Expanding Expressiveness of Diffusion Models with Limited Data via Self-Distillation based Fine-Tuning

Jiwan Hur, Jaehyun Choi, Gyojin Han, Dong-Jae Lee, and Junmo Kim
School of Electrical Engineering, KAIST, South Korea
{jiwan.hur, chlwogus, gjhan0820, jhtwosun, junmo.kim}@kaist.ac.kr

Abstract

Training diffusion models on limited datasets poses challenges in terms of limited generation capacity and expressiveness, leading to unsatisfactory results in various downstream tasks utilizing pretrained diffusion models, such as domain translation and text-guided image manipulation. In this paper, we propose Self-Distillation for Fine-Tuning diffusion models (SDFT), a methodology to address these challenges by leveraging diverse features from diffusion models pretrained on large source datasets. SDFT distills more general features (shape, colors, etc.) and less domain-specific features (texture, fine details, etc) from the source model, allowing successful knowledge transfer without disturbing the training process on target datasets. The proposed method is not constrained by the specific architecture of the model and thus can be generally adopted to existing frameworks. Experimental results demonstrate that SDFT enhances the expressiveness of the diffusion model with limited datasets, resulting in improved generation capabilities across various downstream tasks.

1. Introduction

Recently, Diffusion Models [13, 31] (DMs) have emerged as a powerful family of generative models due to their diverse and high-quality image generation capability. While generative adversarial networks (GANs) [10] show powerful generating capabilities in synthesizing high-quality images, they are known to have poor mode coverage [29,43]. On the other hand, diffusion models are formulated to approximate the data distribution through likelihood estimation with a denoising score matching [35], and diffusion models trained on large datasets such as ImageNet [7] outperform state-of-the-art GAN-based methods [2], in image generation in terms of image fidelity and diversity [9].

However, despite intensive research on diffusion models using large-scale datasets, there has been relatively little focus on training diffusion models on limited datasets. Limited datasets, compared to large datasets, are more susceptible to bias and often lack diversity in terms of the range of images and attributes they contain. For instance, MetFaces [14], containing approximately 1K faces from medieval artworks, lacks diversity in facial attributes when compared to FFHQ [15], which contains 70K real human faces as shown in Fig. 1. In each domain, both datasets exhibit some shared facial attributes (region ①). However, different from the FFHQ, MetFaces does not contain more various facial attributes (region ②) such as skin and hair colors, facial expressions, accessories, *etc.* Consequently, diffusion models trained on limited datasets may have limited *expressiveness*, generating less diverse outputs and exhibiting biases in their representation.

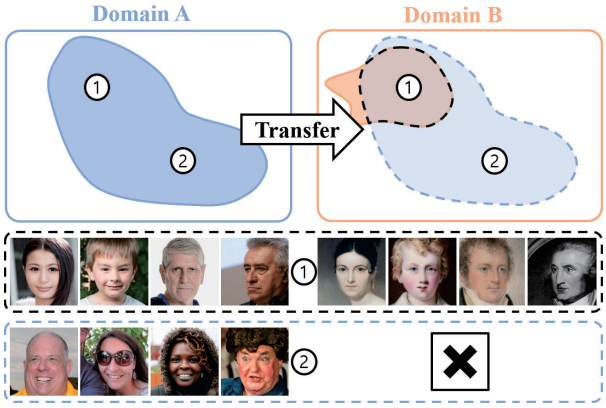

Figure 1. Compared to the large datasets on Domain A (e.g. FFHQ [15]), limited datasets on domain B (e.g. MetFaces [14]) constrain the ability for diverse image generation and manipulation in Domain B. The goal of this paper is to utilize a model pretrained on Domain A (source), and effectively transfer the diverse knowledge while training on Domain B (target).

The limited expressiveness of diffusion models not only hampers the generation capability of the model but also results in unsatisfactory outputs in various downstream tasks such as domain translation [3,42,42] and text-guided image manipulation [17, 19] since these methods heavily depend

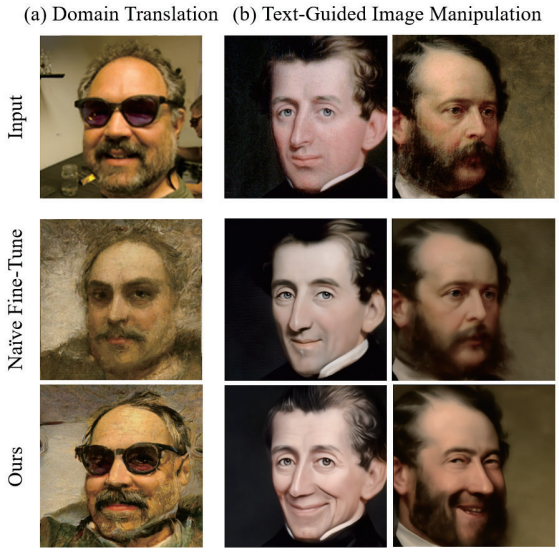

Figure 2. Results of downstream tasks which utilize diffusion models trained on MetFaces, such as (a) domain translation [23] and (b) text-guided image manipulation [19] (script: *smiling*). Since their performance is highly affected by the expressiveness of the model, naïvely fine-tuning the model pretrained on FFHQ [15] results in the loss of crucial attributes (sunglasses) in (a) and restricted facial expression in (b) (smiling). With our proposed fine-tuning method, SDFT, the model can inherit diverse attributes from the source model, effectively resolving these problems.

on diffusion models to generate plausible outputs.

To mitigate the aforementioned problems, in this paper, we aim to utilize diverse knowledge from the source diffusion model which is trained on large diverse datasets. However, naïvely fine-tuning the model on limited datasets can lead to a well-known catastrophic forgetting problem, where the model loses diverse knowledge during the fine-tuning process. To this end, we propose Self-Distillation for Fine-Tuning diffusion models (SDFT), which leverages the diverse features from the diffusion models pretrained on large source datasets. Specifically, to successfully transfer diverse knowledge from the source model without disturbing the training process on target datasets, SDFT prioritizes distilling general features (shape, color, *etc.*) from the source model, while less emphasizing domain-specific features (texture, fine details, *etc.*) from the source model. Furthermore, we propose an auxiliary input to effectively transfer more diverse information from the source model with limited datasets. It is worth noting that SDFT is not constrained by the specific architecture of the model, thus it can be generally adopted by existing frameworks.

Throughout various experiments, we show that the enhanced expressiveness of diffusion models by the SDFT helps to generate more diverse attributes in various downstream tasks even though they are not contained in the target limited datasets, such as sunglasses and wide smile in Fig. 2. Furthermore, we present that enhanced expressiveness also can be beneficial to unconditional image generation, as it helps to generate diverse and high-fidelity images.

2. Background

2.1. Diffusion Models

Diffusion Models (DMs) [13, 31] perturb the complex data with the tractable noise, and aim to recover the data from the noise. Specifically, *forward process* is a Markov chain that gradually perturbs the input data $\boldsymbol{x}_0$ with Gaussian noise $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ and DMs ϵ_θ learn the inversion of the forward process, called *reverse process*, where the joint distribution $p_\theta(\boldsymbol{x}_{0:T})$ denotes the reverse process running from timestep T to 0. Then, DMs are trained to predict noise given time t with an objective function

$$\mathcal{L}(\epsilon_\theta) := \sum_{t=1}^{T} w_t \left[\frac{\beta_t}{(1-\beta_t)(1-\bar{\alpha}_t)} \|\epsilon_\theta(\boldsymbol{x}_t, t) - \epsilon\|_2^2 \right], \tag{1}$$

where β_t is a predefined noise schedule, which is a strictly decreasing function of time $t \in [0, T]$ and $\bar{\alpha}_t := \prod_{i=1}^{t} \alpha_i = \prod_{i=1}^{t} 1 - \beta_t$. Ho et al. [13] empirically found that loss function with $w_t = (1-\beta_t)(1-\bar{\alpha}_t)/\beta_t$ shows better results. Recently, Choi et al. [4] propose a weighting scheme, perception prioritized weighting (P2 weighting), which considers a signal-to-noise ratio (SNR) [18], such that

$$w'_t = \frac{w_t}{(k + \mathrm{SNR}(t))^\gamma}, \tag{2}$$

where SNR is strictly decreasing function of time t and can be defined as $\mathrm{SNR}(t) = \bar{\alpha}_t/(1-\bar{\alpha}_t)$. P2 weighting weights more in time steps with large SNR. This allows the model to focus more on high-level context. From the trained diffusion model, the realistic image $\boldsymbol{x}_0$ can be sampled from the initial noise $\boldsymbol{x}_T$ via stochastic Markovian sampling process [13, 31]. However, it takes several hundred to thousand sampling steps to generate an image. Song et al. [34] propose denoising diffusion implicit models (DDIMs), which break the Markov chain and allow the generation of reasonable samples with few generative steps. Additionally, DDIM provides deterministic sampling from the initial noise under the specific hyperparameter setting.

2.2. Image Translation in DMs

Earlier works on image translation have been studied using GANs and shown promising results [5,20,26,38]. However, they often fail to deal with various real-world images due to their limited mode coverage [17,33]. Since this drawback prevents a range of real-world applications, DMs have received great attention for various image translation tasks.

Domain Translation. The key concept behind domain translation in DMs [3, 23, 42] lies in leveraging the powerful generalization ability and expressiveness of the model ϵ^{trg} which is trained solely on the target datasets. Stochastic Differential Editing (SDEdit) [23] demonstrates that by leveraging noise-perturbed data from the other domain, $\boldsymbol{x}_t$, DMs can effectively translate $\boldsymbol{x}_t$ to the target domain $\boldsymbol{x}_0^{trg}$ through iterative denoising using ϵ^{trg}. Energy-Guided Stochastic Differential Equation (EGSDE) [42] further utilizes domain-specific and domain-independent energy functions during the sampling process for the more realistic and faithful translation, namely, domain classifier and low-pass filter, respectively. While these approaches have shown remarkable results, their abilities to capture and translate diverse attributes totally depend on the pretrained DMs.

Text-Guided Image Translation. Text-guided image translation in DMs [17, 19] mostly aims to fine-tune the pretrained DMs with a CLIP [27]. DiffusionCLIP [17] proposed to fine-tune the whole diffusion model with CLIP loss for robust image editing. Recently, the work by Kwon et al. [19], namely, Asyrp, demonstrates that fine-tuning only the deepest layer of the U-Net, instead of the entire diffusion model, leads to a more scalable, robust, and efficient fine-tuning process. However, since these methods fine-tune the diffusion model without further training datasets, their expressiveness also relies on the pretrained DMs.

2.3. Fine-Tuning Unconditional Diffusion Models

To our knowledge, fine-tuning unconditional DMs has not been widely researched yet. Recently, Moon et al. [24] introduced a fine-tuning method to prevent overfitting during training on limited datasets. They fix the pretrained model and introduce a learnable time-aware adapter that fine-tunes the attention block of the diffusion model. However, as their primary objective is to prevent overfitting, they do not consider transferring diverse knowledge from the source domain to the target domain.

3. Methods

3.1. Problem Definition

In this paper, we consider the training of diffusion models on limited datasets. As previously noted in the introduction and illustrated in Fig. 1, limited datasets tend to exhibit a reduced degree of diversity and inherent biases compared to large datasets. As a remedy to this issue, we fine-tune diffusion model ϵ_ϕ^{trg} which is initialized with source model ϵ_θ^{src} pretrained on a large source dataset. To successfully inherit the diverse information, while avoiding catastrophic forgetting, we distill the knowledge from fixed source diffusion model ϵ_θ^{src} during the training of the target diffusion model ϵ_ϕ^{trg}. In the following sections, we provide a comprehensive explanation of the proposed distillation scheme.

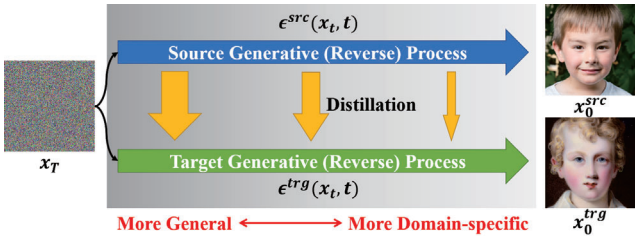

Figure 3. We employ weighted distillation approach which prioritizes the general features (color, shape, pose, *etc*.), while allocating lesser emphasis on domain-specific features (texture, fine-details, *etc*.). This allows the target model ϵ^{trg} to inherit various features from the source model ϵ^{src}, ensuring that the training process within the target domain remains undisturbed.

To simplify the presentation of our formulas, we omit the parameters of source and target diffusion models, θ and ϕ.

3.2. Fine-Tuning Diffusion Models with Distillation

In this section, we describe our approach to effectively distilling diverse information from the source model ϵ^{src} to the target model ϵ^{trg}, while ensuring the training on target datasets remains undisturbed. More specifically, we desire to distill more general features (color, shape, *etc*.) and less the domain-specific features (texture, fine details, *etc*.) from the source model ϵ^{src}.

For a similar objective, prior research on GANs has selected specific feature spaces of the generator and distilled the information in feature spaces [20]. However, their applicability may be constrained by the specific model architecture, such as StyleGAN2 [16] and careful design and analysis on the feature space need to be preceded [15, 16, 37].

To avoid these limitations and establish a more general methodology for DMs, we choose to distill the prediction of each time step. From the perturbed input data in target domain $\boldsymbol{x}_t^{trg}$, we distill the knowledge by matching each prediction of source and target diffusion models:

$$\mathcal{L}^{distill}(\epsilon^{trg}) = \sum_{t=1}^{T} \left[w_t^{distill} \cdot \|\epsilon^{src}(\boldsymbol{x}_t^{trg}, t) - \epsilon^{trg}(\boldsymbol{x}_t^{trg}, t)\|_2^2 \right], \quad (3)$$

where $w_t^{distill}$ denotes the distillation weight on time t.

Before deciding the $w_t^{distill}$, we emphasize that distinct from other generative models, DMs are known to generate images by the iterative reverse process where coarse features are generated initially, and fine details are integrated later [4]. In other words, during the reverse process $p_\theta(\boldsymbol{x}_{0:T})$, DMs synthesize general features in low SNR (large t) and gradually synthesize more domain-specific outputs in large SNR (small t). Thus we can set $w_t^{distill}$

to inversely proportional to the SNR(t)

$$w_t^{distill} = \frac{w_t}{(k + \text{SNR}(t))^{\gamma^{distill}}}, \tag{4}$$

which has a same formulation with P2 weighting in Eq. (2). However, we note that while P2 weighting aims to help diffusion models focus more on perceptually rich contents of the datasets, our proposed distillation weighting scheme is designed to distill more general features from the source model to the target model. With $w_t^{distill}$, the target model ϵ^{trg} can preserve general diverse features from source model ϵ^{src}. Moreover, it can be universally applied to any diffusion model, as it does not rely on the specific architecture of the model. Fig. 3 shows the overall illustration of the proposed distillation method.

However, while the distillation loss is beneficial in preserving the diversity from the source model, there still remain several considerations to be addressed. In the following sections, we tackle several expected problems and provide solutions if necessary.

3.3. Does $\epsilon^{src}(\boldsymbol{x}_t^{trg}, t)$ Generate Meaningful Output?

In Eq. (3), the source model ϵ^{src} generates output from the input $\boldsymbol{x}_t^{trg}$, which is a target domain sample perturbed by the Gaussian noise. Since ϵ^{src} is only trained with source datasets, $\boldsymbol{x}_t^{trg}$ can be considered as out-of-domain data, which may result in the failure to generate meaningful outputs from ϵ^{src}, leading to unsuccessful distillation.

Insights into this issue can be gained from previous research. Deja et al. [6] found that DMs have generalizability on other data distributions. Furthermore, studies on domain translation methods such as SDEdit [23] described in Sec. 2.2 demonstrated that adding noise to images from a similar but distinct domain and then running the reverse diffusion process can yield reasonable in-domain outputs. Thus, even with out-of-domain inputs, diffusion models can effectively convert into meaningful in-domain outputs, which enables successful distillation using Eq. (3). Moreover, SDEdit [23] demonstrates that running the reverse process in small t only changes domain-specific features (e.g. texture) from the input image while maintaining general features (e.g. silhouette). This means that the reverse process of DMs in small t generates domain-specific features while that in large t generates general features, which gives more justification for using $w_t^{distill}$ to prioritize general features over domain-specific features in Eq. (4).

3.4. Distilling More Diverse Features

Expanding on the issues outlined in Sec. 3.3, using $\boldsymbol{x}_t^{trg}$ as input introduces an additional unresolved problem. Remember that we consider the target datasets which include limited samples with reduced diversity compared to the source datasets. As a result, we can not fully extract the diverse features inherent within ϵ^{src} using limited input $\boldsymbol{x}_t^{trg}$.

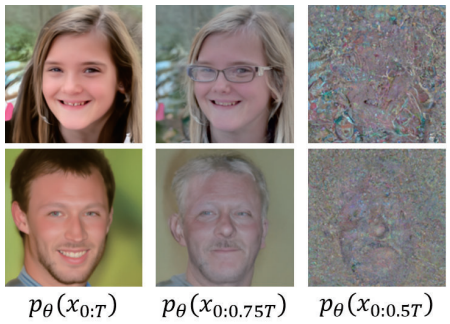

Figure 4. Images in each row are sampled from the same initial noises $p(\boldsymbol{x}_{T'}) = \mathcal{N}(0, \mathbf{I})$, but they take different *partial reverse process* $p_\theta(\boldsymbol{x}_{0:T'})$. It shows that when the T' is large, diffusion models can generate reasonable outputs from pure noise.

To address this issue, we propose to use auxiliary inputs to distill more diverse features from ϵ^{src}, without accessing the source or additional datasets. Using proxy inputs to transfer diverse knowledge from the teacher network has been widely explored in data-free knowledge distillation [39, 41]. They demonstrated that the utilization of synthesized input data effectively facilitates the transfer of diverse knowledge from the teacher network, even though these data significantly differ from the original source data. Inspired by this, we propose an auxiliary input specifically designed for the diffusion models to distill more diverse features from ϵ^{src}. Notably, we discovered that diffusion models possess the capability to generate plausible outputs from the pure noise $\boldsymbol{x}_T$, without requiring a perturbed image $\boldsymbol{x}_t$, in the initial reverse process. That means, diffusion models can generate meaningful outputs from pure noise in the initial reverse process. Fig. 4 shows sampled outputs obtained from the same initial noise in each row, but applying various *partial reverse process* $p_\theta(\boldsymbol{x}_{0:T'})$, defined in Sec. 2.1($(T' < T)$). In the second row, even though the initial reverse process ranging from T to 0.75T is skipped, diffusion models can generate reasonable outputs, albeit with some color degradation. To this end, we propose additional loss $\mathcal{L}^{aux}$ that utilizes pure noise as an auxiliary input for a diverse feature distillation:

$$\mathcal{L}^{aux}(\epsilon^{trg}) = \sum_{t=1}^{T} \left[w_t^{aux} \cdot \|\epsilon^{src}(\boldsymbol{x}_T, t) - \epsilon^{trg}(\boldsymbol{x}_T, t)\|_2^2 \right], \tag{5}$$

where w_t^{aux} is a same weighted distillation scheme in Eq. (4), but using different hyperparameter γ^{aux}. As depicted in the third column of Fig. 4, since the output drastically collapses as T' decreases, we set high γ^{aux} to make w_t^{aux} nearly 0 for a small t.

3.5. Total Loss Function with SDFT

To summarize, the proposed fine-tuning method with distillation uses the loss function:

$$\mathcal{L}^{total} = \mathcal{L}^{diffusion} + \lambda^{distill}\mathcal{L}^{distill} + \lambda^{aux}\mathcal{L}^{aux}, \quad (6)$$

where $\mathcal{L}^{diffusion}$ is a objective function of base diffusion model and $\lambda^{distill}$ and λ^{aux} is a hyperparameter.

4. Experiments

4.1. Experimental Setup

In this section, we verify the effectiveness of SDFT for fine-tuning DMs to inherit expressiveness from the source models when the target datasets have limited samples and attributes. We present that the enhanced expressiveness of DMs results in improvements in various downstream image translation tasks including domain translation and text-guided image manipulation. To our knowledge, as the open source implementation of the fine-tuning method for unconditional diffusion model [24] is not publicly available, we compare SDFT with a model trained from scratch and naïvely fine-tuned model, referred to as the Scratch model and Naïve Fine-Tune model, respectively. For simplicity, we also name the model trained with SDFT as SDFT.

Datasets. We use FFHQ [15] for the source dataset, which has 70K real faces with various attributes. For the target limited datasets, we utilize MetFaces [14], which has 1,336 high-quality portraits. Due to the limited samples and inherent biases, MetFaces do not or scarcely contain diverse facial attributes (e.g. smiling with teeth, sunglasses, various hairstyles *etc.*). We further utilize the AAHQ [22] for the target dataset, which has 25K high-quality artistic faces. For the limited dataset, we select images from AAHQ using CLIP following the nie et al. [25]. As a result, we utilize 1,437 images that contain expressionless males without glasses. All images are resized to 256×256 resolution. A detailed explanation for preparing the dataset is provided in the supplementary material.

Implementation detail. For all experiments, we utilize officially implemented ADM architecture [9] and public checkpoint pretrained on FFHQ for the source model [4]. For efficiency, we use 40-step deterministic DDIM sampling [34] for all experiments. We train all models until the 80k training iterations and report the best results. The hyperparameters can be found in supplementary material.

4.2. Results in Domain Translation

We present that the enhanced expressiveness of diffusion models can greatly improve the performance of domain translation tasks. For successful domain translation, the translated image should be *realistic* to fit the style of the target domain and *faithful* to ensure that the various attributes from the input image are accurately preserved.

Qualitative Comparison. Fig. 5 shows the qualitative comparison between scratch, naïve fine-tune, and ours (SDFT) with domain translation methods, SDEdit [42] and EGSDE [42]. Since these methods utilize diffusion models trained on the target domain, the expressiveness of the model determines the success of the translation. Domain translation outputs with the scratch and Naïve fine-Tune model show less *faithful* outputs due to the limited training datasets, failing to translate *unseen* attributes from the training data such as glasses, and facial expressions. Furthermore, the translated outputs show biased representations such as translating baby and female to male in 3rd row. With a proposed fine-tuning approach, SDFT resolves the above problems and shows more *realistic* translated outputs, while showing reasonable *faithful*.

Quantitative Comparison. To measure the *faithfulness*, we evaluate the similarity between input-output pairs using the peak signal-to-noise ratio (PSNR), structural similarity index (SSIM), and learned perceptual image patch similarity (LPIPS) [40]. We randomly select 10K images from FFHQ and generate 10K translated outputs. For the *realism*, we report the widely used Fréchet inception distance (FID) [12] and kernel inception distance (KID) [1] where the latter is known to be unbiased, thus more proper to the limited datasets. The metrics for *faithful* are calculated using translated outputs and paired FFHQ inputs and metrics for *realism* are calculated using translated outputs and entire target datasets. For experiments on AAHQ, we use entire datasets to calculate FID and KID.

Tab. 1 shows the quantitative results. In each domain translation method, SDFT shows the most *faithful* results, achieving the highest PSNR and SSIM, and the lowest LPIPS, except for PSNR in AAHQ. However, SDFT is reported as less *realistic* than Naïve Fine-Tune model in SDEdit of MetFaces since SDFT translates attributes that are not contained in the MetFaces, as depicted in Fig. 5. However, using EGSDE [42], SDFT achieves the lowest KID, showing the most *realistic* outputs since the domain-specific energy function, a classifier between source and target domain, naturally drives the translated outputs to the target domain, without harming the *realism*. In limited AAHQ in which we calculate FID and KID using the entire AAHQ dataset, SDFT shows the most *realistic*, indicating that SDFT can generate high fidelity and diversity samples in the target domain with limited target datasets.

4.3. Results in Text-Guided Image Translation

We further show that the improved expressiveness of the diffusion model by the SDFT can be helpful in the more complex downstream tasks, such as text-guided image translation. For the text-guided image translation method, we choose Asyrp [19] since it fully utilizes the expressiveness embedded in the pretrained diffusion model. It only

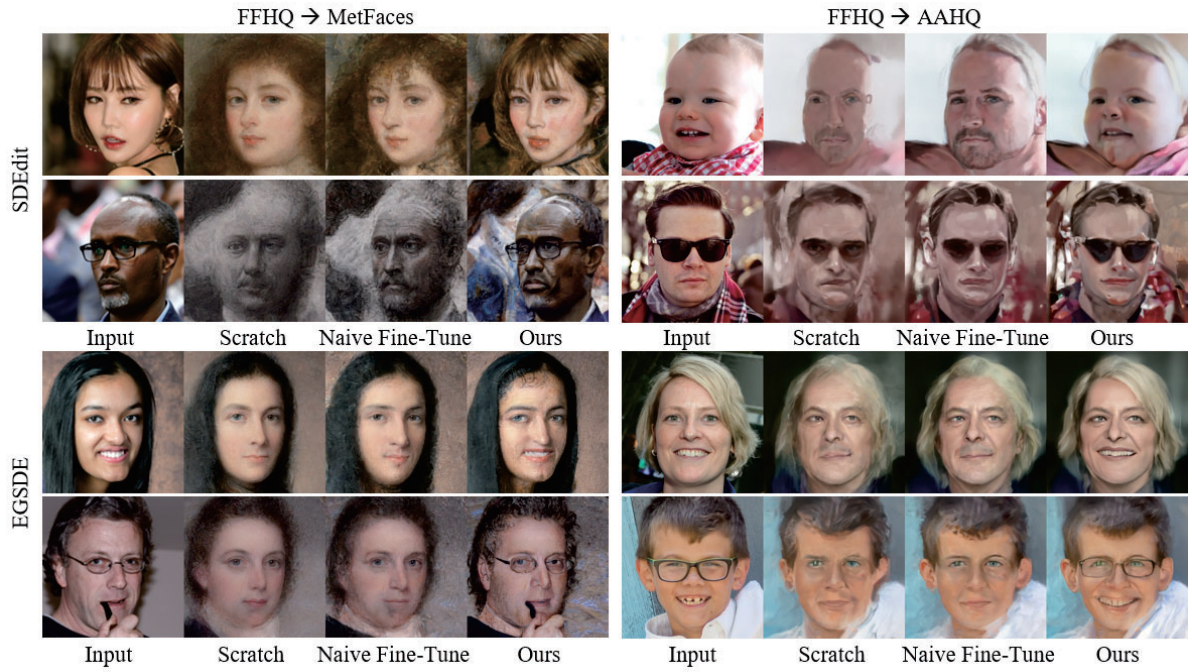

Figure 5. The generated images sampled utilizing the domain translation method denoted above.

Dataset	Translation Method	Training Method	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓	KID ($\times 10^3$) ↓
MetFaces [14]	SDEdit [23]	Scratch	13.8	0.31	0.579	74.63	52.45
		Naïve Fine-Tune	14.95	0.309	0.533	**56.42**	**37.00**
		Ours (SDFT)	**16.44**	**0.353**	**0.481**	65.18	40.95
	EGSDE [42]	Scratch	15.15	0.31	0.539	89.67	71.14
		Naïve Fine-Tune	16.03	0.333	0.509	**70.02**	50.52
		Ours (SDFT)	**17.42**	**0.392**	**0.440**	70.78	**43.84**
AAHQ	SDEdit [23]	Scratch	14.26	0.362	0.521	65.54	63.42
		Naïve Fine-Tune	**14.60**	0.374	0.489	54.75	48.69
		Ours (SDFT)	14.59	**0.369**	**0.486**	**51.12**	**44.46**
	EGSDE [42]	Scratch	15.91	0.407	0.464	82.00	78.26
		Naïve Fine-Tune	**16.18**	0.422	0.430	65.14	57.30
		Ours (SDFT)	16.13	**0.421**	**0.423**	**60.49**	**52.24**

Table 1. Quantitative results of domain translation methods using diffusion models trained with various methods.

trains a small module that translates the deepest layer of U-net, while keeping all parameters unchanged during the fine-tuning with CLIP [27]. Please note that we introduce a fine-tuning method, SDFT, for the unconditional diffusion model on limited datasets and Asyrp is a fine-tuning method for the text-guided image translation using a trained unconditional diffusion model.

Qualitative Comparison. Fig. 6 shows the input images and translated images with various text guidance. Since naïve fine-tuning on the MetFaces have a limited expressiveness due to the limited and biased samples, it can not generate diverse facial attributes and often fails to maintain identities. Contrarily, SDFT can express more diverse facial attributes, while successfully preserving the identity of the face. Even though the training datasets do not have diverse facial attributes, SDFT can express through enhanced expressiveness which is inherited from the source model. Notably, even though we transfer the knowledge from the FFHQ, the enhanced expressiveness affects not only the outputs close to the source dataset but also the more distant outputs, such as monochrome paintings.

Quantitative Comparison. We evaluate the successful text-guided manipulation in two metrics: Directional CLIP similarity ($\mathcal{S}_{\text{dir}}$) [17] and face identity similarity (ID). $\mathcal{S}_{\text{dir}}$ measure the successful manipulation of input image given text guidance using pretrained CLIP [27] embedding and ID

Figure 6. Results of Asyrp [19] from Naïve Fine-Tune model and Ours (SDFT). SDFT can express more diverse facial attributes.

	MetFaces		AAHQ	
	$\mathcal{S}_{\text{dir}}\uparrow$	ID$\uparrow$	$\mathcal{S}_{\text{dir}}\uparrow$	ID$\uparrow$
Naïve Fine-Tune	0.060	0.760	0.133	0.318
Ours (SDFT)	**0.081**	**0.765**	**0.143**	**0.452**

Table 2. Quantitative results on text-guided image manipulation.

measures the preservation of identity using pretrained face recognition models [8]. Tab. 2 shows the results using 5 text prompts (*smiling, sad, angry, young and old*). SDFT outperforms the naïve fine-tune model, demonstrating superior semantic manipulation capabilities given text while preserving various identities.

4.4. Effect on Unconditional Image Generation

Finally, we present that the benefits of SDFT are also helpful for generating diverse and high-fidelity images. By expanding the dual diffusion implicit bridges (DDIBs) [36] theorem, the DDIM sampling between the source model and target model generates semantically aligned outputs from the same initial noise. A more comprehensive analysis of DDIBs and their connection to unconditional generation is provided in the supplementary material. Fig. 7 shows the results of unconditional generation from the source model trained on FFHQ and from the various models fine-tuned on

Figure 7. Generated samples from unconditional image generation. Images in each row are generated from the same initial noise.

MetFaces. Scratch and Naïve Fine-Tune model shows semantically aligned images, but they fall short in generating a range of diverse attributes. However, SDFT can generate more semantically aligned images, preserving more diverse attributes that are not included in the target datasets such as the smiling face (1st row) and people with dark skin (2nd row). Tab. 3 shows the measured FID and KID using 10K generated samples with MetFaces and the entire AAHQ dataset. The perceptual distance (LPIPS) is measured between the 2K generated samples between the

	FID↓	KID ($\times 10^3$)↓	LPIPS↓
	MetFaces		
Scratch	65.91	44.79	0.488
Naïve Fine-Tune	43.45	22.81	0.474
Ours (SDFT)	**35.11**	**17.14**	**0.46**
	AAHQ		
Scratch	62.48	48.67	0.575
Naïve Fine-Tune	64.08	56.70	0.562
Ours (SDFT)	**42.54**	**33.75**	**0.469**

Table 3. Quantitative results on unconditional image generation.

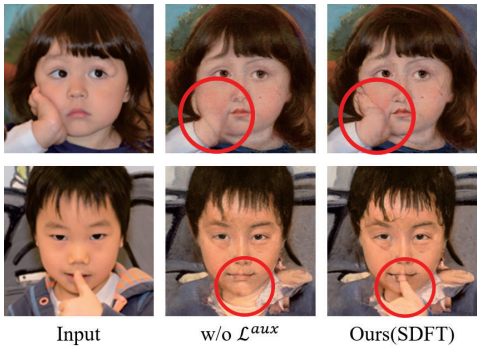

Figure 8. $\mathcal{L}^{aux}$ helps DMs deal with more diverse inputs.

source and target model from the same initial noises. From the results, the proposed SDFT shows the capability to generate diverse and high-fidelity images from limited datasets while preserving the diversities from the source model.

4.5. Ablation Study

In this section, we present the effectiveness of the proposed methods using domain translation from face to portrait using MetFaces [14] and EGSDE [42].

Auxiliary Input We present the effectiveness of the auxiliary inputs for distilling more diverse features described in Sec. 3.4. Fig. 8 illustrates the results of domain translation on out-of-domain inputs. By utilizing the auxiliary inputs during the training, the more diverse features that are not included in the target datasets can be transferred and help the model deal with out-of-domain samples successfully.

Different Weighting Scheme We also compare the proposed weighting scheme for distillation with other weightings which have been used for the training of DMs such as constant weighting ($w_t^{distill} = 1 \cdot w_t$) [13] and a Min-SNR weighting ($w_t^{distill}$ =min{SNR(t),γ} $\cdot w_t$) where w_t is defined in Sec. 2.1 and γ is set to 5 following Hang et al. [11]. Fig. 9 shows that other methods fail to transfer attributes from source images such as teeth. However, the proposed weighting scheme prioritizes the general features while discarding domain-specific features from the source model, leading to successful domain translation. For unconditional generation in MetFaces, Tab. 4 shows that SDFT can generate more realistic (low FID and KID) images while preserving more semantics from the source model (low LPIPS) compared to other weighting strategies.

	FID↓	KID ($\times 10^3$)↓	LPIPS↓
Constant	45.21	28.42	0.494
Min-SNR-5	44.75	27.96	0.476
Ours (SDFT)	**35.11**	**17.14**	**0.46**

Table 4. Ablation studies on different weighting strategies.

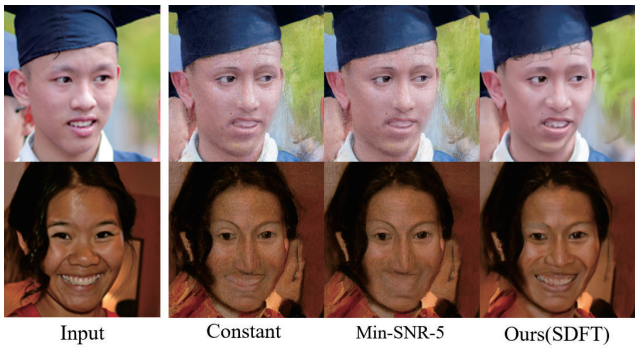

Figure 9. Effect of various distillation weights for fine-tuning.

5. Conclusion and Future Works

In this paper, we propose a self-distillation-based fine-tuning method for training diffusion models in limited datasets by leveraging the diverse knowledge from the source model trained on large datasets. Experimental results demonstrate that SDFT can effectively enhance the expressiveness of the diffusion models, leading to improved performance in various downstream tasks. **Future works.** Recently, rather than training full parameters, parameter-efficient fine-tuning (PEFT) can bring efficient and promising results [28, 32]. Since the SDFT can be orthogonally combined with these methods, we leave it for future work to investigate the advantage of combining SDFT with PEFT. Lastly, we only consider the diffusion models as noise predictors, but recent studies on distilling diffusion models found that utilizing velocity can bring effective knowledge transfer [21, 30]. Combining SDFT with various diffusion parametrizations is also an interesting future work.

Acknowledgements. This work was partly supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.2021-0-02068, Artificial Intelligence Innovation Hub) and the Engineering Research Center Program through the National Research Foundation of Korea (NRF) funded by the Korean Government MSIT (NRF-2018R1A5A1059921)

References

[1] Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*, 2018. 5

[2] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018. 1

[3] Jooyoung Choi, Sungwon Kim, Yonghyun Jeong, Youngjune Gwon, and Sungroh Yoon. Ilvr: Conditioning method for denoising diffusion probabilistic models. *arXiv preprint arXiv:2108.02938*, 2021. 1, 3

[4] Jooyoung Choi, Jungbeom Lee, Chaehun Shin, Sungwon Kim, Hyunwoo Kim, and Sungroh Yoon. Perception prioritized training of diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11472–11481, 2022. 2, 3, 5

[5] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8789–8797, 2018. 2

[6] Kamil Deja, Anna Kuzina, Tomasz Trzcinski, and Jakub Tomczak. On analyzing generative and denoising capabilities of diffusion-based deep generative models. *Advances in Neural Information Processing Systems*, 35:26218–26229, 2022. 4

[7] 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. Ieee, 2009. 1

[8] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019. 7

[9] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021. 1, 5

[10] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 1

[11] Tiankai Hang, Shuyang Gu, Chen Li, Jianmin Bao, Dong Chen, Han Hu, Xin Geng, and Baining Guo. Efficient diffusion training via min-snr weighting strategy. *arXiv preprint arXiv:2303.09556*, 2023. 8

[12] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 5

[13] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 1, 2, 8

[14] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. *Advances in neural information processing systems*, 33:12104–12114, 2020. 1, 5, 6, 8

[15] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 1, 2, 3, 5

[16] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020. 3

[17] Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2426–2435, 2022. 1, 2, 3, 6

[18] Diederik Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models. *Advances in neural information processing systems*, 34:21696–21707, 2021. 2

[19] Mingi Kwon, Jaeseok Jeong, and Youngjung Uh. Diffusion models already have a semantic latent space. *arXiv preprint arXiv:2210.10960*, 2022. 1, 2, 3, 5, 7

[20] Dongyeun Lee, Jae Young Lee, Doyeon Kim, Jaehyun Choi, Jaejun Yoo, and Junmo Kim. Fix the noise: Disentangling source feature for controllable domain translation. *arXiv preprint arXiv:2303.11545*, 2023. 2, 3

[21] Yanyu Li, Huan Wang, Qing Jin, Ju Hu, Pavlo Chemerys, Yun Fu, Yanzhi Wang, Sergey Tulyakov, and Jian Ren. Snapfusion: Text-to-image diffusion model on mobile devices within two seconds. *arXiv preprint arXiv:2306.00980*, 2023. 8

[22] Mingcong Liu, Qiang Li, Zekui Qin, Guoxin Zhang, Pengfei Wan, and Wen Zheng. Blendgan: Implicitly gan blending for arbitrary stylized face generation. *Advances in Neural Information Processing Systems*, 34:29710–29722, 2021. 5

[23] Chenlin Meng, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021. 2, 3, 4, 6

[24] Taehong Moon, Moonseok Choi, Gayoung Lee, Jung-Woo Ha, and Juho Lee. Fine-tuning diffusion models with limited data. In *NeurIPS 2022 Workshop on Score-Based Methods*. 3, 5

[25] Weili Nie, Arash Vahdat, and Anima Anandkumar. Controllable and compositional generation with latent-space energy-based models. *Advances in Neural Information Processing Systems*, 34:13497–13510, 2021. 5

[26] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2085–2094, 2021. 2

[27] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning

transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 3, 6

[28] Simo Ryu. Low-rank adaptation for fast text-to-image diffusion fine-tuning, 2023. 8

[29] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016. 1

[30] Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. *arXiv preprint arXiv:2202.00512*, 2022. 8

[31] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015. 1, 2

[32] Kihyuk Sohn, Nataniel Ruiz, Kimin Lee, Daniel Castro Chin, Irina Blok, Huiwen Chang, Jarred Barber, Lu Jiang, Glenn Entis, Yuanzhen Li, et al. Styledrop: Text-to-image generation in any style. *arXiv preprint arXiv:2306.00983*, 2023. 8

[33] Haorui Song, Yong Du, Tianyi Xiang, Junyu Dong, Jing Qin, and Shengfeng He. Editing out-of-domain gan inversion via differential activations. In *European Conference on Computer Vision*, pages 1–17. Springer, 2022. 2

[34] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 2, 5

[35] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. 1

[36] Xuan Su, Jiaming Song, Chenlin Meng, and Stefano Ermon. Dual diffusion implicit bridges for image-to-image translation. *arXiv preprint arXiv:2203.08382*, 2022. 7

[37] Zongze Wu, Yotam Nitzan, Eli Shechtman, and Dani Lischinski. Stylealign: Analysis and applications of aligned stylegan models. *arXiv preprint arXiv:2110.11323*, 2021. 3

[38] Shuai Yang, Liming Jiang, Ziwei Liu, and Chen Change Loy. Pastiche master: Exemplar-based high-resolution portrait style transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7693–7702, 2022. 2

[39] Hongxu Yin, Pavlo Molchanov, Jose M Alvarez, Zhizhong Li, Arun Mallya, Derek Hoiem, Niraj K Jha, and Jan Kautz. Dreaming to distill: Data-free knowledge transfer via deepinversion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8715–8724, 2020. 4

[40] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 5

[41] Yiman Zhang, Hanting Chen, Xinghao Chen, Yiping Deng, Chunjing Xu, and Yunhe Wang. Data-free knowledge distillation for image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7852–7861, 2021. 4

[42] Min Zhao, Fan Bao, Chongxuan Li, and Jun Zhu. Egsde: Unpaired image-to-image translation via energy-guided stochastic differential equations. *arXiv preprint arXiv:2207.06635*, 2022. 1, 3, 5, 6, 8

[43] Shengjia Zhao, Hongyu Ren, Arianna Yuan, Jiaming Song, Noah Goodman, and Stefano Ermon. Bias and generalization in deep generative models: An empirical study. *Advances in Neural Information Processing Systems*, 31, 2018. 1