# Conditional Distribution Modelling for Few-Shot Image Synthesis with Diffusion Models

Parul Gupta[1], Munawar Hayat[1], Abhinav Dhall[2], and Thanh-Toan Do[1]

[1] Monash University, Clayton VIC 3168, Australia
{parul,munawar.hayat,toan.do}@monash.edu
[2] Flinders University, Adelaide SA 5042, Australia
abhinav.dhall@flinders.edu.au

**Abstract.** Few-shot image synthesis entails generating diverse and realistic images of novel categories using only a few example images. While multiple recent efforts in this direction have achieved impressive results, the existing approaches are dependent only upon the few novel samples available at test time in order to generate new images, which restricts the diversity of the generated images. To overcome this limitation, we propose *Conditional Distribution Modelling (CDM)* – a framework which effectively utilizes Diffusion models for few-shot image generation. By modelling the distribution of the latent space used to condition a Diffusion process, CDM *leverages the learnt statistics of the training data* to get a better approximation of the unseen class distribution, thereby removing the bias arising due to limited number of few shot samples. Simultaneously, we devise a *novel inversion based optimization strategy* that further improves the approximated unseen class distribution, and ensures the fidelity of the generated samples to the unseen class. The experimental results on four benchmark datasets demonstrate the effectiveness of our proposed CDM for few-shot generation.

**Keywords:** Few-shot image generation · Diffusion models

## 1 Introduction

In this paper, we tackle few-shot image synthesis i.e., given only a few samples of a novel category, our goal is to generate diverse and realistic images of this new concept. Few shot synthesis can be effectively used to generate rarely occurring objects (such as rare species of birds or animals) and help to overcome the class imbalance issue in the naturally occurring datasets [26] by generating more examples of the minority categories. However, few-shot synthesis is a challenging task because the state of the art generative models such as GANs [13] and VAEs [23] need a large amount of data to learn any concept. Even with sufficient amount of data, due to their inherent nature (involving adversarial learning) adapting GANs for few samples is unstable to converge [46]. On the other hand, the sample quality achieved by VAEs is not as good as GANs even though their training is stable [36]. The scarcity of the available novel examples in few-shot

learning makes it difficult to estimate the distribution of the new categories which is disjoint from the training distribution. This often leads to overfitting on the few available test samples and lack of diversity in the generated examples.

Most of the existing few-shot synthesis approaches employ GAN-based architectures with carefully formulated auxiliary loss functions. We can broadly categorize these methods into 3 types: fusion based, optimization based and transformation based. Fusion based approaches [2, 14, 19, 21] fuse the latent features of a set of same-class images and then decode the fused feature into a new image to generate sample of the same class. However in addition to needing at least two novel examples, the generated images lack diversity, as they are similar to the few available examples and can't cover the entire distribution of the novel class. Optimization based approaches [5, 24, 47] introduce meta-learning, where they first learn a base model and then fine tune it for each novel category to generate new images. But the images generated by these methods are often blurry and of low quality. Transformation based approaches [1, 20] learn the intra-category transformations in the training data and apply them on the novel samples to generate new examples. Our method is a combination of the transformation and optimization based approaches, having the merits of both, *i.e.* good quality samples belonging to unseen class but with added diversity in the generated images, owing to our modelling the entire distribution of the novel class in the conditional latent space, which yields efficient and meaningful augmentation of the novel class samples (as explained in detail in Sec. 3.2).

We use a recently popular class of generative methods called Denoising Diffusion Probabilistic Models [17] for few shot image generation. These models can learn to generate meaningful images belonging to the training distribution from pure noise and enjoy the benefit of a stabilized training process (no adversarial learning) while maintaining the sample quality. They have shown remarkable success in various applications such as text-conditioned image generation [30,34], image super-resolution [35] and have outperformed GANs in class-conditioned image generation [7]. Since diffusion models are essentially probability density estimators, their application to few shot generation is non-trivial. Unlike VAEs, their latent space is unstructured – the latents are just training images disturbed by adding varying levels of Gaussian noise. Therefore, applying the ideas from few-shot GAN-based approaches [14, 20] and fusing or transforming the latent representations of diffusion models is not intuitively expected to yield any meaningful results. Diffusion-conditioning mechanisms have been developed for controlling simpler concepts such as low-shot attribute generation [36], class-conditional synthesis (for seen classes during training) [28], artistic domain-adaptation [49] and test-time adaptation [4], but not yet explored to generate entirely novel complex classes, from only a few available examples, using a limited amount of training data. Our approach successfully generates new samples of unseen concepts by modelling the distribution of the space used to condition the diffusion process. In order to do so, we develop *Conditional Distribution Modelling* (see Sec. 3.2) where we estimate the probability distribution of samples belonging to each class. This enables us to augment the conditional vectors

for novel class by simply sampling from its approximated distribution, whose statistics are borrowed from the closest class in the training set and optimized using the few novel class samples. These augmented unseen conditionals when passed through the conditional diffusion process in turn give us a set of diverse and realistic new samples. While the latent space distribution of training classes can be accurately estimated using the abundant samples available, using only the small number of samples of the never seen concepts at test time results in poor approximation of their distribution statistics. To counter this, for each unseen class, we propose to transfer the distribution statistics from the closest seen classes [44] to get an initial estimate and optimize them using the available samples. Thus, the sample diversity in our approach comes from two sources – firstly, the conditional space modelling helps us generate more conditionals for the diffusion process and secondly the diffusion model learns valid intra-class transformations (from the conditional latent to the target latent).

In summary, our major contributions are:

- We propose **a novel diffusion model-based framework for few-shot image generation** that effectively captures the diversity while maintaining the distinctive characteristics of unseen classes by modelling their distributions in its conditional space.
- Specifically, instead of relying only on the few-shot examples to generate new samples, we develop a principled approach that leverages the learned statistics from the neighbouring seen classes to approximate the unseen class distribution and **faithfully capture the unseen class diversity**.
- Further, we propose **a novel inversion based optimization to refine the unseen class distribution** which in turn ensures the fidelity of the generated samples to the unseen class.

## 2 Related Work

The existing literature on few shot image generation can be broadly divided into three categories - fusion-based, optimization-based and transformation-based methods.

**Fusion based approaches** produce new images of unseen classes by *fusing* the latent features of the example images in some manner and passing the fused representation through a decoder, e.g. GMN [2] and Matching-GAN [19] combine the Matching Network [42] (used for few-shot classification) with VAEs and GANs respectively. F2GAN [21] enhances the fusion of high level features by filling low level details from the example images using a Non-local Attention module. Similarly, LoFGAN [14] is based upon a learnable Local feature Fusion module (LFM) combined with GANs. WaveGAN [43] adapts Haar Wavelet transform to capture features at different frequencies and fuses them. The images generated using these methods often lack in diversity due to their dependence on the few unseen examples available for the fusion operation.

**Optimization based approaches**, e.g., FIGR [5] and DAWSON [24] use adversarial learning (GANs) combined with meta-learning methods, e.g., Reptile [27]

(used by [5]) and MAML [10] (used by [24]) to generate new images. LSO [47] adapts the latent space learnt using StyleGAN [22] (an approach for training GANs with limited data) for each unseen class using optimization and semantic stabilization losses. Exposing the model [47] to the few unseen samples at test time helps it to capture the class-specific characteristics, allowing the generated samples to have high fidelity with the unseen class; however optimization needs to be performed carefully to ensure that the model does not over-fit on the few samples that can take away the generation quality and diversity.

**Transformation based approaches** (*e.g.*, AGE [8])learn the pattern of transformations between different pairs of the same class during training and use these transformations to generate new samples of novel categories from the available samples, e.g., in DAGAN [1] any sample passed though the encoder gives a feature containing its class-level information. This feature along with a random vector is passed through the decoder to generate a different sample of the same class. DeltaGAN [20] learns to generate transformations called *sample-specific deltas* which represent the information needed to convert the sample to another image of the same class. Given a sample, different plausible deltas can be generated based upon a random vector input. However, the current transformation-based approaches require end-to-end training of transformation and generation that can be unstable resulting in generation of low quality images.

A common limitation of most of the existing approaches is their lack of generation capability on coarse-grained datasets such as CIFAR-100 or ImageNet, due to the inherently higher inter-category variance. *Due to this, currently, we keep the scope of our approach limited to fine-grained datasets only.*

While Diffusion Models are rarely explored for few-shot image synthesis, several works focus on few-shot generative *adaptation*. These methods pre-train the model on a large-scale dataset from a related source domain and adapt it to the target domain. For example, D2C [36] jointly trains a VAE and a Diffusion model in the VAE's latent space on a large dataset, then uses a few labeled examples to train a classifier in that space– e.g., they use the CelebA-64 dataset [25] for initial training and learn a binary classifier to predict attributes like blond/female using just 100 labeled examples. To generate images with these attributes, they produce VAE latents from random noise, select those with high classifier scores, and decode them via the VAE. Similarly, DDPM-PA [49] employs a pairwise similarity loss during domain adaptation to maintain relative distances between generated samples, enhancing diversity in the target domain. In a related direction, there are several *personalisation* techniques being developed over the large-scale pre-trained *foundation* Diffusion models, wherein limited number of images of a particular object/person can be used to generate the same object/person in arbitrary contexts using inversion [11] or fine-tuning [33]. However, the objects being generated in these approaches aren't exactly *unseen* to the Diffusion models, which have been trained on internet-scale data. Hence, there is no apples-to-apples comparison of these approaches with the existing methods for few-shot image generation. To our knowledge, diffusion models for few-shot synthesis have only been investigated in FSDM [12], utilising a DDPM to generate images of

different classes through a fusion-based approach. It obtains a class-wise set-level context by passing the set of images along with the timestep embedding through a Vision Transformer [9] and uses the context to condition the generative path (reverse path) of the DDPM. However, its diffusion operates in image space, limiting scalability to high-resolution images due to memory and inference time constraints, evaluated mainly on small-scale $32 \times 32$ datasets. In contrast, our approach ensures diversity and fidelity in the synthesized images by adequately modelling the distribution of the few shot samples in the conditional space of the diffusion process. This provides us with the flexibility to sample new and varied conditionals which are then effectively mapped to the image space. Since the diffusion itself happens in a regularized latent space, we are able to efficiently generate high-resolution and realistic novel class images.

## 3 Method

### 3.1 Preliminaries

**Problem Definition** Given a dataset $\mathbb{D} = (\mathbf{x}_j, y), y \in [1, \mathbb{C}], j \in [1, n_y]$ with $\mathbb{C}$ classes and $n_y$ images in each class, we divide the dataset into seen classes $\mathbb{C}_s$ and unseen classes $\mathbb{C}_u$ where $\mathbb{C}_s \cap \mathbb{C}_u = \phi$. Only the images from seen classes can be used while training. The task of $K-$ shot generation involves generating new images of any class from $\mathbb{C}_u$, using only $K$ images of this class.
Below, for the sake of completeness, we briefly revisit diffusion models [17] and latent diffusion models [31], followed by the description of our proposed conditional distribution modelling approach for few-shot synthesis (Sec. 3.2)
**Diffusion Models** [17] are a special class of generative models which configure the data distribution as a reverse (or *denoising*) process of iteratively adding noise to the data. Thus, the forward *diffusion* process converts the structured data ($\mathbf{x}_0$) into pure noise ($\mathbf{x}_T$) in $T$ timesteps by adding small amounts of Gaussian noise each time. The amount of noise $\epsilon$ added at each step is controlled by a non-decreasing variance schedule $[\beta_t \in (0,1)]_{t=1}^{T}$. Thus at each step $t$, we sample $\epsilon \sim \mathcal{N}(0, I)$ and then get $\mathbf{x}_t = \sqrt{1-\beta_t}\mathbf{x}_{t-1} + \sqrt{\beta_t}\epsilon$. After $T$ steps through the encoder, the information in the data point is destroyed completely and $\mathbf{x}_T$ becomes random Gaussian noise. The Denoising Diffusion Probabilistic Model (DDPM) [17] learns the reverse, where we start from pure noise and recover a point from the original data distribution through $T$ steps of a decoder (denoted by $p_\theta$). Given the noisy data point $\mathbf{x}_t$ and the time step $t$ as input, the decoder predicts the noise ($\epsilon$) added to the original data point that lead to this noisy input. The architecture most commonly used for $p_\theta$ is a time-conditioned UNet [32]. The training objective ($L_{vlb}$) comes by minimizing the variational bound on the negative log-likelihood of the data distribution and is used in combination with the mean square error ($L_{simple} = E_{t,x_0,\boldsymbol{\epsilon}}[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)\|^2]$) between the true noise and predicted noise. A detailed derivation of the training objective can be found in [17].
**Latent Diffusion Model** (LDM) [31] offers an efficient setup for high-resolution image synthesis, leveraging the diversity and high quality of Diffusion models

without sacrificing computational efficiency by operating in the latent space of a pretrained Variational Autoencoder [23] (with Encoder $\mathcal{E}$ and Decoder $\mathcal{D}$). Pretraining the VAE creates a compressed yet meaningful latent space, allowing the diffusion process to model the semantic aspects of the data more efficiently than in the high-dimensional image space. Additionally, LDM introduces a Cross-Attention mechanism [41] to condition diffusion on signals from other modalities. We adapt this conditional latent diffusion pipeline for our use case.

### 3.2 Conditional Distribution Modelling (CDM)

As depicted in Fig. 1, Conditional Distribution Modelling involves five stages, each of which are described below.

***A. ResNet and VAE Training.*** (Fig. 1 (A)) We initially train a vector-quantization [39] regularized Variational Autoencoder (VAE) on the seen data (denoted by $\mathbb{C}_s$), following the LDM architecture. The LDM is trained at a later stage in this VAE's latent space, denoted by $\mathbf{z}$. Simultaneously, we also train a simple ResNet-based classifier on the seen data using Cross-Entropy loss and choose the output of its penultimate layer (denoted by $f$) as the conditioning space for our LDM.

***B. Distribution Modelling of seen and unseen classes.*** (Fig. 1 (B)) We obtain class-wise latent mean $\mu^y$ and latent variance $\left(\sigma^2\right)^y$ on the seen data as follows–

$$\mu^y = \frac{1}{n_y} \sum_{j=1}^{n_y} f_j^y \tag{1}$$

$$\left(\sigma^2\right)^y = \frac{1}{n_y - 1} \sum_{j=1}^{n_y} \left(f_j^y - \mu^y\right)^2 \tag{2}$$

$$\mathbb{D}^y = \mathcal{N}\left(\mu^y, \left(\sigma^2\right)^y\right) \tag{3}$$

where $n_y$ is the number of samples in seen class $y$ and each dimension in the latent space $f$ is considered to be *uncorrelated*. We assume that every dimension in the latent vector $f_j^y$ follows a Gaussian distribution (denoted by $\mathbb{D}_y$, $f_{seen} = \{\mathbb{D}_y\}_{y \in \mathbb{C}_s}$) and observe that similar classes usually have similar statistics (mean and variance) of the feature representations. This allows us to transfer the mean and variance statistics across similar classes, i.e., from seen classes for which we have a better approximation of these statistics to the unseen classes for which we only have a few examples, which are insufficient to approximate the underlying distribution. Therefore, given $K$ support samples of an unseen class $c(\in \mathbb{C}_u)$, we first obtain the latent vectors $f_k^c$ $(1 \leq k \leq K)$ for each of these $K$ samples using the ResNet from stage (A). Then, we augment the latent space $f^c$ by calibrating the mean and variance statistics from the seen classes which are nearest to this unseen class. The nearest seen classes are the ones with the minimum Euclidean distance between their mean latent $\mu^y$ and the mean of the unseen latent $\mu^c$.

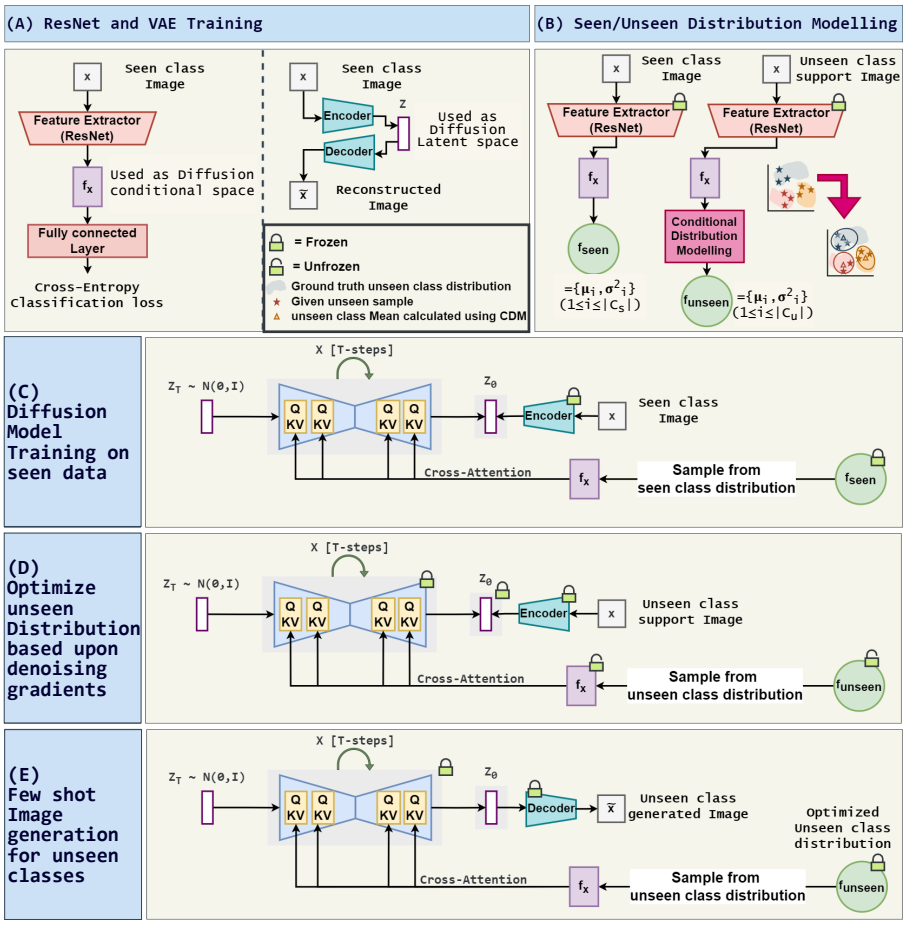


**Fig. 1: CDM Pipeline: (A)** First, we train a ResNet classifier and a Variational Autoencoder (VAE) on the seen data (denoted by the set $\mathbb{C}_s$). The ResNet's penultimate layer's output ($f$) is to be used as a conditional input to the Diffusion model later. **(B)** We calculate the class-wise means ($\mu_i$) and variances ($\sigma_i^2$) of the seen data in the latent space $f$. This collection of Gaussian distributions is denoted by $f_{seen}$. Now, for each unseen class (belonging to the set $\mathbb{C}_u$), we have $K$ support samples for $K$-shot image generation task. The seen classes whose distributions $(\mu_i, \sigma_i^2)$ are closest to this unseen class are used to estimate its distribution $(\mu_i, \sigma_i^2)$ through Conditional Distribution Modelling. This process is shown in the two plots having 3 unseen classes, where 3 support samples per class and $f_{seen}$ are used approximate the unseen class distributions. **(C)** We train a Diffusion model in the VAE latent space, conditioned upon the samples obtained from the seen class distributions ($f_{seen}$). **(D)** We use inversion based optimization to improve the unseen class distributions ($f_{unseen}$) using the denoising gradients from the support samples. **(E)** The optimized unseen class distributions are used to generate new samples from the Diffusion Model.

The mean of an unseen class distribution is computed from latent vectors $f_k^c$ of support samples of that class:

$$\mu^c = \frac{\sum_{k=1}^{K} f_k^c}{K} \tag{4}$$

Let the set of nearest seen classes be denoted by $\mathbb{S}_N$, the calibrated variance $\left(\sigma^2\right)^c$ of that unseen class distribution is given by:

$$\left(\sigma^2\right)^c = \frac{\sum_{y \in \mathbb{S}_N} \left(\sigma^2\right)^y}{|\mathbb{S}_N|} \tag{5}$$

**Algorithm 1** Training LDM with CDM

**Require:** Trained Feature Extractor and VAE
**Require:** Training data $D = (\mathbf{x}_j, y), y \in \mathbb{C}_s, j \in [1, n_y]$
**Require:** Seen classes' statistics $\{\mu^y, \left(\sigma^2\right)^y\}, y \in \mathbb{C}_s$ obtained as per Eq. (1) and Eq. (2)
**for** $m = 1, ..., \#epochs$ **do**
  **for** $y = 1, ..., |\mathbb{C}_s|$ **do**
    **for** $j = 1, ..., n_y$ **do**
      Obtain the latent representation $\mathbf{z}$ for $\mathbf{x}_j$ by passing it through VAE Encoder
      Sample a timestep $t \in [1, T]$ uniformly at random and get the noised version of $\mathbf{z}$ denoted by $\mathbf{z}_t$
      Sample a latent $f$ for class $y$ from the gaussian distribution $\mathbb{D}^y$ as per Eq. (3)
      Use $f$ as a conditional input to the LDM to denoise $\mathbf{z}_t$ into $\mathbf{z}$ and update the UNet parameters based upon the loss $\mathcal{L}$ as per Eq. (6).
    **end for**
  **end for**
**end for**

***C. Diffusion Model Training.*** (Fig. 1 (C)) In this stage, we train a Latent Diffusion Model (LDM) on the seen data in the $z$-space of the stage (A) VAE. This LDM is conditioned using samples from the seen data distributions $f_{seen}$ defined in (B). Thus, given an image $\mathbf{x}$ belonging to a seen class $y$, we first sample a latent $f^y \sim \mathbb{D}^y$ (the corresponding distribution obtained as per Eq. (3)) and also get the perceptually compressed representation $\mathbf{z}$ corresponding to $\mathbf{x}$ from the pretrained VAE. Now, the LDM learns to denoise the noisy version $\mathbf{z}_t$ of $\mathbf{z}$, conditioned on $f^y$ and time-stamp $t$ which can be shown as the overall training objective [28],

$$\mathcal{L} := \mathcal{L}_{simple} + \lambda \mathcal{L}_{vlb}, \tag{6}$$

where $\mathcal{L}_{vlb}$ is the variational lower bound based loss and $\mathcal{L}_{simple}$ is the mean squared error between the true noise and predicted noise and is given by

$$\mathcal{L}_{simple} := \mathbb{E}_{\mathcal{E}(\mathbf{x}), y, \epsilon \sim \mathcal{N}(0,I), t} \left[ \|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{z}_t, t, f^y)\|_2^2 \right] \tag{7}$$

We further employ classifier-free guidance proposed by [18] in order to enhance the sample diversity. In summary, while training the LDM, we ensure that the model learns to *transform* the vectors $f^y$ that are sampled from the class distribution $\mathbb{D}^y$ into the latent representations $\mathbf{z}$ of the same class. Algorithm 1 describes the LDM training process with the proposed CDM.

**Algorithm 2** Inversion based optimization of unseen class distributions

**Require:** $K$ unseen class samples $\{\mathbf{x}_i, c\}_{i=1}^{K}, c \in \mathbb{C}_u$
**Require:** Trained VAE and LDM models
**Require:** Initial unseen classes' statistics $\{\mu^c, \left(\sigma^2\right)^c\}, c \in \mathbb{C}_u$ obtained as per Eq. (4) and Eq. (5)
**for** $m = 1, ..., \#optimization\ steps$ **do**
  **for** $c = 1, ..., |\mathbb{C}_u|$ **do**
    **for** $i = 1, ..., K$ **do**
      Obtain the latent representation $\mathbf{z}$ for $\mathbf{x}_i$ by passing it through VAE Encoder
      Sample a timestep $t \in [1, T]$ uniformly at random and get the noised version of $\mathbf{z}$ denoted by $\mathbf{z}_t$
      Sample a latent $f^c$ for class $c$ by first sampling a latent $\boldsymbol{\epsilon}$ from the standard normal distribution $\boldsymbol{\epsilon} \sim \mathcal{N}(0, I)$ and then getting $f^c = \mu^c + \sigma^c * \boldsymbol{\epsilon}$
      Use $f^c$ as conditional input to the LDM to denoise $\mathbf{z}_t$ into $\mathbf{z}$ and update $\mu^c, \left(\sigma^2\right)^c$
    **end for**
  **end for**
**end for**

***D. Inversion based optimization of unseen class distributions.*** (Fig. 1 (D)) Once the diffusion model is trained, we can use the unseen class distributions defined in stage (B) (denoted by $f_{unseen}$) to generate unseen class samples from the model, however, we observe that the samples generated using these distributions are more similar to the seen classes which are in the neighbourhood of the unseen classes. We hypothesize that the few conditionals $f_k^c$ corresponding to the support samples of unseen class $c$ (derived from a classifier which is trained to differentiate only amongst the *seen* classes) are not able to properly capture all the characteristics of the unseen class and hence, the unseen class distribution ends up being very close to the neighbouring seen class distributions. Hence, we propose to refine the unseen distributions using *inversion* based optimization, i.e., given a support image $\boldsymbol{s}$ from an unseen class $c$, we aim to find the conditional $f_{\boldsymbol{s}}^c$ that results in the construction of $\boldsymbol{s}$ from our frozen LDM. Since the conditional $f_{\boldsymbol{s}}^c$ is sampled from the Gaussian distribution of class $c$, we back-propagate the gradients to optimize the mean $\mu^c$ and variance $\left(\sigma^2\right)^c$ as well. Our optimization goal can therefore be defined as

$$(\mu^c)^*, \left(\left(\sigma^2\right)^c\right)^* = \underset{\mu^c, (\sigma^2)^c}{\operatorname{argmin}} \mathbb{E}_{\mathcal{E}(\boldsymbol{s}), c, \epsilon \sim \mathcal{N}(0,I), t} \left[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{z}_t, t, f_{\boldsymbol{s}}^c)\|_2^2\right] \tag{8}$$

Algorithm 2 describes the inversion based optimization of unseen class distributions.

***E. Few-shot Image generation using optimized unseen distributions.*** (Fig. 1 (E)) Once our unseen class distributions are optimized, we can generate diverse samples for each class using the LDM by sampling $\mathbf{z}_T \sim \mathcal{N}(0, I)$ and $f^c \sim \mathcal{N}\left(\mu^c, \left(\sigma^2\right)^c\right)$.

| Method | Shot | Flowers | | Animal Faces | | VGGFace | | NABirds | |
|---|---|---|---|---|---|---|---|---|---|
| | | FID(↓) | LPIPS(↑) | FID(↓) | LPIPS(↑) | FID(↓) | LPIPS(↑) | FID(↓) | LPIPS(↑) |
| DAGAN [1] | 3 | 151.21 | 0.0812 | 155.29 | 0.0892 | 128.34 | 0.0913 | 159.69 | 0.1405 |
| | 1 | *179.59* | *0.0496* | *185.54* | *0.0687* | *134.28* | *0.0608* | *183.57* | *0.0967* |
| MatchingGAN [19] | 3 | 143.35 | 0.1627 | 148.52 | 0.1514 | 118.62 | 0.1695 | 142.52 | 0.1915 |
| F2GAN [21] | 3 | 120.48 | 0.2172 | 117.74 | 0.1831 | 109.16 | 0.2125 | 126.15 | 0.2015 |
| LoFGAN [14] | 3 | 112.55 | 0.2687 | 116.45 | 0.1756 | 106.24 | 0.2096 | 124.56 | 0.2041 |
| DeltaGAN [20] | 3 | 104.62 | 0.4281 | 87.04 | 0.4642 | 78.35 | **0.3487** | 95.97 | 0.5136 |
| | 1 | *109.78* | ***0.3912*** | *89.81* | ***0.4418*** | *80.12* | ***0.3146*** | *96.79* | ***0.5069*** |
| LSO [47] | 3 | **47.34** | 0.3805 | 43.29 | 0.4446 | **4.77** | 0.2835 | **21.67** | 0.5347 |
| | 1 | ***55.79*** | *0.2721* | ***64.35*** | *0.2230* | ***5.88*** | *0.1650* | ***25.23*** | *0.3318* |
| FSDM [12] | 3 | 63.87 | **0.5219** | 74.94 | **0.6309** | - | - | 73.95 | **0.7359** |
| CDM (Ours) | 3 | 77.26 | 0.4034 | **40.04** | 0.4459 | 12.77 | 0.2029 | 41.48 | 0.5406 |
| | 1 | *83.97* | *0.3728* | *64.88* | *0.3077* | *12.61* | *0.1240* | *45.53* | *0.4958* |
| F2DGAN* [48] | 3 | 38.26 | 0.4325 | 25.24 | 0.5463 | 4.25 | 0.3521 | - | - |
| CDM*(Ours) | 3 | - | - | 36.56 | 0.5224 | - | - | - | - |

**Table 1:** FID (↓) and LPIPS (↑) of images generated by different methods for unseen categories on four datasets in 1 and 3-shot setting. The best results for each shot are in bold and the second best results have been underlined. 1-shot results are displayed in italics. We quote the results of the baseline methods from the DeltaGAN paper [20]. The LSO [47] and FSDM [12] results are obtained by running their official code using our setting. * refers to evaluations using the F2DGAN setting. CDM is able to maintain a good balance between sample diversity and generation quality for all the datasets.

## 4 Experiments

### 4.1 Evaluation Setup

For our evaluation, we follow the setup in recent state-of-the-art approach [20]. Given a dataset $\mathbb{D} = (\mathbf{x}_j, y), y \in [1, \mathbb{C}], j \in [1, n_y]$ with $\mathbb{C}$ classes and $n_y$ images in each class, we split it into $\mathbb{C}_u$ unseen and $\mathbb{C}_s$ seen categories. After training on $\mathbb{C}_s$, at test time for $K$-shot generation, the model uses $K$ images and generates 128 fake images of each unseen class giving a set $S_{fake}$ of $128|\mathbb{C}_u|$ fake images. These fake images are generated by considering 2 nearest classes corresponding to each image for conditional space modelling (stage (B) in Sec. 3) and using 25 steps of DDIM sampling [37] in the diffusion model. The remaining $n_y - K$ images are combined from each unseen category to obtain a set $S_{real}$ of real images. For quantitative comparisons, we calculate the Fréchet Inception Distance (FID) [16] between $S_{real}$ and $S_{fake}$ and the Learned Perceptual Image Patch

Similarity (LPIPS) [45] for $S_{fake}$. FID score measures the distance between the latent feature distributions of real and generated unseen images, indicating both fidelity and diversity. The latent space used to calculate the distance is from an Inception-V3 [38] architecture pretrained on the Imagenet dataset [6]. LPIPS indicates the diversity in the generated images by calculating the distance between all the same class image pairs and then averaging it for each class followed by measuring the average over all the unseen classes. We use the following four datasets in order to compare our method with the existing approaches:
**Flowers** [29] dataset has total 102 categories with number of images in each class varying from 40 to 258. We split it into 85 seen and 17 unseen classes resulting in 7121 seen and 680 unseen images. **Animal Faces** [6] dataset contains 149 classes. We split it into 119 seen categories having total 96621 images and 30 unseen categories with 3000 images. In **VGGFace** [3] dataset, we select 2299 classes and choose 100 images for each class. 1802 classes out of these are used for training and the remaining 497 for evaluation. **NABirds** [40] dataset has 555 classes out of which 444 are used for training and 111 for evaluation. This gives a total of 38306 seen images and 10221 unseen images.

### 4.2 Implementation Details

We use a downsampling factor of 4 for the Variational Autoencoder (VAE), transforming $\mathbf{x} \in \mathbb{R}^{128\times128\times3}$ to $\mathbf{z} \in \mathbb{R}^{32\times32\times3}$. A codebook $\mathbb{Z}$ of size 2048 aids vector-quantization regularization. The batch size is 16 for all other datasets, except VGGFace (128), and we train the VAE for 100 epochs. For the conditional space, ResNet-50 [15] is trained with Cross-Entropy Loss and SGD optimizer over 500 epochs on the seen data, yielding a latent space of $4 \times 4 \times 512$ from the last latent convolution layer. The LDM uses 1000 timesteps with a linear noise schedule; and $\lambda = 0.001$ in the objective Eq. (6) following previous works [28]. It is trained with batch size 128 for 200 epochs, with a linear warmup schedule for the learning rate, increasing from 0 to $2 \times 10^{-4}$ after 50K steps. Training the diffusion model on a Nvidia GeForce RTX 3060 takes $\approx$ 24 hours for all other datasets, except for VGGFace, which requires 70 hours on an Nvidia RTX A6000 GPU. Inversion-based optimization runs at learning rate $2 \times 10^{-4}$ for 50K steps.

### 4.3 Quantitative Evaluation

We compare our proposed approach against several methods, including DAGAN [1], MatchingGAN [19], F2GAN [21], LoFGAN [14], DeltaGAN [20], LSO [47], FSDM [12] and F2DGAN [48] in both 3-shot and 1-shot settings. For fusion-based methods, we only report 3-shot results, as they can't be evaluated in 1-shot settings. Transformation-based methods, needing one conditional image, generate $S_{fake}$ in 3-shot setting using 3 images per episode, choosing one randomly. Results summarized in Tab. 1 on two datasets show our method performs comparably to GAN-based approaches and better than previous Diffusion based method (FSDM) in both sample quality and diversity. For the Flowers dataset, our model ranks second in both FID and LPIPS scores in 1-shot setting. While

| Method | Flowers | | AnimalFaces | |
|---|---|---|---|---|
| | 1-shot | 5-shot | 1-shot | 5-shot |
| MatchingGAN [19] | - | 74.09 | - | 70.89 |
| LoFGAN [14] | - | 75.86 | - | 73.43 |
| DeltaGAN [20] | **61.23** | 77.09 | **60.31** | **74.59** |
| LSO [47] | 57.42 | **79.41** | 32.91 | 47.01 |
| CDM (Ours) | 60.12 | 78.99 | 47.96 | 74.33 |

**Table 2:** Accuracy(%) of different methods on two datasets in few-shot classification setting (10-way 1/5-shot) averaged over 10 episodes. Results of prior methods are as per DeltaGAN [20]. The results of LSO [47] have been calculated using their official code in our setting.

| Score | ✗ Inversion | ✓ Inversion |
|---|---|---|
| FID(↓) | 85.42 | 40.04 |
| LPIPS(↑) | 0.5684 | 0.4459 |

| #seen(⟶) | 0 | 1 | 2 | 3 | 5 |
|---|---|---|---|---|---|
| FID(↓) | 42.20 | 41.89 | **40.04** | 42.79 | 43.34 |
| LPIPS(↑) | 0.4304 | 0.4370 | **0.4459** | 0.4396 | 0.4302 |

**Table 3:** **(Above)** 3-shot FID(↓) & LPIPS(↑) scores on AnimalFaces dataset, when applying CDM with and without inversion based optimization of the unseen class distributions.
**(Below)** 3-shot FID(↓) and LPIPS(↑) on AnimalFaces dataset, when utilising different number of neighbouring seen classes to calculate the unseen class distributions in CDM.

LSO has the best sample quality, its generated samples lack diversity. On the AnimalFaces dataset, we achieve the best FID in 3-shot setting. The LSO method uses a StyleGAN2-Ada model with data augmentation applied only for the discriminator, thus improving training data size without compromising fidelity. In contrast, Diffusion models reflect any augmentation applied to the training data in generated samples which compromises the sample fidelity. Hence, their effective training data size gets limited due to lack of augmentations. Consequently, LSO has better FID scores than CDM due to superior augmentation, but CDM samples are more diverse and better amalgamate input support samples, unlike LSO, which shows lower LPIPS scores. DeltaGAN samples, while varied, don't align well with unseen class distribution, resulting in worse FID scores than CDM.

**Few-Shot Classification** To further evaluate the representativeness of the generated samples, we present the results of Few-shot classification performed on a held-out test set comprising of unseen classes. For $N$-way $C$-shot classification, we randomly choose $N$ unseen classes and use $C$ images of each class to generate 512 fake images. A ResNet18 pretrained on seen data is used as a feature extractor and the final linear layer is trained using $N \times (C + 512)$ images. The rest of the unseen images are used for testing. As Tab. 2 shows, the models trained using CDM-generated samples are able to classify the test set images with performance comparable to state-of-the-art GAN-based methods. Therefore, the generated samples match the characteristics of the unseen classes.

### 4.4 Qualitative Results

For qualitative comparison, we show some generated samples using LSO [47] in 3-shot setting in Fig. 2 on Flowers, VGGFace and NABirds datasets. For all the datasets, the images generated by CDM are comparable to the ones produced by the state-of-the-art LSO in terms of quality as well as diversity. We observe changes in pose/orientation and colour transfers when compared to the conditional (support) images for both CDM and LSO. Rather, for VGGFace dataset, we can observe a doppelganger kind of effect on the generated samples for LSO, i.e., the generated faces don't necessarily have the same identity as the input faces (evident in rows 4 and 6). This is not the case with CDM, where the gen-

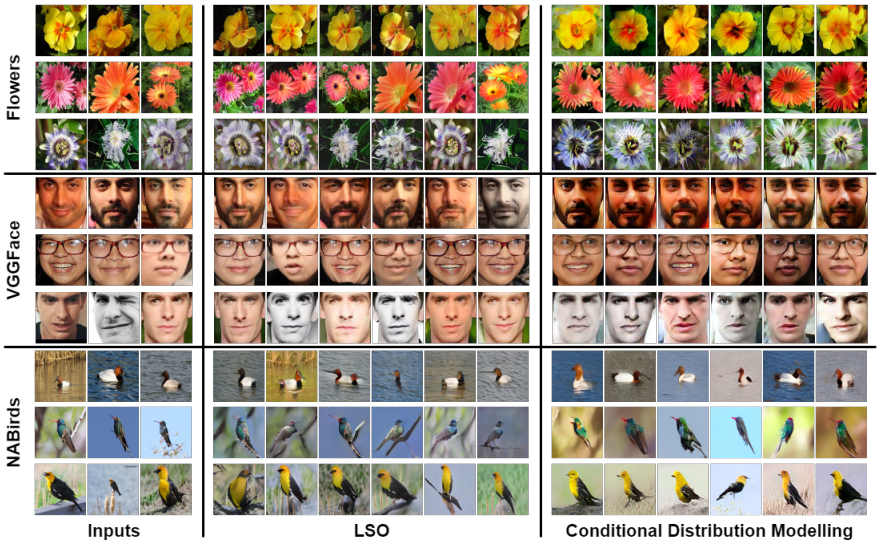


**Fig. 2:** Comparison based on images generated by LSO [47] and our Conditional Distribution Modelling in 3-shot setting on Flowers, VGGFace and NABirds datasets. The conditional images are in the input columns.

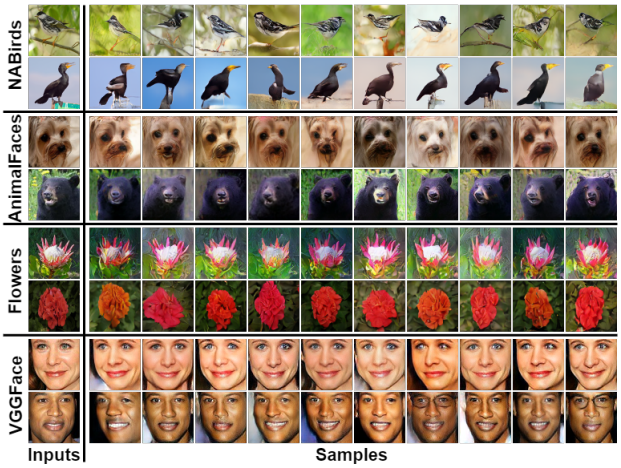


**Fig. 3:** Samples generated in 1-shot setting using CDM on four datasets. The conditional images are in the input column.

erated faces are considerably diverse, yet preserve the input identity. The same effect is visible in the NABirds dataset samples too, in row 8 (column 5) and row 9 (column 4), where the generated bird appears to be from a different (yet considerably similar) category than the input images, which does not happen with CDM. Perhaps the CDM samples lag in terms of FID scores, primarily because of poorer background generation as compared to LSO, and not due to the quality of birds themselves. We also compare CDM samples with LoFGAN

samples in supplemetary.
In Fig. 3, we show some samples generated in 1-shot setting on the NABirds, AnimalFaces, Flowers and VGGFace datasets using our approach. The generated samples display a wide range of poses/orientations/expressions, while maintaining the definitive characteristics of the input sample for all four datasets.

### 4.5 Ablation Experiments

**Varying the number of neighboring seen classes to calculate the initial unseen class distributions affects sample quality.** As shown in Tab. 3, the optimal number for the AnimalFaces dataset is 2, leading to the best generation quality. Using more than 2 seen classes includes dissimilar classes, disturbing distribution variance and resulting in suboptimal FID and LPIPS scores. Similarly, using fewer than 2 classes also yields suboptimal results.
To demonstrate the **importance of inversion-based optimization** for unseen class distributions, we evaluate FID and LPIPS scores for samples generated with and without inversion on the AnimalFaces dataset in 3-shot setting (Tab. 3). Results show significant improvements in image quality with inversion, though there is a slight decrease in sample diversity. This enhancement is due to improved fidelity, avoiding similar samples from neighboring seen classes, as shown in Fig. 4 (row 3). For a specific fox species, CDM without inversion produces a different (seen) species, which is corrected after applying inversion.

## 5 Conclusion and Future Directions

In this work, we have proposed *Conditional Distribution Modelling (CDM)* - a framework that successfully employs Diffusion models for few-shot image synthesis on large scale fine-grained datasets and achieves state-of-the-art results. We show how taking advantage of the neighbouring seen class statistics at test time can greatly benefit the image generation diversity. As a future direction, we aim to extend our approach to few-shot image generation on coarse grained datasets.

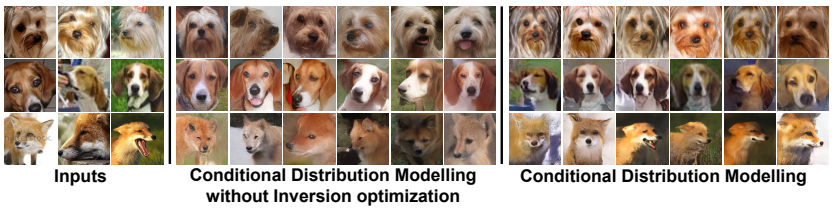


**Fig. 4:** We show the images generated by CDM without and with inversion based optimization for comparison on the AnimalFaces Dataset in 3-shot setting. Inversion optimization improves the fidelity of the generated samples to the unseen class.

## References


1. Antoniou, A., Storkey, A., Edwards, H.: Data augmentation generative adversarial networks (2017). https://doi.org/10.48550/ARXIV.1711.04340, https://arxiv.org/abs/1711.04340
2. Bartunov, S., Vetrov, D.: Few-shot generative modelling with generative matching networks. In: Storkey, A., Perez-Cruz, F. (eds.) Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics. Proceedings of Machine Learning Research, vol. 84, pp. 670–678. PMLR (09–11 Apr 2018), https://proceedings.mlr.press/v84/bartunov18a.html
3. Cao, Q., Shen, L., Xie, W., Parkhi, O.M., Zisserman, A.: Vggface2: A dataset for recognising faces across pose and age. In: 2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018). pp. 67–74. IEEE (2018)
4. Choi, J., Kim, S., Jeong, Y., Gwon, Y., Yoon, S.: Ilvr: Conditioning method for denoising diffusion probabilistic models. arXiv preprint arXiv:2108.02938 (2021)
5. Clouâtre, L., Demers, M.: Figr: Few-shot image generation with reptile (2019). https://doi.org/10.48550/ARXIV.1901.02199, https://arxiv.org/abs/1901.02199
6. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
7. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. CoRR **abs/2105.05233** (2021), https://arxiv.org/abs/2105.05233
8. Ding, G., Han, X., Wang, S., Wu, S., Jin, X., Tu, D., Huang, Q.: Attribute group editing for reliable few-shot image generation. In: CVPR (2022)
9. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. CoRR **abs/2010.11929** (2020), https://arxiv.org/abs/2010.11929
10. Finn, C., Abbeel, P., Levine, S.: Model-agnostic meta-learning for fast adaptation of deep networks. CoRR **abs/1703.03400** (2017), http://arxiv.org/abs/1703.03400
11. 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 (2022). https://doi.org/10.48550/ARXIV.2208.01618, https://arxiv.org/abs/2208.01618
12. Giannone, G., Nielsen, D., Winther, O.: Few-shot diffusion models (2022). https://doi.org/10.48550/ARXIV.2205.15463, https://arxiv.org/abs/2205.15463
13. Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks (2014). https://doi.org/10.48550/ARXIV.1406.2661, https://arxiv.org/abs/1406.2661
14. Gu, Z., Li, W., Huo, J., Wang, L., Gao, Y.: Lofgan: Fusing local representations for few-shot image generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8463–8471 (2021)
15. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. CoRR **abs/1512.03385** (2015), http://arxiv.org/abs/1512.03385
16. 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. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. p. 6629–6640. NIPS'17, Curran Associates Inc., Red Hook, NY, USA (2017)

17. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. CoRR **abs/2006.11239** (2020), https://arxiv.org/abs/2006.11239
18. Ho, J., Salimans, T.: Classifier-free diffusion guidance (2022). https://doi.org/10.48550/ARXIV.2207.12598, https://arxiv.org/abs/2207.12598
19. Hong, Y., Niu, L., Zhang, J., Zhang, L.: Matchinggan: Matching-based few-shot image generation. CoRR **abs/2003.03497** (2020), https://arxiv.org/abs/2003.03497
20. Hong, Y., Niu, L., Zhang, J., Zhang, L.: Deltagan: Towards diverse few-shot image generation with sample-specific delta. In: ECCV (2022)
21. Hong, Y., Niu, L., Zhang, J., Zhao, W., Fu, C., Zhang, L.: F2GAN: fusing-and-filling GAN for few-shot image generation. CoRR **abs/2008.01999** (2020), https://arxiv.org/abs/2008.01999
22. 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)
23. Kingma, D.P., Welling, M.: Auto-encoding variational bayes (2013). https://doi.org/10.48550/ARXIV.1312.6114, https://arxiv.org/abs/1312.6114
24. Liang, W., Liu, Z., Liu, C.: DAWSON: A domain adaptive few shot generation framework. CoRR **abs/2001.00576** (2020), http://arxiv.org/abs/2001.00576
25. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. CoRR **abs/1411.7766** (2014), http://arxiv.org/abs/1411.7766
26. Liu, Z., Miao, Z., Zhan, X., Wang, J., Gong, B., Yu, S.X.: Large-scale long-tailed recognition in an open world. CoRR **abs/1904.05160** (2019), http://arxiv.org/abs/1904.05160
27. Nichol, A., Achiam, J., Schulman, J.: On first-order meta-learning algorithms. CoRR **abs/1803.02999** (2018), http://arxiv.org/abs/1803.02999
28. Nichol, A., Dhariwal, P.: Improved denoising diffusion probabilistic models. CoRR **abs/2102.09672** (2021), https://arxiv.org/abs/2102.09672
29. Nilsback, M.E., Zisserman, A.: Automated flower classification over a large number of classes. In: 2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing. pp. 722–729. IEEE (2008)
30. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents (2022). https://doi.org/10.48550/ARXIV.2204.06125, https://arxiv.org/abs/2204.06125
31. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models (2021)
32. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. pp. 234–241. Springer (2015)
33. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation (2022)
34. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., Ghasemipour, S.K.S., Ayan, B.K., Mahdavi, S.S., Lopes, R.G., Salimans, T., Ho, J., Fleet, D.J., Norouzi, M.: Photorealistic text-to-image diffusion models with deep language understanding (2022). https://doi.org/10.48550/ARXIV.2205.11487, https://arxiv.org/abs/2205.11487
35. Saharia, C., Ho, J., Chan, W., Salimans, T., Fleet, D.J., Norouzi, M.: Image super-resolution via iterative refinement (2021). https://doi.org/10.48550/ARXIV.2104.07636, https://arxiv.org/abs/2104.07636

36. Sinha, A., Song, J., Meng, C., Ermon, S.: D2C: diffusion-denoising models for few-shot conditional generation. CoRR **abs/2106.06819** (2021), https://arxiv.org/abs/2106.06819
37. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. CoRR **abs/2010.02502** (2020), https://arxiv.org/abs/2010.02502
38. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: Computer Vision and Pattern Recognition (CVPR) (2015), http://arxiv.org/abs/1409.4842
39. Van Den Oord, A., Vinyals, O., et al.: Neural discrete representation learning. Advances in neural information processing systems **30** (2017)
40. Van Horn, G., Branson, S., Farrell, R., Haber, S., Barry, J., Ipeirotis, P., Perona, P., Belongie, S.: Building a bird recognition app and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 595–604 (2015)
41. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
42. Vinyals, O., Blundell, C., Lillicrap, T.P., Kavukcuoglu, K., Wierstra, D.: Matching networks for one shot learning. CoRR **abs/1606.04080** (2016), http://arxiv.org/abs/1606.04080
43. Yang, M., Wang, Z., Chi, Z., Feng, W.: Wavegan: Frequency-aware gan for high-fidelity few-shot image generation. In: European Conference on Computer Vision. pp. 1–17. Springer (2022)
44. Yang, S., Liu, L., Xu, M.: Free lunch for few-shot learning: Distribution calibration. arXiv preprint arXiv:2101.06395 (2021)
45. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 586–595. IEEE Computer Society, Los Alamitos, CA, USA (jun 2018). https://doi.org/10.1109/CVPR.2018.00068, https://doi.ieeecomputersociety.org/10.1109/CVPR.2018.00068
46. Zhao, S., Liu, Z., Lin, J., Zhu, J.Y., Han, S.: Differentiable augmentation for data-efficient gan training. Advances in Neural Information Processing Systems **33**, 7559–7570 (2020)
47. Zheng, C., Liu, B., Zhang, H., Xu, X., He, S.: Where is my spot? few-shot image generation via latent subspace optimization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3272–3281 (2023)
48. Zhou, Y., Ye, Y., Zhang, P., Wei, X., Chen, M.: Exact fusion via feature distribution matching for few-shot image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8383–8392 (June 2024)
49. Zhu, J., Ma, H., Chen, J., Yuan, J.: Few-shot image generation with diffusion models. arXiv preprint arXiv:2211.03264 (2022)