yu-Chih Chen
berriechen@utexas.edu

Krishna Srikar Durbha
krishna.durbha@utexas.edu

Abstract

Increasing image resolution is a critical task in computer vision, and various deep learning-based methods have been proposed to address it, including CNNs, GANs, and Diffusion Probabilistic Models. However, Diffusion Probabilistic Models often suffer from high computation costs, and slow inference times, which limit their practical applicability. In this work, we explore the potential of pre-trained diffusion models, specifically zero-shot and fine-tuning approaches, to overcome these challenges and evaluate their generalization ability with limited time steps, iterations, and data samples. We also qualitatively evaluate the zero-shot approach. Our results demonstrate that these approaches can enhance the generalization ability of diffusion models and reduce the need for extensive training from scratch.

1. Introduction

Inspired by non-equilibrium thermodynamics, diffusion models were initially proposed in 2015 [19] and then have rapidly emerged as a potent class of deep generative models that have advanced the state-of-the-art image generation and image-to-image translation tasks. In low-level vision tasks such as Single-Image Super-Resolution (SISR), diffusion models aim to enhance the resolution of the input low-resolution image to a higher resolution, by learning the underlying probability distribution of high-resolution images given low-resolution images as input, while preserving its visual details and content.

This is a challenging task that requires the recovery of high-frequency details using sophisticated algorithms that exploit statistical and signal-processing techniques. SISR becomes even more challenging for larger magnifications, while multiple high-resolution images could be consistent with corresponding low-resolution input. While deep generative models, such as GANs, have shown remarkable progress in modeling statistical patterns and regularities [3, 6], and have been applied to SISR task [10, 22], they can be unstable during training, leading to issues such as mode collapse, vanishing gradients, training instabilities and converge failure, which limits their diversity [12, 20].

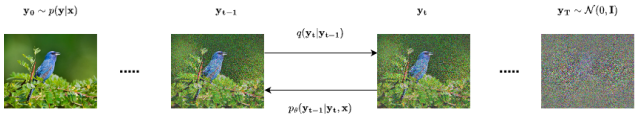

Figure 1. The forward and reverse diffusion processes

In contrast, diffusion probabilistic models utilize an iterative denoising process to learn the data distribution and sample images. These models have shown significant promise in image-to-image translation tasks [23], and low-level computer-vision tasks, including image super-resolution [15, 18]. Diffusion models can capture and generate highly complex details, making them uniquely suited for large magnification SISR tasks. However, this new line of work comes with challenges such as high computation cost, slow inference, hallucinations, and mode collapse in certain data distributions. Additionally, diffusion models require larger training sets and may perform poorly with out-of-distribution (OOD) data. While randomness may benefit diversity, it can harm output consistency.

In this work, we aim to address some of these challenges by exploring the usefulness of pre-trained diffusion models. Specifically, we will investigate the efficacy of zero-shot and fine-tuning approaches using pre-trained models. By leveraging the power of pre-trained models, we aim to reduce the computational cost and training time, while improving the accuracy and stability of the models. The results of our study will provide valuable insights into the effectiveness of diffusion models for image super-resolution tasks and could pave the way for future research in this area.

2. Conditional Denoising Diffusion Model

Unlike traditional Denoising Diffusion Probabilistic Models (DDPMs) which generate high-resolution images from noise, our task requires conditioning on given low-resolution images. Therefore, we use a conditional DDPM and adopt the iterative refinement approach which was proposed in [18] with modifications.

Consider an image super-resolution dataset of input-output pairs denoted by $\mathcal{D} = \{\mathbf{x_i}, \mathbf{y_i}\}_{i=1}^{i=N}$ where $\mathbf{y_i}$ is the

reference or target image that needs to be generated and $\mathbf{x_i}$ is an image of lower resolution. A conditional DDPM generates its target $\mathbf{y_0} \sim p(\mathbf{y}|\mathbf{x})$ in $\mathbf{T}$ refinement steps from noise by being conditioned on input $\mathbf{x}$. The intermediate states $(\mathbf{y_{T-1}}, \mathbf{y_{T-2}}, ..., \mathbf{y_2}, \mathbf{y_1})$ of the process which is a first-order Markov-chain have the dimensions similar to the final refined output image or target image.

Starting with pure Gaussian noise image $\mathbf{y_T} \sim \mathcal{N}(0, \mathbf{I})$, each step of the process iteratively refines the image according to learned conditional transition distribution $p_\theta(\mathbf{y_{t-1}}|\mathbf{y_t}, \mathbf{x})$. Similar to a normal diffusion model, the distributions of intermediate state images in the chain are defined by the forward diffusion process which gradually adds gaussian noise to the signal defined by the process $q(\mathbf{y_t}|\mathbf{y_{t-1}})$ and our goal is to reverse the process by recovering the signal from noise through reverse Markov chain by conditioning on $\mathbf{x}$. The reverse Markov chain is modeled as a denoising deep learning model f_θ. The forward and reverse diffusion processes are shown in Figure 1.

2.1. Diffusion Process

Let $\mathbf{y_0}$ be the high-resolution image, then we define the forward first-order Markovian diffusion process q as:

$$q(\mathbf{y_t}|\mathbf{y_{t-1}}) = \mathcal{N}(\mathbf{y_t}; \sqrt{\alpha_t}\mathbf{y_{t-1}}, (1-\alpha_t)\mathbf{I}) \tag{1}$$

$$q(\mathbf{y_{1:T}}|\mathbf{y_0}) = \prod_{t=1}^{t=T} q(\mathbf{y_t}|\mathbf{y_{t-1}}) \tag{2}$$

where $\alpha_{1:T}$ are hyper-parameters such that $0 < \alpha_i < 1$ and $(\mathbf{y_1}, \mathbf{y_2}, ..., \mathbf{y_{T-2}}, \mathbf{y_{T-1}})$ are the intermediate state images. α_t controls the variance of the noise and $\sqrt{\alpha_t}$ makes sure the variance of random variables doesn't explode as $t \to \infty$. By marginalizing, the distribution of any intermediate state image $\mathbf{y_t}$ given $\mathbf{y_0}$ can be written as:

$$q(\mathbf{y_t}|\mathbf{y_0}) = \mathcal{N}(\mathbf{y_t}; \sqrt{\gamma_t}\mathbf{y_0}, (1-\gamma_t)\mathbf{I}) \tag{3}$$

$$\gamma_t = \prod_{i=1}^{i=t} \alpha_i \tag{4}$$

The posterior distribution of $\mathbf{y_{t-1}}$ given $(\mathbf{y_t}, \mathbf{y_0})$ can be written as:

$$q(\mathbf{y_{t-1}}|\mathbf{y_t}, \mathbf{y_0}) = q(\mathbf{y_t}|\mathbf{y_{t-1}}, \mathbf{y_0}) \frac{q(\mathbf{y_{t-1}}|\mathbf{y_0})}{q(\mathbf{y_t}|\mathbf{y_0})} \tag{5}$$

On solving the above equation, we get:

$$q(\mathbf{y_{t-1}}|\mathbf{y_t}, \mathbf{y_0}) = \mathcal{N}(\mathbf{y_{t-1}}; \mu_t, \sigma_t^2\mathbf{I}) \tag{6}$$

$$\mu_\mathbf{t} = \frac{\sqrt{\gamma_{t-1}}(1-\alpha_t)}{1-\gamma_t}\mathbf{y_0} + \frac{\sqrt{\alpha_t}(1-\gamma_{t-1})}{1-\gamma_t}\mathbf{y_t} \tag{7}$$

$$\sigma_t^2 = \frac{(1-\gamma_{t-1})(1-\alpha_t)}{1-\gamma_t} \tag{8}$$

We tried to maintain the notation the same as [18] to explain the difference in approaches simpler.

2.2. Training the Denoising Model

We use a neural-network f_θ to denoise the image in the reverse diffusion process. The objective of this denoising model is to estimate the noise in the image. To make sure we don't lose sight of our super-resolution task we condition the task by providing additional information i.e. the input image of the process $\mathbf{x}$ and the amount of noise present in the image given as input to the denoising model. So, our training objective is to minimize:

$$\mathbb{E}_{(\mathbf{x},\mathbf{y})}\mathbb{E}_{(\epsilon,\gamma)}\|f_\theta(\mathbf{x}, \tilde{\mathbf{y}}, \gamma) - \epsilon\|^2 \tag{9}$$

$$\tilde{\mathbf{y}} = \sqrt{\gamma}\mathbf{y_0} + (1-\gamma)\epsilon \tag{10}$$

$$\epsilon \sim \mathcal{N}(0, \mathbf{I}) \tag{11}$$

where $\tilde{y}$ is obtained by applying the reparameterization trick on Eq. 3 and its represents a noisy input i.e $\mathbf{y_t}$ with $\gamma \sim p(\gamma)$.

2.3. Inference

Similar to [18], we shall perform iterative refinement. During inference which is a reverse Markovian process, we start from the pure Gaussian noise and work our way to high-resolution image and the parameterized process can be modeled as:

$$p_\theta(\mathbf{y_{t-1}}|\mathbf{y_t}, \mathbf{x}) = \mathcal{N}(\mathbf{y_{t-1}}; \mu_\theta(\mathbf{x}, \mathbf{y_t}, \gamma_t), \sigma_t^2 I) \tag{12}$$

During inference we are conditioning the reverse Markov process such that is it close to the true reverse of the forward Markov diffusion process. So, by applying reparameterization-trick and rearranging the terms of Eq. 3 we can estimate $\mathbf{y_0}$:

$$\hat{\mathbf{y}}_\mathbf{0} = \frac{1}{\sqrt{\gamma_t}}\left(\mathbf{y_t} - \sqrt{1-\gamma_t} f_\theta(\mathbf{x}, \mathbf{y_t}, \gamma_t)\right) \tag{13}$$

By the substituting the estimated $\hat{\mathbf{y}}_\mathbf{0}$ in Eq. 7 and applying reparameterization-trick on Eq. 6 we get,

$$\mu_\theta(\mathbf{x}, \mathbf{y_t}, \gamma_t) = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{y_t} - \frac{1-\alpha_t}{\sqrt{1-\gamma_t}} f_\theta(\mathbf{x}, \mathbf{y_t}, \gamma_t)\right) \tag{14}$$

$$\mathbf{y_{t-1}} = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{y_t} - \frac{1-\alpha_t}{\sqrt{1-\gamma_t}} f_\theta(\mathbf{x}, \mathbf{y_t}, \gamma_t)\right) + \sqrt{1-\alpha_t}\epsilon_\mathbf{t} \tag{15}$$

$$\epsilon_\mathbf{t} \sim \mathcal{N}(0, \mathbf{I}) \tag{16}$$

2.4. Zero-Shot Inference a Null-Space Model

The Range-Null space decomposition, proposed in [23], offers a new perspective on the relationship between realness and data consistency. By modeling and finding the appropriate null-space contents, diffusion models satisfy realness while ensuring that data consistency, which is only related to range-space contents is strictly guaranteed.

Assume that the image down-sampling task is a matrix multiplication for simplicity. Let $\mathbf{A}$ denote the linear degradation or down-sampling operator, $\mathbf{y_0}$ denotes the high-resolution image corresponding to $\mathbf{x_0}$ which is the low-resolution image. So, the process of down-sampling can be modeled as:

$$\mathbf{x_0} = \mathbf{A}\mathbf{y_0} \tag{17}$$

Let the pseudo inverse of $\mathbf{A}$ be $\mathbf{A^{-1}}$ which can be calculated using SVD etc. The properties of $\mathbf{A}$ and $\mathbf{A^{-1}}$ can be leveraged to model any sample $\mathbf{y_0}$. Any sample can be decomposed as:

$$\mathbf{y_0} = \mathbf{A^{-1}}\mathbf{A}\mathbf{y_0} + (\mathbf{I} - \mathbf{A^{-1}}\mathbf{A})\mathbf{y_0} \tag{18}$$

The same approach of iterative refinement is used for refining null-space components using the reverse diffusion process. The process of reverse diffusion can be modeled as follows:

$$\mathbf{y_{0|t}} = \frac{1}{\sqrt{\gamma_t}}\left(\mathbf{y_t} - \sqrt{1-\gamma_t} f_\theta(\mathbf{y_t}, \gamma_t)\right) \tag{19}$$

$$\hat{\mathbf{y}}_\mathbf{0} = A^{-1}\mathbf{x_0} + (\mathbf{I} - \mathbf{A^{-1}}\mathbf{A})\mathbf{y_{0|t}} \tag{20}$$

$$\mathbf{y_{t-1}} \sim p_\theta(\mathbf{y_{t-1}}|\mathbf{y_t}, \hat{\mathbf{y}}_\mathbf{0}) \tag{21}$$

$$\tag{22}$$

where $\mathbf{y_{0|t}}$ is initial estimate of $\mathbf{y_0}$ and it is refined later using null-space contents. And finally, the iterative refinement is used to generate a super-resolution or high-resolution output of input $\mathbf{x_0}$.

Although SR3 [15] and DDNM (Denoising diffusion null-space model) [23] use iterative refinement, the procedure is used to estimate different terms in each of them. SR3 [15] uses iterative refinement to slowly remove the noise contents at a time-step t from the input and it takes both the noisy image $\mathbf{y_t}$ and input $\mathbf{x}$. In SR3 [15], to condition the model on the input $\mathbf{x}$, it is interpolated to higher resolution and concatenated with $\mathbf{y_t}$ along channel dimension and passed as input to the model Eq.16. But in DDNM [23], the input $\mathbf{x}$ is not passed as input to the model Eq.22 which estimates null-space contents but is later used to refine the initial estimate $\mathbf{y_{0|t}}$. And also a key difference is a conditional DDPM is used in SR3 [15] and an unconditional DDPM is used in DDNM [23].

3. Related Work

3.1. Diffusion Models

Diffusion models are a type of generative model that has gained attention in recent years due to their high-quality and diverse image generation capabilities, as demonstrated in notable works such as [5, 13, 17, 19, 23]. These models have been successfully applied to various computer vision tasks including image generation, SISR, and image inpainting.

Ho et al. [5] proposed a cascaded design for the SISR task. Their method first generates a low-resolution image, which is then fed to subsequent models to obtain a high-resolution version. Benny et al. [1] conducted a study on predicting sample images or noise using diffusion models. They observed that returning noise and image during reverse diffusion could benefit some problems.

Saharia et al. [16] proposed a novel design for image-to-image translation using diffusion models and showed that L_2 loss and self-attention can improve the quality and detail of the output. Wang et al. [21] used pre-trained models from GLIDE [11] which are adjusted for conditional input and fine-tuned to get semantic latent space for specific image generation tasks. DDPMs learn the dataset distribution by iteratively denoising samples from a noise distribution, resulting in high-quality results for various image generation tasks. However, they require many iterations and are computationally expensive, prompting recent work to focus on reducing inference time [13].

3.2. Super Resolution

Generative models, such as Variational Autoencoders (VAEs), have been used to learn the underlying distribution of high-resolution images and generate SR outputs, as deep neural networks have suffered from poor details, over-smoothed regions, and overfitting. However, VAEs struggle with producing high-quality results due to the limitations of the mean-field approximation used in training. To address these limitations, diffusion models have emerged as a promising approach for generative modeling. While diffusion models have been successful in image-to-image translation tasks, limited research has been done for specific SISR tasks.

SR3 [18] is a diffusion model that utilizes a U-Net model [14] trained on denoising at various noise levels. This model is simple to train because it minimizes a well-defined loss function without regularization and optimization tricks as GANs do. SR3 achieves conditional image generation by adapting DDPMs [4, 19] through a stochastic denoising process. However, it exhibits limitations when handling OOD data, particularly images captured in the wild with unknown types of degradation. To address this issue, SR3+ [15] was presented, which trains with a combination of composite, parameterized degradations in the data augmentation pipeline for self-supervised training, and noise-conditioning augmentation, which was first utilized in cascaded diffusion models [5], achieving robustness to OOD inputs.

Cascading diffusion pipelines [5] benefit from conditioning augmentation, which improves sample quality without using image classifiers. The cascaded model comprises multiple diffusion models, each using a U-Net architecture trained on denoising at different noise levels [4, 18].

Latent Diffusion Models (LDMs) [13] have been proposed as a solution to reduce the complexity of training diffusion models, which require significant computational resources, especially when training on high-resolution images. LDMs train a universal autoencoding stage to provide a low-dimensional space that can then be reused for multiple diffusion model training or to explore different tasks, resulting in less computation. Furthermore, LDMs combine the power of Diffusion Models with Generative Modeling of Latent Representations, which enhances their performance for super-resolution. LDMs can also be trained to perform super-resolution directly by conditioning low-resolution images through concatenation, making them a flexible and efficient solution for this task.

SRDiff [8] is a diffusion model with fast training speed using residual prediction. It employs a pre-trained encoder to transform the low-resolution input image into hidden conditions, followed by a conditional noise predictor to generate the high-resolution image through a Markov chain. SRDiff is stable and easy to train and has shown greater performance compared to GAN-based methods.

4. Methods

Current diffusion model-based solutions for SISR often suffer from feature hallucination, consistent artifacts, and poor preservation of high-level image attributes. To evaluate the generalization ability of these models on OOD datasets, we propose to use SR3 as our baseline model for SISR and investigate its performance on OOD data. We further explore the impact of fine-tuning variables, such as time steps, iterations, and dataset size, on the performance of the model. Additionally, we examine the zero-shot learning capability of large diffusion models for natural image super-resolution, by comparing pre-trained guided diffusion models from OpenAI. Finally, we train SR3 from scratch on the Imagenet-1K dataset to achieve results on par with zero-shot learning on pre-trained models.

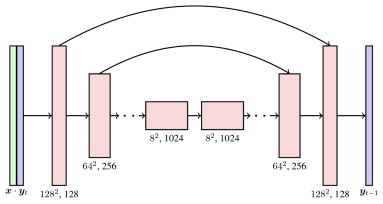

Figure 2. SR3 utilizes the modified U-Net architecture with skip connections.

SR3: SR3 utilizes a modified U-Net architecture, as illustrated in Figure 2, with self-attention layers for improving image quality. In the models' training pipeline, low-resolution images are first interpolated to the target high resolution and then concatenated with noisy high-resolution images. For 16x16 to 128x128 tasks, the total number of parameters amounts to 550M million, making it computationally expensive, especially for larger image sizes.

Fine-tuning: Although SR3 has shown impressive performance for super-resolution tasks, it has limitations when handling OOD data, as shown in Figure 3. To improve its generalization ability, we performed fine-tuning on the pre-trained model. The details of the fine-tuning settings will be introduced in Section 5.2.

Zero-shot learning approach: Zero-shot learning methods for super-resolution often struggle to balance realness and data consistency. The Range-Null space decomposition offers a new perspective on the relationship between these two factors, where data consistency is only related to the range-space contents, which can be calculated. Diffusion models are ideal tools for solving this problem, and thus can be combined with null-space decomposition to create a Denoising Diffusion Null-Space Model, as introduced in Sec. 2.4. Our approach refines only the null-space contents during the reverse diffusion sampling, resulting in realistic and data-consistent results using only a pre-trained diffusion model.

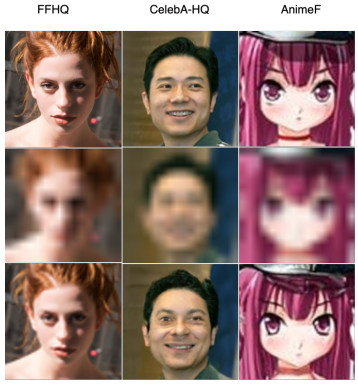

Figure 3. Representative output of implemented and fine-tuned version of SR3. The top row shows the high-resolution image, the middle row shows the low-resolution input image, and the bottom row shows generated super-resolved image at 16x16 to 128x128.

5. Experiments

In our work, first, we focus on fine-tuning SR3 [15] as our base model for our experiments, which was trained on FFHQ [7] from scratch until convergence for 8x upsampling from 16x16 to 128x128. We also trained our implemented version on Imagenet-1K [2] from scratch for the same task due to resource constraints. Additionally, we used a pre-trained model from guided diffusion for zero-shot learning experiments. We divided our experiments into two categories: 1) face datasets containing human and anime faces,

Figure 4. Results of fine-tuning SR3 pretrained on CelebA dataset on AnimeF dataset with a limited number of time steps.

Figure 5. Results of fine-tuning SR3 pre-trained on CelebA dataset on AnimeF dataset with limited iterations.

and 2) natural images. Our fine-tuning experiments are divided into three categories: limited-time steps, limited iterations, and limited data samples.

5.1. Datasets

We trained our model on different datasets for different experiments. For face super-resolution, we trained our model on Flickr-Faces-HQ (FFHQ) [7], while for natural image super-resolution, we used Imagenet-1K [2]. We also used AnimeF and DF2K-OST [25] datasets for our experiments.

FFHQ [7] and CelebA-HQ [9] are human-face datasets, FFHQ is two times larger than CelebA-HQ, which has around 30,000 samples. We used the AnimeF dataset for our fine-tuning experiments, which consists of around 63,565 total samples, and we used a 70-30 train-test split. The DF2K-OST dataset is a combination of Div2K, Flick2K, and OST datasets, each having 800, 2650, and 300 samples.

5.2. Fine Tuning Details

We trained and fine-tuned our models on TACC with a batch size of 16, using the Adam optimizer with a fixed learning rate of 1e-4, consistent with the original SR3 method. To condition low-resolution input, we resized them to input high-resolution with bicubic interpolation, as done in SR3. Additionally, we used a simple linear noise schedule for α during our experiments.

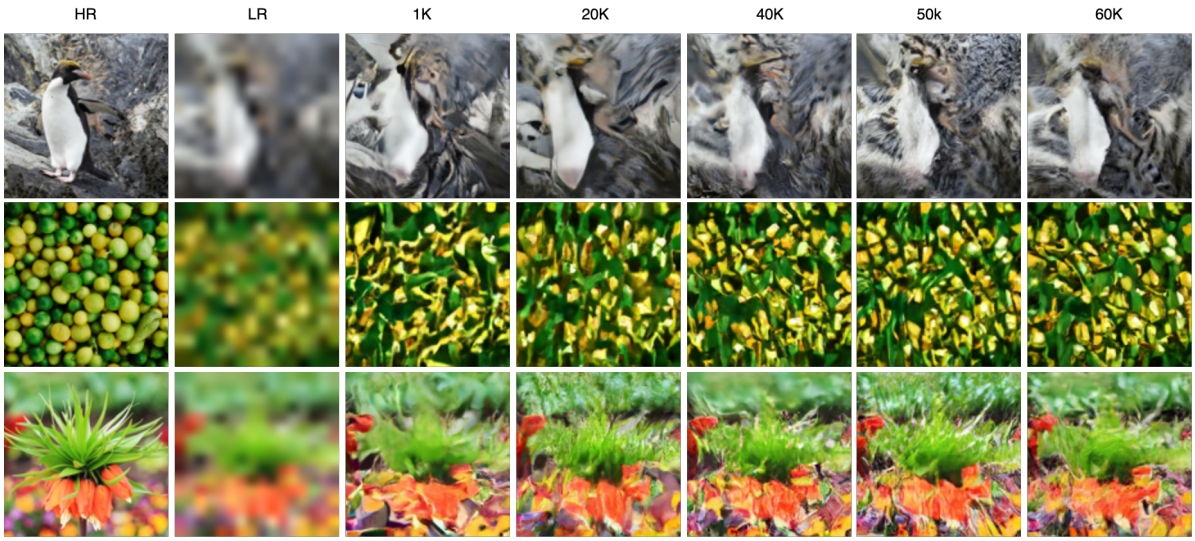

Figure 6. SR results on DF2K-OST dataset using pre-trained FFHQ model.

Figure 7. Results of fine-tuning SR3 pretrained on CelebA dataset on AnimeF dataset with limited numbers of data samples.

Evaluation Metrics: Our evaluation of the performance of all models is based on standard image quality estimation techniques, specifically, PSNR and SSIM [24]. These metrics have been widely used in the industry for the past few decades and are the go-to image quality evaluation metrics. While PSNR does not incorporate human perceptual factors and correlates less with human judgments, SSIM [24] was designed with human perception in mind and, as a result, correlates more with human judgments.

In the following two subsections, we do a detailed discussion about qualitative and quantitative results. We approach each subsection by dividing it into four main categories, three for fine-tuning.

5.3. Qualitative Results

Time-Steps in Diffusion Models: During the training process, diffusion models learn across all time-steps to generate the final image output using the reverse Markov chain process. It is important to note that these models learn the noise added to the image following a specific distribution, which implies that providing an OOD sample might not yield satisfactory results. Moreover, the reverse diffusion process generates prominent shapes and structures at the initial time-steps and progressively adds high-frequency details at later stages. Consequently, fine-tuning only the first few hundred time-steps out of a total of 2000 should suffice in producing decent results, as this would essentially reduce

the domain gap. To validate our hypothesis, we fine-tuned our SR3 model, which had been pre-trained on the FFHQ dataset, for various time-steps, specifically: 0-200, 0-500, 0-1000, 1000-1500, 1500-2000, and 0-2000.

Figure 4 shows that satisfactory reconstruction with sharp details can be achieved even when fine-tuning is limited to specific time-steps. The results presented in the figure encompass up to 4000 fine-tuning iterations and are comparable to, or even surpass, those achieved by fine-tuning across all time-steps. It is postulated that as more time-steps are fine-tuned, additional iterations are necessary to diminish the domain gap, and thus, for a fixed number of iterations generated images with more time-steps exhibit fewer details when compared to generated images with less time-steps.

Fine-tuning diffusion models across a reduced number of time-steps allows for faster training and convergence, which can be beneficial in scenarios with limited computational resources or when quick experimentation is desired. However, limiting the number of time-steps being fine-tuned may result in generated images lacking desired levels of detail and realism. Therefore, it is important to find a balance between computational efficiency and capturing high-frequency details and subtle nuances present in the target domain. Overall, fine-tuning diffusion models across a limited range of time-steps has been shown to be effective in reducing the domain gap and generating plausible images.

Fine-tuning iterations: We fine-tuned our pre-trained model on approximately 45,000 samples from the AnimeF dataset for more than 100,000 iterations, reserving the remaining samples for evaluation purposes. Qualitative results in Figure 5 demonstrate that increasing the number of iterations enhances the quality of the generated images. However, it is important to strike a balance between improving image quality and maintaining consistent performance on the source dataset. Figure 5 shows that satisfactory results can be obtained at 4,000 iterations, but increasing the number of iterations may lead to diminishing sharpness and even overfitting on the target dataset.

We also fine-tuned the same FFHQ [7] pre-trained model on DF2K-OST [25], a dataset with completely unrelated distributions. However, the results, as shown in Figure 6, were not satisfactory, indicating a large domain gap. Although training the model for longer iterations produced some recognizable reconstructions, it was still not comparable to the level of generalization achieved in the AnimeF/similar domain experiments.

Limited Dataset: When adapting a pre-trained model on the target domain, the total number of samples is generally lesser in the scales of magnitudes. To simulate this scenario, we restricted the training samples for AnimeF datasets to only 10% of the original, approximately 4500 samples. Figure 7 shows the results of fine-tuning on the limited dataset. We found that achieving results comparable to fine-tuning on all samples for 4000 iterations, only after 10,000 iterations. This is because super-resolution gives better results with more data from the target domain and fewer details with less data.

Time-Steps	PSNR	SSIM
0-200	**18.972**	**0.49085**
0-500	18.604	0.47024
0-1000	**18.943**	**0.50175**
1000-1500	18.214	0.45565
1500-2000	18.089	0.44747
0-2000	18.728	0.47721

(a) Different time steps.

Iterations	PSNR	SSIM
4K	18.728	0.47721
20K	18.778	0.48047
30K	19.186	0.50921
50K	**19.593**	**0.51913**
70K	19.453	0.51075
100K	19.174	0.51628

(b) Different fine-tuning iterations.

Table 1. Quantitative results on different experiments.

To justify the use of the pre-trained model, we also trained our model on AnimeF from scratch on all training samples. As shown in Figure 8, our model generated good results even after 140,000 iterations but failed to generate the appropriate colors and gave inconsistent shifted hues. In contrast, we obtained consistent results with the pre-trained model with just under 4000 iterations.

Zero-shot learning: We used zero-shot learning for natural images, where the pre-trained model guided diffusion model was employed, with weights publicly available at https://github.com/openai/guided-diffusion/tree/22e0df8183507e13a7813f8d38d51b072ca1e67c. Figure 9 shows the results from the guided diffusion model pre-trained unconditionally on the ImageNet dataset. However, the results lacked sharpness due to the limitation in the generative capacity of the imagenet pre-trained model. Instead of relying on a different model, we also trained our SR3 implementation version from scratch on the imagenet dataset. Despite initial hue shift issues, the model stabilized after 150,000 iterations. Due to time constraints, the training was not further pursued, leaving it for future tasks. Figure 10 displays the results for training over iterations.

5.4. Quantitative Evaluation

Fine-tuning with limited time-steps: Table 1a presents the quantitative evaluation of different fine-tuning time steps using PSNR and SSIM as the evaluation metrics. The results indicate that the model performs best when fine-tuned in the early time steps, particularly in time steps 0-200 and 0-1000. Specifically, the model fine-tuned in time steps 0-200 achieves the highest PSNR score, while the model fine-tuned in time-steps 0-1000 has a better SSIM score. As a reduced number of time steps leads to faster training and convergence, 0-200 is considered the optimal number of time steps for this case.

Fine-tuning with limited iterations: In this experiment, we evaluate the performance of fine-tuning a pre-

Figure 8. We trained the SR3 model from scratch using around 45,000 samples from the AnimeF dataset.

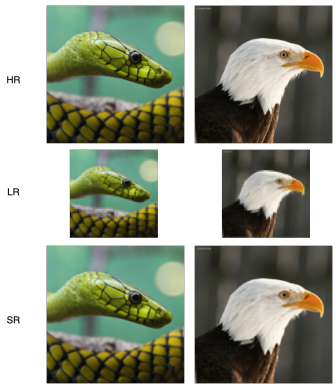

Figure 9. Samples from zero-shot learning using an ImageNet pre-trained model from guided diffusion.

trained model with a limited number of iterations. Table 1b shows the quantitative evaluation on different fine-tuning iterations using PSNR and SSIM as evaluation metrics. As we can observe from the table, both PSNR and SSIM values increase as the number of iterations increases. The model achieves the highest performance when the number of iterations reaches 50K. Beyond 50K iterations, we observe that the PSNR score starts to slightly decrease while the SSIM score has a different trend of decreasing. This indicates that the model has reached a point of convergence where further iterations may not necessarily result in better performance.

5.5. Catastrophic forgetting

In the previous sections, we demonstrated that good SR results can be achieved on new domain data with limited fine-tuning steps in the previous sections. However, it is important to investigate whether the fine-tuned model still performs well on the original dataset, in this case, the CelebA-HQ dataset. Unfortunately, we observed catastrophic forgetting when the CelebA-HQ pre-trained SR3 model was fine-tuned on the AnimeF dataset for multiple iterations. As the model was fine-tuned on a new data distribution, it forgot the previously learned information from CelebA-HQ after a certain number of iterations. As shown in Figure 11, more distortions were observed in the SR images when fine-tuning more iterations. While the fine-tuned model achieved the highest performance on the AnimeF dataset when the number of iterations reached 50K, the results in Figure 12 demonstrate the catastrophic forgetting results on the original CelebA-HQ dataset. This suggests that a fine-tuned model on a new dataset may not perform as well on the original dataset, and further investigation is needed to mitigate the issue of catastrophic forgetting.

6. Conclusion and Future Works

In this work, we proposed an approach to address the challenges of training diffusion models for super-resolution tasks by exploring the usefulness of pre-trained models, specifically zero-shot and fine-tuning approaches. We demonstrated that these approaches can increase the generalization ability and reduce the effort of training from scratch. However, we also observed the potential issue of catastrophic forgetting when fine-tuning a new data distribution, which is an important challenge to overcome in future research.

Our work opens up new avenues for exploring the use-

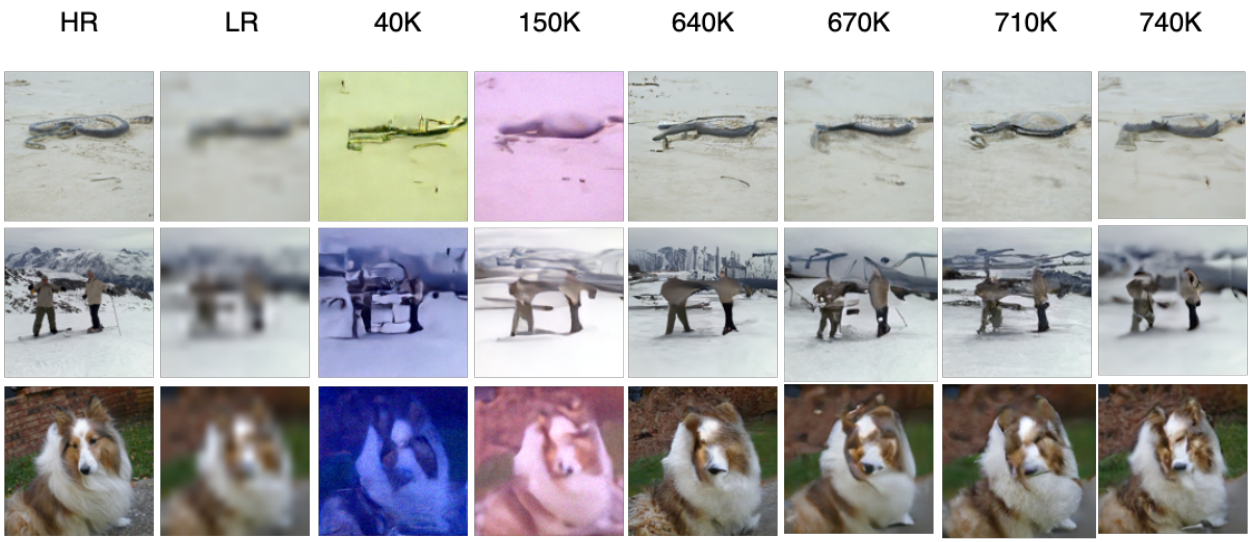

Figure 10. We trained the SR3 model from scratch using Imagenet-1K dataset.

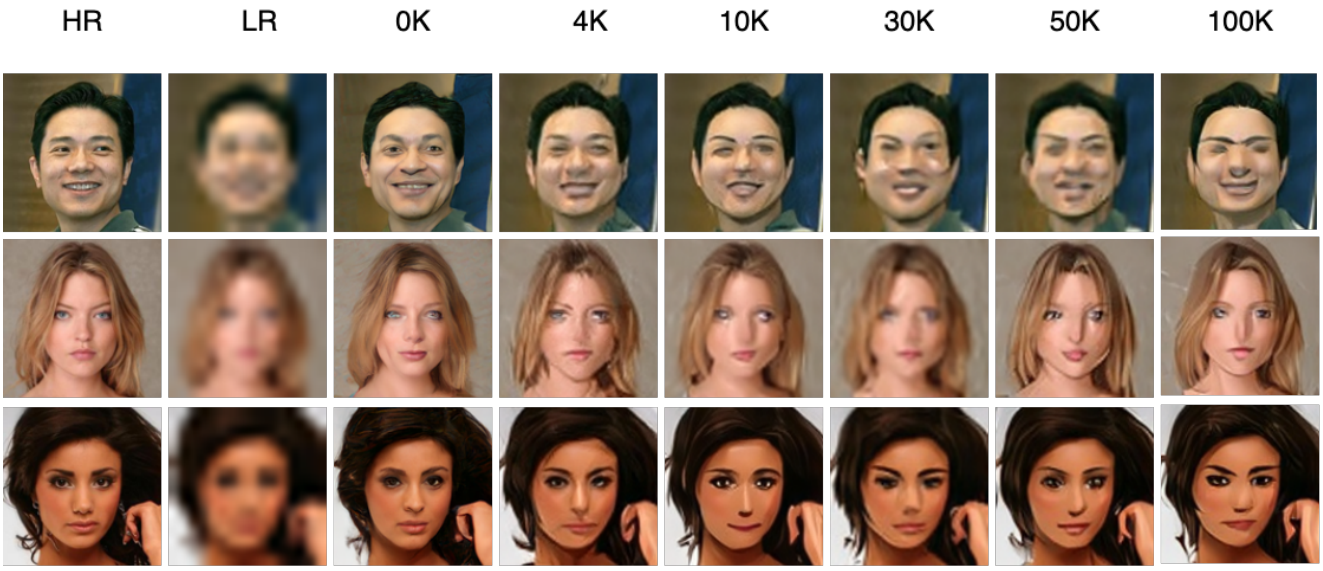

Figure 11. Results on CelebA-HQ dataset when CelebA-HQ dataset pre-trained SR3 model is fine-tuned on AnimeF dataset.

fulness of pre-trained models in other domains, such as face and OOD data. We plan to conduct more experiments to investigate the effectiveness of zero-shot conditional diffusion models on other datasets and domains.

7. Teamwork Assignment

Shreshth Saini: Coding, experiments related to fine-tuning, training from scratch, presentation, and report.

Yu-Chih Chen: Coding, all zero-shot related experiments, presentation, and report.

Krishna Srikar Durbha: Coding, experiments related to fine-tuning, code debugging, presentation, and report.

References

[1] Yaniv Benny and Lior Wolf. Dynamic dual-output diffusion models. *2022 IEEE/CVF Conference on Computer Vision*

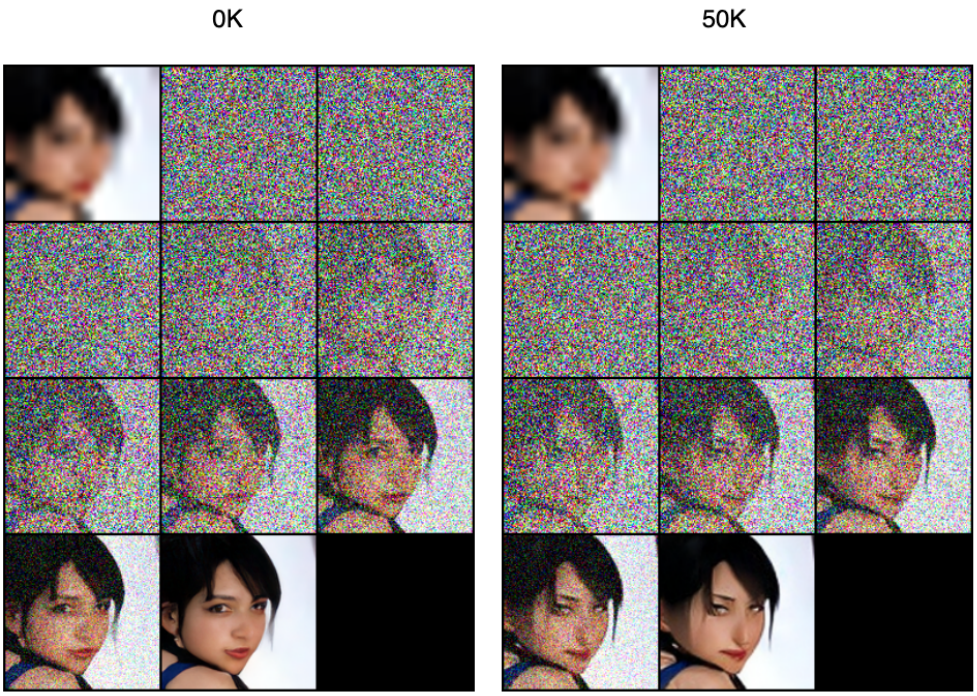

Figure 12. Catastrophic forgetting results from 0 to 50K iterations on original CelebA-HQ after fine-tuning on AnimeF.

and Pattern Recognition (CVPR), pages 11472–11481, 2022. 3

[2] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 4, 5

[3] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. 1

[4] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 3

[5] Jonathan Ho, Chitwan Saharia, William Chan, David J. Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *J. Mach. Learn. Res.*, 23:47:1–47:33, 2021. 3

[6] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *International Conference on Learning Representations*, 2018. 1

[7] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. *CoRR*, abs/1812.04948, 2018. 4, 5, 7

[8] Haoying Li, Yifan Yang, Meng Chang, Shiqi Chen, Huajun Feng, Zhihai Xu, Qi Li, and Yueting Chen. Srdiff: Single image super-resolution with diffusion probabilistic models. *Neurocomputing*, 479:47–59, 2022. 4

[9] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015. 5

[10] Sachit Menon, Alexandru Damian, Shijia Hu, Nikhil Ravi, and Cynthia Rudin. Pulse: Self-supervised photo upsampling via latent space exploration of generative models. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2434–2442, 2020. 1

[11] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models, 2021. 3

[12] Suman V. Ravuri and Oriol Vinyals. Classification accuracy score for conditional generative models. *ArXiv*, abs/1905.10887, 2019. 1

[13] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021. 3, 4

[14] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham, 2015. Springer International Publishing. 3

[15] Hshmat Sahak, Daniel Watson, Chitwan Saharia, and David Fleet. Denoising diffusion probabilistic models for robust image super-resolution in the wild, 2023. 1, 3, 4

[16] Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 Conference Proceedings*, SIGGRAPH '22, New York, NY, USA, 2022. Association for Computing Machinery. 3

[17] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding, 2022. 3

[18] 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*, pages 1–14, 2022. 1, 2, 3

[19] Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, page 2256–2265. JMLR.org, 2015. 1, 3

[20] Hoang Thanh-Tung and T. Tran. Catastrophic forgetting and mode collapse in gans. *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–10, 2020. 1

[21] Tengfei Wang, Ting Zhang, Bo Zhang, Hao Ouyang, Dong Chen, Qifeng Chen, and Fang Wen. Pretraining is all you need for image-to-image translation, 2022. 3

[22] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 1905–1914, 2021. 1

[23] Yinhuai Wang, Jiwen Yu, and Jian Zhang. Zero-shot image restoration using denoising diffusion null-space model, 2022. 1, 2, 3

[24] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. 6

[25] Kai Zhang, Jingyun Liang, Luc Van Gool, and Radu Timofte. Designing a practical degradation model for deep blind image super-resolution. In *IEEE International Conference on Computer Vision*, pages 4791–4800, 2021. 5, 7