Exact Fusion via Feature Distribution Matching for Few-shot Image Generation

Yingbo Zhou Yutong Ye Pengyu Zhang Xian Wei Mingsong Chen*
MoE Eng. Research Center of SW/HW Co-Design Tech. and App., East China Normal University

Abstract

*Few-shot image generation, as an important yet challenging visual task, still suffers from the trade-off between generation quality and diversity. According to the principle of feature-matching learning, existing fusion-based methods usually fuse different features by using similarity measurements or attention mechanisms, which may match features inaccurately and lead to artifacts in the texture and structure of generated images. In this paper, we propose an exact **F**usion via **F**eature **D**istribution matching **G**enerative **A**dversarial **N**etwork (**F2DGAN**) for few-shot image generation. The rationale behind this is that feature distribution matching is much more reliable than feature matching to explore the statistical characters in image feature space for limited real-world data. To model feature distributions from only a few examples for feature fusion, we design a novel variational feature distribution matching fusion module to perform exact fusion by empirical cumulative distribution functions. Specifically, we employ a variational autoencoder to transform deep image features into distributions and fuse different features exactly by applying histogram matching. Additionally, we formulate two effective losses to guide the matching process for better fitting our fusion strategy. Extensive experiments compared with state-of-the-art methods on three public datasets demonstrate the superiority of F2DGAN for few-shot image generation in terms of generation quality and diversity, and the effectiveness of data augmentation in downstream classification tasks*[1].

1. Introduction

Relying on substantial data, deep generative models [3, 12, 14, 25] have led to a series of breakthroughs in synthesizing high-quality and diverse images. Unfortunately, there are many scenarios in the real world where it is impossible to obtain large-scale samples of one category for model training. To adapt to such practical applications, a variety of few-shot image generation methods have been introduced as data augmentation techniques [19, 24, 50] to produce more new images with a glance at a few samples of one category. However, without sufficient training data, few-shot image generation is still challenging to yield realistic images with considerable diversity.

In general, prevailing few-shot image generation methods commonly suppose that trained models with the seen data have a good generalization ability towards unseen samples. Based on this assumption, optimization-based methods [7, 30] are adopted to seek proper initial parameters to generate images, which neglect to integrate the knowledge of unseen categories and result in a distorted generation without fine-grained details. To effectively capture the information of unseen samples, transformation-based methods [9, 28, 51] utilize pre-trained latent spaces of style-series models on seen data to invert the available unseen images into the latent space. However, since complicated image transformations are not end-to-end training processes, the generated images usually tend to be biased by the distribution of the seen classes. Consequently, the generation with little category information of unseen data has subtle benefits to downstream visual applications [16, 38, 39]. To let the model "learn to learn" in the training process, fusion-based methods [2, 15, 20, 21, 53] design different fusion modules with the episodic training mechanism [41]. Technically, current fusion strategies are implemented either on image pixels or feature representations, which match features inaccurately and result in unsatisfactory images with artifacts and limited diversity.

To fuse the visual semantics for high-quality image generation, it is crucial to make the model well-informed about the consistency of global distribution in the fusion process. In this paper, we concentrate on the shortcomings of current fusion strategies and raise the question, "can we fuse different features via feature distribution matching for few-shot image generation?" Theoretically, one can fuse different features to match feature mean and standard deviation by assuming that features follow Gaussian distribution. However, since training one model to learn feature distributions from a few samples of one category is difficult, the learned feature mean and standard deviation are not representative to match different features exactly. Motivated by EFDM [49], which applies the exact histogram matching

[*] Corresponding author (mschen@sei.ecnu.edu.cn)
[1] Code is available at: https://github.com/ZYBOBO/F2DGAN

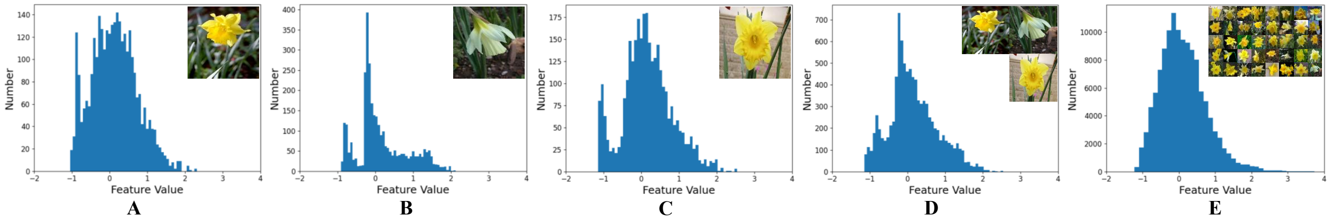

Figure 1. Histograms of feature values of different images in a randomly selected channel, where features are computed from the first residual block of ResNet-18 [17] pre-trained on ImageNet [8]. A, B, and C are different feature distributions from one single image, and D and E denote the cumulative feature distribution of three images and all the images in the same category, respectively.

algorithm to produce more stable style-transfer images, we visualize the histograms of feature values to get an intuition of the characteristics of feature distribution in one category. As shown in Figure 1, one can clearly observe that different images of the same class have different feature distributions, and the accumulation of feature values of different images can better reflect the feature distribution of their categories.

Based on the above insights, we propose a novel exact Fusion via **F**eature **D**istribution matching **G**enerative **A**dversarial **N**etwork (**F2DGAN**) for few-shot image generation. To be concrete, we design a lightweight but effective feature distribution matching module (FDMM) to fuse different image features, which can obtain more consistent structures and textures with an image-matching reconstruction loss. Moreover, we devise a simple but ingenious variational feature learning module (VFLM) to produce more diverse feature representations, which can transform high-level semantics into distributions smoothly with a feature reconstruction loss. Regarding FEMM and VFLM as our variational feature distribution matching fusion, F2DGAN shows significant superiority over the state-of-the-art methods on three well-known benchmarks.

Our contributions can be summarized as follows:

- We propose a novel fusion-based framework called **F2DGAN** for few-shot image generation, which can achieve better generation quality and diversity by our variational feature distribution matching fusion module.
- We design a feature distribution matching module with an image-matching reconstruction loss, which can fuse consistent semantics exactly with histogram matching.
- We devise a variational feature learning module with a feature reconstruction loss, which further guarantees the diversity of fused deep semantics in feature space.

Comprehensive experiments show that our proposed F2DGAN can not only achieve stable few-shot image generation with state-of-the-art quality and diversity, but also significant improvements in the accuracy for downstream classification tasks.

2. Related Work

Few-shot Image Generation. Given a few samples from one novel category, few-shot image generation aims to produce more new images with high quality and various diversity. Existing methods are classified as optimization-based, fusion-based, and transformation-based. Optimization-based methods [7, 30] introduce meta-learning to explore a set of initial parameters with good generalization for different tasks, which can hardly generate realistic images. Fusion-based methods [2, 20, 21] adopt feature-matching learning to obtain fused features with an attention mechanism or similarity measurement, which are prone to yield limited-diversity images with artifacts in detail. While transformation-based methods [9, 28, 51] can learn more differences in the unseen-specific features via intra-category transformations, the generated images have little category information preserved for the unseen images. To be concrete, AGE [9] generates images with an intra-category transformation from the ”editing-based“ perspective, LSO [51] introduces latent space optimization to model the distribution of unseen samples with category-specific features, and HAE [28] captures the hierarchy among images in hyperbolic space to control the semantic diversity of the generated images. Despite the above-mentioned transformation-based approaches producing leading results in diversity, their roles as image augmentation techniques for downstream classification tasks are inferior to the latest fusion-based method WaveGAN [45]. In this work, we build a novel fusion-based framework by employing variational feature learning and feature distribution matching to generate more diverse images and further improve the accuracy for downstream classification tasks. To the best of our knowledge, we are the first to introduce variational feature learning and feature distribution matching into few-shot image generation.

Feature Distribution Matching. In the arbitrary style transfer (AST) [13, 22] field, image styles can be interpreted as feature distributions and style transfer can be achieved by cross-distribution feature matching. The most common approach for feature distribution matching is to match feature mean and standard deviation with Gaussian distribution [29, 33]. Unfortunately, the feature distributions of limited real-world data are usually too difficult to be sampled by Gaussian. To minimize the feature distribution divergence for AST, EFDM [49] introduces an exact histogram-

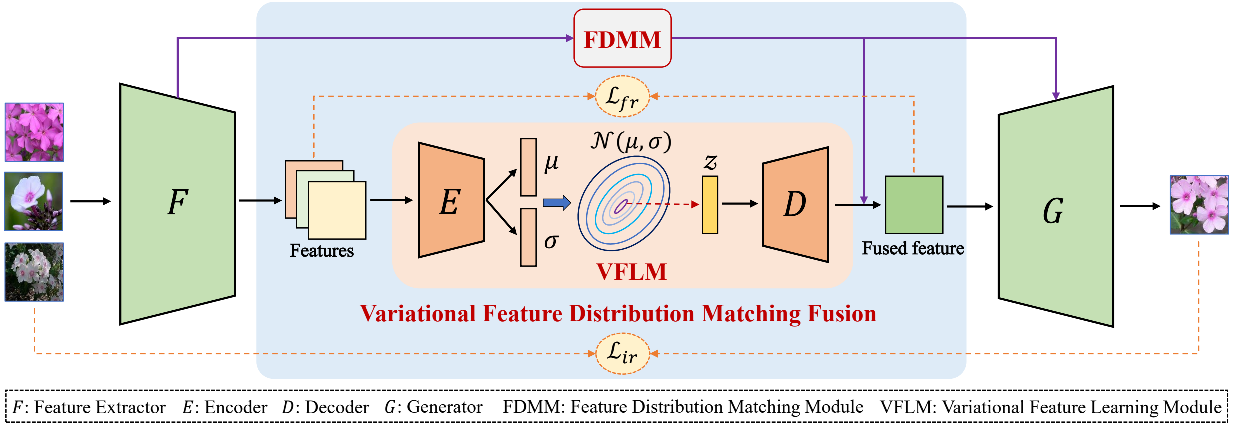
FDMM
$\mathcal{L}_{fr}$
F
Features
E
μ
σ
$\mathcal{N}(\mu,\sigma)$
z
D
Fused feature
G
VFLM
Variational Feature Distribution Matching Fusion
$\mathcal{L}_{ir}$
F: Feature Extractor E: Encoder D: Decoder G: Generator FDMM: Feature Distribution Matching Module VFLM: Variational Feature Learning Module

Figure 2. Overview of our framework. Given a few samples of a novel category, we can generate new images belonging to the same category by our variational feature distribution matching fusion strategy. Note that we do not visualize the discriminator branch for simplicity.

matching algorithm to match the empirical cumulative distribution function of image features, which can result in matched feature distributions with exact mean, standard deviation, and high-order statistics. Considering that image features can be distinguished by equivalent pixel values either randomly or according to their local mean, in this paper, we propose a more accurate fusion strategy based on the matching of histograms for few-shot image generation.

Variational Feature Learning. Variational autoencoders (VAEs) [11, 23], as one common type of variational feature learning model, can transform an image or feature into a distribution, which has been used in various tasks, such as zero/few-shot learning [26, 47], metric learning [10, 44], and disentanglement learning [31]. For fusion-based few-shot image generation, feature inconsistency makes it hard to correctly fuse different features by matching the semantically nearest local representations or global vectors. Even worse, the existing fusion strategies can be roughly regarded as point estimate approaches, which are sensitive to sample noise from random selection and inductive bias from scarce data. To alleviate the inconsistency between different features, in this paper, we utilize VAEs to model class distributions in image feature space, which are robust to limited data than point estimates.

3. Methodology

3.1. Overview

Problem Definition. Given k images sampled of the same category, we intend to generate realistic and diverse images for this category, defined as a k-shot image generation task. To handle this task, a dataset usually is split into two subsets, i.e., the seen categories $\mathcal{C}_s$ and the unseen categories $\mathcal{C}_u$, where $\mathcal{C}_s \cap \mathcal{C}_u = \emptyset$. In the training phase, hundreds of k-shot image generation tasks sampled from $\mathcal{C}_s$ are fed into the model to learn transferable generation ability. In the test phase, we utilize the well-trained model to generate new images for $\mathcal{C}_u$.

Overall Framework. As shown in Figure 2, the overall framework of F2DGAN consists of four basic components, i.e., one feature extractor F, one encoder E, one decoder D, and one generator G. The key design of our method is the variational feature distribution matching fusion strategy, which is composed of the feature distribution matching module (FDMM) and the variational feature learning module (VFLM). Given a few images as inputs, we first obtain deep features by our feature extractor f. Then, our VFLM transforms these features into an approximate distribution based on variational inference. Finally, we regard feature maps at different scales as feature distributions to fuse different semantics with their category distributions based on exact histogram matching. During the whole training process, we formulate two effective losses (i.e., feature reconstruction loss $\mathcal{L}_{fr}$ and image-matching reconstruction loss $\mathcal{L}_{ir}$) to guide the model training. Note that we omit the discriminator branch in our overall framework due to our fusion strategy mainly involving the generator part.

3.2. Feature Distribution Matching Module

Empirical Cumulative Distribution Function. Empirical cumulative distribution function (eCDF) is one nonparametric method to describe the distribution of sample data in statistics. Given the data points x of samples X, the cumulative probability of each point in the overall distribution is defined as:

$$\hat{F}_X(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{x_i \leq x}, \tag{1}$$

where $\mathbf{1}_A$ is the indicator of event A and x_i is the i-th element of x. Based on the Glivenko–Cantelli theorem [40], the eCDF asymptotically converges to the CDF when the number of samples approaches infinity [42].

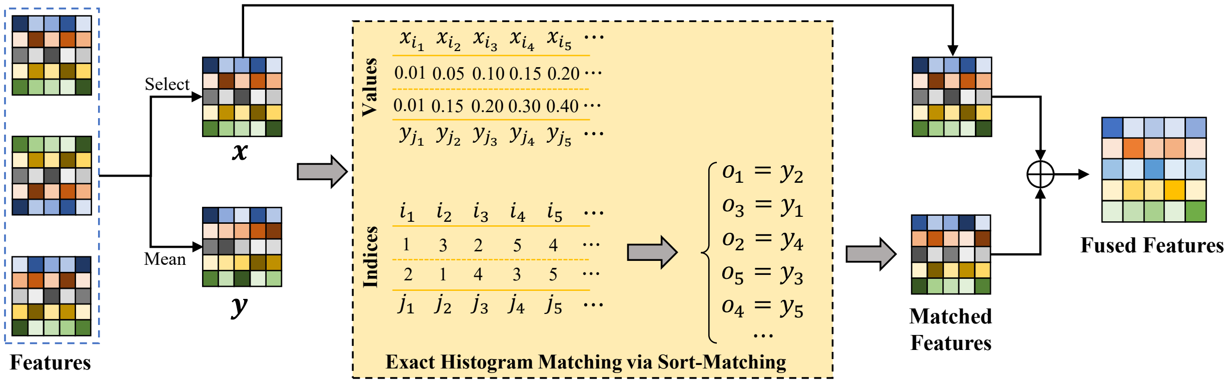
Features
Select
Mean
x
y
Values
x_{i_1} x_{i_2} x_{i_3} x_{i_4} x_{i_5} ⋯
0.01 0.05 0.10 0.15 0.20 ⋯
0.01 0.15 0.20 0.30 0.40 ⋯
y_{j_1} y_{j_2} y_{j_3} y_{j_4} y_{j_5} ⋯
Indices
i_1 i_2 i_3 i_4 i_5 ⋯
1 3 2 5 4 ⋯
2 1 4 3 5 ⋯
j_1 j_2 j_3 j_4 j_5 ⋯
$o_1 = y_2$
$o_3 = y_1$
$o_2 = y_4$
$o_5 = y_3$
$o_4 = y_5$
⋯
Exact Histogram Matching via Sort-Matching
Matched Features
Fused Features

Figure 3. An illustration of feature distribution matching module. We utilize feature values and indices to implement the exact histogram matching via the sort-matching algorithm, resulting in matched features and fused features.

Histogram Matching. The goal of histogram matching is to transform an input vector x into an output vector o, whose eCDF matches the target eCDF of a target vector y. According to the Eq. (1), we can find the y_i satisfies $\hat{F}_X(x_i) = \hat{F}_Y(y_i)$ for each element x_i of the input x, resulting in the transformation function: $H(x_i) = y_i$. Ideally, histogram matching could exactly match eCDFs of image features using bins of infinitesimal width. However, due to the finite number of bits to represent features in discrete feature space [5], histogram matching is hard to equal to eCDFs when there exist equivalent feature values in inputs.

Fusion with Feature Distribution Matching. To match histograms of image pixels exactly, we apply an exact histogram-matching algorithm to distinguish equivalent pixel values and apply an element-wise transformation to merge equivalent values. Given a few features $X \in \mathbb{R}^{B\times k\times C\times H\times W}$ of the same category, we randomly select the features with the index of $b \in \{1, 2, \cdots, k\}$ as x and compute the average of all the image features as y. Then, we adopt the sort-matching algorithm [37] to match two sorted feature vectors for feature distribution matching fusion as shown in Figure 3. Flattening feature maps into feature vectors, the feature indexes are illustrated as:

$$\begin{aligned} \mathbf{x} : i &= (i_1 \;\; i_2 \;\; i_3 \;\; \cdots \;\; i_n), \\ \mathbf{y} : j &= (j_1 \;\; j_2 \;\; j_3 \;\; \cdots \;\; j_n), \end{aligned} \tag{2}$$

where $\{x_{i_m}\}_{m=1}^{n}$ and $\{y_{j_m}\}_{m=1}^{n}$ are sorted values of x and y in an ascending order. In other words, $x_{i_1} = \min(x)$, $x_{i_n} = \max(x)$, and $x_{i_m} \leq x_{i_n}$ if $m < n$. y_{j_m} has a similar properties as x_{i_m}. Based on the definition in Eq. (2), the matched feature vectors o with its i_m-th element o_{i_m} as:

$$\mathbf{o}_{i_m} = \mathbf{y}_{j_m}. \tag{3}$$

To enable the gradient back-propagation in the model without the interference of non-parametric matching, we practical perform the feature distribution matching for our fusion strategy by modifying Eq. (3) as:

$$Match(x, y) : \mathbf{M}_{i_m} = \mathbf{x}_{i_m} + \mathbf{y}_{j_m} - \langle \mathbf{x}_{i_m} \rangle, \tag{4}$$

where $\langle\cdot\rangle$ denotes the stop-gradient operation. In this way, we can get the fused features as follows:

$$Fusion(x, y) : \mathbf{F}_{i_m} = \mathbf{x}_{i_m} + \mathbf{M}_{j_m}. \tag{5}$$

To measure the distribution divergence more exactly, we regard the input images $I = \{I_1, I_2, \cdots, I_k\}$ as distributions to formulate the image-matching reconstruction loss:

$$\mathcal{L}_{ir} = \Phi(I_b, \tfrac{1}{k}\textstyle\sum_{i=1}^{k} I_i), \tag{6}$$

where $\Phi(\cdot)$ represents the mean square error (MSE) loss in a channel-wise manner following [22, 52].

3.3. Variational Feature Learning Module

Although we can use feature values as feature distributions to promote the learning ability of the model, the values of deep features are too few to accumulate histograms as feature distributions effectively. To alleviate this problem, we transform deep features into category distributions with VAEs. Since the estimated distribution is not biased toward specific examples, features sampled from Gaussian noises have a better consistency in high-level semantics. In this way, we can match class-level features for robust feature fusion with feature distribution matching.

Variational Feature Learning with VAEs. Formally, we aim to transform deep features into a class distribution $\mathcal{N}$, and sample the variational feature z from $\mathcal{N}$ for feature matching as shown in Figure 2. Suppose that the features s follow a conditional distribution $p(s|z)$, where z is sampled from a prior distribution $p(z)$, we can learn the distribution of z based on Bayes rule $p(z|s) = p(z)p(s|z)/p(s)$. However, the process involves a computation intractable integral with an unknown variable z. For easy to compute, we model the posterior distribution with variational inference. More concretely, we approximate the true posterior distribution $p(z|s)$ with another distribution $q(z|s)$ by minimizing the Kullback-Leibler (KL) divergence:

$$\mathcal{D}_{\mathrm{KL}}(q(z|s)\|p(z|s)) = \int q(z|s)\log\frac{q(z|s)}{p(z|s)}, \tag{7}$$

which is equivalent to maximizing the evidence lower bound (ELBO) [27]:

$$\mathrm{ELBO} = \mathbb{E}_{q(z|s)}[\log p(s|z)] - \mathcal{D}_{\mathrm{KL}}(q(z|s)\|p(z)). \tag{8}$$

According to Eq. (8), a KL-divergence between the posterior and prior distributions needs to be calculated. We assume the prior distribution z is a centered isotropic multivariate Gaussian, $p(z) = \mathcal{N}(0, I)$, and the posterior distribution $q(z|s)$ is a multivariate Gaussian with diagonal covariance, i.e., $q(z|s) = \mathcal{N}(\mu, \sigma)$. To obtain the parameters of μ and σ, we use an encoder E to encode the category feature $s = \frac{1}{k}\sum_{i=1}^{k} s_i$, where s_i are a few deep features of the same category. Then, the variational feature z can be implemented by the parameterization trick [27], where $z = \mu + \sigma \cdot \epsilon$, and $\epsilon \sim \mathcal{N}(0, I)$. In this way, the first term of Eq. (8) is simplified as a reconstruction loss between the input s and the decoded feature s' in the L2 distance:

$$\mathcal{L}_{rec} = \|s - s'\| = \|s - D(z)\|, \tag{9}$$

where $D(\cdot)$ denotes a feature decoder. Meanwhile, the second term of Eq. (8) is directly minimized as:

$$\begin{aligned}\mathcal{L}_{\mathrm{KL}} &= \mathcal{D}_{\mathrm{KL}(\mathcal{N}(\mu,\sigma))\|\mathcal{N}(0,\mathrm{I})} \\ &= \frac{1}{2}(\mu^2 + \sigma - \log\sigma - 1),\end{aligned} \tag{10}$$

which forces the variation feature z to learn a class-specific distribution with a normal distribution.

Fusion with Variational Feature Learning. To combine variational feature learning and feature distribution matching, we flatten deep features $s_i \in \mathbb{R}^{B\times C\times H\times W}$ generated from the last layer of the feature extractor F into $s_i \in \mathbb{R}^{B\times CHW}$, where $i \in \{1, 2, \cdots, k\}$, which can preserve the values of features in the process of variational inference. By optimizing $\mathcal{L}_{rec}$ and $\mathcal{L}_{\mathrm{KL}}$, we can fuse the variational features s' with the selected features s_b via feature distributions matching. Based on Eq. (4) and (5), the process of fused deep features with variational feature learning can be written as in the following equation:

$$Fusion(s_b, s') : F^{s}_{i_m} = 2 * (s_b)_{i_m} + s'_{j_m} - \langle s_{i_m}\rangle . \tag{11}$$

To further ensure that the fused deep features are semantically consistent in category information, we define an orthogonal projection loss within a mini-batch as follows:

$$\mathcal{L}_{opt} = 1 - \sum_{i=j}^{i,j\in B}\frac{s_b^i\cdot s_j'}{\|s_b^i\|\cdot\|s_j'\|} + \left|\sum_{i\neq j}^{i,j\in B}\frac{s_b^i\cdot s_j'}{\|s_b^i\|\cdot\|s_j'\|}\right|, \tag{12}$$

where B denotes mini-batch size, $\|\cdot\|$ and $|\cdot|$ refer to the L2 distance and the absolute value operator, respectively. In Eq. (12), our objective is to ensure the clustering of the same class semantics and the orthogonality of different class information, which can obtain more robust category distributions. Adding the orthogonal projection loss as a regularization term to the reconstruction loss, our feature reconstruction loss is established as:

$$\mathcal{L}_{fr} = \mathcal{L}_{rec} + \lambda_1\mathcal{L}_{\mathrm{KL}} + \lambda_2\mathcal{L}_{opt}, \tag{13}$$

where λ_1 and λ_2 are the hyper-parameters balancing the weight for the three different constraints.

3.4. Optimization Objective

Given k images $X = \{x_1, x_2, \cdots, x_k\}$ of a novel category C, our goal is to generate new image $\hat{x}$ with the same label. With hundreds of k-shot image generation tasks as inputs, our model is updated in an end-to-end episodic training mechanism, where the generator G and the discriminator D_{is} are optimized alternatively by adversarial training. Apart from the proposed $\mathcal{L}_{ir}$ and $\mathcal{L}_{fr}$, we also use adversarial loss and classification loss to constrain our objective.

Adversarial Loss. To make the generated image $\hat{x}$ close to real inputs X, we employ a discriminative model D_{is} like LoFGAN [15], which utilizes the hinge version of adversarial loss [46] to train the model. The loss functions for the discriminator and the generator are expressed as follows:

$$\begin{aligned}&\mathcal{L}^{D_{is}}_{adv} = \max(0, 1 - D(X)) + \max(0, 1 + D(\hat{x})), \\ &\mathcal{L}^{G}_{adv} = -D(\hat{x}).\end{aligned} \tag{14}$$

Classification Loss. To ensure that the generated image $\hat{x}$ has the same label as X, we follow ACGAN [36] and construct an auxiliary classifier in the discriminator. In this way, we can guide the generator to produce images while maintaining the category information and constrain the discriminator to classify images with the following loss:

$$\begin{aligned}\mathcal{L}^{D_{is}}_{cls} &= -\log P(C|X), \\ \mathcal{L}^{G}_{cls} &= -\log P(C|\hat{x}).\end{aligned} \tag{15}$$

Therefore, the overall optimization objective of our method can be formulated as:

$$\begin{aligned}\mathcal{L}_{D_{is}} &= \mathcal{L}^{D_{is}}_{adv} + \lambda_{cls}\mathcal{L}^{D_{is}}_{cls}, \\ \mathcal{L}_{G} &= \mathcal{L}^{G}_{adv} + \lambda_{cls}\mathcal{L}^{G}_{cls} + \lambda_{ir}\mathcal{L}_{ir} + \lambda_{fr}\mathcal{L}_{fr},\end{aligned} \tag{16}$$

where λ_{cls}, λ_{ir}, and λ_{fr} are weight coefficients controlling the whole optimization process.

4. Experiments

In this section, We evaluate the superiority of our approach on two few-shot image tasks, i.e., few-shot image generation and low-data image classification.

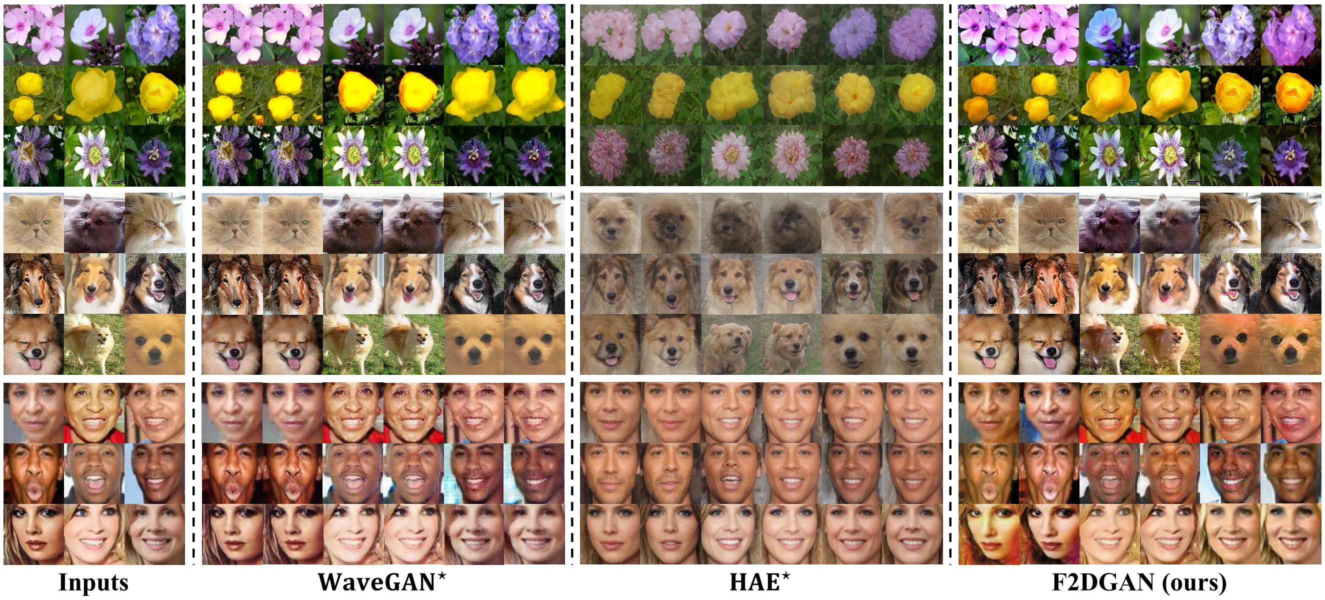

Figure 4. Qualitative comparison with state-of-the-art methods on Flowers, Animal Faces, and VGGFaces. Please zoom in for more details.

4.1. Implementation Details

Network Details. Our feature extractor F consists of five convolutional blocks, each of which has one convolution, Leaky-ReLU activation [34], and batch normalization [4] operations. The generator G is symmetric to the structure of the extractor F, and G has extra skip operations in upsampling convolutional blocks. We use the feature distribution matching module (FDMM) to fuse features at different scales in each convolution block of the F and add the fused features to the generator G by skip operations, which are implemented with element-wise additions. As for the variational feature learning module (VFLM), we implement it based on a light encoder E and decoder D, where E contains a linear layer and two parallel linear layers to produce μ and σ and D consists of two linear layers to generate the variational features s'. Note that our VFLM only works on 8×8 deep feature maps, and it incorporates the FDMM to get the fused deep features. Additionally, our discriminator D_{is} adopts a similar network in [15], which has four residual blocks and two linear layers to accomplish the classification and discrimination of inputs.

Hyper-parameter Settings. In the training stage, we use the Adam optimizer to optimize our network with $100,000$ iterations, where the learning rate α is fixed as $1e-4$ in the first $50,000$ iterations and linearly decayed to 0 in the second $50,000$ iterations. We set the mini-batch to 8, which means that we randomly sample eight k-shot image generation tasks in each iteration. The selection ratios of λ_1 and λ_2 in the feature reconstruction loss $\mathcal{L}_{fr}$ are set to 0.0025 and 2, respectively. For our overall optimization objective function, we set $\lambda_{cls} = \lambda_{ir} = 1$, and $\lambda_{fr} = 10$.

4.2. Baselines and Metrics

Baselines. We compare our F2DGAN with several few-shot image generation methods, including the optimization-based FIGR [7] and DAWSON [30], the fusion-based MatchingGAN [20], F2GAN [21], LoFGAN [15] and WaveGAN [45], the transformation-based AGE [9], LSO [51] and HAE [28]. To ensure a fair comparison, we reproduce LoFGAN, WaveGAN and HAE under the same experimental environment, denoted as LoFGAN*, WaveGAN* and HAE*, respectively.

Metrics. We use Fréchet Inception Distance (FID) [18] and Learned Perceptual Image Patch Similarity (LPIPS) [48] as the metrics for quantitative evaluation. FID is used to measure the distance between real image vectors and generated image vectors, where the smaller distance indicates the higher quality of the generated images. LPIPS is widely employed to evaluate the consistency between paired input and output in the image-to-image translation field [1, 43]. In contrast, we utilize LPIPS to measure the diversity of the generated images, where the higher LPIPS means the better diversity of the generated images.

4.3. Evaluation Datasets

We conduct experiments on three commonly used benchmarks, namely Flowers [35], Animal Faces [32], and VGGFaces [6] following the settings described in [15].

Flowers. We randomly split Flowers into 85 seen categories for training and 17 unseen categories for testing. All the images are collected with the resolution of 128×128.

Animal Faces. We select 119 classes as seen classes for training and 30 classes as unseen classes for evaluation. All

Table 1. Quantitative comparison results between our model and the baselines. The best and the second-best results are **highlighted** and underlined, respectively. The symbol ↓ indicates that lower is better while the symbol ↑ indicates that higher is better.

Method	Type	Flowers [35]		Animal Faces [32]		VGGFaces [6]	
		FID ↓	LPIPS ↑	FID ↓	LPIPS ↑	FID ↓	LPIPS ↑
FIGR [7]	O	190.12	0.0634	211.54	0.0756	139.83	0.0834
DAWSON [30]	O	188.96	0.0583	208.68	0.0642	137.82	0.0769
MatchingGAN [20]	F	143.35	0.1627	148.52	0.1514	118.62	0.1695
F2GAN [21]	F	120.48	0.2172	117.74	0.1831	109.16	0.2125
LoFGAN* [15]	F	78.83	0.3869	113.01	0.4894	20.52	0.2921
WaveGAN* [45]	F	42.17	0.3868	30.35	0.5076	4.96	0.3255
AGE [9]	T	45.96	0.4305	28.04	0.5575	34.86	0.3294
LSO [51]	T	**34.59**	0.3914	**23.67**	0.5198	**3.98**	0.3344
HAE* [28]	T	50.10	**0.4739**	26.33	**0.5636**	35.93	**0.5919**
F2DGAN (Ours)	F	38.26	0.4325	25.24	0.5463	4.25	0.3521

the images have a resolution of 128×128.

VGGFaces. We divide VGGfaces into two subsets, where the seen subset has 1802 categories for training and the unseen subset has 552 categories for testing. All the images are collected with the resolution of 64×64.

4.4. Few-shot Image Generation

To highlight the superiority of our method in generation performance, we carry out the following experiments to qualitatively and quantitatively evaluate F2DGAN compared with the state-of-the-art methods.

Qualitative Evaluation. We qualitatively compare our F2DGAN with LoFGAN [15], WaveGAN [45] and HAE [28]. The visualization of synthesis results on Flowers, Animal Faces, and VGGFaces are presented in Figure 4, where the leftmost three columns are the inputs for each method. From Figure 4, we can observe that F2DGAN can produce higher-quality images with fine-grained details while maintaining more category-related diversity than others. For instance, the flowers generated by WaveGAN* tend to be homogeneous, while F2DGAN can yield new flowers with petal color and texture variations. Although animal faces generated by HAE* have more semantic diversity in appearances and poses, they preserve little category-related information of the inputs, which look less natural than the animal faces generated by F2DGAN. Moreover, the gender and outline of faces generated by HAE are inconsistent with those of real images, while F2DGAN can produce identity-preserving faces with skin color and shadow change. In a nutshell, F2DGAN can also achieve more visually pleasing and diverse results than WaveGAN* and HAE*.

Quantitative Evaluation. For quantitative comparison, we randomly sample 128 generated images for each unseen category and calculated the FID and LPIPS to evaluate the fidelity and diversity of the generation following 3-shot settings in [15]. As shown in Table 1, we present our quantitative results with 9 baselines on Flowers, Animal Faces, and VGGFaces, where "O", "F" and "T" in the second column denote optimization-based, fusion-based and transformation-based method, respectively. Specifically, our method significantly outperforms all the fusion-based methods on FID and LPIPS scores over four datasets, which indicates the effectiveness of our fusion strategy. Although transformation-based methods (i.e., AGE, LSO, and HAE) show slight advantages on FID or LPIPS, their generation plays limited roles in downstream classification tasks due to the absence of category-specific information of unseen images. In the following subsection, we will elaborate on the performance of the generated images as a means of data augmentation on low-data image classification tasks.

Table 2. Top-1 accuracy (%) comparison on low-data image classification over Flowers, Animal Faces, and VGGFaces.

	Flowers	Animals	VGGFaces
Baseline	58.09	32.98	60.35
LoFGAN [15]	66.10	35.71	67.89
AGE [9]	68.94	43.14	64.38
HAE [28]	70.23	45.28	63.35
WaveGAN [45]	79.24	55.27	72.40
F2DGAN (ours)	**84.02**	**60.76**	**77.92**

4.5. Low-data Image Classification

To investigate the performance of generated images on downstream image classification, we use few-shot generation models as data augmentation techniques to produce more new images for unseen categories. Specifically, we split the unseen categories of each dataset into $\mathbb{D}_{tr}$, $\mathbb{D}_{va}$ and $\mathbb{D}_{te}$, where the ratios are $10 : 10 : 15$ for Flowers while $30 : 35 : 35$ for Animal Faces and VGGfaces, respectively. Following [15], we first initialize a ResNet-18 backbone [17] based on the seen categories. Then, we train a new classifier with the augmentations to evaluate the resulting classification. Note that we regard the new well-trained classifier using $\mathbb{D}_{tr}$ without any augmentation as the

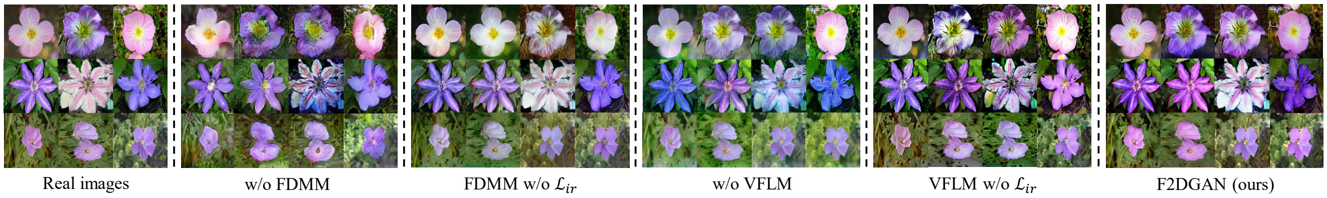

Figure 5. Visual comparison of ablation studies with different fusion components on Flowers. Please zoom in for more details.

Baseline. Using each few-shot image generation model as a data augmentation technique, we augment 30 images for each unseen category in Flowers, while 50 images for each unseen category in Animal Faces and VGGFaces.

The results are presented in Table 2. Compared with the *baseline*, all the few-shot image generation models can improve the accuracy for low-data image classification tasks on four benchmarks. Furthermore, the images generated by transformation-based AGE and HAE bring fewer benefits than those of WaveGAN, which indicates that transformation-based methods have poor generalization abilities in capturing the category information of unseen images. More importantly, the images generated by our proposed method as the data augmentation achieve the best accuracy on all three datasets. Such performance in classification tasks demonstrates that F2DGAN can effectively learn the unseen category distributions and generate new images with unseen category information.

4.6. Ablation Studies

Fusion Components. Our fusion strategy includes the feature distribution matching module (FDMM) with the image-matching reconstruction loss $\mathcal{L}_{ir}$ and the variational feature learning module (VFLM) with the feature reconstruction loss $\mathcal{L}_{fr}$. To verify the effectiveness of our fusion components. we disable each component for ablation studies and evaluate each component with FID, LPIPS, and accuracy gain as shown in Table 3. In this Table, we can find that each component of our proposed fusion strategy plays a positive role in the generation quality and diversity. Moreover, the satisfying gain in the classification task demonstrates that fusing different features with the FDMM and FVLM can effectively capture the real unseen distributions. In addition, we visualize our ablation study results in Figure 5, which further demonstrates the effectiveness of our variational feature distribution matching fusion.

k **Shots.** To explore the generation performance under different k-shot image generation settings, we train our F2DGAN with $k \in \{2, 3, 5, 7, 9\}$ on Flowers. As seen in Figure 6, we use $\lg(FID)$ and e^{LPIPS} to substitute FID and LPIPS scores, respectively. As k increases, $\lg(FID)$ scores vary from 1.56 to 1.61, and e^{LPIPS} scores vary from 1.52 to 1.58, which manifests our method has good generalization ability for few-shot image generation tasks.

Table 3. Quantitative results of ablation studies on generation performance and low-data image classification over Flowers.

Conditions		Flower		
FDMM	VFLM	FID ↓	LPIPS ↑	Accuracy (%)
w/o FDMM		58.48	0.3915	77.84
FDMM w/o $\mathcal{L}_{ir}$		46.42	0.4129	80.19
w/o VFLM		44.96	0.4083	72.88
VFLM w/o $\mathcal{L}_{fr}$		40.87	0.4234	80.96
F2DGAN (ours)		38.26	0.4325	84.02

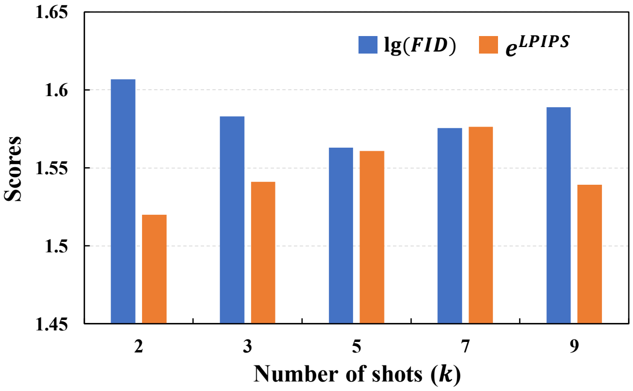

Figure 6. Scores under different k-shot settings on Flowers.

5. Conclusion

In this paper, we delve into fusion-based few-shot image generation from a new perspective of feature distribution matching. To obtain an exact fusion via feature distribution matching, we propose a variational feature distribution matching fusion strategy, which can fuse different features consistently and maintain category-specific semantics unchangeably for unseen categories. Qualitative and quantitative results on three well-known datasets demonstrate the superiority and robustness of our method, which brings considerable benefits for downstream classification tasks.

Acknowledgements

This work was supported by the Natural Science Foundation of China (62272170), and “Digital Silk Road” Shanghai International Joint Lab of Trustworthy Intelligent Software (22510750100).

References

[1] Kyungjune Baek, Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Hyunjung Shim. Rethinking the truly unsupervised image-to-image translation. In *IEEE/CVF International Conference on Computer Vision, ICCV*, pages 14134–14143, 2021. 6

[2] Sergey Bartunov and Dmitry P. Vetrov. Few-shot generative modelling with generative matching networks. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 670–678, 2018. 1, 2

[3] Ankan Kumar Bhunia, Salman H. Khan, Hisham Cholakkal, Rao Muhammad Anwer, Jorma Laaksonen, Mubarak Shah, and Fahad Shahbaz Khan. Person image synthesis via denoising diffusion model. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*, pages 5968–5976, 2023. 1

[4] Johan Bjorck, Carla P. Gomes, Bart Selman, and Kilian Q. Weinberger. Understanding batch normalization. In *Advances in Neural Information Processing Systems, NeurIPS*, pages 7705–7716, 2018. 6

[5] Wilhelm Burger and Mark James Burge. *Digital Image Processing - An Algorithmic Introduction, Third Edition*. Springer, 2022. 4

[6] Qiong Cao, Li Shen, Weidi Xie, Omkar M. Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *13th IEEE International Conference on Automatic Face & Gesture Recognition, FG*, pages 67–74, 2018. 6, 7

[7] Louis Clouâtre and Marc Demers. FIGR: few-shot image generation with reptile. *arXiv preprint arXiv:1901.02199*, 2019. 1, 2, 6, 7

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE/CVF Computer Society Conference on Computer Vision and Pattern Recognition CVPR*, pages 248–255, 2009. 2

[9] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*, pages 11184–11193, 2022. 1, 2, 6, 7

[10] Zheng Ding, Yifan Xu, Weijian Xu, Gaurav Parmar, Yang Yang, Max Welling, and Zhuowen Tu. Guided variational autoencoder for disentanglement learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*, pages 7917–7926, 2020. 3

[11] Carl Doersch. Tutorial on variational autoencoders. *arXiv eprint arXiv:1606.05908*, abs/1606.05908, 2016. 3

[12] Zhengcong Fei, Mingyuan Fan, Li Zhu, Junshi Huang, Xiaoming Wei, and Xiaolin Wei. Masked auto-encoders meet generative adversarial networks and beyond. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*, pages 24449–24459, 2023. 1

[13] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*, pages 2414–2423, 2016. 2

[14] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems, NeurIPS*, pages 2672–2680, 2014. 1

[15] Zheng Gu, Wenbin Li, Jing Huo, Lei Wang, and Yang Gao. LoFGAN: fusing local representations for few-shot image generation. In *IEEE/CVF International Conference on Computer Vision, ICCV*, pages 8443–8451, 2021. 1, 5, 6, 7

[16] Jiaming Han, Yuqiang Ren, Jian Ding, Ke Yan, and Gui-Song Xia. Few-shot object detection via variational feature aggregation. In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI*, pages 755–763, 2023. 1

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*, pages 770–778, 2016. 2, 7

[18] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems, NeurIPS*, pages 6626–6637, 2017. 6

[19] Minui Hong, Jinwoo Choi, and Gunhee Kim. Stylemix: Separating content and style for enhanced data augmentation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 14862–14870, 2021. 1

[20] Yan Hong, Li Niu, Jianfu Zhang, and Liqing Zhang. Matchinggan: Matching-based few-shot image generation. In *International Conference on Multimedia and Expo (ICME)*, pages 1–6, 2020. 1, 2, 6, 7

[21] Yan Hong, Li Niu, Jianfu Zhang, Weijie Zhao, Chen Fu, and Liqing Zhang. F2GAN: fusing-and-filling GAN for few-shot image generation. In *International Conference on Multimedia (MM)*, pages 2535–2543, 2020. 1, 2, 6, 7

[22] Xun Huang and Serge J. Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *IEEE/CVF International Conference on Computer Vision, ICCV*, pages 1510–1519, 2017. 2, 4

[23] Ananya Harsh Jha, Saket Anand, Maneesh Singh, and V. S. R. Veeravasarapu. Disentangling factors of variation with cycle-consistent variational auto-encoders. In *European Conference on Computer Vision, ECCV*, pages 829–845, 2018. 3

[24] Liming Jiang, Bo Dai, Wayne Wu, and Chen Change Loy. Deceive D: adaptive pseudo augmentation for GAN training with limited data. In *Advances in Neural Information Processing Systems, NeurIPS*, pages 21655–21667, 2021. 1

[25] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*, pages 8107–8116, 2020. 1

[26] Junsik Kim, Tae-Hyun Oh, Seokju Lee, Fei Pan, and In So Kweon. Variational prototyping-encoder: One-shot learning with prototypical images. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*, pages 9462–9470, 2019. 3

[27] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *2nd International Conference on Learning Representations, ICLR*, 2014. 5

[28] Lingxiao Li, Yi Zhang, and Shuhui Wang. The euclidean space is evil: Hyperbolic attribute editing for few-shot image generation. In *IEEE/CVF International Conference on Computer Vision, ICCV*, pages 22714–22724, 2023. 1, 2, 6, 7

[29] Pan Li, Lei Zhao, Duanqing Xu, and Dongming Lu. Optimal transport of deep feature for image style transfer. In *Proceedings of the 4th International Conference on Multimedia Systems and Signal Processing, ICMSSP 2019*, pages 167–171, 2019. 2

[30] Weixin Liang, Zixuan Liu, and Can Liu. DAWSON: A domain adaptive few shot generation framework. *arXiv preprint arXiv:2001.00576*, 2020. 1, 2, 6, 7

[31] Xudong Lin, Yueqi Duan, Qiyuan Dong, Jiwen Lu, and Jie Zhou. Deep variational metric learning. In *European Conference on Computer Vision, ECCV*, pages 714–729, 2018. 3

[32] Ming-Yu Liu, Xun Huang, Arun Mallya, Tero Karras, Timo Aila, Jaakko Lehtinen, and Jan Kautz. Few-shot unsupervised image-to-image translation. In *IEEE/CVF International Conference on Computer Vision, ICCV*, pages 10550–10559, 2019. 6, 7

[33] Ming Lu, Hao Zhao, Anbang Yao, Yurong Chen, Feng Xu, and Li Zhang. A closed-form solution to universal style transfer. In *IEEE/CVF International Conference on Computer Vision, ICCV*, pages 5951–5960, 2019. 2

[34] Andrew L Maas, Awni Y Hannun, and Andrew Y Ng. Rectifier nonlinearities improve neural network acoustic models. In *Proceedings of International Conference on Machine Learning, ICML*, page 3, 2013. 6

[35] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Sixth Indian Conference on Computer Vision, Graphics & Image Processing, ICVGIP*, pages 722–729, 2008. 6, 7

[36] Augustus Odena, Christopher Olah, and Jonathon Shlens. Conditional image synthesis with auxiliary classifier gans. In *Proceedings of the 34th International Conference on Machine Learning, ICML*, pages 2642–2651, 2017. 5

[37] Jannick P. Rolland, V. Vo, B. Bloss, and Craig K. Abbey. Fast algorithms for histogram matching: Application to texture synthesis. *J. Electronic Imaging*, 9(1):39–45, 2000. 4

[38] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip H. S. Torr, and Timothy M. Hospedales. Learning to compare: Relation network for few-shot learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*, pages 1199–1208, 2018. 1

[39] Hao Tang, Xingwei Liu, Shanlin Sun, Xiangyi Yan, and Xiaohui Xie. Recurrent mask refinement for few-shot medical image segmentation. In *IEEE/CVF International Conference on Computer Vision, ICCV*, pages 3898–3908, 2021. 1

[40] Aad W Van der Vaart. *Asymptotic statistics*. Cambridge university press, 2000. 3

[41] Oriol Vinyals, Charles Blundell, Tim Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. Matching networks for one shot learning. In *Advances in Neural Information Processing Systems, NeurIPS*, pages 3630–3638, 2016. 1

[42] Pierre Wilmot, Eric Risser, and Connelly Barnes. Stable and controllable neural texture synthesis and style transfer using histogram losses. *arXiv preprint arXiv:1701.08893*, 2017. 3

[43] Shaoan Xie, Yanwu Xu, Mingming Gong, and Kun Zhang. Unpaired image-to-image translation with shortest path regularization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*, pages 10177–10187, 2023. 6

[44] Jingyi Xu, Hieu Le, Mingzhen Huang, ShahRukh Athar, and Dimitris Samaras. Variational feature disentangling for fine-grained few-shot classification. In *IEEE/CVF International Conference on Computer Vision, ICCV*, pages 8792–8801, 2021. 3

[45] Mengping Yang, Zhe Wang, Ziqiu Chi, and Wenyi Feng. Wavegan: Frequency-aware GAN for high-fidelity few-shot image generation. In *European Conference on Computer Vision, ECCV*, pages 1–17, 2022. 2, 6, 7

[46] Han Zhang, Ian J. Goodfellow, Dimitris N. Metaxas, and Augustus Odena. Self-attention generative adversarial networks. In *Proceedings of the 36th International Conference on Machine Learning, ICML*, pages 7354–7363, 2019. 5

[47] Jian Zhang, Chenglong Zhao, Bingbing Ni, Minghao Xu, and Xiaokang Yang. Variational few-shot learning. In *IEEE/CVF International Conference on Computer Vision, ICCV*, pages 1685–1694, 2019. 3

[48] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*, pages 586–595, 2018. 6

[49] Yabin Zhang, Minghan Li, Ruihuang Li, Kui Jia, and Lei Zhang. Exact feature distribution matching for arbitrary style transfer and domain generalization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*, pages 8025–8035, 2022. 1, 2

[50] Shengyu Zhao, Zhijian Liu, Ji Lin, Jun-Yan Zhu, and Song Han. Differentiable augmentation for data-efficient GAN training. In *Advances in Neural Information Processing Systems, NeurIPS*, 2020. 1

[51] Chenxi Zheng, Bangzhen Liu, Huaidong Zhang, Xuemiao Xu, and Shengfeng He. Where is my spot? few-shot image generation via latent subspace optimization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*, pages 3272–3281, 2023. 1, 2, 6, 7

[52] Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Domain generalization with mixstyle. In *9th International Conference on Learning Representations, ICLR*, 2021. 4

[53] Yingbo Zhou, Zhihao Yue, Yutong Ye, Pengyu Zhang, Xian Wei, and Mingsong Chen. Eqgan: Feature equalization fusion for few-shot image generation. *arXiv preprint arXiv:2307.14638*, 2023. 1