Few-Shot Image Generation by Conditional Relaxing Diffusion Inversion

Yu Cao and Shaogang Gong

Queen Mary University of London, UK
{yu.cao, s.gong}@qmul.ac.uk

Abstract. In the field of Few-Shot Image Generation (FSIG) using Deep Generative Models (DGMs), accurately estimating the distribution of target domain with minimal samples poses a significant challenge. This requires a method that can both capture the broad diversity and the true characteristics of the target domain distribution. We present *Conditional Relaxing Diffusion Inversion* (CRDI), an innovative 'training-free' approach designed to enhance distribution diversity in synthetic image generation. Distinct from conventional methods, CRDI does not rely on fine-tuning based on only a few samples. Instead, it focuses on reconstructing each target image instance and expanding diversity through a few-shot learning. The approach initiates by identifying a *Sample-wise Guidance Embedding* (SGE) for the diffusion model, which serves a purpose analogous to the explicit latent codes in certain Generative Adversarial Network (GAN) models. Subsequently, the method involves a scheduler that progressively introduces perturbations to the SGE, thereby augmenting diversity. Comprehensive experiments demonstrate that our method outperforms GAN-based reconstruction techniques and achieves comparable performance to state-of-the-art (SOTA) FSIG methods. Additionally, it effectively mitigates overfitting and catastrophic forgetting, common drawbacks of fine-tuning approaches. Code is available at GitHub.

Keywords: Few-shot Learning · Diffusion Model · Implicit Latent Space

1 Introduction

Deep Generative Model (DGM) has been developed for generating images [13, 18, 42], audio [22, 33] and point clouds [24, 56]. A notable limitation, however, is their dependency on large-scale datasets and substantial computational resources for optimal performance. In many practical applications, only a few samples, sometimes a single sample, are available, *e.g.* photos of rare animal species and some medical images, in which case conventional DGM models are significantly limited [2, 32]. To overcome this problem, Few-Shot Image Generation (FSIG) methods have been proposed [30, 52, 61] to generate sufficiently high-quality and diverse images with only a few samples as training data, *e.g.* 10 samples. A natural way to achieve this goal is to transform the problem into a few-shot 'style' transformation, adapt prior knowledge from generative models built on larger

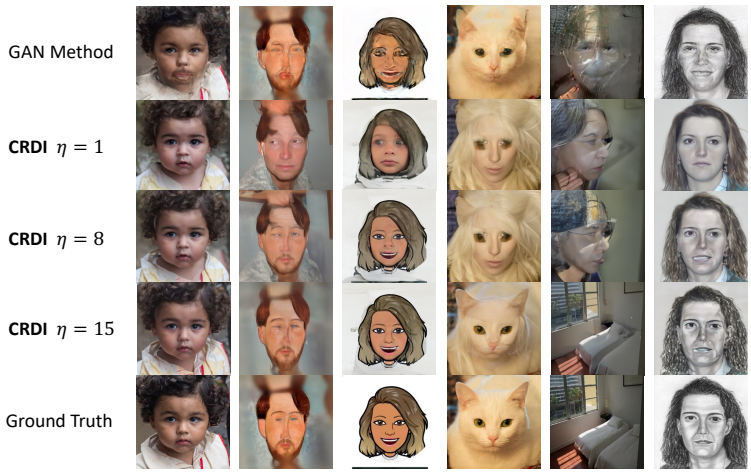


Fig. 1: Comparison of reconstruction results using Image2StyleGAN (GAN Based) [1] and our proposed method CRDI (Diffusion Based) with varying η values. Both source models are pre-trained on FFHQ [18]. We used the fast sampling method DDIM [44] with total 25 inference steps. We further set $\eta = 1, 8, 15$, whilst larger value means a stricter diffusion time-dependent SGE.

but 'similar' source datasets [30]. Generative Adversarial Network (GAN) [10] is the most widely used method due to its high quality generation. However, if only a few samples are available for learning the underlying distribution of a target domain, standard knowledge transfer approach used in GANs such as fine-tuning suffers from overfitting, mode collapse and catastrophic forgetting [21, 37, 43].

More recently, diffusion models (DMs) [14, 45] have demonstrated remarkable success, surpassing GANs in image generation [7]. In particular, their stochastic processes and probabilistic nature make diffusion models inherently well-suited for tasks such as image generation, text-to-image translation [38, 39, 42], and image editing [28]. It is attractive to consider if DMs can also be developed for FSIG to provide a better solution than the existing methods dominated by GANs. However, directly applying existing adaptation techniques used in GANs, including regularization [25, 32] and modulation [61] to the DMs not only fails to solve the problems faced by GANs, but also makes overfitting and catastrophic forgetting problems even worse due to the significantly larger number of parameters of DMs [2]. In parallel, it has been shown recently that high-fidelity StyleGAN2 [19] can represent a latent space for 'accommodating' a vast array of out-of-distribution data, as shown in Fig. 1 [1, 30]. This has triggered a desire to discover and learn a GAN latent space for the target domain, therefore enabling image generation by sampling latent codes from the latent space [1, 30], effectively mitigating the pitfalls of overfitting and catastrophic forgetting associated with fine-tuning pre-trained models. However, this approach introduces challenges, such as data leakage, and fails to apply directly to DMs, which lack a deterministic explicit latent code. Current latent space analysis of DMs fo-

cus primarily on semantic understanding, relying heavily on extensive data to identify a latent space for controlling the output, through whether learning an embedding space via VAEs [20] or geometric methods, hence incompatible with few-shot settings [23, 34, 46, 64].

To address this problem, we introduce a novel 'training-free' approach, *Conditional Relaxing Diffusion Inversion* (CRDI), to maximize distribution diversity in a diffusion generative process, leveraging a *Sample-wise Guidance Embedding* (SGE) to enhance diversity for a target domain. In particular, we discard the traditional concept of *style* transformation and solve the FSIG problem from the perspective of improving diversity. This can be further decomposed into two steps, Reconstruction and Diversity Enhancement. Specifically, **first**, we discover and estimate an SGE for the diffusion model, which serves as a guide for inference process, enabling reconstruction of a given target sample. Crucially, we allow flexibility in the intermediate noisy states, introducing a conditional relaxing process that enables a more robust reconstruction with initial variability. However, we find that the diversity enhancement from this step alone is limited. Inspired by Sadat et al. [41], we identify that the core diversity issue in diffusion models stems from their tendency to consistently associate identical labels with specific regions in the distribution space through label embeddings. Consequently, what should be stochastic variability is reduced to mere slight perturbations, rendering desirable randomness in diversity to an almost negligible factor. To address this limitation and further boost diversity, we introduce a crucial second step in our approach. **Second**, by manipulating the SGE through an annealing noise perturbation scheduler, we enhance the diversity of these reconstructions, fulfilling the FSIG objectives without the need for additional training or fine-tuning within the target domain. This dual-step approach ensures stable estimation of the target domain distribution while enhancing diversity for FSIG tasks. To the best of our knowledge, this work is the first to successfully adapt DMs for few-shot domains, bypassing the traditional model fine-tuning and mitigating the associated risk of overfitting to limited samples.

Our contributions are as follows: (1) Introduce and formulate a novel Sample-wise Guidance Embedding (SGE) as a dynamic *guidance* mechanism for diffusion models, enabling reconstruction within specific domains. We further show both theoretically and experimentally that this SGE possesses comparable functionalities to the explicit latent codes of GANs. (2) Propose a novel approach to FSIG by replacing the conventional style transformation with two separate processes utilizing SGE, consisting of a per target instance reconstruction and a few-shot target domain diversity expansion, without any additional training. (3) Explore the correlation between the rigidity and diversity of SGE to quantitatively control and provide insight into its effectiveness across different target domains.

2 Related Works

Few Shot Image Generation (FSIG) The objective of FSIG is to generate samples that are both high-quality and varied within a novel domain [52].

Conventional approaches typically apply fine-tuning a Generative Model pre-trained on a large dataset of a similar domain [5, 6, 52]. However, fine-tuning a full generative network mostly results in overfitting [16]. In practice, fine-tuning only updates a part of a model, *e.g.* BSA [31] and FreezeD [29]. To further improve the effectiveness of fine-tuning given a few shots, EWC [25], AdAM [61] and RICK [62] exploited some kernel methods by identifying important weights from the source model using Fisher Information [26] and preserve those knowledge while fine-tuning. In addition, [32] and RSSA [55] introduce additional loss functions to keep the structure of the generated target domain distribution close to that of the source domain. Hu *et al.* [16] modified these loss functions in order for a Large Multimodal Model (LMM) such as CLIP [36] to be able to apply to a Diffusion Model. In parallel, a representation learning method GenDA [30] was introduced for FSIG by constructing a manifold in a latent space.

Foundation Models Because of the high scalability of a diffusion model, many Large Multimodal Models (LMMs) such as DALL-E [38] and Stable Diffusion [39] have gained significant attention for zero-shot (prompt) generalization. Several few-shot adaptation methods based on these foundation models have emerged as a potential new solution for FSIG, such as DreamBooth [40], LoRA [15], and Textual-Inversion [9]. Although these methods can generate samples from a few shots, they are limited to adapting at the subject level. For optimal performance, the categories of the provided samples must be familiar to the model. Due to computational resource constraints, it is impractical to endlessly expand datasets to cover all target domain categories. Moreover, a fundamental premise of FSIG is that the target and source domains must not overlap [2,25,51,61], however, using foundation models may have exposed target domain in training, voiding FSIG assumptions. Therefore, the research on FSIG remains uniquely challenging.

Generative Model Latent Space It has been shown that high-fidelity model such as StyleGAN2 [19] can capture a latent space for accommodating out-of-distribution image generations [1], see examples in the first row of Fig. 1. To explore this idea, GenDA [30] was proposed to explicitly construct a target data manifold in a GAN latent space in order to generate images by sampling latent codes from this discovered manifold. However, extending this concept to the diffusion model presents greater challenges due to its iterative nature and the absence of a deterministic explicit latent space [14, 35, 54]. The current latent space analysis of diffusion methods is all semantically manipulable using inversion techniques [44, 49]. These techniques can be summarized in two directions: Some methods [34, 46, 57] use a variational autoencoder (VAE) [20] to construct implicit latent code and disentangle the desired feature for downstream tasks. Other methods [23, 64] use geometric analysis or a pre-trained LMM such as CLIP to influence the original source domain latent space. These approaches seek to establish an extensive semantic embedding space, necessitating large datasets. Our approach is designed to overcome the inherent limitations of constructing a diffusion FSIG model from a small sample size target domain, capable of simultaneously achieving diverse image generations beyond categories closely similar to the source domain, and sufficiently robust tractable model behavior.

3 Methodology

3.1 Preliminaries

Overview Our concept of decomposing FSIG into discrete steps is inspired by the Latent Diffusion Model [39], showing the advantage of training a separate compression network and a latent code sampler in isolation. This divide-and-conquer principle not only provides a clearer understanding on which components are limiting generation quality and diversity but also simplifies the overall diversification discovery with only a few-shot from the target domain.
Diffusion Model Background Generating noise from data is a 'simple' process, which can be described by the Stochastic Differential Equation (SDE):

$$d\mathbf{x}_t = \boldsymbol{f}\left(\mathbf{x}_t, t\right) dt + \boldsymbol{g}\left(\mathbf{x}_t, t\right) d\mathbf{w}_t, \quad t \in [0, T] \tag{1}$$

where $\mathbf{w}$ is the standard Wiener process, $\boldsymbol{f} : \mathbb{R} \to \mathbb{R}$ and $\boldsymbol{g} : \mathbb{R} \to \mathbb{R}$ are scalar drift and diffusion coefficients, respectively, with continuous time variable $t \in [0, T]$. Song *et al.* [45] shows that the SDE in Eq.(1) can be converted to a generative model by first sampling $x_T \sim \mathcal{N}\left(\mathbf{0}, \mathbf{I}\right)$ and then reversing the process through a given SDE:

$$d\mathbf{x}_s = \left[-\boldsymbol{f}\left(\mathbf{x}_s, s\right) + \boldsymbol{g}\left(\mathbf{x}_s, s\right) \boldsymbol{g}\left(\mathbf{x}_s, s\right)^{\top} \nabla_{\mathbf{x}_s} \log p_s\left(\mathbf{x}_s\right)\right] dt + \boldsymbol{g}\left(\mathbf{x}_s, s\right) d\mathbf{w}_s \tag{2}$$

which is the reverse-time Eq.(1) [3,11,45], where $x_s := x_{T-t}$ and $\nabla_{\mathbf{x}_s} \log p_s\left(\mathbf{x}_s\right)$ is the score function of the marginal distribution over x_s. Correspondingly, $\boldsymbol{f}$ and $\boldsymbol{g}$ are chosen to satisfy $x_T \sim \mathcal{N}\left(\mathbf{0}, \mathbf{I}\right)$. Further more, thanks to Tweedie's formula [8,47], score network $\mathbf{s}_{\boldsymbol{\theta}}\left(\mathbf{x}_t, t\right)$ can be proven to be equivalent to a noise network $\epsilon_{\boldsymbol{\theta}}(\mathbf{x}_t, t)$ introduced in Denoising Diffusion Probabilistic Model (DDPM) [14], which describes diffusion process from probability viewpoint. The reverse process equation using the noise prediction network is given by:

$$x_{t-1} = \sqrt{\bar{\alpha}_{t-1}} \left(\frac{x_t - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta(x_t, t)}{\sqrt{\bar{\alpha}_t}} \right) + \sqrt{1-\bar{\alpha}_{t-1}}\epsilon_\theta(x_t, t). \tag{3}$$

Problem Definitions To ensure notation consistency and enhance clarity, we formulate the FSIG task as follows: Consider a pre-trained generative model, the underlying distribution of the source data on which the model was trained is represented as P_S. Given a few samples from a target domain $\mathcal{T}$ with underlying distribution P_T, the goal is to adapt the pre-trained generative model to synthesize samples that follow a distribution approximating the target distribution P_T. Here, $\mathcal{S}$ and $\mathcal{T}$ represent the source domain and target domain respectively. Unless specified otherwise, the number of given samples of $\mathcal{T}$ is set to 10.

3.2 Unseen Target Domain Reconstruction

Unseen target domain reconstruction can be regarded as fine-grained conditional generating. A conditional generative model can be derived as $p_t(\mathbf{x}(t) \mid \mathbf{y})$ where

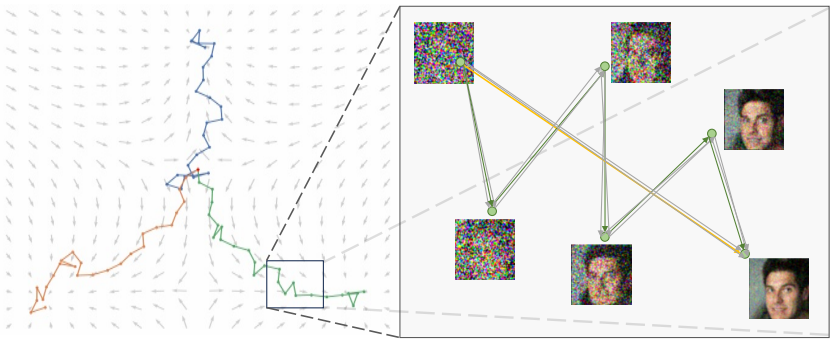


Fig. 2: A visualization (left) of three randomly sampled trajectories (blue, orange and green), all originating from the same initial point (red) and generated with Langevin dynamics. The green dots (right) represent the intermediate state x_t. A time-independent SGE is learned from one direct trajectory (yellow), which can be regarded as a directional path from x_α to x_β. The SGEs used to guide generation are perturbed by noise (gray) as defined in Sec. 3.3. Note that the right corner does not represent x_0.

$\mathbf{y}$ is the condition. Per Bayes' theorem, $p_t(\mathbf{x}(t) \mid \mathbf{y}) \propto p_t(\mathbf{x}(t))p(\mathbf{y} \mid \mathbf{x}(t))$, express this relationship to a score equation, a conditional DM is described as:

$$\nabla_{\mathbf{x}} \log p_t(\mathbf{x}(t) \mid \mathbf{y}) = \nabla_{\mathbf{x}} \log p_t(\mathbf{x}(t)) + \nabla_{\mathbf{x}} \log p(\mathbf{y} \mid \mathbf{x}(t)) \tag{4}$$

where $\nabla_{\mathbf{x}} \log p_t(\mathbf{x}(t))$ and $\nabla_{\mathbf{x}} \log p(\mathbf{y} \mid \mathbf{x}(t))$ are respectively the scores of a unconditional diffusion model and a time-dependent classifier [45]. Eq.(4) offers a transformation from unconditional to conditional sampling. Rather than using a module with simpler structures trained by few target domain samples to act as a 'classifier', we consider that the latter form does not necessarily need to be a trainable module, instead, it can be a fixed sample-wise tensor, which guides the generating process, we call it a Sample-wise Guidance Embedding (SGE). Fig. 2 shows a diagram of the proposed SGE. Since the classifier (Eq.(4)) is time-dependent, we further classify the SGE tensor into two forms: time-dependent $G_\theta(t)$ and, time-independent G_θ which is a special case of $G_\theta(t)$.

Degree of Rigidity For a diffusion model with T inference steps, we introduce a parameter η, which can take any integer value from 1 to T. We define $\eta = T$ as a strict time-dependent SGE and, similarly, $\eta = 1$ as a time-independent SGE. We call η as a *degree of rigidity*. In experiments, we observed that the minimum value of η varies across different target domains during reconstruction (Fig. 1). For those images similar to the source domain requires only a small η for successful reconstruction. Conversely, for images completely distinct from the source domain, e.g., reconstructing bedrooms with a model trained on a human face dataset such as FFHQ, even with $\eta = T$ is insufficient for reconstruction. This observation inspired us to consider η as a crucial metric for assessing the applicability of the knowledge learned on the source domain to a given target domain. Moreover, from a gradient perspective, as the value of η decreases from

Algorithm 1 Proposed Method Pseudo Code ($\eta = 1$)

Input: Target Domain $\mathcal{T}$ (given samples), Time Parameter α and β, Randomly Initialized SGE G_θ^i for $i \in [1, N]$, a Frozen Noise Network ϵ_θ and Learning Rate ν.
while not converge **do**
  for i, x_0 in *enumerate*($\mathcal{T}$) **do**
    Sample t uniformly from $[\beta, \alpha]$
    Given $x_{t-1} \leftarrow$ sample from $\sqrt{\bar{\alpha}_{t-1}}x_0 + \sqrt{1-\bar{\alpha}_{t-1}}\epsilon, \quad \epsilon \sim \mathcal{N}(0, \mathbf{I})$
    $\hat{\epsilon} \leftarrow \epsilon_\theta(x_t) - \sqrt{1-\bar{\alpha}_t} G_\theta^i$
    $x_0' \leftarrow \frac{x_t - \sqrt{1-\bar{\alpha}_t}\hat{\epsilon}}{\sqrt{\bar{\alpha}_t}}$
    $x_{t-1}' \leftarrow \sqrt{\bar{\alpha}_{t-1}}\left(\frac{x_t - \sqrt{1-\bar{\alpha}_t}\hat{\epsilon}}{\sqrt{\bar{\alpha}_t}}\right) + \sqrt{1-\bar{\alpha}_{t-1}}\hat{\epsilon}$
    $G_\theta^i \leftarrow G_\theta^i - \nu \nabla_{G_\theta^i} \mathcal{L}$
    where $\mathcal{L} = \|x_0 - x_0'\|^2 + \|x_{t-1} - x_{t-1}'\|^2 + ||G_\theta^i - \frac{1}{N}\sum_{j=1}^{N} G_\theta^j||^2$
return G_θ

T to 1, the role of our SDE shifts from dictating the pixel values to influencing the pixel evolution process. More analysis is given in Sec. 4.2.

Per Instance Conditional Relaxing Reconstruction Pixel-level reconstruction using SGE involves a fixed noisy state x_t obtained by adding noise to x_0 according to Eq.(1). However, this condition can be relaxed by allowing the inherent randomness in Eq.(1) to generate different x_t while finding the SGE, thereby enhancing reconstruction diversity. During the generation process, DMs predict the previous state x_{t-1} from the current state x_t using Eq.(3) with a noise prediction network [14, 45]. Additionally, Ho *et al.* [14] provide a direct estimate of x_0 from x_t as shown in Eq.(6). In line with our SGE principle, the sample-wise model adaptation can be effectively performed using Eq.(6). Nonetheless, during the generation phase, the SGE with $\eta > 1$ not only aids in estimating the initial state x_0 (Fig. 2) but also facilitates the generation of the preceding state x_{t-1}. Therefore, our loss function is formulated as Eq.(5):

$$\mathcal{L} = \|x_0 - x_0'\|^2 + \|x_{t-1} - x_{t-1}'\|^2 \quad (5) \qquad x_0 = \frac{x_t - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta(x_t, t)}{\sqrt{\bar{\alpha}_t}} \quad (6)$$

where x_0' and x_{t-1}' are derived using Eq.(6) and Eq.(3), respectively.

3.3 One-Shot Diversity Enhancement

Diversity is defined as the capacity of a model to generate a variety of outputs for a given condition [41]. As demonstrated by Song *et al.* [44], abandoning the constraint of a strict Markovian process enables the diffusion process to employ fast sampling by skipping certain steps. This is achieved through a sub-sequence τ drawn from the sequence $[0, ..., T]$. However, a shorter τ corresponds to fewer steps in the diffusion process, resulting in a reduced diversity of the stochastic processes. Similarly, in our method, varying η leads to an equal division of $[0, ..., T]$ into intervals, and for those diffusion processes that need greater η to reconstruct a target sample, their diversity in the target domain would be negatively affected. To reduce the negative impact of the high degree of rigidity

on diversity, we utilize a annealing schedule function $\lambda(t)$ designed by Sadat *et al.* [41] which corrupts a given condition y based on:

$$\hat{\boldsymbol{y}} = \sqrt{\gamma(t)}\boldsymbol{y} + s\sqrt{1-\gamma(t)}\boldsymbol{\epsilon} \quad (7) \qquad \gamma(t) = \begin{cases} 1 & t \leq \beta \\ \frac{\beta - t}{\beta - \alpha} & \beta < t < \alpha \\ 0 & t \geq \alpha \end{cases} \quad (8)$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}(0, 1)$, s is initial noise scale and α and β are the time parameters defining respectively the beginning and the end of the noise scaling interval. In our case, the SGE is the condition $\boldsymbol{y}$ in Eq.(7).

Fig. 3: Generated Babies facial images with different η, slightly source domain leakage problem (orange box) when $\eta = 1$.

However, the SGE in our method is not adaptable based on the current state x_t, making the condition scale inapplicable in our case. Taking this into account, we can rewrite Eq.(7) as: $\hat{\boldsymbol{y}} = \lceil\sqrt{\gamma(t)}\rceil\boldsymbol{y} + s\sqrt{1-\gamma(t)}\boldsymbol{\epsilon}$. Moreover, $\lambda(t) \to 0$ as $t \to T$, then $\hat{y} = s \cdot \epsilon$ is a scaled normal distribution independent of y. This allows us to streamline the discovery of the SGE from encompassing all timesteps to focusing on a specific sub-process. Specifically, the training scheme can be described as: choosing a starting point α and an ending point $\beta \in [0, T]$, for any given sample $x \sim \mathcal{T}$, we first calculate x_α by Eq.(1), then we learn an SGE using pre-trained diffusion model for every $t \in [\beta, ..., \alpha]$. Finally, we add noise perturbation $s\sqrt{1-\gamma(t)}\boldsymbol{\epsilon}$ based on Eq.(8) to our SGE, as shown by the gray line in Fig. 2. However, employing this method on SGE with a low η could potentially lead to data leakage problem (Fig. 3) due to the indiscriminate application of noise perturbation. We further discuss this phenomenon in Sec. 4.2

3.4 Synergy Effect with a Theoretical Analysis

In Sec. 3.2 and 3.3, we have described how to construct Sample-wise Guidance Embedding (SGE) and to increase diversity by Noise Perturbation. Here we explain how we extend this method to the few-shot setting and give a theoretical analysis. From a Score-based Diffusion viewpoint, the goal is to find an explicit solution to solve the reverse-time SDE given by Eq.(2). Specifically, in the probability distribution space $\mathcal{P}$, consider three distributions: the initial distribution $P_I \sim \mathcal{N}(0, I)$; and two distinct final distributions – a source distribution P_S and a target distribution P_T. The transformation from P_I to P_S within the space $\mathcal{P}$ is characterized by Eq.(2) with a pre-trained score network $\mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}_t, t)$.

In FSIG task, we want a score network $\mathbf{s}'_{\boldsymbol{\theta}}(\mathbf{x}_t, t)$ which 'transforms' P_I to the target distribution P_T. Let X_t represents the state of such a stochastic variable

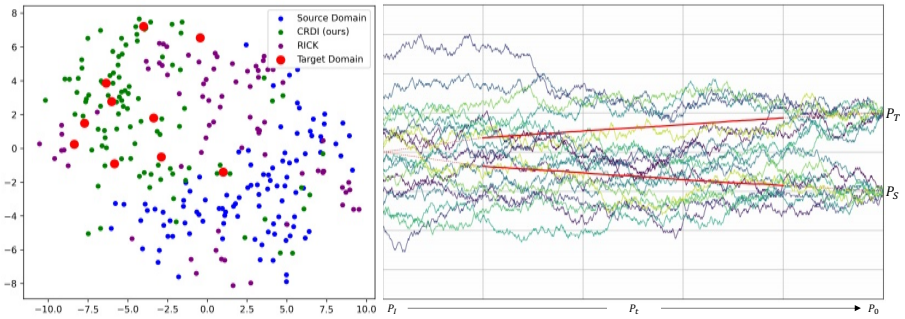


Fig. 4: Left: t-SNE results of given samples from Target Domain (Babies) (red), Source Domain (FFHQ) (blue), our generated samples (green), RICK [30] generated samples (purple). We show that our generated samples are more align with given target domain samples over RICK. Right: A simulation depicting two SDE transitions from P_I to the P_S and P_T. The two solid red lines illustrate the mean trajectories towards the Source and Target Domains, while the red dashed line indicates their extension.

at time t. Accordingly, for α and $\beta \in [0, T]$ and $\beta < \alpha$, Eq.(2) forges a dynamic bridge from X_α to X_β. Despite the absence of $\mathbf{s}'_{\boldsymbol{\theta}}(\mathbf{x}_t, t)$, we can utilize Eq.(1) to sample infinite intermediate state X_t for any given sample and hence optimize an SGE as described in Sec. 3.2. In essence, our SGE can be viewed as an individual trajectory guidance from P_I to P_T. However, it is critical to recognize that learning with a single sample may lead to substantial bias, potentially leading to generated samples falling outside the target domain (implausible diversion). To address this, we employ a mean across all the sample-wise guidance embeddings to serve as a penalty loss. The aggregate mean of the SGE across all given samples serves as an approximation of $\mathbf{s}'_{\boldsymbol{\theta}}(\mathbf{x}_t, t)$, providing a general guidance for the transition process. This strategy offers a practical solution by transitioning from a sample-wise perspective to a set-wise perspective. Consequently, this 'set-wise' guidance embedding ensures a more stable learning process by providing a robust and generalized direction for the transformation. Fig. 4 illustrates an overview of this concept and Algo. 1 summarizes the overall model learning process.

4 Experiments

Datasets Following previous work [25,30,32], we used Flickr Faces HQ (FFHQ) [18] as source domain datasets. We constructed a FSIG diffusion model to adapt to the following common target domains for comparisons with existing FSIG methods: (1) FFHQ-Babies [32], (2) FFHQ-Sunglasses [32], (3) Face Sketches [50], (4) Emoji Faces from bitmoji.com API [48], (5) MetFaces [17], (6) portrait paintings from the artistic faces dataset [58].

Metrics and Baseline For the reconstruction task, we calculate the SSIM (Structural similarity index measure) as quantitative metrics. FID (Fréchet inception distance) [12] and Intra-LPIPS (Intra-cluster pairwise Learned Percep-

Table 1: SSIM [53] Score (↑) of Image2StyleGAN [1] and CRDI (ours) with vary η, quantifying the reconstruction effectiveness.

Method	Backbone	Babies	Amedeo	Bitmoji	Cat	Bedroom	Sketches
Image2Style	StyleGAN2	0.57	0.76	0.68	0.73	0.52	0.43
CRDI $\eta = 1$	DDPM	0.55	0.66	0.67	0.65	0.58	0.48
CRDI $\eta = 8$	DDPM	0.68	0.75	0.81	0.77	0.67	0.63
CRDI $\eta = 15$	DDPM	**0.74**	**0.84**	**0.84**	**0.83**	**0.84**	**0.71**

tual Image Patch Similarity) [32, 60] are the most commonly used metrics in FSIG tasks, quantitatively measuring how closely the generated samples match the target domain in terms of quality and diversity, respectively. We further propose a new metric, MC-SSIM (Mode Coverage Structural Similarity Index Measure), which calculates the average of the top n SSIM scores for each generated image against the given set of target samples, a higher MC-SSIM score indicates superior mode coverage. We compared our proposed method against 11 FSIG models including TGAN [52], TGAN+ADA [17], BSA [31], FreezeD [29], EWC [25], CDC [32], RSSA [55], DDPM-PA [63] AdAM [61], RICK [62] and GenDA [30]. RICK and GenDA are considered the SOTA methods for fine-tuning and representation learning, respectively.

Implementation Details For the source model, we used Guided Diffusion [7] and checkpoint from Baranchuk *et al.* [4] for FFHQ at 256×256. We further utilized DDIM [44] and set the inference step at 25. Model learning was performed on A100 & H100 GPU with batch size 10. We considered 10 randomly sampled target samples, same as in existing methods for fair comparison, unless otherwise specified. For more details refer to the supplementary material.

4.1 Results

Per Target Instance Reconstruction We performed a comparative analysis, both qualitatively and quantitatively, of our reconstruction results against Image2StyleGAN [1] (widely used by other FSIG methods), as shown in Fig. 1 and Tab. 1. The comparison was carried out on distinct images from six domains with different similarity to the source domain FFHQ (examples can be found in Fig. 5). Additionally, in line with the methodology described in Sec. 3.2, we also compared the outcomes with different values of η. It is evident that our method consistently outperformed Image2StyleGAN in all six domains. Qualitatively, our technique excelled, especially in the Babies and Bitmoji categories, reconstruction can be achieved even when $\eta = 1$ without introducing artifacts. For domains that differ significantly from source domain, such as Bedrooms [59] and Amedeo paintings [58], a larger η is required for reconstruction. Surprisingly, although the Sketches [50] appear to be similar to source domain, they cannot be fully reconstructed, even with strict time-dependent SGE.

FSIG Qualitative Evaluation We show examples of the generated images of our method across two target domains (Babies and MetFaces), which vary

Fig. 5: We present the generated samples and Intra-LPIPS (↑) for our method alongside four other high performance methods across Babies ($\mathcal{T}_1$) and MetFaces ($\mathcal{T}_2$) with different degrees of similarity to the source domain ($\mathcal{S}$). While not consistently the best in Intra-LPIPS (↑), the quality and mode coverage (red box) of our samples is superior, characterized by fewer artifacts and an absence of noticeable overfitting phenomena. Best in **bold** and the second best in **underline with bold**. For more visual examples, please refer to supplementary material.

in their degree of similarity to the source domain, as in the top and bottom of Fig. 5, respectively. It can be observed that the fine-tuning approaches (FreezeD, RSSA, RICK) exhibit artifacts and overfitting in the generation within both target domains, while the representation learning approach (GenDA) results in images with limited diversity (low Intra-LPIPS). Our approach surpasses these methods by minimizing visual artifacts through reconstruction and substantially enhancing image diversity with progressively noise perturbation. However, in the Babies category, there remains a diversity gap between our method and RICK. We would like to emphasize that in many cases, guaranteeing the generation quality of the resulting image and aligning it with the target domain is more important than diversity, while over-increasing diversity can be dangerous. To verify that our method outperforms existing methods on this point, we employ a t-SNE [27] analysis against RICK on Babies in Fig. 4, it can be seen that our generated distribution shows a higher level of alignment with the samples

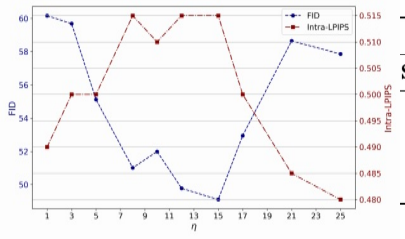
Table 2: Comparing FID (↓) Scores and MC-SSIM (↑) (for MetFaces only introduced in Sec. 4) between 11 different methods and our proposed method (CRDI). Best in **bold** and the second best in **underline with bold**.

		Babies	Sunglasses	MetFaces	
Methods	**Backbone**	FID ↓	FID↓	FID↓	MC-SSIM↑
TGAN [52]	StyleGAN	104.79	55.61	76.81	0.61
TGAN+ADA [17]	StyleGAN	101.58	53.64	75.82	0.61
BSA [31]	StyleGAN	140.34	76.12	–	0.69
FreezeD [29]	StyleGAN	110.92	51.29	73.33	0.64
EWC [25]	StyleGAN	87.41	59.73	62.67	0.64
CDC [32]	StyleGAN2	74.39	42.13	65.45	**0.70**
RSSA [55]	StyleGAN2	75.67	44.35	72.63	0.68
DDPM-PA [63]	DDPM	48.92	34.75	–	–
AdAM [61]	StyleGAN2	48.83	28.03	**51.34**	0.65
RICK [62]	StyleGAN2	**39.39**	**25.22**	**48.53**	0.69
GenDA [30]	StyleGAN2	63.31	35.64	104.48	0.35
CRDI (Ours)	DDPM	**48.52**	**24.62**	94.86	**0.78**

from the target domain and a reduced alignment with the source domain. This underscores the superior controllability of our method in the generation process.
FSIG Quantitative Evaluation In Tab. 2, we present complete FID scores, highlighting the performance of our method as superior against that of other representation learning techniques across all three target domains. Our approach outperforms the SOTA fine-tuning method RICK in the Sunglasses category, but faces challenges in the Babies and MetFaces. The notable discrepancy in MetFaces arises from its inherent variety, including sketches, ceramics, and ancient paintings, with an uneven distribution of these sub-domains in the full dataset. While other methods may achieve lower FID scores by excelling in dominant subdomains, they fail to capture the full range of MetFaces variety. Our approach, CRDI, however, consistently generates samples across all sub-domains within MetFaces (shown in Fig.5). To verify this, we utilize MC-SSIM, which assesses the distribution of generated samples across all target sub-domains. The results in Table 2 indicate a clear domain coverage advantage of our method over others. Despite not always achieving the lowest FID scores, broader coverage highlights its effectiveness in handling complex, diverse domains such as MetFaces.

4.2 Further Analysis and Discussion

Evaluation on Degree of Rigidity In Sec. 3.2, we theoretically analyzed the impact of η on the quality and diversity of diffusion generation. In Sec. 4.1, we qualitatively and quantitatively analyzed the impact of η on reconstruction. Here, we further explore the effects of varying η on the generation quality and diversity on Babies, the quantitative results shown in Fig. 6 (left). For both FID and Intra-LPIPS, we observe an initial optimization as η increases, while the



Comps.		FID Score (↓)		
SGE	Ptb.	Babies	Sunglasses	MetFaces
✗	✗	161.63	110.00	136.39
✗	✓	160.58	108.43	136.05
✓	✗	179.30	124.02	158.58
✓	✓	**48.52**	**24.62**	**94.86**

Fig. 6: Left: FID (↓) (blue) and Intra-LPIPS (↑) (red) for different degree of rigidity (η value) in target domain Babies. Right: Evaluating the impact of removing each components (Comps.), SGE and noise perturbation (Ptb.) by calculating the FID(↓) across three target domains.

quality of the generated images remains consistent across different η values, with minimal artifacts. This pattern can be attributed to a slight data leakage issue when η is low, which diminishes as η increases (shown in Fig. 3), allowing both FID and Intra-LPIPS to reach their optimal values. However, further increasing η imposes a stricter constraint, making the generated images align more with the given samples. This leads to a divergence from the test set distribution, resulting in an increased FID and reduced diversity. This property allows us to adjust between quality and diversity based on actual needs.

Moreover, as shown in Tab. 1 and Fig. 1, our good performing category, Babies, in terms of FID has the lowest SSIM score during reconstruction. Some fine-grained details (such as the collar) cannot be reconstructed, even when all randomness is removed during generation. We believe that the observed variation can be attributed to the distinct roles that SGE plays in generating samples across various categories, determined by the compatibility of the prior knowledge of the source model with the target domain. For generating images of Babies, SGE serves as a guiding mechanism, as the prior knowledge of the source model aligns with baby images. Conversely, for target domains vastly different from the source domain, it is challenging to apply the prior knowledge learned from the source domain, hence requiring SGE to store more semantic information and thus leaving little room for further diversification in generation.

Component Effectiveness Evaluation We evaluated the impact of removing each components by calculating the FID scores across three target domains, quantitative results are shown in Fig. 6 (Right). It can be observed that the removal of each component has a significant negative effect on the model performance measured in FID. Removing SGE, our model is reduced to an unconditional diffusion model, which can only generate samples from source domain. For SGE only (remove all randomness), it is reduced to the same setting as for reconstruction only. More visual examples refer to supplementary material.

From Few-Shot to One-Shot Till now, our experiments have leveraged 10 images from the target domain for domain adaptation. However, as described in Sec. 3.4, our method is adaptable with just one image. To demonstrate the effectiveness of our model under this extreme setting, we compared our results

Table 3: Comparisons of model performance from few-shot to one-shot given by the k value in k-shot adaptation on generation quality, evaluated by the FID score (↓).

	1-shot		5-shot		10-shot	
Methods	**Babies**	**Sunglasses**	**Babies**	**Sunglasses**	**Babies**	**Sunglasses**
GenDA	105.13	83.70	65.47	45.44	62.14	35.64
Ours	**100.85**	**74.60**	**55.87**	**31.35**	**48.52**	**24.62**

with GenDA [30] (only compatible method) across Babies and Sunglasses in Tab. 3. It is evident that our method outperforms GenDA in both domains.

Comparison with Foundation Model based Adaptation Methods While foundation models like Stable Diffusion [39] can generate diverse images, they lack precise control on producing samples that belong to a specific domain. Adaptation methods based on foundation model such as DreamBooth [40], Textual-Inversion [9] and LoRA [15], despite under few-shot settings, are primarily designed for subject-level image editing, resulting in poor performance on FSIG metrics. Moreover, foundation models may violate FSIG definition due to potential exposure to target domains during training. Despite these limitations, CRDI significantly outperforms these methods as shown in Tab. 4, demonstrating superior capability in generating samples that accurately represent the target domain. Visual examples and implementation detail refer to supplementary material.

Model	**Backbone**	**Babies**	**Sunglasses**	**MetFaces**
LoRa [9]	SD-1.5	143.78	88.38	99.65
DB [40]	SD-2.0	172.89	160.56	187.23
T-I [15]	SD-2.0	348.72	156.99	297.47
CRDI	DDPM	**48.52**	**24.62**	**94.86**

Table 4: Comparing FID (↓) of CRDI (ours) vs. DreamBooth (DB), Textual-Inversion (T-I) and LoRa on Babies, Sunglasses and MetFaces.

5 Conclusion

In this work, we present a novel framework to tackle the FSIG challenge, showing that limited data can be better utilized through distinct reconstruction and diversity enhancement phases. Our approach achieves SOTA performance using diffusion models, bypassing GANs. This represents a crucial advancement in FSIG technology by offering a measurable balance between the quality and diversity of generated images, and directly assessing the transferability of source models to target domains. Additionally, our method is scalable, compatible with current diffusion models, and optimized for efficiency and lightness.

Limitations & Future Work Our model exhibits reduced effectiveness in reconstructing sketches compared to Bedrooms, despite bedrooms being less similar to the RGB facial images (FFHQ). Looking forward, incorporating CLIP [36] as an additional gray-scale image guidance could allow our SGE to focus more on semantic information relevant to the target domain, potentially improving performance across diverse domains without compromising FSIG constraints.

Acknowledgments

This work was partially supported by Veritone, Adobe, and has utilized Queen Mary's Apocrita HPC facility from QMUL Research-IT.

References

1. Abdal, R., Qin, Y., Wonka, P.: Image2stylegan: How to embed images into the stylegan latent space? In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4432–4441 (2019)
2. Abdollahzadeh, M., Malekzadeh, T., Teo, C.T., Chandrasegaran, K., Liu, G., Cheung, N.M.: A survey on generative modeling with limited data, few shots, and zero shot. arXiv preprint arXiv:2307.14397 (2023)
3. Anderson, B.D.: Reverse-time diffusion equation models. Stochastic Processes and their Applications **12**(3), 313–326 (1982)
4. Baranchuk, D., Rubachev, I., Voynov, A., Khrulkov, V., Babenko, A.: Label-efficient semantic segmentation with diffusion models. arXiv preprint arXiv:2112.03126 (2021)
5. Bartunov, S., Vetrov, D.: Few-shot generative modelling with generative matching networks. In: International Conference on Artificial Intelligence and Statistics. pp. 670–678. PMLR (2018)
6. Clouâtre, L., Demers, M.: Figr: Few-shot image generation with reptile. arXiv preprint arXiv:1901.02199 (2019)
7. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Advances in neural information processing systems **34**, 8780–8794 (2021)
8. Efron, B.: Tweedie's formula and selection bias. Journal of the American Statistical Association **106**(496), 1602–1614 (2011)
9. Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A.H., Chechik, G., Cohen-Or, D.: An image is worth one word: Personalizing text-to-image generation using textual inversion. arXiv preprint arXiv:2208.01618 (2022)
10. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. Advances in neural information processing systems **27** (2014)
11. Haussmann, U.G., Pardoux, E.: Time reversal of diffusions. The Annals of Probability pp. 1188–1205 (1986)
12. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems **30** (2017)
13. Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., Lerchner, A.: beta-vae: Learning basic visual concepts with a constrained variational framework. In: International conference on learning representations (2016)
14. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
15. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685 (2021)
16. Hu, T., Zhang, J., Liu, L., Yi, R., Kou, S., Zhu, H., Chen, X., Wang, Y., Wang, C., Ma, L.: Phasic content fusing diffusion model with directional distribution consistency for few-shot model adaption. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2406–2415 (2023)

17. Karras, T., Aittala, M., Hellsten, J., Laine, S., Lehtinen, J., Aila, T.: Training generative adversarial networks with limited data. Advances in neural information processing systems **33**, 12104–12114 (2020)
18. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4401–4410 (2019)
19. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of stylegan. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8110–8119 (2020)
20. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
21. Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A.A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., et al.: Overcoming catastrophic forgetting in neural networks. Proceedings of the national academy of sciences **114**(13), 3521–3526 (2017)
22. Kong, Z., Ping, W., Huang, J., Zhao, K., Catanzaro, B.: Diffwave: A versatile diffusion model for audio synthesis. arXiv preprint arXiv:2009.09761 (2020)
23. Kwon, M., Jeong, J., Uh, Y.: Diffusion models already have a semantic latent space. arXiv preprint arXiv:2210.10960 (2022)
24. Li, C.L., Zaheer, M., Zhang, Y., Poczos, B., Salakhutdinov, R.: Point cloud gan. arXiv preprint arXiv:1810.05795 (2018)
25. Li, Y., Zhang, R., Lu, J., Shechtman, E.: Few-shot image generation with elastic weight consolidation. arXiv preprint arXiv:2012.02780 (2020)
26. Ly, A., Marsman, M., Verhagen, J., Grasman, R.P., Wagenmakers, E.J.: A tutorial on fisher information. Journal of Mathematical Psychology **80**, 40–55 (2017)
27. Van der Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research **9**(11) (2008)
28. Meng, C., He, Y., Song, Y., Song, J., Wu, J., Zhu, J.Y., Ermon, S.: Sdedit: Guided image synthesis and editing with stochastic differential equations. arXiv preprint arXiv:2108.01073 (2021)
29. Mo, S., Cho, M., Shin, J.: Freeze the discriminator: a simple baseline for fine-tuning gans. arXiv preprint arXiv:2002.10964 (2020)
30. Mondal, A.K., Tiwary, P., Singla, P., Prathosh, A.: Few-shot cross-domain image generation via inference-time latent-code learning. In: The Eleventh International Conference on Learning Representations (2022)
31. Noguchi, A., Harada, T.: Image generation from small datasets via batch statistics adaptation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2750–2758 (2019)
32. Ojha, U., Li, Y., Lu, J., Efros, A.A., Lee, Y.J., Shechtman, E., Zhang, R.: Few-shot image generation via cross-domain correspondence. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10743–10752 (2021)
33. Oord, A.v.d., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., Kavukcuoglu, K.: Wavenet: A generative model for raw audio. arXiv preprint arXiv:1609.03499 (2016)
34. Pandey, K., Mukherjee, A., Rai, P., Kumar, A.: Diffusevae: Efficient, controllable and high-fidelity generation from low-dimensional latents. arXiv preprint arXiv:2201.00308 (2022)
35. Preechakul, K., Chatthee, N., Wizadwongsa, S., Suwajanakorn, S.: Diffusion autoencoders: Toward a meaningful and decodable representation. In: Proceedings

of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10619–10629 (2022)
36. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
37. Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434 (2015)
38. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I.: Zero-shot text-to-image generation. In: International Conference on Machine Learning. pp. 8821–8831. PMLR (2021)
39. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
40. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22500–22510 (2023)
41. Sadat, S., Buhmann, J., Bradely, D., Hilliges, O., Weber, R.M.: Cads: Unleashing the diversity of diffusion models through condition-annealed sampling. arXiv preprint arXiv:2310.17347 (2023)
42. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-to-image diffusion models with deep language understanding. Advances in Neural Information Processing Systems **35**, 36479–36494 (2022)
43. Saito, M., Matsumoto, E., Saito, S.: Temporal generative adversarial nets with singular value clipping. In: Proceedings of the IEEE international conference on computer vision. pp. 2830–2839 (2017)
44. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)
45. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. arXiv preprint arXiv:2011.13456 (2020)
46. Song, Y., Keller, A., Sebe, N., Welling, M.: Flow factorized representation learning. Advances in Neural Information Processing Systems **36** (2024)
47. Stein, C.M.: Estimation of the mean of a multivariate normal distribution. The annals of Statistics pp. 1135–1151 (1981)
48. Taigman, Y., Polyak, A., Wolf, L.: Unsupervised cross-domain image generation. arXiv preprint arXiv:1611.02200 (2016)
49. Wallace, B., Gokul, A., Naik, N.: Edict: Exact diffusion inversion via coupled transformations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22532–22541 (2023)
50. Wang, X., Tang, X.: Face photo-sketch synthesis and recognition. IEEE transactions on pattern analysis and machine intelligence **31**(11), 1955–1967 (2008)
51. Wang, Y., Gonzalez-Garcia, A., Berga, D., Herranz, L., Khan, F.S., Weijer, J.v.d.: 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. pp. 9332–9341 (2020)

52. Wang, Y., Wu, C., Herranz, L., Van de Weijer, J., Gonzalez-Garcia, A., Raducanu, B.: Transferring gans: generating images from limited data. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 218–234 (2018)
53. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing **13**(4), 600–612 (2004)
54. Xiang, W., Yang, H., Huang, D., Wang, Y.: Denoising diffusion autoencoders are unified self-supervised learners. arXiv preprint arXiv:2303.09769 (2023)
55. Xiao, J., Li, L., Wang, C., Zha, Z.J., Huang, Q.: Few shot generative model adaption via relaxed spatial structural alignment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11204–11213 (2022)
56. Yang, G., Huang, X., Hao, Z., Liu, M.Y., Belongie, S., Hariharan, B.: Pointflow: 3d point cloud generation with continuous normalizing flows. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4541–4550 (2019)
57. Yang, Y., Wang, R., Qian, Z., Zhu, Y., Wu, Y.: Diffusion in diffusion: Cyclic one-way diffusion for text-vision-conditioned generation. arXiv preprint arXiv:2306.08247 (2023)
58. Yaniv, J., Newman, Y., Shamir, A.: The face of art: landmark detection and geometric style in portraits. ACM Transactions on graphics (TOG) **38**(4), 1–15 (2019)
59. Yu, F., Zhang, Y., Song, S., Seff, A., Xiao, J.: Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. arXiv preprint arXiv:1506.03365 (2015)
60. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018)
61. Zhao, Y., Chandrasegaran, K., Abdollahzadeh, M., Cheung, N.M.M.: Few-shot image generation via adaptation-aware kernel modulation. Advances in Neural Information Processing Systems **35**, 19427–19440 (2022)
62. Zhao, Y., Du, C., Abdollahzadeh, M., Pang, T., Lin, M., Yan, S., Cheung, N.M.: Exploring incompatible knowledge transfer in few-shot image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7380–7391 (2023)
63. Zhu, J., Ma, H., Chen, J., Yuan, J.: Few-shot image generation with diffusion models. arXiv preprint arXiv:2211.03264 (2022)
64. Zhu, Y., Wu, Y., Deng, Z., Russakovsky, O., Yan, Y.: Boundary guided mixing trajectory for semantic control with diffusion models. arXiv preprint arXiv:2302.08357 (2023)