Fast Adaptive Meta-Learning for Few-shot Image Generation

Aniwat Phaphuangwittayakul, Yi Guo, and Fangli Ying

***Abstract*—Generative Adversarial Networks (GANs) are capable of effectively synthesising new realistic images and estimating the potential distribution of samples utilising adversarial learning. Nevertheless, conventional GANs require a large amount of training data samples to produce plausible results. Inspired by the capacity for humans to quickly learn new concepts from a small number of examples, several meta-learning approaches for the few-shot datasets are presented. However, most of meta-learning algorithms are designed to tackle few-shot classification and reinforcement learning tasks. Moreover, the existing meta-learning models for image generation are complex, thereby affecting the length of training time required. Fast Adaptive Meta-Learning (FAML) based on GAN and the encoder network is proposed in this study for few-shot image generation. This model demonstrates the capability to generate new realistic images from previously unseen target classes with only a small number of examples required. With 10 times faster convergence, FAML requires only one-fourth of the trainable parameters in comparison baseline models by training a simpler network with conditional feature vectors from the encoder, while increasing the number of generator iterations. The visualisation results are demonstrated in the paper. This model is able to improve few-shot image generation with the lowest FID score, highest IS, and comparable LPIPS to MNIST, Omniglot, VGG-Faces, and *mini*ImageNet datasets.**

***Index Terms*—Meta-Learning, Few-shot Image Generation, Generative Adversarial Network, Unsupervised Learning.**

I. Introduction

In recent years, the development of deep generative models has led to rapid progress in the field of image generation [1]–[7]. Generative Adversarial Networks (GANs) [8] are commonly used in the research community, especially for generative tasks since they are useful for bridging the gap between the natural human learning process and artificial intelligence. They are also capable of understanding images and generating new samples based on a random latent vector, subsequently creating a variety of different images. Despite demonstrating significant processing ability, GAN models require a far greater amount of training data points in comparison to humans [9]. By applying the previous experience, the natural human learning process recognises not only the existing knowledge but also new knowledge. In order to deal with a small amount of data called few-shot data, GANs have been redesigned to perform few-shot image generation tasks [10], [11].

Meta-learning is mainly applied in the design of models that are capable of developing new skills or rapidly adapting the existing knowledge to new environments with only a small number of training examples required [12]. At the present time, Model Agnostic Meta Learning (MAML) [13] is a widely used approach to few-shot meta-learning. Nevertheless, MAML is more suitable for addressing supervised learning problems since the quality of the model is determined by the direction of loss function for each task. It remains a challenge to train MAML in unsupervised learning problems. Based on embedded clustering method, CACTUs-MAML [14] is designed to perform unsupervised meta-learning in the absence of labelled data.

The biased initial model could be encountered in the MAML-based approaches during meta-training phase towards some tasks. Task-Agnostic Meta-Learning (TAML) [15] avoids bias and improves the generalisability of a meta-learner by introducing the concepts of entropy and inequality-minimisation. Even though TAML can produce satisfactory results on a wide range of classification tasks, modifications are still required for other tasks, such as generation. Meta-SGD [16] is the first approach to adopt Stochastic Gradient Descent (SGD) as a meta-leaner for carrying out few-shot learning in regression, classification, and reinforcement learning tasks. Despite the effectiveness of learning for regression, classification, and reinforcement learning tasks, the above-mentioned meta-learning approaches are unfit for few-shot image generation. To apply SGD as a meta-learner with the GAN model, Few-shot Image Generation with Reptile (FIGR) [17] has been introduced as a baseline model for few-shot learning on image generation through meta-learning. Although FIGR is capable of generating plausible images from an unseen class, training the model is a time-consuming and expensive process due to the complexity of the network. Besides, this approach is limited and can produce only a few varieties of images due to mode collapse [18].

In order to address the aforementioned drawbacks, Fast Adaptive Meta-Learning (FAML) based on the GAN model is proposed for few-shot image generation, since it learns

A. Phaphuangwittayakul is now with the Department of Computer Science and Engineering, East China University of Science and Technology, Shanghai, 200237, China (email: aniwat.pha@gmail.com).

Y. Guo is now with the Department of Computer Science and Engineering, East China University of Science and Technology, Shanghai, 200237, China, Business Intelligence and Visualisation Research Center, National Engineering Laboratory for Big Data Distribution and Exchange Technology, Shanghai, 200436, China, and also with Shanghai Engineering Research Center of Big Data & Internet Audience, Shanghai, 200272, China (email: guoyi@ecust.edu.cn).

F. Ying is now with Department of Computer Science and Engineering, East China University of Science and Technology, Shanghai, 200237, China, and also with State Key Laboratory of Bioreactor Engineering, East China University of Science and Technology, Shanghai, China (email: yfangli@ecust.edu.cn).

Aniwat Phaphuangwittayakul and Fangli Ying contributed equally to this work. Corresponding author: Yi Guo

to generate reasonable output images significantly faster than the baseline models using only a few image examples. The FAML model is redesigned using two latent vectors, based on whether the discriminator or generator of the GAN model are established and the number of iterations. This model not only emphasises the generation of images from the current task but also samples the images with a wider range of diversity from unseen classes. The model in this study can not only adapt to the binary image datasets (MNIST and Omniglot) but also can be applied to colour image datasets (VGG-Faces, and *mini*ImageNet), and the results measured both quantitatively and qualitatively. The main contributions of this study are summarised as follows:

- A novel generative adversarial approach called FAML is proposed. The model is intended for few-shot image generation based on meta-learning. It is over ten times faster for model convergence and requires a much smaller number of parameters than baseline methods.
- An unsupervised meta-learning model for high dimensional few-shot datasets (VGG-Faces and *mini*ImageNet) is designed in this work. The model can retain the original input information without any additional labels by conditioning the feature vectors from the encoder to GAN model.
- The FAML outperforms existing methods in both quality and diversity of generated output images as demonstrated by extensive experiments.

II. Related Work

A. Few-Shot Learning

Artificial Intelligence (AI) is powerful, but is unable to adapt or transfer prior knowledge to a new task swiftly and efficiently like human intelligence. It is necessary to bridge the gap between AI and human-like learning processes, representing a new research direction for the advancement of AI technology. Considering the limitations of supervised annotations, Few-Shot Learning (FSL), also known as one-shot learning, was proposed by [19] to solve the problem of learning from a limited number of examples. The concept of FSL has been studied from the perspective of both supervised and unsupervised learning. Furthermore, FSL is capable of learning a new task with limited information by utilising prior knowledge [20]. Over the past few years, few-shot learning has been applied to image recognition [21] and classification [22], [23] to determine whether it is similar to the learning process of humans.

B. Meta-Learning

The meta-learning algorithm is designed for use on a batch of training tasks $\{t_i\}$ for solving a problem t_{test}. In other words, the training experience acquired by the algorithm is used to undertake the task t_{test}. To solve an *N*-way *K*-shot classification problem in a task t_{test} with the meta-training set *D*, the procedure involves a certain number of episodes. An episode is composed of a classification task t_i similar to task t_{test} requiring a solution. As a result, the meta-learning algorithm learns to map between the support set, query set, and label, while the traditional learning classification algorithm learns to map between the image and label. Notably, meta-learning can be adapted using the experience acquired by the meta-knowledge extracted from a previous learning episode in a single dataset, and/or from different domains or problems [24].

A variety of theories behind meta-learning on few-shot learning have been proposed and categorised depending on their learning algorithms. (1) Memory-augmented network [25]–[27] is based on the idea of using stored information from a previous image classification. This method relies on Recurrent Neural Networks to learn and maintain the relevant information in the historical data. (2) Metric learning classifies query images by comparing their embedding to that computed from support set images. The Siamese network algorithm [28] marks the first attempt at performing few-shot classification with the metric calculation of a distance function over objects. Additionally, other approaches based on metric learning [29]–[31] have been proposed. (3) Gradient-based meta-learners [13] learn an efficient way to fine-tune a convolutional neural network in the support set to accurately classify the query set. (4) Data generation [32] relies on techniques such as the augmented metric learning algorithm or applied generative model to generate additional data.

C. Generative Adversarial Network

Generative Adversarial Networks or GANs [8] are classed as deep neural network architectures consisting of two neural networks: generator *G* and discriminator *D*. Trained in an adversarial manner, whereby two neural networks play a game against each other to generate data for the simulation of large scale data distribution, GANs have achieved significant progress in several generative tasks such as image translation [33], [34] and super-resolution [35], [36]. This work differs from others involving classical GANs in that the generator is trained with two random noise vectors, conditioned with a feature vector extracted from the encoder. However, training GANs with conditional contexts as inputs generally suffers from the mode collapse issue [37]. Various methods have been proposed to solve this problem [38]–[40]. In this work, mode seeking regularisation [41] is adopted using two random noise vectors. With this technique, the model can prevent mode collapse. Furthermore, instead of modifying the number of discriminator iterations as in the previous work on RaGAN [42], the quality and diversity of samples generated from the model in this study is controlled by the number of generator iterations and two conditional latent vectors, respectively.

D. Few-Shot Image Generation

Most of the current meta-learning applications address classification problems by creating a classifier. The classifier learns the training data and recognises unseen data using small labelled examples [25], [29]–[31], [43]–[46], with meta-learning based on the few-shot technique applied to undertake generation tasks. The first effective approach [47] was proposed and illustrated by training a set of binary font samples

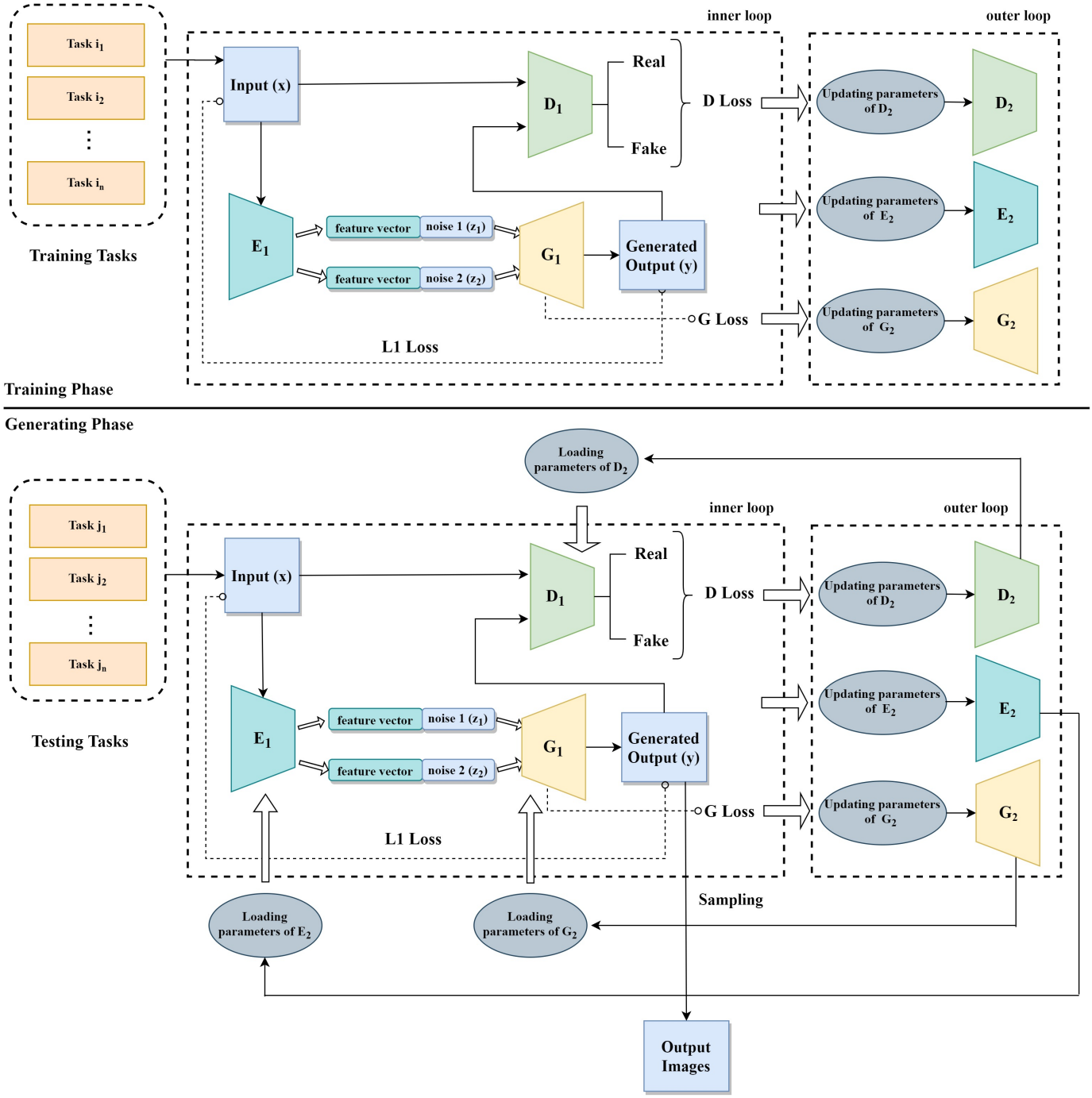


Fig. 1. The main process of our proposed FAML method. The upper part represents the training phase and the lower part refers to generating phase. The diagram consists of two independent neural networks in inner loop and outer loop with the same network topology.

with their strokes. The initial approach was improved using FIGR [17], training GAN model with a new meta-learning technique without additional stroke information. Despite producing remarkable results on binary images, it incurred high computation and was time-consuming for data training. The semi few-shot attribute translation approach [12] successfully carried out meta-learning in an attribute transfer generative network on a CelebA dataset. DAGAN [32] generated images by conditioning one image in the generator network with random noise. Even though DAGAN produced slightly diverse images, one conditioned image could not retrieve the information from multiple images in the same category. It can be observed that GANs can potentially be used in meta-learning for few-shot image generation, since they are capable of extracting key features from a few examples to perform the new generation tasks. In this work, a novel framework to capture the key features with only few examples required in meta-learning manner and learn to generate diverse realistic images with less training parameters, while preventing the mode collapse is proposed. The details of method are

explained in the following section. Unlike the existing works, this current study involves two independent neural networks with the same network topology in outer loop and inner loop, two conditional latent vectors, and a modified number of generator iterations.

III. Proposed Method

The FAML algorithm is proposed based on an encoder network and GANs. Figure 1 shows the flowchart of the meta-learning model in this study. The model is split into two main phases, the first of which is the training phase, and the second phase is the generating phase. Original images are sampled into training tasks τ_i and testing tasks τ_j with the same number of batch sizes prior to the start of the processes. The testing tasks are excluded from the training tasks. The training and generating phases are commonly used and consist of inner loop and outer loop. During the inner loop of the training phase, the encoder E_1 projects the input x from the training task τ_i, downsampled into a feature vector r, $r = E_1(x)$. The inputs of generator network G_1 are the two noise vectors (z_1 and z_2) concatenated with the extracted feature vector r from encoder E_1. The generator network outputs are the generated images, $y_1 = G_1(z_1, r)$ and $y_2 = G_1(z_2, r)$. The discriminator network D_1 classifies the real images (input x) and fake images (generated sample y) from generator network G_1. The global parameters of discriminator D_2, generator G_2, and encoder E_2 are updated in the outer loop by setting the gradient of Φ to be equal to $\Phi - W_\tau$. The outer loop will be put into operation once the k inner loop has been completed. During the generating phase, testing task τ_j is trained by adapting the discriminator D_1, generator G_1, and encoder E_1 parameters from the training phase to the model. It is intended that the model learns from a small amount of data and then applies it to a new task following the concept of meta-learning. The overall process of the meta-learning model in this study is described according to the pseudo-code in Algorithm 1.

The FAML meta-training algorithm is required during the training phase. The pseudo-code of the FAML meta-training algorithm is explained in Algorithm 2.

The FAML generation algorithm generates output images from the testing tasks and unseen data once the training process is completed. The FAML generation algorithm is described in Algorithm 3.

For every task τ, I images are sampled from input x and output y. Objective functions of the discriminator and generator are calculated as equations (1) and (2).

$$\mathcal{L}_D = -\mathbb{E}_{x_i \sim x}[\log(\tilde{D}(x_i)] - \mathbb{E}_{y_i \sim y}[\log(1 - \tilde{D}(y_i)] \quad (1)$$

$$\mathcal{L}_G = -\mathbb{E}_{y_i \sim y}[\log(\tilde{D}(y_i)] - \mathbb{E}_{x_i \sim x}[\log(1 - \tilde{D}(x_i)] \quad (2)$$

Where

$$\tilde{D}(x) = sigmoid(C(x) - \mathbb{E}_{y_i \sim y}\, C(y_i)) \quad (3)$$

$$\tilde{D}(y) = sigmoid(C(y) - \mathbb{E}_{x_i \sim x}\, C(x_i)) \quad (4)$$

i is the image index of I sampled images, $i \in [0, I]$. $C(a)$ is defined as a measure of how convincingly realistic the input a is. If the value is negative, this means the input data appears

Algorithm 1 Overall FAML process

The parameters of the implemented model in the paper are defaulted as: n_{iter} = 100,000

Require: n_{iter} number of global iterations, Φ_d the discriminator parameter vector, Φ_g the generator parameter vector, Φ_e the encoder parameter vector, W_d the discriminator weights, W_g the generator weights, W_e the encoder weights
Initialise Φ_d
Initialise Φ_g
Initialise Φ_e
for $i < n_{iter}$ **do**
    Make a copy of Φ_d resulting in D_2
    Make a copy of Φ_g resulting in G_2
    Make a copy of Φ_e resulting in E_2
    FAML meta-training algorithm
    if $mod(1000) == 0$ **then**
        Using W_d, a copy of the meta-trained Φ_d
        Using W_g, a copy of the meta-trained Φ_g
        Using W_e, a copy of the meta-trained Φ_e
        FAML generation algorithm
    end if
end for

Algorithm 2 FAML meta-training

The parameters of the implemented model in the paper are defaulted as: $n_D = 1, n_G = 5, m = 4, k = 10$

Require: n_D number of discriminator iterations, n_G number of generator iterations, m batch size of images, k number of inner loop iterations
Sample training task τ
Sample m images as x from X_τ
for $i < k$ **do**
    # Train discriminator D_1
    for $j < n_D$ **do**
        Generate random noise vectors z_1, z_2
        Generate fake images y_1, y_2 with z_1, z_2 and W_g
        Perform step of Adam update on W_d with discriminator loss between x and y_1, y_2
    end for
    # Train generator G_1 and encoder E_1
    for $j < n_G$ **do**
        Generate random noise vectors z_1, z_2
        Generate fake images y_1, y_2 with z_1, z_2 and W_g
        Perform step of Adam update on W_g with generator loss between x and y_1, y_2
        Perform step of Adam update on W_e with reconstruction loss between x and y_1, y_2
    end for
end for
Set Φ_d gradient of D_2 to be $\Phi_d - W_d$
Perform step of Adam update on D_2
Set Φ_g gradient of G_2 to be $\Phi_g - W_g$
Perform step of Adam update on G_2
Set Φ_e gradient of E_2 to be $\Phi_e - W_e$
Perform step of Adam update on E_2

Algorithm 3 FAML generation

The parameters of the implemented model in the paper are defaulted as: $n_D = 1, n_G = 5, m = 4, k = 10$

Require: n_D number of discriminator iterations, n_G number of generator iterations, m batch size of images, k number of inner loop iterations
Sample testing task τ
Sample m images as x from X_τ
for $i < k$ **do**
    # Train discriminator D_1
    for $j < n_D$ **do**
        Generate random noise vectors z_1, z_2
        Generate fake images y_1, y_2 with z_1, z_2 and W_g
        Perform step of Adam update on W_d with discriminator loss between x and y_1, y_2
    end for
    # Train generator G_1 and encoder E_1
    for $j < n_G$ **do**
        Generate random noise vectors z_1, z_2
        Generate fake images y_1, y_2 with z_1, z_2 and W_g
        Perform step of Adam update on W_g with generator loss between x and y_1, y_2
        Perform step of Adam update on W_e with reconstruction loss between x and y_1, y_2
    end for
end for
Generate random noise vector z
Generate fake images y

to be fake. The positive value defines greater naturalness of the input data. Since the output is generated according to two random noise vectors z_1, z_2, the loss functions of the discriminator and generator are modified and represented as the following equation.

The Discriminator objective function is:

$$\mathcal{L}_{D1} = -\mathbb{E}_{x_i \sim x}[\log(D_1(x_i)] - \mathbb{E}_{y_{1i} \sim y_1}[\log(1 - D_1(y_{1i})] \quad (5)$$

$$\mathcal{L}_{D2} = -\mathbb{E}_{x_i \sim x}[\log(D_1(x_i)] - \mathbb{E}_{y_{2i} \sim y_2}[\log(1 - D_1(y_{2i})] \quad (6)$$

$$\mathcal{L}_D = \frac{\mathcal{L}_{D1} + \mathcal{L}_{D2}}{2} \quad (7)$$

The Generator objective function is:

$$\mathcal{L}_{G1} = -\mathbb{E}_{y_{1i} \sim y_1}[\log(D_1(y_{1i})] - \mathbb{E}_{x_i \sim x}[\log(1 - D_1(x_i)] \quad (8)$$

$$\mathcal{L}_{G2} = -\mathbb{E}_{y_{2i} \sim y_2}[\log(D_1(y_{2i})] - \mathbb{E}_{x_i \sim x}[\log(1 - D_1(x_i)] \quad (9)$$

$$\mathcal{L}_G = \frac{\mathcal{L}_{G1} + \mathcal{L}_{G2}}{2} \quad (10)$$

The Encoder objective function is:

$$\mathcal{L}_{E1} = \sum_{i=1}^{I} \|x_i - y_{1i}\| \quad (11)$$

$$\mathcal{L}_{E2} = \sum_{i=1}^{I} \|x_i - y_{2i}\| \quad (12)$$

$$\mathcal{L}_E = \frac{\mathcal{L}_{E1} + \mathcal{L}_{E2}}{2} \quad (13)$$

For every sampled task τ including training tasks τ_i and generating tasks τ_j, FAML initialises the weights of the discriminator, generator, encoder while minimising the weights of the discriminator W_d, generator W_g, and encoder W_e in the k inner loop through objective functions $\mathcal{L}_D$, $\mathcal{L}_G$, and $\mathcal{L}_E$, respectively. The local minimum weights of the discriminator, generator, and decoder were achieved by minimising loss $\mathcal{L}_D$, $\mathcal{L}_G$, and $\mathcal{L}_E$ as in the following equations.

$$minimise \sum_{\tau} \sum_{k} \mathcal{L}_{D\tau} \quad (14)$$

$$minimise \sum_{\tau} \sum_{k} \mathcal{L}_{G\tau} \quad (15)$$

$$minimise \sum_{\tau} \sum_{k} \mathcal{L}_{E\tau} \quad (16)$$

The global parameters of discriminator Φ_d, generator Φ_g, and encoder Φ_e are updated by FAML in the outer loop through equation (17) by minimising the distance between the initial global parameters Φ_d, Φ_g, Φ_e and optimised weights W_d, W_g, W_e from the inner loop.

$$minimise \sum_{\tau} (\Phi_d - W_{d\tau}) + (\Phi_g - W_{g\tau}) + (\Phi_e - W_{e\tau}) \quad (17)$$

The total loss function our FAML can be written as

$$\mathcal{L} = \mathcal{L}_D + \mathcal{L}_G + \lambda_E \mathcal{L}_E + \lambda_{ms} \mathcal{L}_{ms} \quad (18)$$

Where $\mathcal{L}_{ms}$ is a mode seeking regularisation term [41] that maximises the ratio of the distance between $G_1(z_1, r)$ and $G_1(z_2, r)$ corresponding to the distance between z_1 and z_2 as equation (19).

$$\mathcal{L}_{ms} = \max_{G_1} \left(\frac{d(G_1(z_1, r)), G_1(z_2, r))}{d(z_1, z_2)} \right) \quad (19)$$

in which $d(\cdot, \cdot)$ is the L_1 norm distance metric. Following [41], we set the hyper-parameters $\lambda_{ms} = 1$ and set $\lambda_E = 1$.

The model is optimised and rapidly converges with a small amount of data with these four objective functions. Moreover, the model is capable of generalisation and it is possible to apply it on various sources of the dataset.

IV. Datasets

A. MNIST

The MNIST [48] dataset contains 10 classes or 10 digits in greyscale format, providing simplicity in the simulation of initial model ideas. Due to its characteristics, MNIST is widely used for the classification and image generation tasks in model experiments. For few-shot image generation, 10 dataset classes are divided into 10 tasks from 0 (τ_0) to 9 (τ_9). From task 0 (τ_0) to task 8 (τ_8) were selected as training tasks. Task 9 (τ_9) is a testing task. A total of images was 60,000. The goal of this study is to obtain a set of parameters Φ from the training tasks (τ_0) to (τ_8) for rapid adaptation into a testing task (τ_9).

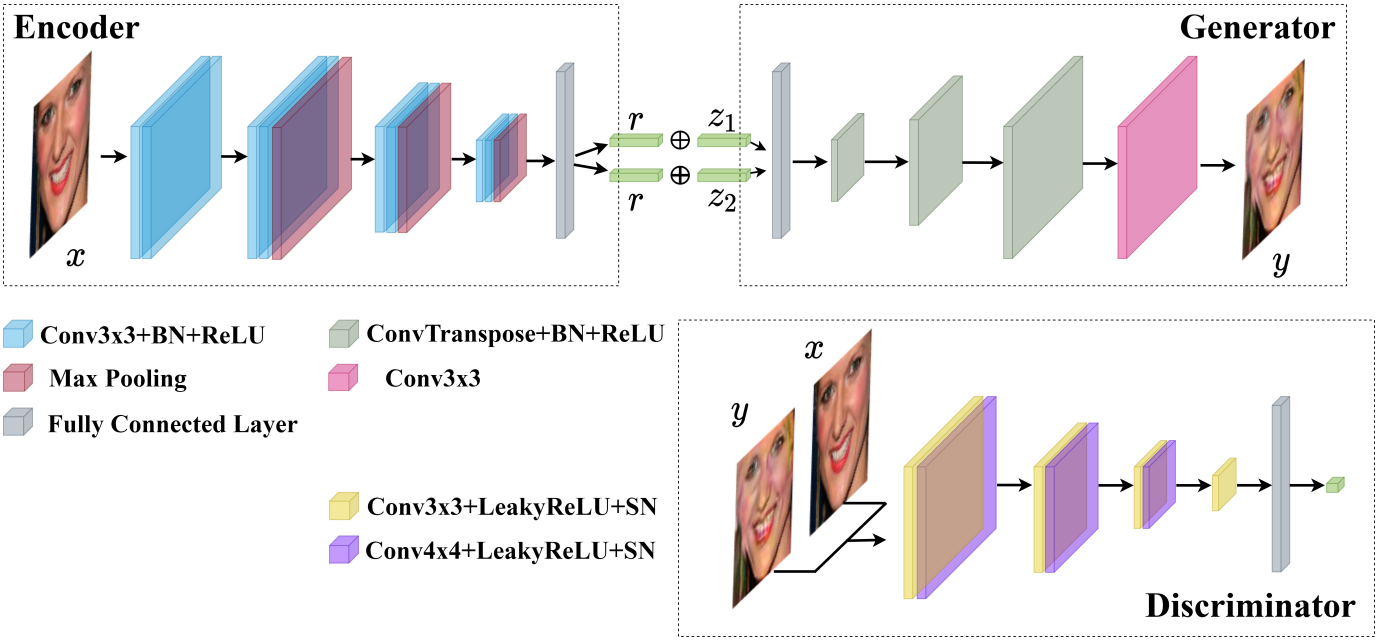


Fig. 2. The implementation of Encoder, Generator, and Discriminator network.

B. Omniglot

The Omniglot [47] dataset consists of 1,623 handwritten characters from 50 different alphabets. The Omniglot dataset is considered as the baseline dataset for few-shot learning in either as an image classification or image generation tasks. In this work, the dataset is divided into training classes with 1,603 characters and testing classes consisting of 20 characters. In contrast to the MNIST dataset, Omniglot contains a wider set of classes and supports the capability of the study model to achieve a highly complex greyscale dataset.

C. VGG-Faces

The VGG-Faces [49] dataset is a large-scale face dataset containing images from various identities with different ethnicities, accents, professions, and ages. From 2,369 classes, 20 samples from each class are randomly selected. The dataset is split into 2,269 training classes and 100 testing classes.

D. miniImageNet

The *mini*ImageNet [29] dataset is used for few-shot learning evaluation. The *mini*ImageNet is part of the full ImageNet [50] dataset. The *mini*ImageNet dataset contains 100 classes with 600 samples in each. The classes are separated into 80 training classes and 20 testing classes.

V. Experiments

A. Implementation

There are three networks in the method proposed in this study, as shown in Figure 2. The GAN model is initially designed based on a Relativistic average GAN (RaGAN) [42] with binary cross-entropy loss and encoder E adopting the structure of the U-Net [51] encoder. The encoder contains four residual blocks. The parameter details of each network can be found in Appendix A. According to the experiments, the output images depend on the number of Generator G iterations. The FAML is trained with one Discriminator D iteration and five Generator G iterations in every episode. Additionally, the network is trained with two random noise vectors z_1 and z_2 instead of one random noise vector following the mode seeking GAN [41]. The random noise vectors z_1, z_2 and feature vector r both have size of 128. There are four default sample images (I=4) in the training model. With this training technique, FAML increases the variation in the generated images and prevents mode collapse. Training with two random noise vectors z_1 and z_2, the loss functions of the network are modified. In every network, Adam optimisers [52] with $\beta_1 = 0.5$ and $\beta_2 = 0.999$ are utilised to update both the inner loop and outer loop gradients. Spectral normalisation (SN) is applied to the discriminator, not only to stabilise the training process but also to improve the quality of generated images. Spectral normalisation supports the model to generate more diverse and complex images than other weight normalisation [53].

The inner and outer learning rates are set at 0.0002 and 0.00001, respectively. All MNIST, Omniglot, VGG-Faces, and *mini*ImageNet training and evaluation images are normalised between the range -1 and 1, with image samples generated every 1,000 episodes (meta-learning steps). Plausible results appear after 5,000 episodes, with all experiments tested on a private cloud using a Tesla M10 graphics processing unit.

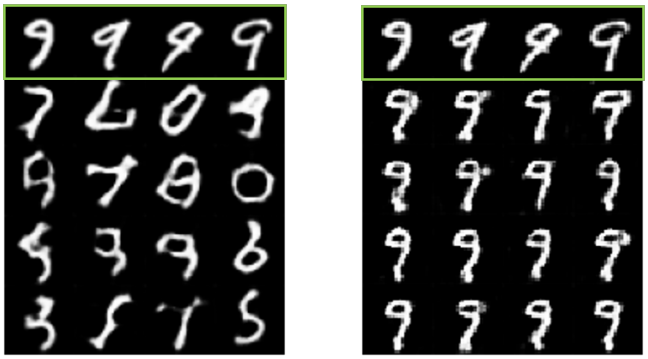
B. Empirical Validation

With the advantages of identified from prior work, the GAN model reveals a crucial knowledge in relation to the number of generator iterations and random noise vectors during the design of the model. For generation tasks in meta-learning, GAN shows more plausible generated images in comparison to

other types of generative models. In this work, it is noticeable that the output of the GAN model is determined by both the number of generator iterations as well as random noise vectors.

The generated images in this section are treated as the results of unseen classes in three datasets. The original images are presented in the first row of each figure and underlined in green. The subsequent four rows are the output images generated by the model after ten gradient steps a batch size of $(m) = 4$

1) MNIST: The MNIST dataset was separated into a training set and testing set. The training set included digit images from numbers 0 to 8. Only images of digit 9 were included in the testing set. Figure 3 provided a comparison between different types of optimisers. The model with the SGD optimiser still preserves information from the observed unsatisfied meta-learning data. As a result, the optimiser was changed from SGD to Adam since the latter can generate output images while maintaining critical information from unseen data.



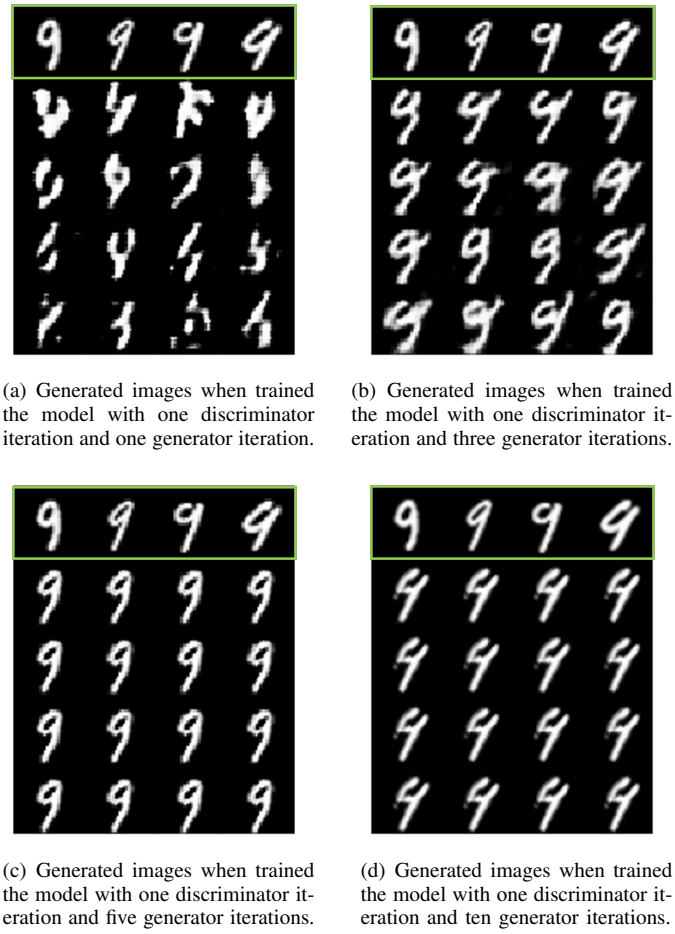
(a) Generated images based on SGD optimiser during training phase.

(b) Generated images based on Adam optimiser during training phase.

Fig. 3. Generated images of MNIST dataset from different types of the optimisers updating with 10 gradient steps of inner loop in 4,000 episodes.

Figure 4 shows the images generated using the model in this study with a varying number of generator iterations. The model was trained on tasks involving the generation of digit images from numbers 0 to 8 before applying the model to generate the image of digit 9. The model was trained in a few-shot learning manner since it has been established that with more generator iterations, the model can fit a small number of input images. A smaller number of iterations have been found to be favourable in maintaining high variation while retaining less information from the training images. In contrast, a large number of iterations reduced variance but increased bias. In other words, increasing the number of iterations in the generator can help to extract the key features from previous tasks where the digit images generated from numbers 0 to 8 focus on the current task of generating number 9.

Moreover, the size of the random noise vector is a crucial factor in generating diverse images. Figure 5 illustrates the generated images from the fine-tuned model in 1,200 episodes using a different number of random noise vectors. The images in Figure 5(b) generated from two random noise vectors show the diversity of generated images. In contrast, the images in Figure 5(a) create a set of similar output images due to mode collapse.



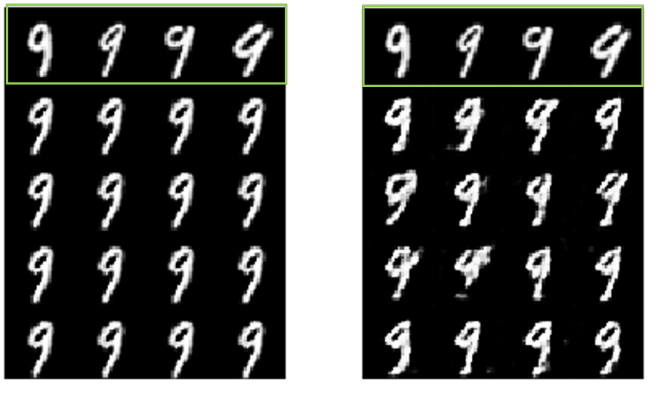
(a) Generated images when trained the model with one discriminator iteration and one generator iteration.

(b) Generated images when trained the model with one discriminator iteration and three generator iterations.

(c) Generated images when trained the model with one discriminator iteration and five generator iterations.

(d) Generated images when trained the model with one discriminator iteration and ten generator iterations.

Fig. 4. The generated images from the different number of generator iterations.



(a) Generated images from the model with a random noise vector z

(b) Generated images from the model with two random noise vectors z_1, z_2.

Fig. 5. The generated MNIST images from different number of random noise vectors with fine-tuned model in 1,200 episodes.

2) Omniglot: The Omniglot data were split into a training set and testing set. The training set was integrated with 1603 classes of character images while the testing set included 20

classes of character images. The study model produced to good results after 8,000 episodes, while the baseline model FIGR generated suitable outputs after 50,000 episodes. Samples of the generated images are presented in Figure 6.

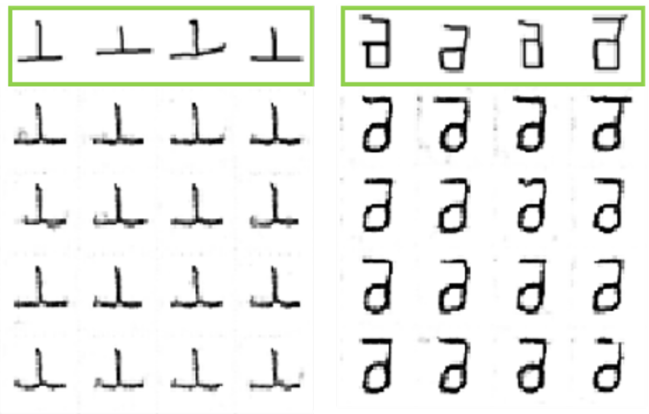


Fig. 6. The samples of generated images from the Omniglot dataset using our method in 8,000 episodes.

3) VGG-Faces: The VGG-Faces dataset was separated into 2,269 classes as a training set and 100 classes as a testing set. The generated images of VGG-Faces dataset from a different number of random noise vectors are presented in Figure 7. With two random noise vectors, the generated images in the last two rows of Figure 7 exhibit high variation compared to the samples with generated using a random noise vector (second row and third row of Figure 7). Therefore, the mode collapse problem can be prevented by using two random noise vectors. This solution is verified in both binary image (MNIST) and colour image (VGG-Faces) dataset. The quality of generated images was evaluated based on different methods. Figure 8 presents the images generated by DCGAN [54], DAGAN [32], FIGR [17], and the method in this study.

For visualisation comparison, the method under study produces more image variation and notable results than the baseline methods. The DCGAN captured the face components but produced blurry images, while DAGAN and FIGR generated only the face structure and without any related colour from the input faces. More visualisation results from VGG-Faces images from the proposed model can be found in Appendix B.

4) miniImageNet: The *mini*ImageNet dataset was categorised into a training set of 80 classes and a testing set of 20 classes. Figure 9 demonstrate the output images from the model without encoder network (second row and third row in Figure 9) and with encoder network (last two rows in Figure 9). With the encoder network, FAML maintains information on input images and significantly improves image quality over the model without the encoder network.

The quality of generated images was evaluated from different methods on the *mini*ImageNet dataset. Figure 10 presents the images generated using DCGAN, DAGAN, FIGR, and FAML. The FAML extracts colour and texture from input images in contrast to the other three methods. More generated images from the proposed model are shown in Appendix B.

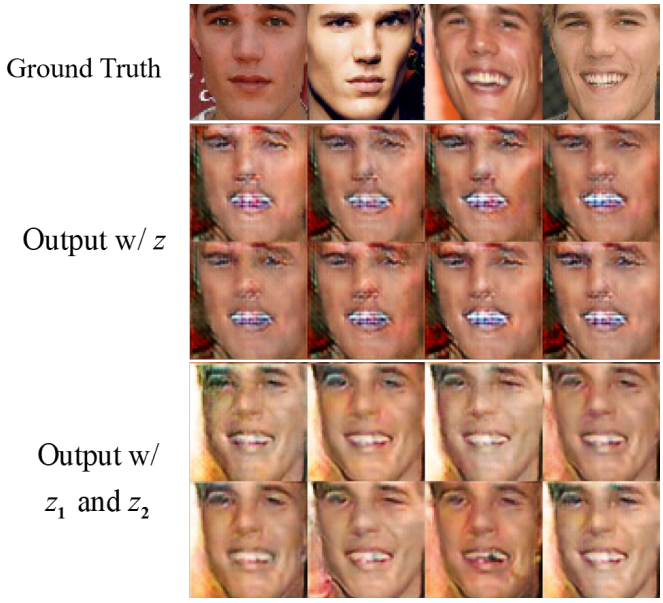


Fig. 7. The generated VGG-Face images from different number of random noise vectors with fine-tuned model in 70,000 episodes.

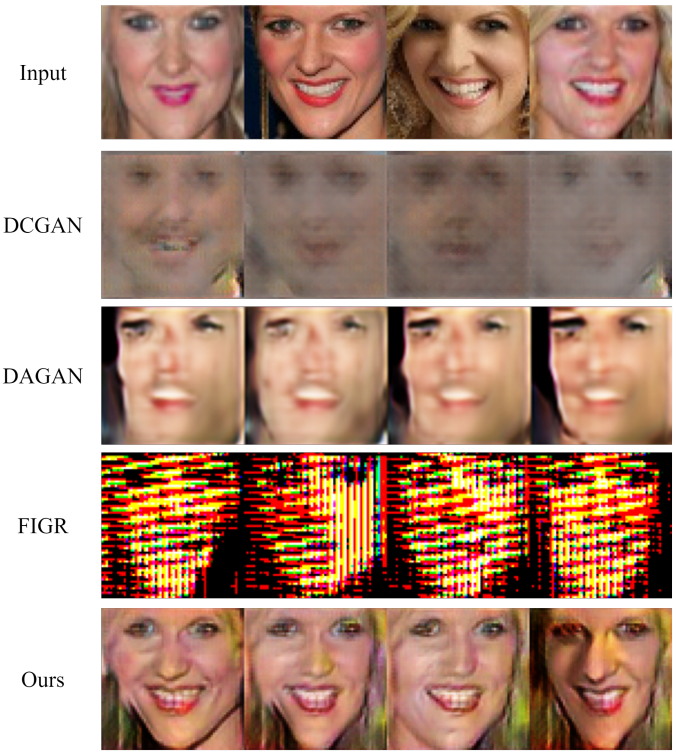


Fig. 8. Results of VGG-Faces generated by DCGAN, DAGAN, FIGR, and our method.

C. Evaluation Metrics

The quality of the generated images was evaluated using Frechet Inception Distance (FID) [55] metrics to measure the similarity between two datasets of images or Evidence Lower Bound (ELBO) in each dataset. The level of similarity indicates the distance between the generated images and the original images based on the feature extraction performed by the Inception Network [56]. A lower FID value suggests

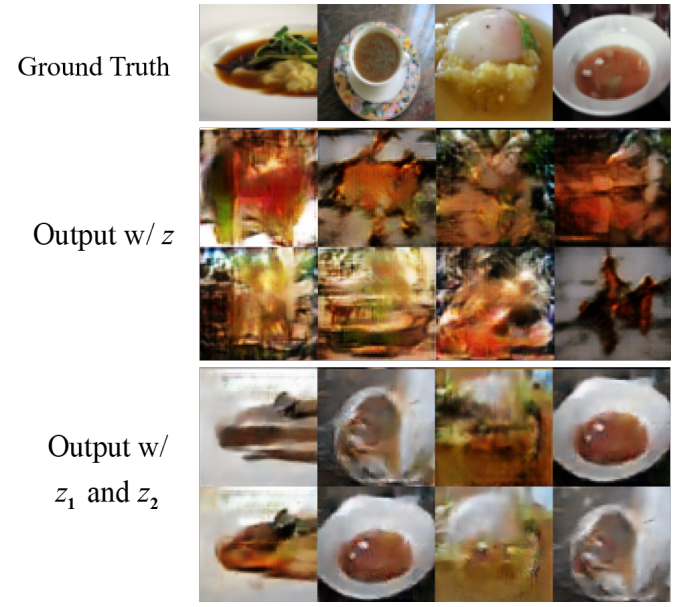


Fig. 9. The generated VGG-Face images from different number of random noise vectors with fine-tuned model in 70,000 episodes.

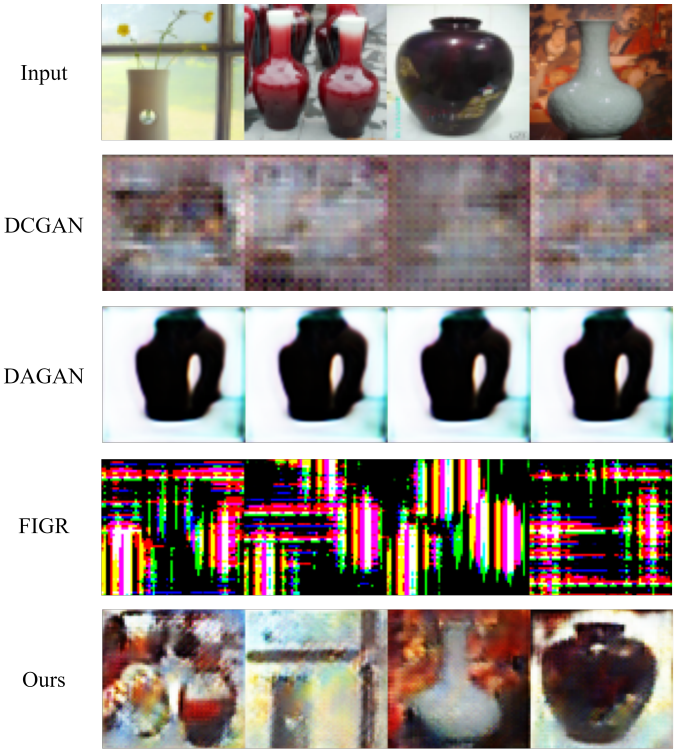


Fig. 10. Results of *mini*ImageNet generated by DCGAN, DAGAN, FIGR, and our method.

a better generated image quality. Moreover, the diversity of generated images was explored using the Inception Score (IS) [57]. The IS uses an Inception network [58] pre-trained on ImageNet and calculates a statistic of the network's outputs to classify the generated images. The higher IS Score, the more capable the model for producing distinctive images. The LPIPS [59] is used to measure the similarity between the original and generated images by calculating the pairwise perceptual distance. Lower LPIPS means greater similarity between two images.

The comparisons drawn between the model proposed in this study and baseline models (DCGAN, DAGAN, and FIGR) are detailed in the following tables. A total of 1,000 images were generated for each unseen testing class, with FID, IS, and LPIPS calculated for each method. The FID scores of the three models trained on the MNIST dataset are listed in Table I. The FID score of the current trained model is lower in comparison to FIGR when training both models for the same amount of time. Furthermore, the current trained model is comparable to the well-known FIGR but is less time-consuming.

TABLE I
FID OF BASELINE MODELS AND OUR MODEL BASED ON GENERATED IMAGES FROM MNIST DATASET.

Model	Time (Hours)	FID (↓)	#Episodes
DCGAN	7	242.44	77,000
	70	163.88	770,000
DAGAN	7	91.67	13,500
	70	83.63	135,000
FIGR	7	105.82	5,000
	70	38.41	50,000
Ours	7	**32.25**	5,500

The FID comparison between the proposed model, and FIGR, DAGAN, DCGAN in the Omniglot dataset are described in Table II. Since the images in the Omniglot dataset are integrated with broader classes and are more challenging for generation tasks than MNIST dataset, the FID is higher than the MNIST dataset. The model proposed in this study retains a lower FID and provides the high quality generated images compared to the baseline model.

TABLE II
FID OF BASELINE MODELS AND OUR MODEL BASED ON GENERATED IMAGES FROM OMNIGLOT DATASET.

Model	Time (Hours)	FID (↓)	#Episodes
DCGAN	7	179.78	77,000
	70	138.75	770,000
DAGAN	7	103.55	13,500
	70	73.37	135,000
FIGR	7	111.82	5,000
	70	87.15	50,000
Ours	7	**83.58**	5,500

The FID and IS comparison between the proposed and model, FIGR, DAGAN, DCGAN in the VGG-Faces dataset are presented in Table III. The traditional inception network was trained on 1,000 classes without human faces. Hence, a pre-trained network of VGG-Faces was used to calculate IS following [60]. The VGG-Faces are colour images and more challenging than MNIST and Omniglot datasets for meta-learning image generation tasks. Table III shows that FAML model achieves lower FID but higher IS than baseline models.

The FID and IS comparison between the proposed model and FIGR, DAGAN, DCGAN in the *mini*ImageNet dataset are demonstrated in Table IV. The *mini*ImageNet enriches the complexity and image variety in comparison to the VGG-Faces dataset. Table IV implies that the FID of the proposed

TABLE III
FID AND IS OF BASELINE MODELS AND OUR MODEL BASED ON GENERATED IMAGES FROM VGG-FACES DATASET.

Model	FID (↓)	IS (↑)	LPIPS (↓)
DCGAN	289.01	2.49	0.9333
DAGAN	196.12	2.22	**0.2674**
FIGR	251.96	3.95	0.5822
Ours	**42.72**	**7.10**	0.4011

method on the *mini*ImageNet dataset is lower and IS higher than baseline models.

TABLE IV
FID AND IS OF BASELINE MODELS AND OUR MODEL BASED ON GENERATED IMAGES FROM *mini*IMAGENET DATASET.

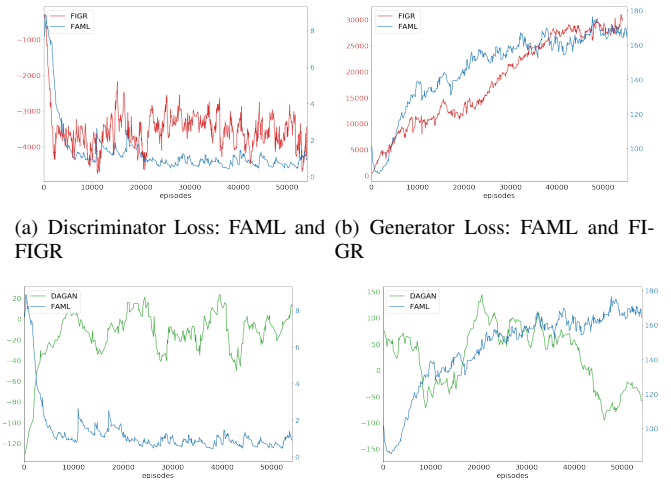
Model	FID (↓)	IS (↑)	LPIPS (↓)
DCGAN	108.20	3.11	0.2260
DAGAN	197.14	1.87	0.4314
FIGR	162.57	3.30	0.2902
Ours	**77.53**	**3.31**	**0.1708**

The results reveal a significant improvement in FAML compared to the baseline FIGR, DAGAN, and DCGAN models. When FAML is applied to the MNIST and Omniglot datasets, FAML converges to generate plausible images more than 10 times faster than FIGR, DAGAN, and DCGAN with the minimum FID. Accordingly, three baseline models and the proposed model were trained on VGG-Faces and *mini*ImageNet for the same length of time. FAML achieved the lowest FID, highest IS, and comparable LPIPS indicating that the proposed method could create more realistic and diverse images compared to baseline methods. Moreover, FAML significantly reduces the time spent training the model. Distinct from FIGR, the method proposed in this study is proven to be workable on colour images and outperforms the baseline methods (see Tables III and IV).

TABLE V
NUMBER OF PARAMETERS IN BASELINE MODEL FIGR AND OUR MODEL FAML WITH RANDOM NOISE VECTOR SIZE = 100.

Model	Networks	#Parameters
DAGAN	Discriminator	7.1M
	Generator	10.5M
	Total	17.6M
FIGR	Discriminator	18M
	Generator	17.5M
	Total	35.5M
Ours	Discriminator	2.9M
	Generator	4.6M
	Encoder	1.7M
	Total	9.2M

Furthermore, the discriminator $\mathcal{L}_D$ and generator loss $\mathcal{L}_G$ were computed as shown in Figure 11. A comparison was performed on the discriminator and generator losses between the competitive baseline models (DAGAN and FIGR) and the proposed FAML model. According to the adversarial theorem of GAN, discriminator loss decreases before stabilising while generator loss increases and remains mostly unchanged after a certain point. As for discriminator loss, FAML converges and



(a) Discriminator Loss: FAML and FIGR

(b) Generator Loss: FAML and FIGR

(c) Discriminator Loss: FAML and DAGAN

(d) Generator Loss: FAML and DAGAN

Fig. 11. The first row indicates the Discriminator loss and Generator loss of both FIGR and FAML models. The second row shows the Discriminator loss and Generator loss of both DAGAN and FAML models. The line and number red indicate the loss value of FIGR model. The line and number in green explain the loss value of DAGAN model. The line and number in blue denote the loss value of our FAML model. The graph was plotted every 100 episodes during model training process.

maintains consistency for longer than FIGR. Nevertheless, in contrast to DAGAN and FIGR, the slope of generator loss in FAML decreases and maintains stability while increasing the training episode.

As displayed in Table V, the number of parameters is calculated to evaluate network complexity. The FAML network reduces the number of parameters to one-fourth of FIGR and half of DAGAN. Thus, FAML requires less computation in each training iteration and generating phase.

VI. CONCLUSION

In this paper, a Fast Adaptive Meta-Learning (FAML) is proposed based on GAN and the encoder network for few-shot image generation. By applying a fast meta-learning algorithm, the model can quickly learn from a small amount of data and extract the key features of real distribution (original images) into feature vectors to generate new images according to the two random noise vectors conditioned using the extracted feature vectors. Furthermore, throughout the training on datasets with similar classes such as MNIST and Omniglot, the model demonstrated its ability to generate images from unseen classes with as few as four samples. In addition to the greyscale image datasets (MNIST and Omniglot), the proposed method is also applicable to high dimensional image datasets (VGG-Faces and *mini*ImageNet). Without hyperparameter tuning and additional labelling, the model can be trained in an unsupervised learning manner.

The proposed model was trained at a much faster pace than the baseline models. The model converged quickly due to the lower network structure complexity, change in learning rate, replacing the SGD with the Adam optimiser, and modification of the objective function. It was observed that the training process with one discriminator iteration and five generator

iterations enabled the proposed model to process a small amount of data and perform meta-learning. Training with a different number of generator iterations increased bias for the meta-learner, helping it to focus on the current task. The variety of generated images increased and the mode collapse problem was addressed.

The quality and diversity of the generated output images were measured through FID, IS, and LPIPS. When the number of training episodes increases, the proposed method achieves the lowest FID score, the highest IS, and comparable LPIPS. Moreover, it converges over ten times faster than the baseline models. According to the number of network parameters, the model required less computation and complexity due to the reduction in parameters to merely one-fourth of FIGR and half of DAGAN.

Even though the model proposed in this study achieves state of the art performance in greyscale and colour image datasets, it still requires further improvement to adapt to higher-quality datasets such as CelebA [61] and ImageNet [50]. Future work should focus on self-supervision [62] and the attention technique [63], which might address this problem.

APPENDIX A
DETAILS OF THE NETWORK ARCHITECTURE

The network architecture details for the encoder, generator, and discriminator are shown in this Appendix. The implementation code will be available on acceptance.

Encoder. There are four residual blocks in the proposed FAML encoder network. Each residual block contains two convolutional layers with batch normalisation and ReLU (Table VII) followed by a max pooling layer for downsampling input. The full encoder architecture is summarised in Table VI.

TABLE VI
THE NETWORK ARCHITECTURE OF OUR FAML ENCODER.

Layer	Resample	Activation	Output Shape
Image x	-	-	64*64*3
Residual Block	-	-	64*64*64
Residual Block	MaxPool	-	32*32*128
Residual Block	MaxPool	-	16*16*256
Residual Block	MaxPool	-	8*8*512
FC	-	Tanh	128
Encoded vector r	-	-	128

TABLE VII
THE NETWORK ARCHITECTURE OF A RESIDUAL BLOCK.

Layer	Norm	Activation
Input	-	-
Conv 3×3 Stride=1 Padding=1	BN	ReLU
Conv 3×3 Stride=1 Padding=1	BN	ReLU
Output	-	-

Generator. The generator and discriminator network are mainly implemented from [42], with the generator consisting of one fully connected (fc) layer followed by three transposed convolutional layers for upsampling and one convolutional layer. The generator input is the concatenation between a feature vector r extracted from the encoder network and random noise vector z. The architecture of the generator is summarised in Table VIII.

TABLE VIII
THE NETWORK ARCHITECTURE OF OUR FAML GENERATOR.

Layer	Norm	Activation	Output Shape
Vector (r, z)	-	-	256
FC	-	-	32,768
Reshape	-	-	8*8*512
ConvTranspose 4×4 Stride=2 Padding=1	BN	ReLU	16*16*256
ConvTranspose 4×4 Stride=2 Padding=1	BN	ReLU	32*32*128
ConvTranspose 4×4 Stride=2 Padding=1	BN	ReLU	64*64*64
Conv 3×3 Stride=1 Padding=1	-	Tanh	64*64*3

Discriminator. To increase the channel size, a convolutional layer is used with kernel size=3 followed by a convolutional layer with kernel size=4 and stride=2 for downsampling. Spectral normalisation and LeakyReLU activation were adopted. The architecture of the proposed discriminator network is summarised in Table IX.

TABLE IX
THE NETWORK ARCHITECTURE OF OUR FAML DISCRIMINATOR.

Layer	Norm	Activation	Output Shape
Image x / Gen Image y	-	-	64*64*3
Conv 3×3 Stride=1 Padding=1	SN	LeakyReLU	64*64*64
Conv 4×4 Stride=2 Padding=1	SN	LeakyReLU	32*32*64
Conv 3×3 Stride=1 Padding=1	SN	LeakyReLU	32*32*128
Conv 4×4 Stride=2 Padding=1	SN	LeakyReLU	16*16*128
Conv 3×3 Stride=1 Padding=1	SN	LeakyReLU	16*16*256
Conv 4×4 Stride=2 Padding=1	SN	LeakyReLU	8*8*256
Conv 3×3 Stride=1 Padding=1	SN	LeakyReLU	8*8*512
FC	-	-	1

APPENDIX B
FURTHER GENERATED RESULTS

As in Section V.B. in the paper, further generated results of the proposed FAML from VGG-Faces and *mini*ImageNet datasets are displayed in Figures 12 and 13, consecutively. Given four images from unseen classes, the model performs image generation based on the feature vectors extracted from the encoder and random noise vectors. The FAML can generate new diverse samples and maintain information on conditional images.

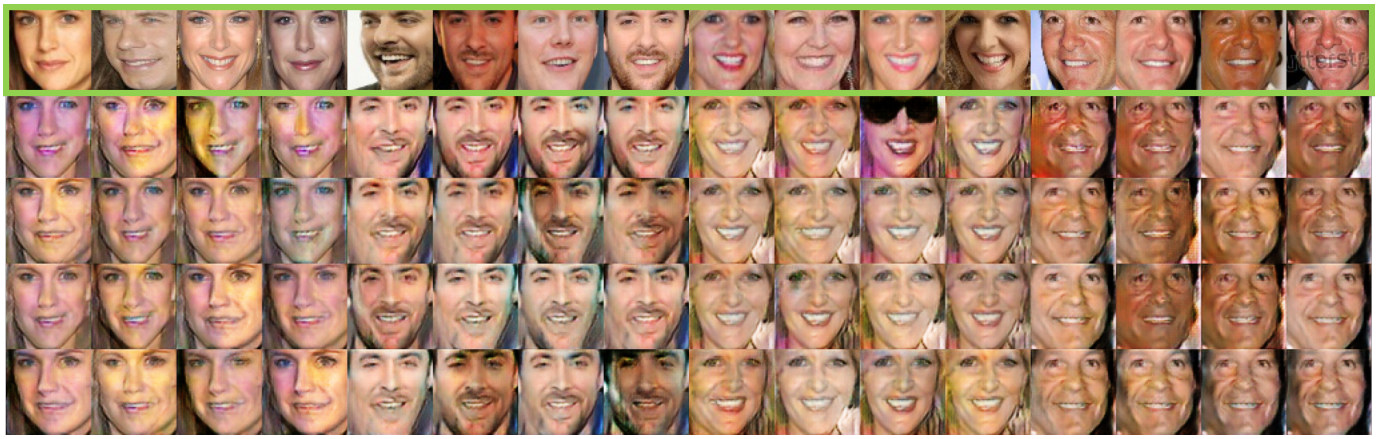


Fig. 12. Images generated by FAML with $m = 4; n_{iter} > 80,000$ from VGG-Faces dataset. There are four sets of input images. The conditional images are in green circle followed by four rows of output images.

Fig. 13. Images generated by FAML with $m = 4; n_{iter} > 80,000$ from *mini*ImageNet dataset. There are four sets of input images. The conditional images are in green circle followed by four rows of output images.

ACKNOWLEDGMENT

This research is financially supported by The National Key Research and Development Program of China (grant number 2018YFC0807105) and Science and Technology Committee of Shanghai Municipality (STCSM) (under grant numbers 17DZ1101003, 18511106602 and 18DZ2252300). Partially Supported by Open Funding Project of the State Key Laboratory of Bioreactor Engineering, East China University of Science and Technology, Shanghai, China; and International College of Digital Innovation (ICDI), Chiang Mai University, Thailand.

REFERENCES

[1] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4401–4410.

[2] Y. Guo, Q. Chen, J. Chen, Q. Wu, Q. Shi, and M. Tan, "Auto-embedding generative adversarial networks for high resolution image synthesis," *IEEE Transactions on Multimedia*, vol. 21, no. 11, pp. 2726–2737, 2019.

[3] X. Xia, R. Togneri, F. Sohel, and D. Huang, "Auxiliary classifier generative adversarial network with soft labels in imbalanced acoustic event detection," *IEEE Transactions on Multimedia*, vol. 21, no. 6, pp. 1359–1371, 2018.

[4] M. Lučić, M. Ritter, M. Tschannen, X. Zhai, O. F. Bachem, and S. Gelly, "High-Fidelity Image Generation With Fewer Labels," in *International Conference on Machine Learning*, 2019.

[5] J.-Y. Zhu, Z. Zhang, C. Zhang, J. Wu, A. Torralba, J. Tenenbaum, and B. Freeman, "Visual object networks: Image generation with disentangled 3D representations," in *Advances in neural information processing systems*, 2018, pp. 118–129.

[6] X. Lin, J. Li, H. Zeng, and R. Ji, "Font generation based on least squares conditional generative adversarial nets," *Multimedia Tools and Applications*, vol. 78, no. 1, pp. 783–797, 2019.

[7] Y. Wu, Z. Liu, and X. Zhou, "Saliency detection using adversarial learning networks," *Journal of Visual Communication and Image Representation*, vol. 67, p. 102761, 2020.

[8] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[9] R. Zhang, T. Che, Z. Ghahramani, Y. Bengio, and Y. Song, "Metagan: An adversarial approach to few-shot learning." *NeurIPS*, vol. 2, p. 8, 2018.

[10] S. Azadi, M. Fisher, V. G. Kim, Z. Wang, E. Shechtman, and T. Darrell, "Multi-content gan for few-shot font style transfer," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7564–7573.

[11] Y. Gao, Y. Guo, Z. Lian, Y. Tang, and J. Xiao, "Artistic glyph image synthesis via one-stage few-shot learning," *ACM Transactions on Graphics*, vol. 38, no. 6, 2019.

[12] R. Durall, F. J. Pfreundt, and J. Keuper, "Semi Few-Shot Attribute Translation," *International Conference Image and Vision Computing New Zealand*, 2019.

[13] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *International Conference on Machine Learning*, vol. 3, 2017, pp. 1856–1868.

[14] K. Hsu, S. Levine, and C. Finn, "Unsupervised Learning via Meta-Learning," in *International Conference on Learning Representations*, 2019.

[15] M. A. Jamal, G. Qi, and M. Shah, "Task-agnostic meta-learning for few-shot learning," *CoRR*, vol. abs/1805.07722, 2018. [Online]. Available: http://arxiv.org/abs/1805.07722

[16] Z. Li, F. Zhou, F. Chen, and H. Li, "Meta-sgd: Learning to learn quickly for few shot learning," *CoRR*, vol. abs/1707.09835, 2017. [Online]. Available: http://arxiv.org/abs/1707.09835

[17] L. Clouâtre and M. Demers, "Figr: Few-shot image generation with reptile," *CoRR*, vol. abs/1901.02199, Jan. 2019.

[18] Y. Saatci and A. G. Wilson, "Bayesian GAN," in *Annual Conference on Neural Information Processing Systems 2017*, 2017, pp. 3622–3631.

[19] L. Fei-Fei, R. Fergus, and P. Perona, "One-shot learning of object categories," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 4, pp. 594–611, 2006.

[20] Y. Wang, Q. Yao, J. Kwok, and L. M. Ni, "Generalizing from a Few Examples: A Survey on Few-Shot Learning," *ACM Computing Surveys (CSUR)*, apr 2019.

[21] Y. Zhu, W. Min, and S. Jiang, "Attribute-guided feature learning for few-shot image recognition," *IEEE Transactions on Multimedia*, 2020.

[22] G.-J. Qi, W. Liu, C. Aggarwal, and T. Huang, "Joint intermodal and intramodal label transfers for extremely rare or unseen classes," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 7, pp. 1360–1373, 2016.

[23] H. Huang, J. Zhang, J. Zhang, J. Xu, and Q. Wu, "Low-rank pairwise alignment bilinear network for few-shot fine-grained image classification," *IEEE Transactions on Multimedia*, 2020.

[24] C. Lemke, M. Budka, and B. Gabrys, "Metalearning: a survey of trends and technologies," *Artificial intelligence review*, vol. 44, no. 1, pp. 117–130, 2015.

[25] A. Santoro, S. Bartunov, M. Botvinick, D. Wierstra, and T. Lillicrap, "Meta-learning with memory-augmented neural networks," in *International conference on machine learning*, 2016, pp. 1842–1850.

[26] G.-S. Xie, L. Liu, X. Jin, F. Zhu, Z. Zhang, J. Qin, Y. Yao, and L. Shao, "Attentive region embedding network for zero-shot learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9384–9393.

[27] G.-S. Xie, L. Liu, F. Zhu, F. Zhao, Z. Zhang, Y. Yao, J. Qin, and L. Shao, "Region graph embedding network for zero-shot learning," in *European Conference on Computer Vision*. Springer, 2020, pp. 562–580.

[28] G. Koch, R. Zemel, and R. Salakhutdinov, "Siamese neural networks for one-shot image recognition," in *ICML deep learning workshop*, vol. 2, 2015.

[29] O. Vinyals, C. Blundell, T. Lillicrap, K. Kavukcuoglu, and D. Wierstra, "Matching networks for one shot learning," in *Advances in Neural Information Processing Systems*, 2016, pp. 3637–3645.

[30] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," in *Advances in Neural Information Processing Systems*, vol. 2017-Decem, 2017, pp. 4078–4088.

[31] F. Sung, Y. Yang, L. Zhang, T. Xiang, P. H. Torr, and T. M. Hospedales, "Learning to Compare: Relation Network for Few-Shot Learning," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1199–1208.

[32] A. Antoniou, A. Storkey, and H. Edwards, "Data augmentation generative adversarial networks," *International Conference on Learning Representations*, 2018.

[33] J. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *IEEE International Conference on Computer Vision, ICCV 2017*. IEEE Computer Society, 2017, pp. 2242–2251.

[34] Y. Viazovetskyi, V. Ivashkin, and E. Kashin, "Stylegan2 distillation for feed-forward image manipulation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, vol. 12367. Springer, 2020, pp. 170–186.

[35] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4681–4690.

[36] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. Change Loy, "Esrgan: Enhanced super-resolution generative adversarial networks," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 0–0.

[37] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans," *arXiv preprint arXiv:1606.03498*, 2016.

[38] Y. Zhao, Z. Jin, G.-j. Qi, H. Lu, and X.-s. Hua, "An adversarial approach to hard triplet generation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 501–517.

[39] G.-J. Qi, L. Zhang, H. Hu, M. Edraki, J. Wang, and X.-S. Hua, "Global versus localized generative adversarial nets," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 1517–1525.

[40] G.-J. Qi, "Loss-sensitive generative adversarial networks on lipschitz densities," *International Journal of Computer Vision*, vol. 128, no. 5, pp. 1118–1140, 2020.

[41] Q. Mao, H. Y. Lee, H. Y. Tseng, S. Ma, and M. H. Yang, "Mode seeking generative adversarial networks for diverse image synthesis," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2019-June, 2019, pp. 1429–1437.

[42] A. Jolicoeur-Martineau, "The relativistic discriminator: A key element missing from standard GaN," in *International Conference on Learning Representations*, 2019.

[43] X. Shu, G.-J. Qi, J. Tang, and J. Wang, "Weakly-shared deep transfer networks for heterogeneous-domain knowledge propagation," in *Proceedings of the 23rd ACM international conference on Multimedia*, 2015, pp. 35–44.

[44] J. Tang, X. Shu, Z. Li, G.-J. Qi, and J. Wang, "Generalized deep transfer networks for knowledge propagation in heterogeneous domains," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 12, no. 4s, pp. 1–22, 2016.

[45] Z. Peng, Z. Li, J. Zhang, Y. Li, G.-J. Qi, and J. Tang, "Few-shot image recognition with knowledge transfer," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 441–449.

[46] Y. Zou, Y. Shi, D. Shi, Y. Wang, Y. Liang, and Y. Tian, "Adaptation-oriented feature projection for one-shot action recognition," *IEEE Transactions on Multimedia*, 2020.

[47] B. Lake, R. Salakhutdinov, J. Gross, and J. Tenenbaum, "One shot learning of simple visual concepts," in *Proceedings of the annual meeting of the cognitive science society*, vol. 33, no. 33, 2011.

[48] Y. LeCun and C. Cortes, "MNIST handwritten digit database," 2010.

[49] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "Vggface2: A dataset for recognising faces across pose and age," in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, 2018, pp. 67–74.

[50] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[51] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[52] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," *International Conference on Learning Representations*, dec 2014.

[53] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, "Spectral normalization for generative adversarial networks," *International Conference on Learning Representations*, feb 2018.

[54] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *International Conference on Learning Representations*, nov 2016.

[55] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," in *Advances in Neural Information Processing Systems*, 2017, pp. 6626–6637.

[56] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[57] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans," in *Advances in neural information processing systems*, 2016, pp. 2234–2242.

[58] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.

[59] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 586–595.

[60] Z. Wang, X. Tang, W. Luo, and S. Gao, "Face aging with identity-preserved conditional generative adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7939–7947.

[61] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 3730–3738.
[62] L. Jing and Y. Tian, "Self-supervised visual feature learning with deep neural networks: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.
[63] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-attention generative adversarial networks," in *International Conference on Machine Learning*, 2019, pp. 7354–7363.

Aniwat Phaphuangwittayakul received B.Eng. from the Faculty of Computer Engineering, Chiang Mai University, Thailand, in 2012, and Master of Science from Beijing Institute of Technology, Beijing, China, in 2017. He is currently a Lecturer in the International College of Digital Innovation, Chiang Mai University as well as a Ph.D. candidate in East China University of Science and Technology, China. His current research interests are Meta-Learning, Deep Generative Model, Computer Vision and Artificial Intelligence.

Yi Guo received his M.Sc. degree in Computer Science from Xidian University, Xi'an, China and Ph.D. degree in Computer Science from Heriot-Watt University, Edinburgh, Scotland in 2005. He is currently a Professor at East China University of Science and Technology. His research concentrates on text mining, information extraction, knowledge discovery and business intelligence analysis. He is the member of Committee Board of National Engineering Laboratory for Big Data Distribution and Exchange Technologies and acts as senior members of IEEE, CMI, IET, BCS and an APMG-MSP/PRINCE2-Practitioner.

Fangli Ying received the B.S. degree from the Department of Software Engineering, Zhejiang University, Hangzhou, China, in 2009, and the Ph.D. degree from the Department of Computer Science, National University of Ireland, Maynooth, in 2014. He is currently a Lecturer in the Department of Computer Science at East China University of Science and Technology and he is also a visiting professor in the International College of Digital Innovation at Chiang Mai University, Chiang Mai, Thailand. His current research interests include computer vision, GIS and IoT for bioprocessing.