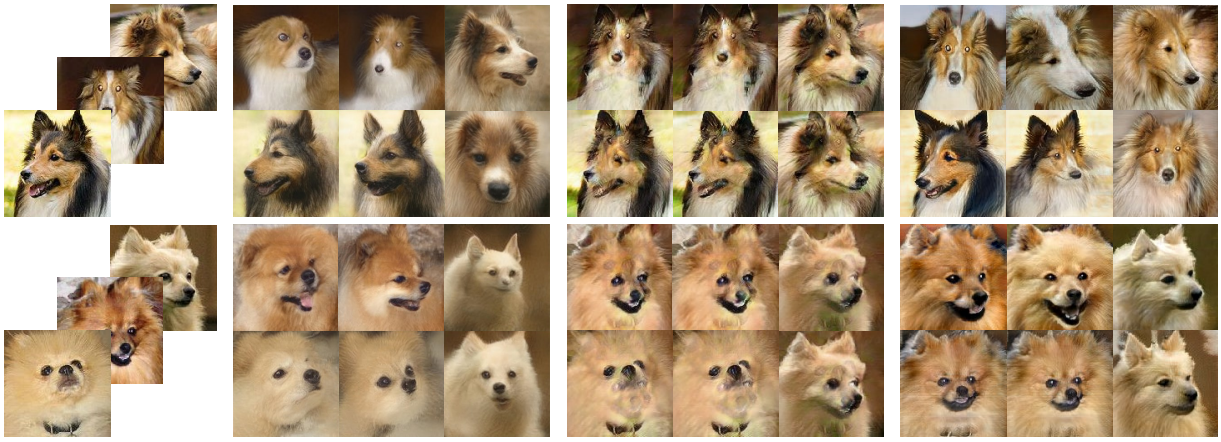
Where is My Spot? Few-shot Image Generation via Latent Subspace Optimization

Chenxi Zheng [1*] Bangzhen Liu [1*] Huaidong Zhang [1†] Xuemiao Xu[1,2,3,4†] Shengfeng He [5]
[1]South China University of Technology [2]State Key Laboratory of Subtropical Building Science
[3]Ministry of Education Key Laboratory of Big Data and Intelligent Robot
[4]Guangdong Provincial Key Lab of Computational Intelligence and Cyberspace Information
[5]Singapore Management University
{cszcx, cs_liubz}@mail.scut.edu.cn, {huaidongz, xuemx}@scut.edu.cn, shengfenghe@smu.edu.sg



(a) 3-shot Input (b) AGE [7] (c) WaveGAN [44] (d) Ours

Figure 1. We propose a new image synthesis approach that allows generating diverse unseen results with only 1- or 3-shot samples. Our key idea is to exploit the continuity of the StyleGAN latent space, and further empower it to generate unseen categories via latent subspace optimization.

Abstract

Image generation relies on massive training data that can hardly produce diverse images of an unseen category according to a few examples. In this paper, we address this dilemma by projecting sparse few-shot samples into a continuous latent space that can potentially generate infinite unseen samples. The rationale behind is that we aim to locate a centroid latent position in a conditional StyleGAN, where the corresponding output image on that centroid can maximize the similarity with the given samples. Although the given samples are unseen for the conditional StyleGAN, we assume the neighboring latent subspace around the centroid belongs to the novel category, and therefore introduce two latent subspace optimization objectives. In the first one we use few-shot samples as positive anchors of the novel class, and adjust the StyleGAN to produce the corresponding results with the new class label condition. The second objective is to govern the generation process from the other way around, by altering the centroid and its surrounding latent subspace for a more precise generation of the novel class. These reciprocal optimization objectives inject a novel class into the StyleGAN latent subspace, and therefore new unseen samples can be easily produced by sampling images from it. Extensive experiments demonstrate superior few-shot generation performances compared with state-of-the-art methods, especially in terms of diversity and generation quality. Code is available at https://github.com/chansey0529/LSO.

1. Introduction

Recent advances in generative models [3, 5, 8, 11, 19, 36] allow synthesizing of high-quality and realistic images with diverse styles. However, the success of these models relies heavily on large-scale data. Preparing new data for a novel class is costly, so it is natural to raise a question, "can we generate high-quality images with a glance at a few images?" This leads to the few-shot image generation problem, where the model is required to generate a novel category with only a few images available. Unfortunately, since the extreme low-shot setting can easily cause catastrophic

*Equal Contributions.
†Corresponding authors.

over-fitting, few-shot image generation is still challenging.

Existing methods commonly suppose that the seen models have implicit generalization ability towards unseen categories. Based on this assumption, task-specific optimization is adopted to seek proper initial parameters, which better generalize to the downstream tasks [6, 25]. Testing phase generation is another solution, which skips integrating the information of unseen category into model weights. Nevertheless, the generated images are either with a lot of class-specific information distortion [7] or fail to restore the detailed features, such as textures [10, 44]. The main assumption of this line of research in model generalization ability is false, and therefore the model trained on seen data cannot extract out-of-domain unseen-specific features without adaptation, *e.g.*, generating a spotted dog via glancing on a golden retriever, which significantly limits their practical usage in real-world scenarios. As a consequence, a key factor to the success of few-shot synthesis is to expose the samples of unseen classes to the model.

One of the major obstacles is the sparsity of the unseen samples. Traditional generative networks require modeling the continuous distribution for generating diverse images with unseen-specific features. However, the discrete data points under the few-shot setting make the model ill-informed about the inner structure of the unseen distribution. On the other hand, the pretrained latent spaces of Style-series models [17–19, 43] are shown to be semantically interpretable and continuous. This property ideally fits our problem. Once the proper latent locations of unseen samples are found, we can complement the marginal region with the hidden semantic information and form a subspace for the unseen category. In this way, diverse unseen images can be generated via sampling from the new subspace.

Based on the above insights, we proposed a novel latent subspace optimization framework for few-shot image generation. The key idea is to search for the optimal sub-distribution of unseen using *latent anchor localization*, and then align the sub-distribution with the input unseen distribution using *latent subspace refinement*. To obtain an unseen correlated semantic region in the latent space, we first locate the subspace of the unseen category by faithful anchor optimization. Specifically, the latent codes of the unseen category are served as reliable latent subspace indicators by inverting the available unseen images into the latent space. Based on these anchors, the coarse centroid of the unseen distribution is pulled to the hypothetical point using a subspace localization loss.

Subsequently, due to the semantic deficiency of few-shot images, distributional shift exists between the resulting distribution of our subspace and the real unseen distribution. To mitigate semantic misalignment, we propose to refine the latent subspace of unseens. We employ an adversarial training scheme to inject the unseen correlated features into the generator. However, the guidance of the adversarial game easily leads to over-emphasis on transferring the low-level features, ignoring the learning of unseen semantics (*e.g.*, fails to generate a wolf but a wolf-like dog). Thus, the generated images may belong to a completely different semantic category, though they contain similar textures with the few-shot examples. To preserve the unseen-specific semantic, we further restrict the latent subspace by a semantic stabilization loss. Once the StyleGAN and its subspace are properly optimized, our framework is able to generate diverse and high-quality unseen images. We compare to state-of-the-art methods extensively on different datasets, and we show significant superiority over them.

In summary, the contribution of this paper is fourfold:

- We delve into few-shot image generation from a novel perspective of exploring the continuity of the latent space for discovering unseen category.
- We propose a novel latent subspace optimization framework to model the distribution of unseen samples, while injecting category-specific features into the generated images.
- Experimental results show that our approach achieves state-of-the-art performances on three datasets, largely reducing the FID scores by 7.58, 4.37, and 0.98 on Flowers, AnimalFaces, and VGGFaces respectively while gaining diversity on most datasets.
- We extend our model to other subfields like image editing and high-resolution image generation with few-shot setting. Additionally, we explore the potential of our framework in few-shot incremental generation.

2. Related Work

Few-shot Image Generation. Few-shot Image Generation aims to generate diverse and high-quality images given a few novel samples. Prevailing methods are summarized as optimization-based, fusion-based, and transformation-based. Optimization-based methods [6, 9, 30] adopt meta-learning to search a set of initial parameters that generalize well for different tasks. Fusion-based methods [2, 10, 12, 14, 44] learn a k-shot fusion strategy for unseen features, while transformation-based methods [1, 7, 13] preserve the unseen-specific features via intra-category transformations. However, the generated images suffer from unseen feature diminishing. WaveGAN [44] adapts Haar wavelet transform to capture high-frequency features to solve this problem. Yet, even without the high-frequency details, WaveGAN still obtains rich details on generating seen samples, which convinces us that the crux of feature diminishing lies in the semantic gap between the seen and the unseen.

Conditional GANs. Broadly defined conditional Generative Adversarial Networks (cGANs) refers to the type

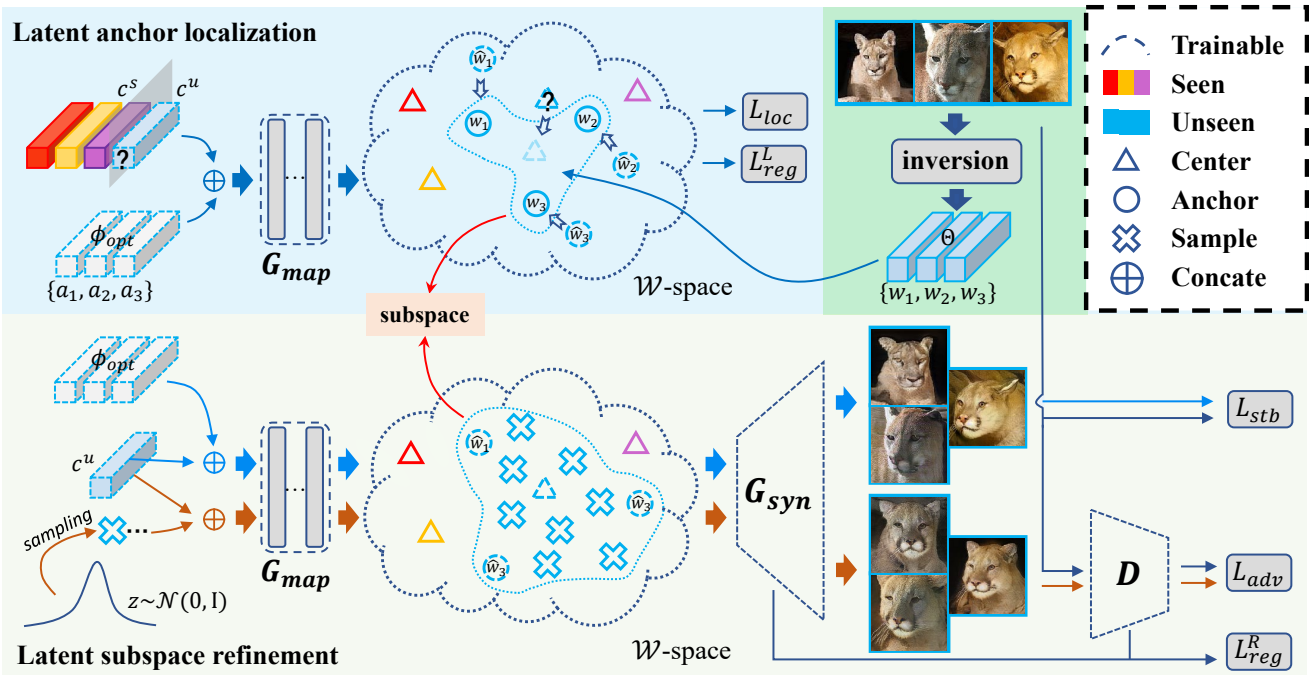


Figure 2. The overview of the proposed two-stage latent subspace optimization framework. In this figure, we take the 3-shot image generation as an example. In stage one, the unseen subspace is localized by optimizing the anchor position of each unseen sample in $\mathcal{W}$-space. In stage two, adversarial loss and semantic stabilization loss work together to refine the latent subspace of the unseen category.

of GANs that use conditional information like image [15], text [45], and audio [39]. In this paper, we focus on cGANs that are conditioned on a category label [24, 27, 28, 33, 34, 41]. This type of cGANs commonly requires combining a category vector and a noise vector as the input of the generator. Several techniques are investigated by previous works to enhance categorical restriction of the conditional discriminator, including input concatenation [27], hidden concatenation [34], label prediction [33], and projection head [3, 16, 28]. In this way, cGANs implicitly disentangle the category-relevant and category-irrelevant information, which fits well with the purpose of few-shot image generation. By incorporating the class-irrelevant information learned by the noise branch, the pretrained cGAN generator can be effectively optimized for an unseen category, thus solving the sparsity issues under the few-shot setting.

Few-shot Generative Domain Adaptation. Few-shot generative domain adaptation [21, 40, 47, 48] aims at transferring a source pretrained model to a specific target domain, with limited target samples. A common approach is finetuning the pretrained model with the few-shot data. However, overfitting happens due to the sparsity of samples, leading to severe mode collapse. Prevailing works apply data augmentation [16], partial parameters freezing [29,32], or contrastive learning [48] to maintain the diversity of generation. In contrast to the above adaption methods, our goal is to generate images for a novel category, which emphasizes the importance of extracting category-specific information rather than adapting the entire feature space.

3. Method

3.1. Preliminaries

3.1.1 Few-shot Image Generation

The goal of few-shot image generation is to generate realistic and diverse images with a few template images. The samples of a given dataset are separated into two subsets, the seens $\mathbb{C}^s$ and the unseens $\mathbb{C}^u$. In the training stage, sufficient images from $\mathbb{C}^s$ are used to train the generator G. During testing, image generation is completed by further optimizing the model with k images of the same categories sampled from $\mathbb{C}^u$.

3.1.2 Structure of Conditional StyleGAN [16]

The generator $G : \mathcal{Z}, \mathcal{Y} \rightarrow \mathcal{X}$ takes a noise $z \in \mathcal{Z}$ and a one-hot label $y \in \mathcal{Y}$ as input to generate an image $\hat{x} \in \mathcal{X}$, where $\mathcal{Z}$, $\mathcal{Y}$, $\mathcal{X}$ represents the noise space, the label space, and the image space respectively.

Concretely, $\mathcal{Z}$ and $\mathcal{Y}$ are first projected to the category-specific feature space $\mathcal{C}$ and the latent space $\mathcal{W}$ by an embedding network $G_{embed} : \mathcal{Y} \rightarrow \mathcal{C}$ and a fully-connected network $G_{map} : \mathcal{Z}, \mathcal{C} \rightarrow \mathcal{W}$ respectively. G_{embed} transforms the sparse label y to a dense category centroid $c \in \mathcal{C}$, which is later mapped as the latent code $w \in \mathcal{W}$ together with the noise z through G_{map}. Afterward, a synthesis network $G_{syn} : \mathcal{W} \rightarrow \mathcal{X}$ decodes w to an image $\hat{x}$. The whole

pipeline can be formulated as:

$$\begin{aligned} c =& G_{embed}(y), \\ w =& G_{map}(z, c), \\ \hat{x} =& G_{syn}(w). \end{aligned} \tag{1}$$

Training a novel category with a pretrained conditional StyleGAN would have required extra parameters for G_{embed}. Nonetheless, we only need a single dense vector c under the few-shot scheme. Thus, we simplify the training of G_{embed} to the procedure of directly optimizing the vector c. In the following sections, we denote the centroid of the unseen category as c^u to represent the category vector in $\mathcal{C}$.

3.2. Latent Subspace Optimization

Under the few-shot setting, the major challenge for image generation comes from the sparsity of the target image space $\mathcal{X}$. The generator either memorizes the patterns of the samples or suffers from severe mode collapse. We attribute the failure of the adversarial training to the model's poor capability to represent the intrinsic features of the unseen categories. The absence of such capability significantly aggravates the few-shot fitting procedure, which requires the generator to capture the structural information and simultaneously extract specific features for the unseen category.

To overcome the above challenges, we propose a novel latent subspace optimization framework. Initially, we optimize the category centroid with anchors to acquire the coarse latent subspace. Latent codes in the coarse subspace are capable of generating structural characteristics but fail to produce unseen-specific features in most cases. Subsequently, we refine the latent space with two objectives for more consistent interaction between the subspace and the synthesis network. For the first one, an adversarial loss is utilized to align the distribution of fake samples and the real unseen samples. For the second, a semantic stabilization loss cooperatively enhances unseen semantics from the view of trainable anchors. In addition, we further apply a regularization loss to maintain the continuity and semantic disentanglement of the latent space. The overview of our method is shown in Fig. 2.

3.2.1 Latent Anchor Localization

Let the conditional StyleGAN pretrained with seen images be G^s, we invert images from a given unseen category $X = \{x_i\}_{i=1}^k$ to the latent space $\mathcal{W}$ of G^s by inversion $I(\cdot,\cdot)$. The set of latent codes is denoted as $\Theta = \{w_i\}_{i=1}^k, \Theta \subset \mathcal{W}$, where $w_i = I(G^s, x_i)$. We also define a set of trainable noise anchors $\Phi_{opt} = \{a_i\}_{i=1}^k \subset \mathcal{Z}$.

To locate the approximate unseen region, we first seek a rough centroid of the unseen category. The procedure can be modeled as generating k latent codes $\{\hat{w}_i\}_{i=1}^k$ with the centroid c^u and the trainable anchor set Φ_{opt}, *i.e.*, $\hat{w}_i = G_{map}(a_i, c^u)$, to approximate the target anchors $w_i \in \Theta$. By jointly optimizing the noise anchor set Φ_{opt} and the category centroid c^u with the approximation loss:

$$\mathcal{L}_{app} = \frac{1}{k}\sum_{i}^{k}(||\hat{w}_i - w_i||_2), \tag{2}$$

we can obtain a subspace defined by c^u. To force the mapping network G_{map} focusing on finding the centroid, we additionally regularize the magnitude of the trainable noise anchors a_i by:

$$\mathcal{L}_{mgt} = \frac{1}{k}\sum_{i}^{k}(||a_i||_2), \tag{3}$$

which avoids the overfitting of anchors during the approximation. We further allow the mapping network G_{map} to be mildly updated to handle the outliers. The overall objective of latent subspace localization is formulated as:

$$\mathcal{L}_{loc} = \mathcal{L}_{app} + \mathcal{L}_{mgt}. \tag{4}$$

The centroid c^u is hauled to the optimal position and forms a subspace so that the latent codes within the subspace can be generated by feeding randomly sampled noises $z \sim \mathcal{N}(0, I)$ and the centroid c^u to the mapping network.

3.2.2 Latent Subspace Refinement

The coarse subspace enables us to generate unseen latent codes with fast noise sampling. However, the subspace is incapable of generating samples with unseen-specific features. The main reason is that the learned unseen centroid locates in the seen domain, which may be beyond the distribution of the real unseen distribution, yet there is no supervision to solve this problem.

To tackle the distributional shift, we propose to align the subspace distribution with the real one via an adversarial game. As the generator progressively refines the original distribution to adapt to the unseen distribution, the generation quality of the subspace is improved. The conditional adversarial loss [27] is formulated as:

$$\begin{aligned} \mathcal{L}_{adv} =& \mathbb{E}_{x\sim P_X}[logD^u(x, y^u)] \\ &+\mathbb{E}_{z\sim P_Z}[log(1 - D^u(G^u(z, c^u), y^u))], \end{aligned} \tag{5}$$

where y^u, P_X and P_Z denote the unseen label, unseen image set, and the normal Gaussian distribution respectively. G^u and D^u denote the conditional generator and the discriminator refined with unseen samples.

Under the guidance of the discriminator, the generator is instructed to learn the features of an unseen category, allowing the subspace to generate samples that conform to the

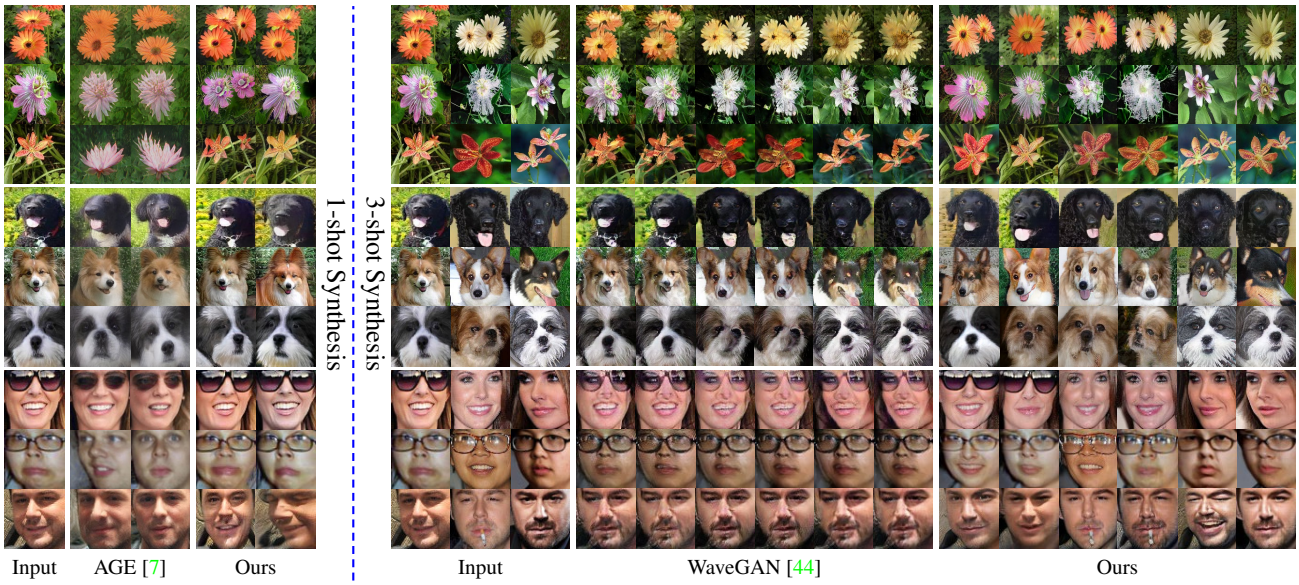


Figure 3. Comparison with state-of-the-art methods under 1-shot and 3-shot settings. The left most of each part are the input few-shot images. The results of our method preserve more category-specific features, while achieving more stable and diverse generation.

real unseen distribution. However, the adversarial supervision may lead to an excessive focus on low-level features, such as colors and fur textures, at the expense of preserving the correct unseen semantics.

To address this issue and avoid the diminishing of unseen semantics during the adversarial game, we introduce a semantic stabilization loss that leverages the correspondence between the trainable anchors $a_i \in \Phi_{opt}$ and the real unseen images $x_i \in X$. Specifically, the proposed loss is composed of a perceptual loss and a magnitude regularization:

$$\mathcal{L}_{stb} = \mathcal{L}_{perc} + \mathcal{L}_{mgt}. \tag{6}$$

To further improve the semantic consistency of the produced image $\hat{x}_i$ and the unseen image x_i, drawing inspiration from [38], we introduce a similarity loss to complement the vanilla perceptual loss [46]. More precisely, we minimize the cosine similarity between the extracted features of inputs. The final perceptual loss can be formulated as:

$$\mathcal{L}_{perc} = \mathcal{L}_{lpips}(\hat{x}_i, x_i) + ||\hat{x}_i - x_i||_2 + \mathcal{L}_{sim}(\hat{x}_i, x_i), \tag{7}$$

where $\hat{x}_i$ represents the image produced by the anchor a_i and unseen centroid c^u, *i.e.*, $\hat{x}_i = G^u(a_i, c^u)$. The magnitude regularization term is defined in Eq. 3. During the optimization, the centroid c^u and the set of noise anchors Φ_{opt} are jointly optimized with the generator.

By jointly optimizing the adversarial loss and semantic stabilization loss, we refine the latent subspace of the unseen category and elevate the generation capability of the latent codes. The overall objective in latent subspace refinement can be formulated as:

$$\mathcal{L}_{ref} = \mathcal{L}_{stb} + \mathcal{L}_{adv}. \tag{8}$$

Table 1. Category Split: the split of seen categories $\mathbb{C}^s$ and unseen categories $\mathbb{C}^u$; Image Split in $\mathbb{C}^u$: the split of the subset for image generation $\mathbb{S}^c_{gen}$ and the subset for metrics evaluation $\mathbb{S}^c_{eval}$ within an unseen category.

Datasets	**Category Split**			**Image Split in $\mathbb{C}^u$**		
	Total	$\mathbb{C}^s$	$\mathbb{C}^u$	**Total**	$\mathbb{S}^c_{gen}$	$\mathbb{S}^c_{eval}$
Flowers	102	85	17	40	10	30
AnimalFaces	149	119	30	100	10	90
VggFaces	2354	1802	552	100	30	70

3.2.3 Regularization with Seen Semantics

One of the premises for latent subspace optimization is to maintain semantic-meaningful directions within the latent space. We expect the current latent space not only fits the unseen distribution but maintains the generation ability toward seen classes. Inspired by [37], we propose to leverage the pretrained latent space to restrict the optimization of the unseen subspace. Specifically, we employ the rich-semantic latent space of seen pretrained StyleGAN as the teacher to preserve semantics for the current latent space.

Suppose the StyleGAN generator and discriminator are $G^{s/u}$ and $D^{s/u}$ respectively, with s/u representing seen pretrained or unseen optimized. The parameters of G^s and D^s are frozen during the regularization. We denote the randomly sampled noise and the seen centroid as z and c^s.

For latent anchor localization, we force the generated latent codes of the optimized mapping network G^u_{map} and the pretrained G^s_{map} to be similar, using the same z and c^s. The regularization loss can be written as:

$$\mathcal{L}^L_{reg} = ||G^u_{map}(z, c^s) - G^s_{map}(z, c^s)||_2. \tag{9}$$

For latent subspace refinement, the regularization is ap-

Table 2. Quantitative comparison with existing competitive methods for few-shot image generation.

Methods	k-shot	Flowers		Animal Faces		VGG Faces	
		FID↓	LPIPS↑	FID↓	LPIPS↑	FID↓	LPIPS↑
DAGAN [1]	1	179.59	0.0496	185.54	0.0687	134.28	0.0608
DeltaGAN [13]	1	109.78	0.3912	89.81	0.4418	80.12	0.3146
AGE [7]	1	45.96	0.4305	28.04	**0.5575**	34.86	0.3294
Ours	1	**35.87**	**0.4338**	**27.20**	0.5382	**4.15**	**0.3834**
FIGR [6]	3	190.12	0.0634	211.54	0.0756	139.83	0.0834
DAWSON [25]	3	188.96	0.0583	208.68	0.0642	137.82	0.0769
GMN [2]	3	200.11	0.0743	220.45	0.0868	136.21	0.0902
MatchingGAN [12]	3	143.35	0.1627	148.52	0.1514	118.62	0.1695
F2GAN [14]	3	120.48	0.2172	117.74	0.1831	109.16	0.2125
LoFGAN [10]	3	79.33	0.3862	112.81	0.4964	20.31	0.2869
WaveGAN [44]	3	42.17	0.3868	30.35	0.5076	4.96	0.3255
Ours	3	**34.59**	**0.3914**	**23.67**	**0.5198**	**3.98**	**0.3344**

plied to both the generator and the discriminator. For the generator, we expect the images produced by G^s and G^u to be visually consistent and semantically correlated. To achieve this, we adopt a perceptual loss on the image pair. For the discriminator, the semantic correlation is preserved by using an L_2 regularization loss on the output features of the last convolution block of D^s_{conv} and D^u_{conv}. The regularization terms during refinement are summarized as:

$$\begin{aligned}\mathcal{L}^R_{reg} =& \mathcal{L}_{perc}(G^u(z,c^s), G^s(z,c^s)) \\ &+||D^u_{conv}(\hat{x}^s) - D^s_{conv}(\hat{x}^s)||_2,\end{aligned} \tag{10}$$

where $\hat{x}^s$ is the fake image generated by G^s.

Final Objective. In conclusion, the losses of our latent subspace optimization are grouped into two parts. For latent anchor localization, the overall loss is formulated as:

$$\mathcal{L} = \lambda_{loc}\mathcal{L}_{loc} + \lambda^L_{reg}\mathcal{L}^L_{reg}. \tag{11}$$

For latent subspace refinement, the loss are denoted as:

$$\mathcal{L} = \lambda_{ref}\mathcal{L}_{ref} + \lambda^R_{reg}\mathcal{L}^R_{reg}. \tag{12}$$

4. Experiments

4.1. Implementation Details

For latent anchor localization, the learning rate of the unseen centroid c^u and the learnable noise anchors Φ_{opt} is set to 0.05, and the parameters of the mapping network G^u_{map} are adjusted with the learning rate of 0.005. λ_{loc} and λ^L_{reg} in Eq. 11 are both set to 1.0.

For latent subspace refinement, the synthesis network G^u_{syn}, the centroid c^u and the noise anchors Φ_{opt} are refined with a learning rate of 0.0025, whereas the mapping network G^u_{map} requires a smaller rate of 0.00025. λ_{ref} and λ^R_{reg} in Eq. 12 are set to 1.0 and 0.02.

We use ADAM [20] for optimization in our experiments. During the training phase and each few-shot optimization procedure, we enable ADA [16] while disabling style mixing regularization [18] and path length regularization [19]. For more details, please refer to the supplementary.

Table 3. Ablation study of the proposed method and its three variants under 3-shot settings. LAL, LSR and Reg represent latent anchor localization, latent subspace refinement and regularization, respectively. We report the cost of time in seconds. Note that the baseline classification accuracy with only real images is 60.98%.

Conditions			Flowers			
LAL	LSR	Reg	FID↓	LPIPS↑	ACC↑	TIME↓
	✓	✓	37.94	0.4173	68.75%	126
✓		✓	53.13	0.3412	54.68%	11
✓	✓		36.22	0.3884	75.39%	116
✓	✓	✓	34.59	39.14%	75.39%	136

4.2. Datasets and Metrics

We evaluate our method on three commonly used benchmarks: **Flowers** [31], **AnimalFaces** [26] and **VggFaces** [4]. The images in the above datasets are collected with the resolution of 128×128, 128×128, and 64×64, respectively. Following [10] [44], each dataset is split into two disjoint parts: the seen categories $\mathbb{C}^s$ and the unseen categories $\mathbb{C}^u$. The images of each unseen category are further separated into two subsets $\mathbb{S}^c_{gen}$ and $\mathbb{S}^c_{eval}$ for image generation and metrics evaluation respectively. The train/test split details of the benchmarks are compared in Tab. 1.

We use FID and LPIPS as the metrics for quantitative evaluation. FID is a common metric to reveal image quality in most image generation tasks. Smaller FID indicates a higher quality of the generated images. LPIPS is widely used in the image-to-image field, like GAN inversion [35] and image translation [26], which is required to be low to maintain consistency between input and output. In contrast, we adopt LPIPS to measure the diversity of the generated unseen images in few-shot image generation.

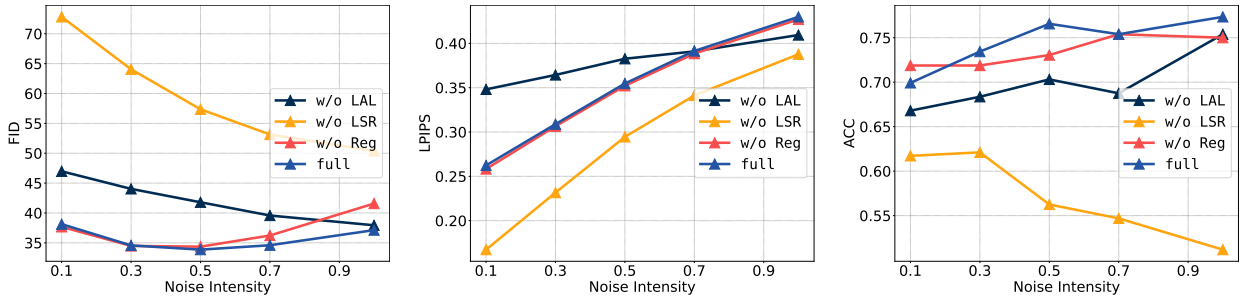


Figure 4. The statistical evaluation of each variant of our proposed methods with respect to different sampling intensities. The three plots are of FID, LPIPS, and ACC respectively from the left to the right.

4.3. Quantitative Evaluation

We randomly sample few-shot generation tasks within the subset $\mathbb{S}^c_{gen}$ and generate a total number of 128 fake images for each unseen category $c \in \mathbb{C}^u$. The produced images are denoted as the synthesis set $\mathbb{S}^c_{fake}$. We calculate the FID between the union of synthesis sets and evaluation sets over the unseen categories, *i.e.*, $\mathbb{S}_{fake} = \bigcup_{c\in\mathbb{C}^u} \mathbb{S}^c_{fake}$ and $\mathbb{S}_{eval} = \bigcup_{c\in\mathbb{C}^u} \mathbb{S}^c_{eval}$. LPIPS is calculated within each fake set $\mathbb{S}^c_{fake}$ to reflect the overall diversity.

We verify the performance of our method under both one-shot and multi-shot settings. The synthesis set under the one-shot setting is generated using only one unseen image. For the multi-shot experiment, we follow the setting in [7, 10] and select $k = 3$ for a fair comparison.

The results are reported in Tab. 2. For the one-shot setting, our method significantly outperforms all the methods on FID and achieves the highest LPIPS on 2 out of 3 benchmarks. Impressive gains superior to the second best method [7] are achieved on VGGFaces. Our method also beats all the methods on the multi-shot generation. With significant gains on FID and LPIPS, our method is qualified for stable, high-quality, and diverse image synthesis.

4.4. Qualitative Evaluation

We compare our method with AGE [7] and WaveGAN [44] under the one-shot and multi-shot settings respectively. The visualization of synthesis results on Flowers, AnimalFaces, and VGGFaces are presented in Fig. 3.

For one-shot image synthesis, the results are shown in the left part of Fig. 3, where the leftmost column is the one-shot input for each method. Compared with the generation results of AGE [7], the images generated by our method preserve more details of the unseen category, such as stamens, fur textures, and wrinkles. In addition, our method can resist the negative impact due to the category gap, *i.e.*, the fuzzy artifacts. Moreover, our method better captures unseen-specific characteristics, such as sunglasses. With the generated images in various shapes, textures, and poses, our method shows great generative capability in most cases, indicating our method's generative diversity.

Under the multi-shot setting, our method effectively improves the diversity of generated images while maintaining high-quality image synthesis. As shown in the right part of Fig. 3, different from WaveGAN [44] that fuses high-frequency features of the given images and tends to generate homogeneous images, our method can produce diverse images that have novel attributes. For instance, our method can synthesize flowers and dogs with diverse appearances and poses. Also, our method can achieve more visually pleasing results. For more examples, please refer to the supplementary material.

Figure 5. Results of image editing on the unseen image. The resulted images are of resolution 64×64. The edited attributes are eyes pose, smile, moustache, and illumination from the top to the bottom. The edit directions are from the left to the right.

4.5. Ablation Studies

The ablation studies are conducted on Flowers [31]. Latent anchor localization, latent subspace refinement, and the regularization are denoted as LAL, LSR, and Reg, respectively. We design three variants to verify the effectiveness of our framework, each without one of the proposed components. We then evaluate each variant with FID, LPIPS, accuracy gain, and the time cost of optimization to demonstrate the comprehensive performance. The accuracy gain, denoted by ACC, proves the semantic consistency between the generated images and the original few-shot category. Following [12], we estimate the gain of the classification accuracy brought by augmenting with the generated images.

The statistical results are compared in Tab. 3. We find

Figure 6. Results of high-quality image generation. On top are the 4-shot examples and the generated images of resolution 512×512 are listed below. Please zoom in for better view.

that the LAL can increase the quality of generated images, but on the contrary, limit the diversity (w/o LAL vs. full). The LSR highly contributes to the high-quality, diversity, and semantic preservation, but it is relatively time-consuming (w/o LSR vs. full). Both LAL and LSR obtain a satisfying gain on the classification task, which shows that our latent subspace is more closed to the real unseen distribution. The introduced regularization maintains the semantic disentanglement and continuity of the latent subspace, therefore improving the image quality and generation diversity (w/o Reg vs. full).

We also plot the ablation results concerning the intensity of sampling noise in Fig. 4. Generally, the intensity of noise sampling controls the variance of samples from the semantic center of the latent subspace. The higher the intensity is, the more disparity of samples we have. As shown in Fig. 4, our model dominates the performance of FID and ACC under various sampling intensities, which shows the stable generation ability of our method. The LPIPS is an exception since Eq. 3 aggregates the latent anchors to form a faithful unseen region, thus impairing the diversity. For more details, please refer to the supplementary material.

4.6. Applications

We further evaluate our method on other image-generative tasks where only extremely low samples are available, *i.e.*, image editing, high-resolution image generation, and few-shot incremental learning.

Image Editing. We manipulate the unseen images with attribute-relevant channels in the style space $\mathcal{S}$ [42] [22] and visualize the results in Fig. 5. As the figure shows, our method shows well editability on unseen images.

Figure 7. Results of few-shot incremental generation. The left most column list the incremental categories, which are provided in a sequential order from the top to the bottom. The incremental generation results are listed on the right.

High-quality Image Generation. We extend our model to high-quality portrait image generation. We adopt the widely used portrait dataset CelebA-HQ [23] following the few-shot setting and collect the images with the image resolution of 512×512. As shown in Fig. 6, our method can generate portrait images with high-resolution and well-preserved details under the 4-shot setting. For more examples, please refer to the supplementary material.

Few-shot Incremental Image Generation. We also conduct experiments with incremental unseen categories on Flowers. We sequentially feed our model with different unseen categories and optimize the model continuously. As shown in Fig. 7, our method successfully generates visual-pleasing results on the new incoming category while preserving the quality of synthesizing the previous categories. This phenomenon reveals that our method can optimize the latent subspace with multiple novel categories.

5. Conclusion

In this paper, we delve into few-shot image generation from a new perspective of latent space continuity. To obtain a proper subspace for the unseen category, we propose a novel latent subspace optimization, which can inject category-specific features into unseen generation. Quantitative and qualitative results demonstrate the robustness and superiority of our method. Our method is also extendable to image editing, high-resolution image generation, and few-shot incremental image generation. We hope that this work will contribute to the community of few-shot image generation, with practical and valuable usage.

Acknowledgements. The work is supported by Guangdong International Technology Cooperation Project(No.2022A0505050009); Key-Area Research and Development Program of Guangdong Province, China (2020B010166003, 2020B010165004); National Natural Science Foundation of China (No. 61972162); Guangdong Natural Science Foundation (No. 2021A1515012625); and Guangdong Natural Science Funds for Distinguished Young Scholar (No. 2023B1515020097).

References

[1] Antreas Antoniou, Amos Storkey, and Harrison Edwards. Data augmentation generative adversarial networks. *ICLR*, 2018. 2, 6

[2] Sergey Bartunov and Dmitry Vetrov. Few-shot generative modelling with generative matching networks. In *AISTATS*, pages 670–678, 2018. 2, 6

[3] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *ICLR*, 2019. 1, 3

[4] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *IEEE FG*, pages 67–74, 2018. 6

[5] Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J Guibas, Jonathan Tremblay, Sameh Khamis, et al. Efficient geometry-aware 3d generative adversarial networks. In *CVPR*, pages 16123–16133, 2022. 1

[6] Louis Clouâtre and Marc Demers. Figr: Few-shot image generation with reptile. *CoRR*, 2019. 2, 6

[7] Guanqi Ding, Xinzhe Han, Shuhui Wang, Shuzhe Wu, Xin Jin, Dandan Tu, and Qingming Huang. Attribute group editing for reliable few-shot image generation. In *CVPR*, pages 11194–11203, 2022. 1, 2, 5, 6, 7

[8] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *CVPR*, pages 12873–12883, 2021. 1

[9] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *ICML*, pages 1126–1135, 2017. 2

[10] Zheng Gu, Wenbin Li, Jing Huo, Lei Wang, and Yang Gao. Lofgan: Fusing local representations for few-shot image generation. In *ICCV*, pages 8463–8471, 2021. 2, 6, 7

[11] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 33:6840–6851, 2020. 1

[12] Yan Hong, Li Niu, Jianfu Zhang, and Liqing Zhang. Matchinggan: Matching-based few-shot image generation. In *ICME*, pages 1–6, 2020. 2, 6, 7

[13] Yan Hong, Li Niu, Jianfu Zhang, and Liqing Zhang. Deltagan: Towards diverse few-shot image generation with sample-specific delta. In *ECCV*, pages 259–276, 2022. 2, 6

[14] Yan Hong, Li Niu, Jianfu Zhang, Weijie Zhao, Chen Fu, and Liqing Zhang. F2gan: Fusing-and-filling gan for few-shot image generation. In *ACM MM*, pages 2535–2543, 2020. 2, 6

[15] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, pages 1125–1134, 2017. 3

[16] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. *NeurIPS*, 33:12104–12114, 2020. 3, 6

[17] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. *NeurIPS*, 34:852–863, 2021. 2

[18] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, pages 4401–4410, 2019. 2, 6

[19] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *CVPR*, pages 8110–8119, 2020. 1, 2, 6

[20] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *ICLR*, 2015. 6

[21] Chaerin Kong, Jeesoo Kim, Donghoon Han, and Nojun Kwak. Few-shot image generation with mixup-based distance learning. In *ECCV*, pages 563–580. Springer, 2022. 3

[22] Oran Lang, Yossi Gandelsman, Michal Yarom, Yoav Wald, Gal Elidan, Avinatan Hassidim, William T Freeman, Phillip Isola, Amir Globerson, Michal Irani, et al. Explaining in style: Training a gan to explain a classifier in stylespace. In *ICCV*, pages 693–702, 2021. 8

[23] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *CVPR*, 2020. 8

[24] Zhansheng Li, Yangyang Xu, Nanxuan Zhao, Yang Zhou, Yongtuo Liu, Dahua Lin, and Shengfeng He. Parsing-conditioned anime translation: A new dataset and method. *ACM TOG*, 42(3):1–14, 2023. 3

[25] Weixin Liang, Zixuan Liu, and Can Liu. Dawson: A domain adaptive few shot generation framework. *CoRR*, 2020. 2, 6

[26] Ming-Yu Liu, Xun Huang, Arun Mallya, Tero Karras, Timo Aila, Jaakko Lehtinen, and Jan Kautz. Few-shot unsupervised image-to-image translation. In *ICCV*, pages 10551–10560, 2019. 6

[27] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014. 3, 4

[28] Takeru Miyato and Masanori Koyama. cgans with projection discriminator. *arXiv preprint arXiv:1802.05637*, 2018. 3

[29] Sangwoo Mo, Minsu Cho, and Jinwoo Shin. Freeze the discriminator: a simple baseline for fine-tuning gans. *CVPRW*, 2020. 3

[30] Alex Nichol and John Schulman. Reptile: a scalable metalearning algorithm. *arXiv preprint arXiv:1803.02999*, 2:4, 2018. 2

[31] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *ICVGIP*, pages 722–729, 2008. 6, 7

[32] Atsuhiro Noguchi and Tatsuya Harada. Image generation from small datasets via batch statistics adaptation. In *ICCV*, pages 2750–2758, 2019. 3

[33] Augustus Odena, Christopher Olah, and Jonathon Shlens. Conditional image synthesis with auxiliary classifier gans. In *ICML*, pages 2642–2651. PMLR, 2017. 3

[34] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *ICML*, pages 1060–1069. PMLR, 2016. 3

[35] Daniel Roich, Ron Mokady, Amit H Bermano, and Daniel Cohen-Or. Pivotal tuning for latent-based editing of real images. *ACM TOG*, 42:1–13, 2022. 6

[36] Haorui Song, Yong Du, Tianyi Xiang, Junyu Dong, Jing Qin, and Shengfeng He. Editing out-of-domain gan inversion via differential activations. In *ECCV*, pages 1–17, 2022. 1

[37] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *NeurIPS*, 30, 2017. 5

[38] Omer Tov, Yuval Alaluf, Yotam Nitzan, Or Patashnik, and Daniel Cohen-Or. Designing an encoder for stylegan image manipulation. *ACM TOG*, 40(4):1–14, 2021. 5

[39] Chia-Hung Wan, Shun-Po Chuang, and Hung-Yi Lee. Towards audio to scene image synthesis using generative adversarial network. In *ICASSP*, pages 496–500. IEEE, 2019. 3

[40] Yaxing Wang, Abel Gonzalez-Garcia, David Berga, Luis Herranz, Fahad Shahbaz Khan, and Joost van de Weijer. Minegan: effective knowledge transfer from gans to target domains with few images. In *CVPR*, pages 9332–9341, 2020. 3

[41] Zongwei Wu, Liangyu Chai, Nanxuan Zhao, Bailin Deng, Yongtuo Liu, Qiang Wen, Junle Wang, and Shengfeng He. Make your own sprites: Aliasing-aware and cell-controllable pixelization. *ACM TOG*, 41(6):1–16, 2022. 3

[42] Zongze Wu, Dani Lischinski, and Eli Shechtman. Stylespace analysis: Disentangled controls for stylegan image generation. In *CVPR*, pages 12863–12872, 2021. 8

[43] Huiting Yang, Liangyu Chai, Qiang Wen, Shuang Zhao, Zixun Sun, and Shengfeng He. Discovering interpretable latent space directions of gans beyond binary attributes. In *CVPR*, pages 12177–12185, 2021. 2

[44] Mengping Yang, Zhe Wang, Ziqiu Chi, and Wenyi Feng. Wavegan: Frequency-aware gan for high-fidelity few-shot image generation. *ECCV*, 2022. 1, 2, 5, 6, 7

[45] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *ICCV*, pages 5907–5915, 2017. 3

[46] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, pages 586–595, 2018. 5

[47] Miaoyun Zhao, Yulai Cong, and Lawrence Carin. On leveraging pretrained gans for generation with limited data. In *ICML*, pages 11340–11351, 2020. 3

[48] Yunqing Zhao, Henghui Ding, Houjing Huang, and Ngai-Man Cheung. A closer look at few-shot image generation. In *CVPR*, pages 9140–9150, 2022. 3