Few-shot Image Generation with Diffusion Models

JingYuan Zhu
Tsinghua University, China
jy-zhu20@mails.tsinghua.edu.cn

Huimin Ma
University of Science and Technology Beijing, China
mhmpub@ustb.edu.cn

Jiansheng Chen
University of Science and Technology Beijing, China
jschen@ustb.edu.cn

Jian Yuan
Tsinghua University, China
jyuan@tsinghua.edu.cn

Abstract

Denoising diffusion probabilistic models (DDPMs) have been proven capable of synthesizing high-quality images with remarkable diversity when trained on large amounts of data. However, to our knowledge, few-shot image generation tasks have yet to be studied with DDPM-based approaches. Modern approaches are mainly built on Generative Adversarial Networks (GANs) and adapt models pre-trained on large source domains to target domains using a few available samples. In this paper, we make the first attempt to study when do DDPMs overfit and suffer severe diversity degradation as training data become scarce. Then we fine-tune DDPMs pre-trained on large source domains to solve the overfitting problem when training data is limited. Although the directly fine-tuned models accelerate convergence and improve generation quality and diversity compared with training from scratch, they still fail to retain some diverse features and can only produce coarse images. Therefore, we design a DDPM pairwise adaptation (DDPM-PA) approach to optimize few-shot DDPM domain adaptation. DDPM-PA efficiently preserves information learned from source domains by keeping the relative pairwise distances between generated samples during adaptation. Besides, DDPM-PA enhances the learning of high-frequency details from source models and limited training data. DDPM-PA further improves generation quality and diversity and achieves results better than current state-of-the-art GAN-based approaches. We demonstrate the effectiveness of our approach on a series of few-shot image generation tasks qualitatively and quantitatively.

1 Introduction

Recent advances in generative models including GANs [11, 4, 26, 27, 25], variational autoencoders (VAEs) [30, 50, 59], and autoregressive models [61, 7, 14] have realized high-quality image generation with great diversity. Diffusion probabilistic models [55] are introduced to match data distributions by learning to reverse multi-step noising processes. Ho et al. [16] demonstrate the capability of DDPMs to produce high-quality results. Following works [57, 9, 44, 28] further optimize the noise addition schedules, network architectures, and optimization targets of DDPMs. Besides, classifier guidance is added to realize DDPM-based conditional image generation [9]. DDPMs have shown excellent generation results competitive with GANs [27, 4] on datasets including CIFAR-10 [32], LSUN [71], and ImageNet [61]. Moreover, DDPMs have also achieved compelling results in generating videos [18, 12, 69, 72], audios [31, 2], point clouds [80, 41, 42, 36], and biological structures [68, 20, 37].

Modern DDPMs depend on large amounts of data to train the millions of parameters in their networks like other generative models, which tend to overfit seriously and fail to produce high-quality images with considerable diversity when training data is limited. Unfortunately, it is not always possible to obtain abundant data under some circumstances. A series of GAN-based approaches [64, 24, 43, 63, 35, 46, 78] have been proposed to adapt models pre-trained on large-scale source datasets to target datasets using a few available training samples (e.g., 10 images). These approaches utilize knowledge from source models to relieve overfitting but can only achieve limited quality and diversity. In addition, the performance of DDPMs trained on limited data and practical DDPM-based few-shot image generation approaches remain to be investigated.
We first evaluate the performance of DDPMs trained on small-scale datasets and show that DDPMs suffer similar overfitting problems to other modern generative models when trained on limited data. Then we fine-tune pre-trained DDPMs on target domains using limited data directly. The fine-tuned DDPMs achieve faster convergence and diverse generated samples compared with DDPMs trained from scratch but still get results sharing similar features and missing high-frequency details. To this end, we introduce the DDPM-PA approach, which keeps the relative distance between generated samples and realizes high-frequency details enhancement during domain adaptation to achieve high-quality few-shot image generation with great diversity.

To sum up, the contributions of our work are:

- We demonstrate the effectiveness of DDPM-PA qualitatively and quantitatively on a series of few-shot image generation tasks and show that DDPM-PA achieves better generation quality and diversity than current state-of-the-art GAN-based approaches.

## 2 Related Work

### 2.1 Diffusion Denoising Probabilistic Models

**DDPMs Formulation** Given training images \( x_0 \) following the distribution \( q(x_0) \), DDPMs define a forward noising (diffusion) process \( q \) adding Gaussian noises with variance \( \beta_t \in (0, 1) \) at diffusion step \( t \) and produce the noised image sequences: \( x_1, x_2, ..., x_T \) as follows:

\[
q(x_1, x_2, ..., x_T|x_0) := \prod_{t=1}^{T} q(x_t|x_{t-1}), \quad (1)
\]

\[
q(x_t|x_{t-1}) := \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I), \quad (2)
\]

We can sample an arbitrary step of the noising process conditioned on \( x_0 \) as follows:

\[
q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\alpha_t}x_0, \sqrt{1-\alpha_t}I), \quad (3)
\]

\[
x_t = \sqrt{\alpha_t}x_0 + \sqrt{1-\alpha_t}e, \quad (4)
\]

where \( \alpha_t := 1 - \beta_t, \alpha_t := \prod_{s=0}^{t} \alpha_s \), and \( e \) represents Gaussian distributions following \( \mathcal{N}(0, I) \). Based on the Bayes theorem, we have the posterior \( q(x_{t-1}|x_t, x_0) \):

\[
q(x_{t-1}|x_t, x_0) = \mathcal{N}(x_{t-1}; \hat{\mu}_t(x_t, x_0), \hat{\beta}_t I), \quad (5)
\]

where \( \hat{\mu}_t(x_t, x_0) \) and \( \hat{\beta}_t \) are defined in terms of \( x_0, x_t \) and the variance as follows:

\[
\hat{\mu}_t(x_t, x_0) := \frac{\sqrt{\alpha_{t-1}}\beta_t}{1 - \beta_t} x_t + \frac{\sqrt{\alpha_t}(1 - \alpha_{t-1})}{1 - \beta_t} x_0, \quad (6)
\]

\[
\hat{\beta}_t := \frac{1 - \alpha_{t-1}}{1 - \beta_t}. \quad (7)
\]

With a large enough \( T \), the noised image \( x_T \) almost follows an isotropic Gaussian distribution. Therefore, we can randomly sample a \( x_T \) from \( \mathcal{N}(0, I) \) and apply the reverse diffusion process to realize sampling following Equation 8. Re-creating DDPMs employs a UNet-based neural network to approximate the reverse distribution \( q(x_{t-1}|x_t) \) as:

\[
p_\theta(x_{t-1}|x_t) := \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \sigma_\theta(x_t, t)). \quad (8)
\]

Naturally, we can train the network to predict \( x_0 \) and apply it to Equation 6 to parameterize the reverse distribution mean \( \mu_\theta(x_t, t) \). Besides, we can also derive the prediction of \( \mu_\theta(x_t, t) \) cominig Equations 4 and 6 as:

\[
\mu_\theta(x_t, t) = \frac{1}{\sqrt{\alpha_t}} (x_t - \frac{\beta_t}{\sqrt{1-\alpha_t}} \epsilon_\theta(x_t, t)), \quad (9)
\]

if we train the network as a function approximator \( \epsilon_\theta(x_t, t) \) to predict the noise \( \epsilon \) in Equation 4. Ho et al. [16] demonstrate that predicting \( \epsilon \) performs well and achieves high-quality results using a reweighted loss function:

\[
L_{\text{simple}} = E_{t, x_0, \epsilon} [||\epsilon - \epsilon_\theta(x_t, t)||^2]. \quad (10)
\]

In Ho et al.’s work [16], the variance \( \Sigma_\theta(x_t, t) \) is fixed as a constant \( \sigma_\theta^2 I \), where \( \sigma_\theta^2 = \beta_t \) and is not learned. The network is only trained to learn the model mean \( \mu_\theta(x_t, t) \) through predicting noises with \( \epsilon_\theta(x_t, t) \). Following works [44, 28] propose to optimize the variational lower bound (VLB) and guide the learning of \( \Sigma_\theta(x_t, t) \) with an additional optimization term \( L_{\text{vib}} \) as follows:

\[
L_{\text{vib}} := L_0 + L_1 + ... + L_{T-1} + L_T, \quad (11)
\]

\[
L_0 := -\log p_\theta(x_0|x_1), \quad (12)
\]

\[
L_{t-1} := D_{KL}(q(x_{t-1}|x_t, x_0) \|| p_\theta(x_{t-1}|x_t)), \quad (13)
\]

\[
L_T := D_{KL}(q(x_T|x_0) \|| p(x_T)), \quad (14)
\]

where \( D_{KL} \) represents the Kullback-Leibler (KL) divergence used to evaluate the distance between distributions.

**Fast Sampling** DDPMs need time-consuming iterative processes to realize sampling following Equation 8. Recent works including DDIM and gDDIM [56, 74] extend
the original DDPMs to non-Markovian cases for fast sampling. DPM-solver [38, 39] presents a theoretical formulation for the solution of probability flow ordinary differential equations (ODEs) and achieves a fast ODE solver needing only 10 steps for DDPM sampling. Diffusion Exponential Integrator Sampler (DEIS) [73] makes use of an exponential integrator to approximately calculate the solution of ODEs. Karras et al. [23] present a design space to identify changes in both the sampling and training processes and achieve state-of-the-art FID [15] on CIFAR-10 [32] under a class-conditional setting.

Applications DDPMs have already been applied to many aspects of applications such as image super-resolution [33, 53, 51, 17], image translation [52, 48], semantic segmentation [3, 1], few-shot generation for unseen classes [10, 54], and natural language processing [2, 34, 6]. Besides, DDPMs are combined with other generative models including GANs [67, 65], VAEs [60, 21, 40], and autoregressive models [49, 19]. Different from existing works, this paper focuses on model-level, unconditional, few-shot image generation with DDPM-based approaches.

2.2 Few-shot Image Generation

Few-shot image generation aims to achieve high-quality generation with great diversity using only a few available training samples. However, modern generative models easily overfit and suffer severe diversity degradation when trained on limited data (e.g., 10 images). They tend to replicate training samples instead of generating diverse images following similar distributions. GAN-based few-shot image generation approaches mainly follow TGAN [64] to adapt GANs pre-trained on large source domains, including ImageNet [61], LSUN [71], and FFHQ [27], to related target domains with limited data. Augmentation approaches [58, 76, 79, 24] also help improve generation diversity. BSA [45] fixes all the parameters except for the scale and shift parameters in the generator. FreezeD [43] freezes parameters in high-resolution layers of the discriminator to relieve overfitting. MineGAN [63] uses additional fully connected networks to modify noise inputs for the generator. EWC [35] makes use of elastic weight consolidation to regularize the generator by making it harder to change the critical weights which have higher Fisher information values. CDC [46] proposes a cross-domain consistency loss and patch-level discrimination to build a correspondence between source and target domains. DCL [78] utilizes contrastive learning to push away the generated samples from real images and maximize the similarity between corresponding image pairs in source and target domains. The proposed DDPM-PA approach follows similar strategies to adapt models pre-trained on large source domains to target domains. Our experiments show that DDPMs are qualified for few-shot image generation tasks and can achieve better results than current state-of-the-art GAN-based approaches in generation quality and diversity.

3 DDPMs Trained on Small-scale Datasets

To evaluate the performance of DDPMs when training data is limited, we train DDPMs on small-scale datasets containing various numbers of images from scratch. We analyze generation diversity qualitatively and quantitatively to study when do DDPMs overfit as training samples decrease.

Basic Setups We sample 10, 100, and 1000 images from FFHQ-babies (Babies), FFHQ-sunglasses (Sunglasses) [46], and LSUN Church [71] respectively as small-scale training datasets. The image resolution of all the datasets is set as $256 \times 256$. We follow the model setups in prior works [44, 9] used for LSUN $256^2$ [71]. The max diffusion step $T$ is set as 1000. We use a learning rate of 1e-4 and a batch size of 48. We train DDPMs for 40K iterations on datasets containing 10 or 100 images and 60K iterations on datasets containing 1000 images empirically.

Qualitative Evaluation In Fig. 1, we visualize the generated samples of DDPMs trained from scratch on few-shot Sunglasses datasets and provide some training samples for comparison (more generated and training samples are added in Appendix F). We observe that DDPMs overfit and tend to replicate training samples when datasets are limited to 10 or 100 images. Since some training samples are flipped in the training process as a step of data augmentation, we can also find some generated images symmetric to the training samples. While for datasets containing 1000 images, DDPMs can generate diverse samples following similar distributions of training samples instead of replicating them. The overfitting problem is relatively alleviated. However, the generated samples are coarse and lack high-frequency details compared with training samples.

Quantitative Evaluation LPIPS [75] is proposed to evaluate the perceptual distances [22] between images. We propose a Nearest-LPIPS metric based on LPIPS to evaluate the generation diversity of DDPMs trained on small-scale datasets. More specifically, we first generate 1000 images randomly and find the most similar training sample having the lowest LPIPS distance to each generated sample. Nearest-LPIPS is defined as the LPIPS distances between generated samples and the most similar training samples in correspondence, averaged over all the generated samples. If a generative model reproduces the training samples exactly, the Nearest-LPIPS metric will have a score of zero. Larger Nearest-LPIPS values indicate lower replication rates and greater diversity relative to training samples.

We provide the Nearest-LPIPS results of DDPMs trained from scratch on small-scale datasets in the top part of Table 1. For datasets containing 10 or 100 images, we have
Figure 1: For small-scale Sunglasses datasets containing 10, 100, and 1000 images, Left: samples picked from the small-scale datasets, Right: samples produced by DDPMs trained on the small-scale datasets from scratch.

Figure 2: DDPM-based image generation samples on 10-shot FFHQ → Babies, FFHQ → Sketches, and LSUN Church → Haunted houses.

Table 1: Nearest-LPIPS (↑) results of DDPMs trained from scratch on several small-scale datasets.

| Number of Samples | Babies | Sunglasses | Church |
|-------------------|--------|------------|--------|
| 10                | 0.2875 | 0.3030     | 0.3136 |
| 100               | 0.3152 | 0.3310     | 0.3327 |
| 1000              | 0.4658 | 0.4819     | 0.5707 |
| 10 (+ flip)       | 0.1206 | 0.1217     | 0.0445 |
| 100 (+ flip)      | 0.1556 | 0.1297     | 0.1177 |
| 1000 (+ flip)     | 0.4611 | 0.4726     | 0.5625 |

Table 1: Nearest-LPIPS (↑) results of DDPMs trained from scratch on several small-scale datasets.

lower Nearest-LPIPS values. While for datasets containing 1000 images, we get measurably improved Nearest-LPIPS values. To avoid the influence of generated images symmetric to training samples, we flip all the training samples as supplements to the original datasets and recalculate the Nearest-LPIPS metric. The results are listed in the bottom part of Table 1. With the addition of flipped training samples, we find apparently lower Nearest-LPIPS values for datasets containing 10 or 100 images. However, we get almost the same Nearest-LPIPS results for DDPMs trained on larger datasets containing 1000 images, indicating that these models can generate diverse samples different from the original or symmetric training samples.

As proved by the results above, it becomes harder for DDPMs to learn the representations of datasets as training data become scarce. When trained on limited data from...
Figure 4: Overview of the DDPM-PA approach. A pairwise similarity loss is introduced to keep the relative pairwise distances of generated samples and their high-frequency details of adapted models similar to source models during adaptation. DDPM-PA guides adapted models to learn more high-frequency details from limited data using the MSE loss between high-frequency details extracted from training data and adapted samples. DDPM-PA generates high-quality and diverse samples sharing similar styles with few-shot training data.

scratch, DDPMs fail to match target data distributions exactly or produce high-quality and diverse samples.

4 Method

To realize DDPM-based few-shot image generation, we first fine-tune DDPMs pre-trained on large-scale source datasets using limited target data directly. The fine-tuned models only need $3K - 4K$ iterations to converge. As shown in the middle row of Fig. 2, they can produce diverse results utilizing only 10 training samples. However, the generated samples lack ample high-frequency details and share similar features like hairstyles and facial expressions, leading to the degradation of generation quality and diversity.

Compared with pre-trained models, the degradation of fine-tuned models mainly comes from the excessively shortened relative distances between generated samples. As shown in Fig. 3, two samples synthesized from fixed noise inputs by the directly fine-tuned DDPM become increasingly similar (e.g., eyes and facial expressions) throughout training, losing various features and high-frequency details. Therefore, we propose to regularize the domain adaptation process by keeping the relative pairwise distances between adapted samples similar to source samples (Sec 4.1). Besides, we guide adapted models to learn high-frequency details from limited data and preserve high-frequency details learned from source domains (Sec 4.2). Our approach fixes source models as reference for adapted models. The weights of adapted models are initialized to the weights of source models and adapted to target domains. An overview of the proposed DDPM-PA approach is illustrated in Fig. 4 using 10-shot FFHQ $\rightarrow$ Babies as an example.

4.1 Relative Distances Preservation

We design a pairwise similarity loss to preserve the relative distances between generated samples during domain adaptation. To construct $N$-way probability distributions for each image, we sample a batch of noised images $\{x_0^n\}_{n=0}^N$ by randomly adding Gaussian noises $\epsilon_\theta$ to training samples $x_0 \sim q(x_0)$ following Equations 1 and 2. Then the source and adapted models are applied to predict the fully denoised images $\{\tilde{x}^n_0\}_{n=0}^N$. We have the prediction of $\tilde{x}_0$ in terms of $x_t$ and $\epsilon_\theta(x_t, t)$ as follows:

$$\tilde{x}_0 = \frac{1}{\sqrt{\alpha_t}} x_t - \frac{\sqrt{1 - \alpha_t} \epsilon_\theta(x_t, t)}{\sqrt{\alpha_t}}$$

which can be derived from Equation 4. Cosine similarity is employed to measure the relative distances between the predicted samples $\tilde{x}_0$. The probability distributions for
expressed as follows:

\[ x \]

scape drawings.

loss for generated images as follows:

function, respectively. Then we have the pairwise similarity

frequency components:

\[ L_{f \rightarrow \text{sou}} = \text{softmax}(\text{sim}(x_i, x_j)) \]

\[ L_{f \rightarrow \text{ada}} = \text{softmax}(\text{sim}(x_i, x_j)) \]

where \( \text{sim} \) and \( \text{softmax} \) denote cosine similarity and softmax

function, respectively. Then we have the pairwise similarity

loss for generated images as follows:

\[ L_{\text{img}}(\epsilon_{\text{sou}}, \epsilon_{\text{ada}}) = E_{t, x_0, \epsilon} \sum_i D_{\text{KL}}(p_i^{\text{ada}} \| p_i^{\text{sou}}) \],

where \( D_{\text{KL}} \) represents KL-divergence.

**4.2 High-frequency Details Enhancement**

To begin with, we employ the typical Haar wavelet trans-
formation [8] to disentangle images into multiple frequency

components. Haar wavelet transformation contains four

kernels including \( LL^T, LH^T, HL^T, HH^T \), where \( L \) and \( H \)

represent low and high pass filters, respectively:

\[ L^T = \frac{1}{\sqrt{2}} [1, 1], \quad H^T = \frac{1}{\sqrt{2}} [-1, 1]. \]

Haar wavelet transformation decomposes inputs into four

frequency components \( LL, LH, HL, \) and \( HH \). \( LL \)

contains fundamental structures of images while other high-

frequency components \( LH, HL, \) and \( HH \) contain rich de-

tails of images. We define \( hf \) as the sum of these high-

frequency components:

\[ hf = LH + HL + HH. \]

We implement high-frequency details enhancement from
two perspectives. Firstly, we use the proposed pairwise sim-
ilarity loss to preserve high-frequency details learned from
source domains. Similarly, the probability distributions for
the high-frequency components of \( x_i \) in the source and adapted models can be
expressed as follows:

\[ p_i^{\text{sou}} = \text{softmax}(\text{sim}(\tilde{x}_0^{\text{sou}}_i, \tilde{x}_0^{\text{sou}}_j)) \]

\[ p_i^{\text{ada}} = \text{softmax}(\text{sim}(\tilde{x}_0^{\text{ada}}_i, \tilde{x}_0^{\text{ada}}_j)) \]

where \( \text{sim} \) and \( \text{softmax} \) denote cosine similarity and softmax

function, respectively. Then we have the pairwise similarity

loss for the high-frequency components in generated samples
are as follows:

\[ p_f^{\text{sou}} = \text{softmax}(\text{sim}(hf(\tilde{x}_0^{\text{sou}}), hf(\tilde{x}_0^{\text{ada}}))) \]

\[ p_f^{\text{ada}} = \text{softmax}(\text{sim}(hf(\tilde{x}_0^{\text{ada}}), hf(\tilde{x}_0^{\text{sou}}))) \]

\[ L_{hf}(\epsilon_{\text{sou}}, \epsilon_{\text{ada}}) = E_{t, x_0, \epsilon} \sum_i D_{\text{KL}}(p_f^{\text{ada}} \| p_f^{\text{sou}}) \]

Secondly, we guide adapted models to learn more high-

frequency details from training data by minimizing the
mean-squared error between the high-frequency compo-
nents in adapted samples \( \tilde{x}_0 \) and training data \( x_0 \) as follows:

\[ L_{hf_{\text{mse}}} = E_{t, x_0, \epsilon} ||hf(\tilde{x}_0) - hf(x_0)||^2. \]

**4.3 Overall Optimization Target**

The overall optimization target of the proposed DDPM-
PA approach combines all the methods stated above:

\[ L = L_{\text{simple}} + \lambda_1 L_{\text{tolb}} + \lambda_2 L_{\text{img}} + \lambda_3 L_{hf} + \lambda_4 L_{hf_{\text{mse}}}. \]

We follow prior works [44] to set \( \lambda_1 \) as 0.001 to avoid
\( L_{\text{tolb}} \) from overwhelming \( L_{\text{simple}} \). \( L_{\text{img}} \) guides adapted
models to preserve the relative distances between adapted
samples. \( L_{hf_{\text{mse}}} \) and \( L_{hf} \) further enhance the preservation
of high-frequency details. We empirically find \( \lambda_2, \lambda_3 \)
ranging between 0.1 and 1.0 and \( \lambda_4 \) ranging between 0.01 and
0.08 to be effective for few-shot adaptation setups.

Figure 5: DDPM-PA samples on 10-shot FFHQ → Sunglasses, FFHQ → Raphael’s paintings, and LSUN Church → Landscape drawings.
Figure 6: 10-shot image generation samples on FFHQ → Sunglasses. All the samples of GAN-based approaches are synthesized from fixed noise inputs (rows 1-7). Samples of the directly fine-tuned DDPM and DDPM-PA are synthesized from fixed noise inputs as well (rows 8-9). Our approach generates high-quality results with fewer blurs and artifacts and achieves considerable generation diversity.

5 Experiments

To demonstrate the effectiveness of DDPM-PA, we evaluate it with a series of few-shot image generation tasks using extremely few training samples (10 images). The performance of DDPM-PA is compared with directly fine-tuned DDPMs and modern GAN-based approaches on generation quality and diversity qualitatively and quantitatively.

5.1 Quality and Diversity Evaluation

Basic Setups We choose FFHQ [27] and LSUN Church [71] as source datasets and train DDPMs from scratch on these two datasets for $300K$ and $250K$ iterations as source models. As for the target datasets, we employ 10-shot Sketches [62], Babies, Sunglasses [27], and face paintings by Amedeo Modigliani and Raphael Peale [70] in correspondence to the source domain FFHQ. Besides, 10-shot Haunted houses and Landscape drawings are used as the target datasets in correspondence to LSUN Church. The model setups are consistent with the experiments on small-scale datasets in Sec. 3. The DDPM-PA models are trained for $3K$-$5K$ iterations with a batch size of 24 on $\times 8$ NVIDIA RTX A6000 GPUs.

Evaluation Metrics We follow Ojha et al.’s work [46] to use Intra-LPIPS for generation diversity evaluation. To be more specific, we generate 1000 images and assign them to one of the training samples with the lowest LPIPS [75] distance. Intra-LPIPS is defined as the average pairwise LPIPS distances within members of the same cluster averaged over all the clusters. If a model exactly replicates training samples, its Intra-LPIPS will have a score of zero. Larger Intra-LPIPS values correspond to greater generation diversity.

FID [15] is widely used to evaluate the generation quality of generative models by computing the distribution distances between generated samples and datasets. However, FID would become unstable and unreliable when it comes to datasets containing a few samples (e.g., 10-shot datasets used in this paper). Therefore, we provide FID evaluation
DDPM-PA adapts source models to target domains naturally and preserves various features different from training data. Samples of people wearing hats can be found when adapting FFHQ to Babies or Sunglasses, which is obviously different from the training samples. The adaptation from LSUN Church to Haunted houses and Landscape drawings retain various architectural structures. Fig. 6 shows samples of GAN-based and DDPM-based approaches on 10-shot FFHQ → Babies and Sunglasses.

Quantitative Evaluation We provide the Intra-LPIPS
results of DDPM-PA under a series of 10-shot adaptation setups in Table 2. DDPM-PA realizes a superior improvement of Intra-LPIPS compared with directly fine-tuned DDPMs. Besides, DDPM-PA outperforms state-of-the-art GAN-based approaches, indicating its strong capability of maintaining generation diversity. Intra-LPIPS results under other adaptation setups are added in Appendix A.

As shown by the FID results in Table 3, DDPM-PA also performs better on learning target distributions from limited data than prior GAN-based approaches.

5.2 Ablation Analysis

We ablate our approach using 10-shot FFHQ → Amedeo’s paintings as an example. With high-frequency details enhancement only, the adapted model cannot match target distributions and produces samples containing unnatural artifacts. With relative distances preservation only, the adapted model produces coarse samples lacking details like hairstyles. DDPM-PA combines both ideas and obtains realistic and diverse results. Ablations of the weight coefficients in Equation 25 are added in Appendix B.

6 Conclusion

This work first introduces diffusion models to few-shot image generation. However, when trained on limited data, DDPMs can only replicate training samples or produce coarse results. Therefore, we propose the DDPM-PA approach to adapt source models to target domains while maintaining diversity and preserving details. DDPM-PA keeps the relative pairwise distances between adapted samples and realizes high-frequency details enhancement during domain adaptation. We demonstrate the effectiveness of DDPM-PA on a series of few-shot image generation tasks. It generates compelling samples with rich details and few blurs, outperforming current state-of-the-art GAN-based approaches on generation quality and diversity. Our work is an essential step toward more data-efficient diffusion models. The limitations are discussed in Appendix C.

References

[1] E. B. Asiedu, S. Kornblith, T. Chen, N. Parmar, M. Minderer, and M. Norouzi. Denoising pretraining for semantic segmentation. In IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2022.

[2] J. Austin, D. D. Johnson, J. Ho, D. Tarlow, and R. van den Berg. Structured denoising diffusion models in discrete state-spaces. Advances in Neural Information Processing Systems, 34:17981–17993, 2021.

[3] D. Baranchuk, I. Rubachev, A. Voynov, V. Khrulkov, and A. Babenko. Label-efficient semantic segmentation with diffusion models. In International Conference on Learning Representations, 2022.

[4] A. Brock, J. Donahue, and K. Simonyan. Large scale GAN training for high fidelity natural image synthesis. In International Conference on Learning Representations, 2019.

[5] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations. In International Conference on Machine Learning, pages 1597–1607. PMLR, 2020.

[6] T. Chen, R. Zhang, and G. Hinton. Analog bits: Generating discrete data using diffusion models with self-conditioning. arXiv preprint arXiv:2208.04202, 2022.

[7] X. Chen, N. Mishra, M. Rohaninejad, and P. Abbeel. Pixelsnail: An improved autoregressive generative model. In International Conference on Machine Learning, pages 864–872. PMLR, 2018.

[8] I. Daubechies. The wavelet transform, time-frequency localization and signal analysis. IEEE Transactions on Information Theory, 36(5):961–1005, 1990.

[9] P. Dhariwal and A. Nichol. Diffusion models beat gans on image synthesis. Advances in Neural Information Processing Systems, 34:8780–8794, 2021.

[10] G. Giannone, D. Nielsen, and O. Winther. Few-shot diffusion models. arXiv preprint arXiv:2205.15463, 2022.

[11] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. Advances in Neural Information Processing Systems, 27, 2014.

[12] W. Harvey, S. Naderiparizi, V. Masrani, C. Weilbach, and F. Wood. Flexible diffusion modeling of long videos. arXiv preprint arXiv:2205.11495, 2022.

[13] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick. Momentum contrast for unsupervised visual representation learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 9729–9738, 2020.

[14] T. Henighan, J. Kaplan, M. Katz, M. Chen, C. Hesse, J. Jackson, H. Jun, T. B. Brown, P. Dhariwal, S. Gray, et al. Scaling laws for autoregressive generative modeling. arXiv preprint arXiv:2010.14701, 2020.
In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 4401–4410, 2019.

[27] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. 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.

[28] D. Kingma, T. Salimans, B. Poole, and J. Ho. Variational diffusion models. Advances in Neural Information Processing Systems, 34:21696–21707, 2021.

[29] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014.

[30] D. P. Kingma and M. Welling. Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114, 2013.

[31] Z. Kong, W. Ping, J. Huang, K. Zhao, and B. Catanzaro. Diffwave: A versatile diffusion model for audio synthesis. In International Conference on Learning Representations, 2021.

[32] A. Krizhevsky, G. Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[33] H. Li, Y. Yang, M. Chang, S. Chen, H. Feng, Z. Xu, Q. Li, and Y. Chen. Srdiff: Single image super-resolution with diffusion probabilistic models. Neurocomputing, 479:47–59, 2022.

[34] X. L. Li, J. Thickstun, I. Gulrajani, P. Liang, and T. B. Hashimoto. Diffusion-lm improves controllable text generation. arXiv preprint arXiv:2205.14217, 2022.

[35] Y. Li, R. Zhang, J. Lu, and E. Shechtman. Few-shot image generation with elastic weight consolidation. Advances in Neural Information Processing Systems, 2020.

[36] X. Liu, L. Wu, M. Ye, and Q. Liu. Let us build bridges: Understanding and extending diffusion generative models. arXiv preprint arXiv:2208.14699, 2022.

[37] T. J. Liu S, Guo H. Molecular geometry pretraining with se (3)-invariant denoising distance matching. arXiv preprint arXiv:2206.13602, 2022.

[38] C. Lu, Y. Zhou, F. Bao, J. Chen, C. Li, and J. Zhu. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. Advances in Neural Information Processing Systems, 2022.
[39] C. Lu, Y. Zhou, F. Bao, J. Chen, C. Li, and J. Zhu. Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models. arXiv preprint arXiv:2211.01095, 2022.

[40] C. Luo. Understanding diffusion models: A unified perspective. arXiv preprint arXiv:2208.11970, 2022.

[41] S. Luo and W. Hu. Diffusion probabilistic models for 3d point cloud generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 2837–2845, 2021.

[42] Z. Lyu, Z. Kong, X. Xu, L. Pan, and D. Lin. A conditional point diffusion-refinement paradigm for 3d point cloud completion. arXiv preprint arXiv:2112.03530, 2021.

[43] S. Mo, M. Cho, and J. Shin. Freeze the discriminator: A simple baseline for fine-tuning gans. In CVPR AI for Content Creation Workshop, 2020.

[44] A. Q. Nichol and P. Dhariwal. Improved denoising diffusion probabilistic models. In International Conference on Machine Learning, pages 8162–8171. PMLR, 2021.

[45] A. Noguchi and T. Harada. Image generation from small datasets via batch statistics adaptation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 2750–2758, 2019.

[46] U. Ojha, Y. Li, J. Lu, A. A. Efros, Y. J. Lee, E. Shechtman, and R. Zhang. Few-shot image generation via cross-domain correspondence. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 10743–10752, 2021.

[47] A. v. d. Oord, Y. Li, and O. Vinyals. Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748, 2018.

[48] M. Özşey, S. U. Dar, H. A. Bedel, O. Dalmaz, Ş. Özturk, A. Güngör, and T. Çukur. Unsupervised medical image translation with adversarial diffusion models. arXiv preprint arXiv:2207.08208, 2022.

[49] K. Rasul, C. Seward, I. Schuster, and R. Vollgraf. Autoregressive denoising diffusion models for variate probabilistic time series forecasting. In International Conference on Machine Learning, pages 8857–8868. PMLR, 2021.

[50] D. J. Rezende, S. Mohamed, and D. Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In International Conference on Machine Learning, pages 1278–1286. PMLR, 2014.

[51] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 10684–10695, 2022.

[52] C. Saharia, W. Chan, H. Chang, C. Lee, J. Ho, T. Salimans, D. Fleet, and M. Norouzi. Palette: Image-to-image diffusion models. In ACM SIGGRAPH 2022 Conference Proceedings, pages 1–10, 2022.

[53] C. Saharia, J. Ho, W. Chan, T. Salimans, D. J. Fleet, and M. Norouzi. Image super-resolution via iterative refinement. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022.

[54] A. Sinha, J. Song, C. Meng, and S. Ermon. D2c: Diffusion-decoding models for few-shot conditional generation. Advances in Neural Information Processing Systems, 34:12533–12548, 2021.

[55] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In International Conference on Machine Learning, pages 2256–2265, 2015.

[56] J. Song, C. Meng, and S. Ermon. Denoising diffusion implicit models. In International Conference on Learning Representations, 2021.

[57] Y. Song and S. Ermon. Improved techniques for training score-based generative models. Advances in Neural Information Processing Systems, 33:12438–12448, 2020.

[58] N.-T. Tran, V.-H. Tran, N.-B. Nguyen, T.-K. Nguyen, and N.-M. Cheung. On data augmentation for gan training. IEEE Transactions on Image Processing, 30:1882–1897, 2021.

[59] A. Vahdat and J. Kautz. Nvae: A deep hierarchical variational autoencoder. Advances in Neural Information Processing Systems, 33:19667–19679, 2020.

[60] A. Vahdat, K. Kreis, and J. Kautz. Score-based generative modeling in latent space. Advances in Neural Information Processing Systems, 34:11287–11302, 2021.

[61] A. Van den Oord, N. Kalchbrenner, L. Espeholt, O. Vinyals, A. Graves, et al. Conditional image generation with pixelcnn decoders. Advances in Neural Information Processing Systems, 29, 2016.
[62] X. Wang and X. Tang. Face photo-sketch synthesis and recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(11):1955–1967, 2008.

[63] Y. Wang, A. Gonzalez-Garcia, D. Berga, L. Herranz, F. S. Khan, and J. v. d. Weijer. Minegan: Effective knowledge transfer from gans to target domains with few images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9332–9341, 2020.

[64] Y. Wang, C. Wu, L. Herranz, J. van de Weijer, A. Gonzalez-Garcia, and B. Raducanu. Transferring gans: Generating images from limited data. In *Proceedings of the European Conference on Computer Vision*, pages 218–234, 2018.

[65] Z. Wang, H. Zheng, P. He, W. Chen, and M. Zhou. Diffusion-gan: Training gans with diffusion. *arXiv preprint arXiv:2206.02262*, 2022.

[66] J. Xiao, L. Li, C. Wang, Z.-J. Zha, and Q. Huang. Few shot generative model adaption via relaxed spatial structural alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11204–11213, 2022.

[67] Z. Xiao, K. Kreis, and A. Vahdat. Tackling the generative learning trilemma with denoising diffusion gans. In *International Conference on Learning Representations*, 2022.

[68] M. Xu, L. Yu, Y. Song, C. Shi, S. Ermon, and J. Tang. Gordiff: A geometric diffusion model for molecular conformation generation. In *International Conference on Learning Representations*, 2022.

[69] R. Yang, P. Srivastava, and S. Mandt. Diffusion probabilistic modeling for video generation. *arXiv preprint arXiv:2203.09481*, 2022.

[70] J. Yaniv, Y. Newman, and A. Shamir. The face of art: Landmark detection and geometric style in portraits. *ACM Transactions on Graphics (TOG)*, 38(4):1–15, 2019.

[71] F. Yu, A. Seff, Y. Zhang, S. Song, T. Funkhouser, and J. Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015.

[72] M. Zhang, Z. Cai, L. Pan, F. Hong, X. Guo, L. Yang, and Z. Liu. Motiondiffuse: Text-driven human motion generation with diffusion model. *arXiv preprint arXiv:2208.15001*, 2022.

[73] Q. Zhang and Y. Chen. Fast sampling of diffusion models with exponential integrator. *arXiv preprint arXiv:2204.13902*, 2022.

[74] Q. Zhang, M. Tao, and Y. Chen. gddim: Generalized denoising diffusion implicit models. *arXiv preprint arXiv:2206.05564*, 2022.

[75] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. 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.

[76] S. Zhao, Z. Liu, J. Lin, J.-Y. Zhu, and S. Han. Differentiable augmentation for data-efficient gan training. *Advances in Neural Information Processing Systems*, 33:7559–7570, 2020.

[77] Y. Zhao, K. Chandrasegaran, M. Abdollahzadeh, and N.-M. Cheung. Few-shot image generation via adaptation-aware kernel modulation. *Advances in Neural Information Processing Systems*, 2022.

[78] Y. Zhao, H. Ding, H. Huang, and N.-M. Cheung. A closer look at few-shot image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9140–9150, 2022.

[79] Z. Zhao, Z. Zhang, T. Chen, S. Singh, and H. Zhang. Image augmentations for gan training. *arXiv preprint arXiv:2006.02595*, 2020.

[80] L. Zhou, Y. Du, and J. Wu. 3d shape generation and completion through point-voxel diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5826–5835, 2021.
A Supplementary DPPM-PA Evaluation

In Table 4, we provide the additional results of Intra-LPIPS [46] on 10-shot FFHQ \( \rightarrow \) Sketches and FFHQ \( \rightarrow \) Amedeo’s paintings as supplements to Table 2. The proposed DPPM-PA approach outperforms modern GAN-based approaches and directly fine-tuned DDPMs on generation diversity under these adaptation setups as well.

FID results of GAN-based baselines in Table 3 are reported in CDC [46] and DCL [78].

B Supplementary Ablation Analysis

In this section, we provide the ablation analysis of the weight coefficients of \( L_{img} \), \( L_{hf} \), and \( L_{hfmse} \) using 10-shot FFHQ \( \rightarrow \) Babies as an example. Intra-LPIPS and FID are employed for quantitative evaluation.

We first ablate \( \lambda_2 \), the weight coefficient of \( L_{img} \). We adapt the source model to 10-shot Babies without \( L_{hf} \) and \( L_{hfmse} \). The quantitative results are listed in Table 5. Corresponding generated samples are shown in Fig. 8. When \( \lambda_2 \) is set as 0.0, the directly fine-tuned model produces coarse results lacking high-frequency details and diversity. With an appropriate choice of \( \lambda_2 \), the adapted model achieves greater generation diversity and better learning of target distributions under the guidance of \( L_{img} \). Too large values of \( \lambda_2 \) make \( L_{img} \) overwhelm \( L_{simple} \) and prevent the adapted model from learning target distributions, leading to degraded generation quality and diversity. The adapted model with \( \lambda_2 \) value of 2.5 gets unnatural generated samples even if it achieves the best FID result. We recommend \( \lambda_2 \) ranging from 0.1 to 1.0 for the adaptation setups used in our paper based on a comprehensive consideration of the qualitative and quantitative evaluation.

Next, we ablate \( \lambda_3 \), the weight coefficient of \( L_{hf} \) with \( \lambda_2 \) set as 0.5. The quantitative results are listed in Table 6. Corresponding generated samples are shown in Fig. 9. \( L_{hf} \) guides adapted models to keep diverse high-frequency details learned from source domains for more realistic results. \( L_{hf} \) helps the adapted model enhance details like clothes and hairstyles and achieves better FID and Intra-LPIPS, indicating improved quality and diversity. Too large values of \( \lambda_3 \) make the adapted model pay too much attention to high-frequency components and fail to produce realistic results following the target distributions. We recommend \( \lambda_3 \) ranging from 0.1 to 1.0 for the adaptation setups used in our paper.

Finally, we ablate \( \lambda_4 \), the weight coefficient of \( L_{hfmse} \) with \( \lambda_2 \) and \( \lambda_3 \) set as 0.5. The quantitative results are listed in Table 7. Corresponding generated samples are shown in Fig. 10. \( L_{hfmse} \) guides the adapted model to learn more high-frequency details from limited training data. Appropriate choice of \( \lambda_4 \) helps the adapted model generate diverse results containing rich details. Besides, the full DPPM-PA approach achieves state-of-the-art results of FID and Intra-LPIPS on 10-shot FFHQ \( \rightarrow \) Babies (see Table 2 and 3). Similar to \( \lambda_2 \) and \( \lambda_3 \), too large values of \( \lambda_4 \) lead to unreasonable results deviating from the target distributions. We
Table 5: Intra-LPIPS (↑) and FID (↓) results of adapted models trained on 10-shot FFHQ → Babies with different $\lambda_2$, the weight coefficient of $L_{img}$.

| $\lambda_2$ | Intra-LPIPS (↑) | FID (↓) |
|-------------|-----------------|---------|
| 0.0         | 0.520 ± 0.026   | 114.95  |
| 0.004       | 0.531 ± 0.031   | 92.87   |
| 0.02        | 0.544 ± 0.026   | 85.11   |
| 0.1         | 0.558 ± 0.033   | 75.17   |
| 0.5         | **0.572 ± 0.027** | 71.77   |
| 1.0         | 0.560 ± 0.034   | 74.68   |
| 2.5         | 0.543 ± 0.038   | **64.08** |
| 5.0         | 0.537 ± 0.028   | 69.18   |

Table 6: Intra-LPIPS (↑) and FID (↓) results of adapted models trained on 10-shot FFHQ → Babies with different $\lambda_3$, the weight coefficient of $L_{h_f}$.

| $\lambda_3$ | Intra-LPIPS (↑) | FID (↓) |
|-------------|-----------------|---------|
| 0.0         | 0.572 ± 0.027   | 71.77   |
| 0.004       | 0.576 ± 0.034   | **66.48** |
| 0.02        | 0.581 ± 0.045   | 72.67   |
| 0.1         | 0.589 ± 0.047   | 70.75   |
| 0.5         | **0.592 ± 0.031** | 70.40   |
| 1.0         | 0.583 ± 0.032   | 68.06   |
| 2.5         | 0.577 ± 0.032   | 71.69   |
| 5.0         | 0.591 ± 0.031   | 71.20   |

Figure 9: Visualized ablations of $\lambda_3$, the weight coefficient of $L_{h_f}$ on 10-shot FFHQ → Babies. Samples of different models are synthesized from fixed noise inputs.

Figure 10: Visualized ablations of $\lambda_4$, the weight coefficient of $L_{h_{mse}}$ on 10-shot FFHQ → Babies. Samples of different models are synthesized from fixed noise inputs.

C Limitations

Despite the compelling results of our approach, it still has some limitations. All the datasets used in this paper have the resolution of 256 × 256. The experiments of DDPM-PA are conducted on ×8 NVIDIA RTX A6000 GPUs (48 GB memory of each). However, the batch size on each GPU is still limited to 3. Therefore, it is challeng-
to expand our approach to larger image resolution. We will work on more lightweight DDPM-based few-shot image generation approaches. Despite that, the datasets used in this paper have larger resolution than many DDPM-based works \cite{10, 44, 2, 6, 28, 74} which use datasets with resolution $32 \times 32$ and $64 \times 64$.

Although we have designed methods for high-frequency details enhancement and achieved realistic results, there still exists room for further improvement, especially when target domains contain apparently more high-frequency components than source domains (e.g., LSUN Church $\rightarrow$ Haunted houses). Besides, DDPM-PA still cannot fully reproduce the styles of some abstract target domains while maintaining generation diversity. Nevertheless, our work first introduces diffusion models to few-shot image generation tasks and has improved generation quality and diversity compared with existing GAN-based approaches. We hope it will be a solid basis for better methods in the future.

### D Inspiration of Loss Design

#### D.1 $\mathcal{L}_{img}$ & $\mathcal{L}_{hf}$

The proposed pairwise similarity loss designed for DDPMs is mainly inspired by the methods in contrastive learning \cite{47, 13, 5}, which build probability distributions based on similarities. Similar methods are applied to GAN-based approaches, including CDC \cite{46} and DCL \cite{78} as well. GAN-based approaches depend on perceptual features in the generator and discriminator to compute similarity and probability distributions. As for the proposed DDPM-PA approach, the predicted input images $\tilde{x}_0$ calculated in terms of $x_t$ and $\epsilon_0(x_t, t)$ (Equation 15) are applied in replacement of perceptual features used for GANs. DDPM-PA directly uses image-level information to preserve the relative pairwise distances between adapted samples during domain adaptation.

We tried to use features in diffusion processes (Design A) and images of several diffusion steps (Design B) for pairwise similarity loss calculation. As shown in Table 8 (FID evaluation on FFHQ $\rightarrow$ Sunglasses, Intra-LPIPS evaluation on 10-shot FFHQ $\rightarrow$ Sunglasses), the proposed loss design is simple, effective, inexpensive, and achieves the best quality and diversity. Here we do not include high-frequency details enhancement for fair comparison.

As illustrated in Sec. 5, DDPM-PA synthesizes more realistic images with fewer blurs and artifacts and achieves better generation diversity than current state-of-the-art GAN-based approaches \cite{46, 78}. We also try to use image-level information to replace the perceptual features for the GAN-based approach CDC \cite{46}. However, we fail to avoid generating artifacts or achieve higher generation quality, as shown in Fig. 11. The proposed pairwise similarity loss matches better with diffusion models than GANs.

#### D.2 $\mathcal{L}_{hfmse}$

DDPMs learn target distributions mainly through mean values of predicted noises using the reweighted loss function (Equation 10). As a result, it is hard for DDPMs to learn high-frequency distributions from limited data, as shown in the smooth samples produced by models trained on limited data from scratch in Fig. 13. Therefore, we propose $\mathcal{L}_{hfmse}$ to strengthen the learning of high-frequency details from limited data during domain adaptation.

### E More Details of Implementation

#### E.1 More Details of DDPMs

We follow the model setups of DDPMs used in prior works \cite{44} for LSUN $256^2$ \cite{71} datasets. All the DDPM-based models used in this paper are implemented based on...
the same codebase [44, 9] and share the same model structure for fair comparison under different adaptation setups and optimization targets. All the source and target datasets are modified to the resolution of \(256 \times 256\). We use a max diffusion step \(T\) of 1000 and a dropout rate of 0.1. The models are trained to learn the variance with \(L_{vlb}\). The Adam optimizer [29] is employed to update the trainable parameters. We set the learning rate as 0.001 and apply the linear noise addition schedule to the noising process. Besides, we use half-precision (FP16) binary floating-point format to save memory and make it possible to use a larger batch size in our experiments (batch size 6 for directly fine-tuned DDPMs and batch size 3 for DDPM-PA per NVIDIA RTX A6000 GPU). All the results produced by DDPM-based models in this paper follow the sampling process proposed in Ho et al.’s work [16] (about 21 hours needed to generate 1000 samples on a single NVIDIA RTX A6000 GPU) without any fast sampling methods [56, 74, 38, 39, 73, 23]. The weight coefficient \(\lambda_2, \lambda_3,\) and \(\lambda_4\) are set as 0.5, 0.5, 0.05 for the results of DDPM-PA listed in Table 2, 3, and 4.

### E.2 More Details of GAN-based Baselines

We employ several GAN-based few-shot image generation approaches as baselines for comparison with the proposed DDPM-PA approach. Here we provide more details of these baselines. We implement all these approaches based on the same codebase of StyleGAN2 [27]. The source models are fine-tuned directly on the target datasets to realize TGAN [64]. TGAN+ADA applies ADA [24] augmentation method to the TGAN baseline. For FreezeD [43], the first 4 high-resolution layers of the discriminator are frozen following the ablations analysis provided in their work. The results of MineGAN [63] and CDC [46] are produced through their official implementation. As for EWC [35] and DCL [78], we implement these approaches following formulas and parameters in their papers since there is no official implementation. These GAN-based approaches are designed for generators [63, 35, 46, 78] and discriminators [24, 43, 78] specially and cannot be expanded to DDPMs directly.

The adapted GANs are trained for \(1K-3K\) iterations. When evaluating generation diversity using Intra-LPIPS [46] for the 10-shot adaptation tasks listed in Tables 2 and 4, we apply fixed noise inputs to different GAN-based ap-
Figure 13: Image samples produced by DDPMs trained from scratch on small-scale datasets, including Babies, Sunglasses, and LSUN Church containing 10, 100, and 1000 images.

Figure 14: Visualized comparison between DDPMs that achieve the best FID results and DDPMs trained for 60K iterations on 1000-shot Babies and LSUN Church. Image samples produced by different models are synthesized from fixed noise inputs.

Table 9: FID (↓) results of DDPMs trained for different iterations from scratch on 1000-shot Babies, Sunglasses, and LSUN Church.

| Iterations (K) | Babies  | Sunglasses | LSUN Church |
|---------------|---------|------------|-------------|
| 0             | 444.35  | 419.75     | 424.53      |
| 10            | 444.91  | 419.38     | 413.68      |
| 20            | 222.81  | 348.48     | 385.25      |
| 30            | **90.16** | 168.62    | 388.81      |
| 40            | 124.97  | 82.48      | 57.68       |
| 50            | 132.33  | 68.26      | **45.43**   |
| 60            | 132.32  | **66.09**  | 69.18       |

Apart from the GAN-based baselines illustrated above, RSSA [66] and AdAM [77] provide different research perspectives for few-shot image generation. RSSA [66] preserves the image structure learned from source domains with a relaxed spatial structural alignment method, which is inappropriate for abstract target domains like artists’ paintings. AdAM [77] focuses on unrelated source/target approaches for fair comparison.
domains with an adaptation-aware kernel modulation approach.

**F DDPMs Trained from Scratch**

In Section 3, we evaluate the performance of DDPMs trained from scratch on small-scale datasets containing 10, 100, and 1000 images. In our experiments, the smaller datasets are included in the larger datasets. For example, 1000-shot Sunglasses includes all the images in 100-shot and 10-shot Sunglasses. Similarly, all the images in 10-shot Sunglasses are included in 100-shot Sunglasses as well. We train DDPMs for $40K$ iterations (about 20 hours on $\times 8$ NVIDIA RTX A6000 GPUs) on datasets containing 10 or 100 images. While for datasets containing 1000 images, DDPMs are trained for $60K$ iterations (about 30 hours on $\times 8$ NVIDIA RTX A6000 GPUs).

We provide several typical examples randomly picked from the small-scale datasets in Fig. 12. Compared with the generated images shown in Fig. 13, it can be seen that DDPMs trained from scratch need enough training samples (e.g., 1000 images) to synthesize diverse results and avoid replicating the training samples. Detailed FID [15] results of DDPMs trained from scratch on Babies, Sunglasses, and LSUN Church containing 1000 images for $60K$ iterations are shown in Table 9. DDPMs trained from scratch on 1000-shot Babies, Sunglasses, and LSUN Church achieve the best FID results at $30K$, $60K$, and $50K$ iterations, respectively.

In Fig. 13, we provide samples generated by DDPMs trained from scratch for $60K$ iterations on all three 1000-shot Babies and LSUN Church datasets. In addition, we add the generated samples of models trained from scratch on 1000-shot Babies and LSUN Church, which achieve the best FID results for comparison in Fig. 14. We do not include samples for all three datasets since the model trained for $60K$ iterations on 1000-shot Sunglasses achieves the best FID result as well. For 1000-shot Babies, the model trained for $60K$ iterations achieves smoother results containing fewer blurs despite its worse FID result. The model trained for $30K$ iterations achieves the best FID result and synthesizes more diverse images. As for 1000-shot LSUN Church, the model trained for $50K$ iterations with the best FID result produces samples containing more detailed structures of churches than the model trained for $60K$ iterations. However, all these results are still coarse and lack

| Approaches | FFHQ       | LSUN Church |
|------------|------------|-------------|
| StyleGAN2  | 0.6688 ± 0.0589 | 0.7159 ± 0.0537 |
| DDPM       | 0.6631 ± 0.0592 | 0.7153 ± 0.0513 |

Table 10: Average pairwise LPIPS ($\uparrow$) results of 1000 samples produced by StyleGAN2 and DDPMs trained on FFHQ $256^2$ and LSUN Church $256^2$.

| Approaches | FFHQ | LSUN Church |
|------------|------|-------------|
| StyleGAN2  | 7.71 | 8.09        |
| DDPM       | 7.00 | 6.06        |

Table 11: FID ($\downarrow$) results of StyleGAN2 and DDPMs trained on FFHQ $256^2$ and LSUN Church $256^2$. 

Figure 15: Image samples produced by DDPMs trained on FFHQ $256^2$ [27] (300$K$ iterations) and LSUN Church $256^2$ [71] (250$K$ iterations).
high-frequency details, indicating the necessity of adapting source models to target domains when training data is limited.

**G DDPM-based Source Models**

We train DDPMs on FFHQ $256^2$ [27] and LSUN Church $256^2$ [71] from scratch for 300K iterations and 250K iterations as source models for DDPM adaptation, which cost 5 days and 22 hours, 4 days and 22 hours on $\times 8$ NVIDIA RTX A6000 GPUs, respectively.

Image samples produced by these two source models can be found in Fig. 15. We randomly sample 1000 images with these two models to evaluate their generation diversity using the average pairwise LPIPS [75] metric, as shown in Table 10. For comparison, we also evaluate the generation diversity of the source StyleGAN2 [27] models used by GAN-based baselines [64, 24, 43, 63, 35, 46, 78]. DDPMs trained on FFHQ $256^2$ and LSUN Church $256^2$ achieve generation diversity similar to the widely-used StyleGAN2 models.

Besides, we sample 5000 images to evaluate the generation quality of the source models using FID [15]. As shown in Table 11, DDPM-based source models achieve FID results similar to StyleGAN2 on the source datasets FFHQ $256^2$ and LSUN Church $256^2$. 
H  DDPM Adaptation Process Analysis

This paper mainly concentrates on the challenging 10-shot image generation tasks. When fine-tuning pre-trained DDPMs on target domains using limited data directly, too many iterations lead to overfitting and seriously degraded diversity. Fine-tuned models trained for about 10K iterations almost exclusively focus on replicating the training samples. Therefore, we train the directly fine-tuned DDPMs for $3K - 4K$ iterations to adapt source models to target domains and maintain diversity. However, the directly fine-tuned DDPMs still generate coarse samples lacking details with reasonable iterations.

In Fig. 16, we provide samples produced by DDPM-PA models trained for different iterations on 10-shot FFHQ → Babies. We apply fixed noise inputs to different models for comparison. As the iterations increase, the styles of the generated images become closer to the training samples. Images synthesized from the same noise inputs as Fig. 3 are included in red boxes. In addition, detailed evaluation of cosine similarity is added in Fig. 17. The source samples are adapted to the target domain while keeping relatively stable cosine similarity. Compared with the directly fine-tuned DDPMs, DDPM-PA has a stronger ability to maintain generation diversity and achieve realistic results containing rich details. Nonetheless, too many iterations still lead to the degradation of quality and diversity and details missing, as shown by the samples of the DDPM-PA model trained
Figure 20: Image samples synthesized from fixed noise inputs by directly fine-tuned DDPMs trained on 1000-shot FFHQ → LSUN Church for different iterations.

Therefore, we recommend choosing suitable iterations for different adaptation setups (e.g., $4K - 5K$ iterations for 10-shot FFHQ → Babies) to adapt the pre-trained models to target domains naturally and guarantee the high quality and great diversity of generated samples.

### I Unrelated Source/Target Domains

In this paper, we mainly focus on related source/target domains for the few-shot image generation tasks, e.g., FFHQ → Sketches and LSUN Church → Haunted houses. This section adds experiments using unrelated source/target domains, including FFHQ → LSUN Church and LSUN Church → Sunglasses. We use target datasets containing 1000 images. Here we only compare directly fine-tuned DDPMs with DDPMs trained from scratch. DDPM-PA is inappropriate for unrelated source/target domains since it is designed to preserve diverse information learned from source domains. FID [75] is employed to evaluate the convergence of models. We aim to find out whether unrelated source models could help fasten the convergence of DDPMs.

As shown in Table 12, fine-tuned DDPMs using unrelated pre-trained models also accelerate convergence compared with DDPMs trained from scratch. Fine-tuned DDPMs achieve the best FID results using $20K$ iterations for the two employed adaptation setups. The models trained from scratch achieve close FID results when trained for a lot more iterations (about $50K-60K$). The adaptation on unrelated source/target domains is still evidently faster than training from scratch. In Fig. 19 and 20, we provide visualized samples of fine-tuned DDPMs throughout training and show that DDPMs are capable of adapting unrelated source images to target domains. However, fine-tuned DDPMs still cannot avoid the generation of coarse samples lacking high-frequency details with limited training data in these cases. Besides, too many iterations may lead to blurred results. We recommend DDPM-PA on related source/target domain.
adaptation to achieve compelling results for few-shot image generation tasks.

**J Harr Wavelet Transformation Examples**

Fig. 21 visualizes several examples of Haar wavelet transformation. The low-frequency components LL contain the fundamental structures of images. High-frequency components including LH, HL, and HH contain rich details in images.

**K Computational Cost**

The time cost of DDPMs and the proposed DDPM-PA approach are listed in Table 13. DDPM-PA costs 24.14% more training time than the original DDPMs. DDPMs trained from scratch need about $40K$ iterations to achieve reasonable results, even if they can only replicate the training samples. DDPM-PA utilizes related pre-trained models to accelerate convergence (about $3K-5K$ iterations) and significantly improve generation quality and diversity. Compared with directly fine-tuned DDPMs, DDPM-PA is not overly time-consuming and achieves more realistic results.

**L Additional Visualized Samples**

We show all the 10-shot datasets used in this paper for few-shot image generation tasks in Fig. 22, including 5 target domains corresponding to FFHQ and 2 target domains corresponding to LSUN Church.
get domains corresponding to the source domain FFHQ [27] and 2 target domains corresponding to LSUN Church [71].

In Fig. 23, we provide image generation samples of DDPM-PA on 10-shot FFHQ → Amedeo’s paintings as supplements to Fig. 2. Compared with directly fine-tuned DDPMs, DDPM-PA generates more diverse samples containing richer details. DDPM-PA encourages adapted models to learn the common features of limited training samples while maintaining considerable generation diversity.

Besides, we provide image generation samples of GAN-based baselines and DDPM-based approaches on 10-shot FFHQ → Babies, FFHQ → Raphael’s paintings, and LSUN Church → Landscape drawings in Fig. 24, 25, and 26 as supplements to Fig. 6. We apply fixed noise inputs to GAN-based approaches and DDPM-based approaches, respectively. DDPMs are more stable and less vulnerable to
Figure 25: 10-shot image generation samples on FFHQ → Raphael’s paintings. All the samples of GAN-based approaches are synthesized from fixed noise inputs (rows 1-7). Samples of the directly fine-tuned DDPM and DDPM-PA are synthesized from fixed noise inputs as well (rows 8-9).

overfitting than GANs. GANs easily overfit and tend to generate samples similar to training samples when training data is limited (see samples of TGAN [64]). Directly fine-tuned DDPMs can still keep a measure of generation diversity under the same conditions. Besides, DDPM-based approaches avoid the generation of blurs and artifacts. However, directly fine-tuned DDPMs tend to produce too smooth results lacking high-frequency details and still face diversity degradation. DDPM-PA generates more realistic results containing richer high-frequency details than GAN-based baselines under all these adaptation setups.

DDPM-PA performs apparently better when we want to keep the fundamental structures of source images and learn the styles of target domains (e.g., LSUN Church → Landscape drawings). As shown in Fig. 26, GAN-based approaches fail to adapt real churches to the drawing style and produce samples containing too many blurs and artifacts. On the other hand, DDPM-PA produces high-quality church drawings and preserves more detailed building structures.
Figure 26: 10-shot image generation samples on LSUN Church → Landscape drawings. All the samples of GAN-based approaches are synthesized from fixed noise inputs (rows 1-7). Samples of the directly fine-tuned DDPM and DDPM-PA are synthesized from fixed noise inputs as well (rows 8-9).