Towards Conditional Image Generation on Fine-Grained Classes

Yiming Zuo
Carnegie Mellon University
yzuo@cs.cmu.edu

Zhipeng Bao
Carnegie Mellon University
zbao@cs.cmu.edu

Abstract

In this report, we target to solve a challenging problem— the conditional image generation on fine-grained classes. We instantiate this framework on a state-of-the-art image generation network, SAGAN [37], which introduces the self-attention mechanism into GAN training. The proposed model contains three key components: the conditioning augmentation module, the category discrimination module and the latent reconstruction loss. We evaluate our methods on two standard fine-grained dayasets, Caltech-UCSD Birds (CUB) [31] and Stanford Dog (DOG) [16] dataset. Experimental results indicate that by introducing these auxiliary modules, our proposed method outperforms the SAGAN baseline on both dataset quantitatively and qualitatively. Additional ablation studies show that all individual components contribute to the full model. We have also discussed the potential research directions for the proposed framework.

1 Introduction

Image generation has been widely studied for decades. It developed quickly in the past few years with the thriving of deep learning and many interesting applications of it have emerged, such as generating a photo-realistic image given the sketch drawn by the user. Not only do people care about the quality of the image generated, but they also care about their control over the content, or the so-called conditional image generation. These conditions can sometimes be very fine-grained. For example, a user may query to generate a car of a specific brand, or a Chihuahua instead of a general dog. In these cases, the general labels are not enough and fine-grained class labels are needed.

The previous works on image generation can mainly be categorized into two sub-classes, i.e. unconditional image generation [37, 15] and conditional image generation [4, 20, 22, 27]. Unconditional image generation, although can genrate with diversity, has little control over the semantic properties of the images being generated [35]. While conditional image generation is much more powerful. It can model a conditional distribution of images over either class labels[4], sketches[14], low-resolution images[17], or even video frames[24]. In this work we mainly study the image generation conditioned on class label, which is the simplest case[21].

Conditioning on fine-grained labels introduces additional problems. The fine-grained labels are expensive to get in general, as they sometimes require the human labelers to possess expert knowledge in order to distinguish the sub-classes accurately, which can be very similar in their appearances. Therefore, the numbers of labeled images are often limited [31, 16]. This falls into the realm of **few-shot learning** [28, 8, 23], where the training set only consists of several images for each class. This problem is especially challenging since the image generation networks heavily rely on the quantity of training data [4]. Several methods based on meta learning or transfer learning have been proposed to tackle this problem [32, 28, 18]. In this project, we adopt a different perspective: designing a robust image-generation architecture that can be applied with few-shot fine-grained settings.

Preprint. Under review.

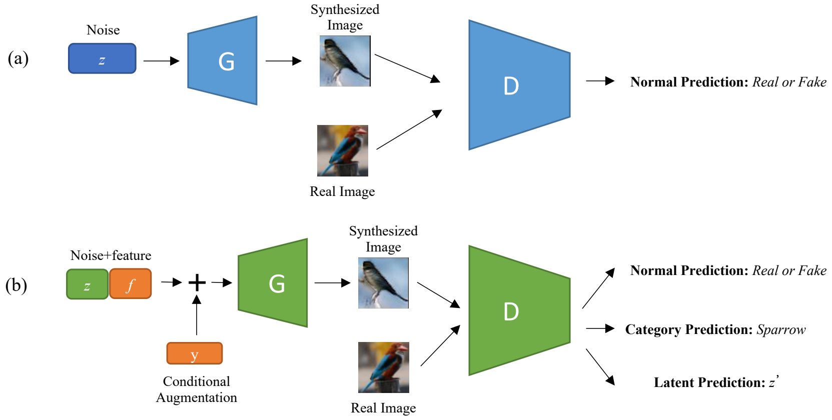


Figure 1: Pipeline of the proposed model. (a) the baseline image generation network; (b) our proposed conditional fine-grained image generation network. We added a Conditioning Augmentation block as the generator inputs, an additional classification loss, and an additional latent vector reconstruction loss.

Inspired by several recent papers, we propose a novel approach to solve the conditional image generation problem on fine-grained classes. We tested our method on two fine-grained dataset and showed that our best approach outperforms a baseline that achieves the state-of-the-art performance on ImageNet [7]. In the midway report we found that introducing an extra classification loss into the discriminator helps improve the quality and diversity of the image being generated. We further show in this report that adding a conditioning augmentation layer as the generator input and adding a reconstruction loss for the noise vector further boost the performance. We show both qualitative and quantitative results and do extensive ablation studies. The pipeline is shown in Figure 1.

2 Related Work

Image Generation There are a wide variety of image generation methods. A certain category is based on auto-regressive methods, such as fully visible belief network. One impressing work of this category is PixelCNN [29], which models the image generation task as a sequential sampling task, and it models the cognitional distribution of a pixel on the pixels that have already been generated, in a raster scan order.

Recently, with the fast development of deep neural networks, conditional image generation models have shown promising results for numerous tasks. Image-to-image translation tasks such as sketches → images[14], low-resolution images → high-resolution images[17], or even video frames → full videos[24]. label-conditioned models [4] have also been shown as a powerful tool for generating images of a specific category.

The most straightforward way to measure the quality of the image being generated is measuring the log-likelihood on the test set. While some methods have a tractable likelihood [22, 29, 26], a large proportion of image generation methods are based on Generative Adversarial Networks (GAN), for which the log-likelihood is hard to measure. Therefore people use other evaluation metrics to mimic human evaluation [39], such as Inception score (IS) [25] and the Frechet Inception distance (FID) [9]. Qualitative evaluation is also commonly used, for example in AMT test [14], where Turkers are asked to choose between a real image and a fake one.

Generative Adversarial Networks (GANs) Another popular image generation regime is based on GANs. GANs are known to have fater evaluation speed (as only one forward pass is needed), and is able to capture high-frequency details, and can generate photo-realistic images under high resolution [4]. The GAN training procedure corresponds to a minimax two-player game between a generator (G) and a discriminator (D), both of which are often parameterized as deep neural networks [10]:

$$\min_G \max_D V(D,G) = \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}(\boldsymbol{x})}[\log D(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{z}}(\boldsymbol{z})}[\log(1 - D(G(\boldsymbol{z})))] \tag{1}$$

The training of GANs are known to be notoriously unstable. And they are also suffering from mode collapse, when the model tends to ignore the diversity of the input noise vector. Following up works on GAN such as WGAN [1] and WGAN-GP [11] studies on how to make the GAN training process stable and robust. Other works such as BicycleGAN [40] focuses on the diversity of the generated images by enforcing a by-directional mapping from embedding to image.

Attention Mechanism The attention mechanism is first introduced in [2]. Despite being invented to solve the long-range dependencies in natural language processing (NLP) problems, it has also been widely applied in the field of computer vision [19, 37, 36]. It is known for its computational efficiency, and is used in the tasks that requires reasoning about global dependencies.

The attention layers usually consists of three separate "heads", namely *Queries(Q), Keys(K)* and *Values(V)*. Each of the head is a convolution layer with 1×1 kernel. The input features are transformed by the heads and the output feature is computed as the following:

$$\text{Attention}\ (Q, K, V) = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2}$$

Where Q, K, V are the features transformed using the corresponding head, and d_k is the embedding dimension of the *Keys*.

Fine-grained Image Generation With the development of image generation techniques, a more challenging task, fine-grained image generation, has raised attention from public [34]. This task aims at synthesizing images on fine-grained categories such as faces of a specific person or objects in a subordinate category. CVAE-GAN [3] first combines a variational auto-encoder with a generative adversarial network under a conditional generative process to tackle this problem. It achieves this goal by varying the fine-grained category fed into the resulting generative model. Several following generative networks also work on this sub-area [37]. The main challenge of this task is to detect and keep the key distinguishing features of each category, as well as dealing with the problem that only a limited number of samples are available at training time, which are also the main problems this project focuses on.

Few-shot Learning Few-shot learning studies the cases when target samples are limited [32]. These kind of methods are usually combined with transfer learning and meta learning since most of the time we need special algorithm design to deal with few-shot problems [28]. In the recent years, few-shot learning has shown its great ability of generalization for several fully supervised tasks such as classification and segmentation [8, 23]. Most recently, researchers also tried to combine few-shot learning with generative models. Wang *et al.*proposed to use a "hallucinator" to generate imaginary samples to help with few-shot classification [33]. MetaGAN [39] could help most few-shot learning problems by introducing fake images as a new category. Generative version of Matching Nets (GMN) extended the capability and robustness of the Matching Nets [30]. Recently, Liu *et al.* proposed FUNIT [18], a framework performing multi-domain image-to-image translation in the few-shot setting to ease the need for huge datasets in this task. They adopted a generator that takes in both a content image and a small set of style images, as well as a multi-class discriminator that outputs scores for multiple domains. In the inference time, the generator is fed with a content image and a small set of style images of a novel domain, and their results show that the generator can achieve the image translation to unseen domain well. However, the time and resources required to train their network is prohibitive.

3 Methods

We first explain the SAGAN [37] on which our method is based in section 3.1. In section 3.2 and section 3.3 we introduce several sub-modules that we adapted from. And finally in section 3.4 we propose a novel method.

3.1 Baselines

In this report, we use the Self-Attention GAN (SAGAN) [37] as the baseline, which is a widely used unconditional image generating network and is proved effective and robust. It achieved the state-of-the-art image generation results on ImageNet [7]. SAGAN itself is originally an unconditional generative network, but the authors have also made some initial implementations towards conditional image generation[1]. The resolution for the synthesized image is 64 at primary setting.

For the **Generator**, SAGAN is based on the resent-18 [12] backbone. It additionally inserts one self-attention layer between the original convolutional layers, where the features are firstly transformed into feature space f and g by linear projection, followed by

$$\beta_{j,i} = \frac{\exp\left(s_{ij}\right)}{\sum_{i=1}^{N} \exp\left(s_{ij}\right)}, \text{ where } s_{ij} = \boldsymbol{f}\left(\boldsymbol{x_i}\right)^T \boldsymbol{g}\left(\boldsymbol{x_j}\right) \tag{3}$$

and $\beta_{j,i}$ indicates the extent to which the model attends to the i^{th} location when synthesizing the j^{th} region.

As for conditioning on the class labels, the authors use conditional batch normalization modules [6] to replace the original batch normalization modules [13]. The conditional batch normalization module has the following form:

$$CBN\left(F_{i,c,h,w} \mid \gamma_c, \beta_c\right) = \gamma_c \frac{F_{i,c,w,h} - \mathrm{E}_{\mathcal{B}}\left[F_{\cdot,c,\cdot,\cdot}\right]}{\sqrt{\mathrm{Var}_{\mathcal{B}}\left[F_{\cdot,c,\cdot,\cdot}\right] + \epsilon}} + \beta_c \tag{4}$$

where γ_c and β_c are two parameters controlling the mean and variance of the normalization.

The **Discriminator** is also a multi-layer CNN followed by two fully connected layers. The adversarial loss function is the hinge loss:

$$\ell_{hinge}(y) = \max(0, 1 - t \cdot y) \tag{5}$$

3.2 Conditioning Augmentation

To solve the sparsity of training data problem in the few-shot learning setup, we adapted a method from a recent paper StackGAN [38]. In this paper, the authors aim to solve the text to image translation problem in a generative adversarial manner. Specifically, they use a stack of two generators: the first one focuses on capturing the low resolution information, while the second one refines the output of the first one.

In StackGAN the authors state that the limited number of training pairs results in sparsity in the text conditioning manifold. This is similar to the problem we are facing, where the number of training examples per class is very limited (on average 30 examples per class in the CUB dataset [31], in contrast to 500 examples per class in ImageNet [7]).

The author resolves this problem by introducing **Conditioning Augmentation (CA)**, which encourages smoothness in the latent conditioning manifold. This technique not only makes the training of GAN easier, but also increases the diversity of the generated image.

Specifically, the text embedding φ_t (or class embedding in our setup) is fed into a MLP to produce a mean and a diagonal variance matrix. An embedding vector is then sampled from the predicted Gaussian Distrinution $\mathcal{N}\left(\mu\left(\varphi_t\right), \Sigma\left(\varphi_t\right)\right)$. This variational setup introduces smoothness in the embedding space, and an additional loss term is applied to encourage the distribution to be similar to the standard Gaussian Distribution:

$$D_{KL}\left(\mathcal{N}\left(\mu\left(\varphi_t\right), \Sigma\left(\varphi_t\right)\right) \| \mathcal{N}(0, I)\right) \tag{6}$$

The sampled embedding vector is further concatenated with the noise vector, and they together serve as the input to the generator. Figure 2 illustrates how conditioning augmentation works.

3.3 Discriminator Design

In order to enable the unconditional image generation model to work under the conditional fine-grained generation scenarios, we propose two modifications on the discriminator.

[1] See `https://github.com/brain-research/self-attention-gan`

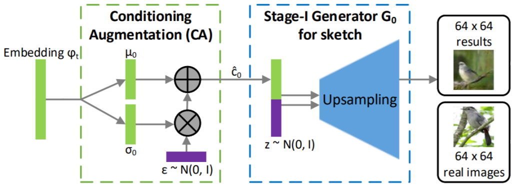


Figure 2: The Conditioning Augmentation layer. Figure directly adapted from StackGANs [38]
.

Firstly, We add a classification output for the discriminator to explicitly force it preserves categorical information. The classification loss is used to update both generator and discriminator. We use cross entropy loss function for such a category loss.

Furthermore, inspired by Chen *et.al.*[5], we introduce an identity regularizer to ensure that the synthesis images preserve the identity for the latent noisy input z so that can better keep the diversity of the synthesized images. We add an extra output layer that predicts z', which is the reconstruction of z. Then we minimize the difference between the real and the reconstructed input as

$$\mathcal{L}_{\text{id}}(G, H) = \mathbb{E}_z \|z - z'\|^2 \tag{7}$$

Here H is the hidden latent discriminator which shares the majority of the convolution layers of the discriminator D.

The adversarial loss is the hinge loss, same as original SAGAN design.

3.4 The Proposed Model

The proposed model is mainly based on the SAGAN [37]. We made the following 3 changes: 1) we added a conditioning augmentation block at the generator inputs 2) we added an additional classification loss 3) we added an additional latent vector reconstruction loss. The overall pipeline is shown in Figure 1.

4 Results

4.1 Implementation Details

Our code is based on the Pytorch Implementation of the SAGAN[2]. The Generator is based on resnet-18 [12] with one self-attention block inserted. The Discriminator has a similar resnet architecture. As for the conditioning augmentation layer, we pass the one-hot label into an embedding layer, followed by a fully connected layer and a leaky-ReLU activation, which is turned into mean and log-stddev of the Gaussian Distribution. We train each model for 50k steps, which takes about 8 hours on one RTX 2080Ti GPU.

4.2 Datasets

We conducted our experiments on two fine-grained datasets: **Caltech-UCSD Birds (CUB)** dataset [31] and **Stanford Dog (DOG)** dataset [16]. **CUB** dataset contains 6,033 images that belongs to 200 classes. **DOG** dataset contains 20,580 images belonging to 120 classes, most of which are from ImageNet [7]. Sample images from these two dataset are shown in Figure 3.

4.3 Evaluation Metrics

In this report we use the following two metrics to measure the quality of the images:

Inception Score is an objective metric for evaluating the quality of generated images. It seeks to capture both the quality and diversity of a collection of generated images [25]. The first term of IS mainly measures the quality of the generated images and the second term measures the diversity.

[2] See `https://github.com/voletiv/self-attention-GAN-pytorch`

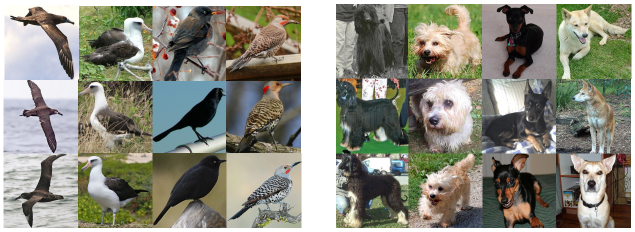


Figure 3: Sampled images from CUB and DOG dataset. Left: sampled images from CUB dataset; Right: sampled images from DOG dataset. Images in the same column belong to the same category.

FID Score is calculated by computing the Fréchet distance [9] between two Gaussians fitted to feature representations of the Inception network. It measures the quality of generated images and robust to noises.

4.4 Primary Results

We first compare our model with the baseline— SAGAN model. For the two compared models, we sampled 20 images for each class during all the experiments. We also report the results for *real images* which is the performance upper bound. The results for our full model and the baseline model on the two datasets are reported in Table 1.

	CUB			DOG		
Model	*Real Images*	SAGAN	Ours	*Real Images*	SAGAN	Ours
IS (↑)	4.56 ± 0.58	3.21 ± 0.42	$\mathbf{3.81 \pm 0.50}$	7.36 ± 0.78	2.14 ± 0.06	$\mathbf{3.64 \pm 0.32}$
FID (↓)	0.0	132.70	**74.05**	0.0	264.83	**221.76**

Table 1: Inception score and FID for SAGAN baseline and our model on CUB and DOG dataset. ↑ means larger is better and ↓ means smaller is better. Compared with SAGAN, the proposed model has a significantly better results for both metrics, indicating the design of our model is effective.

Furthermore, to have a comprehensive understanding of our model, we also sampled some generated images and compared the qualitative results with the baseline model, and the visualization is shown in Figure 4. Besides, Figure 5 further shows the generated images of some randomly picked categories. From these figures, we can see that the proposed model can generate significantly higher-quality images and it can maintain the category information for the challenging fine-grained classes.

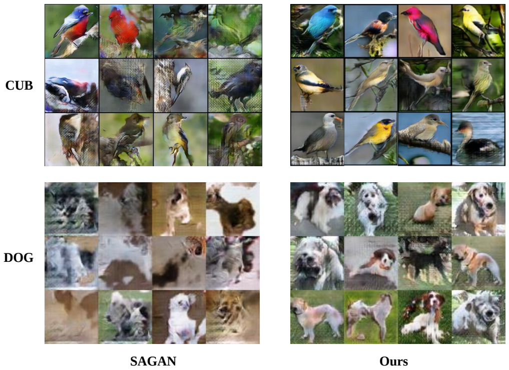


Figure 4: Comparison of the baseline method (SAGAN) and the proposed method. Results are shown on both **DOG** and **CUB** datasets.

.

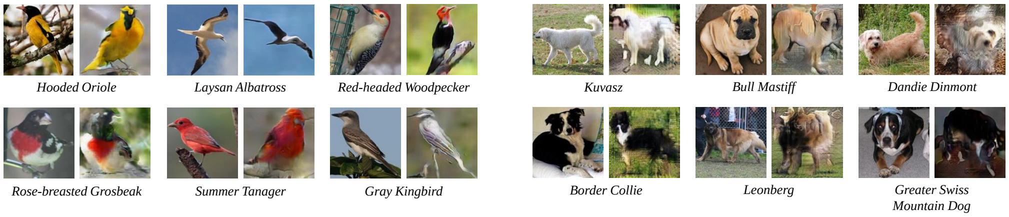


Figure 5: The image generated by our proposed method. For each class, one randomly picked real image is shown on the left, and one fake image is shown on the right. This figure shows that our model can maintain category information for fine-grained classes.

.

4.5 Ablation Studies

To understand the role of each components, we add each proposed components separately to the baseline model and measured their performances. **SAGAN** is the baseline model; **SAGAN-*cls*** is the baseline model equipped with the individual category discrimination module; **SAGAN-*latent*** is the baseline model equipped with the individual latent discrimination module; **SAGAN-*cat*** is the baseline model with a concatenation input embedding strategy; and **SAGAN-*ca*** is the baseline model equipped with the conditional augmentation embedding strategy. The quantitative results of these comparable models are reported in Table 2. We only did ablation studies on the **CUB** dataset where each class contains less images and we consider it a more challenging scenario under our setup.

Model	SAGAN	SAGAN-*cls*	SAGAN-*latent*	SAGAN-*cat*	SAGAN-*ca*	Ours
IS (↑)	3.21 ± 0.42	3.37 ± 0.47	3.21 ± 0.55	3.27 ± 0.06	3.29 ± 0.10	$\mathbf{3.81 \pm 0.50}$
FID (↓)	132.70	104.66	125.25	105.80	102.13	**74.05**

Table 2: Inception score and FID for different variance of the proposed models. ↑ means larger is better and ↓ means smaller is better. Compared with baseline, each individual component can bring improvements for both metrics.

Class Label Embedding Strategies We compare the influence of different class label embedding strategies on the quality of the generated images. **SAGAN-*cat*** has the same embedding layers as **SAGAN-*ca*** but we only take the mean output and ignore the variance part. We compare **SAGAN**, **SAGAN-*cat***, and **SAGAN-*ca*** qualitatively and quantitatively. Results are displayed in Table 2 and Figure 6. Results show that the variational method helps generate higher quality images.

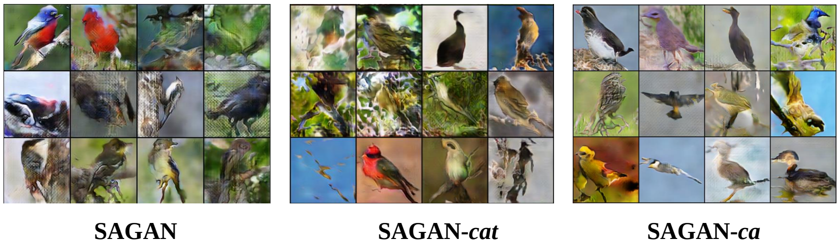
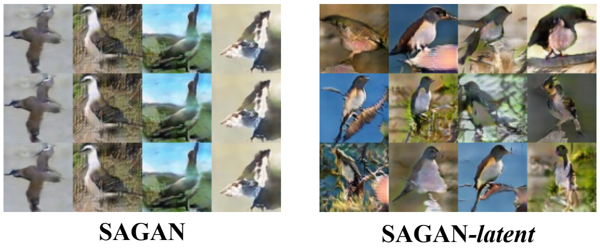


Figure 6: Sampled images generated by **SAGAN**, **SAGAN-*cat***, and **SAGAN-*ca***. The image generated by **SAGAN-*ca*** has the highest quality, whereas the images generated by the other two methods have obvious blurring and artifacts.

Discriminator Auxiliary losses By comparing the performance of **SAGAN-*cls*** and **SAGAN-*latent*** in Table 2, we conclude these two components can both improve the performance of the basic SAGAN model.

Moreover, when further visualizing the synthesized images of the baseline model and these two models, we observe an interesting fact. We find although the basic SAGAN model can do conditional image generation, it seems that it only remembers some frequent patterns of the training images and the whole set of synthesized images is lack of diversity. However, when we introduce the latent discrimination module, the model can generative images of higher diversity. The compared images are shown in Fig. 7.



SAGAN **SAGAN-*latent***

Figure 7: Generated images from SAGAN and SAGAN-*latent*. The basic SAGAN only remembers some common patterns of the training images. In comparison, by introducing latent discrimination module, the generated images are of high diversity.

5 Discussion and Analysis

Section 4 shows the comparison of our model and the baselines. We show that our model is able to generate photo-realistic images with high quality and diversity on two challenging fine-grained dataset. Under the matrices of FID and IS, we show that our method outperforms the baseline by a large margin. Ablation studies on CUB dataset further show that the modifications we made all contribute to the good performance. Furthermore, the extra layers we added are light-weighted and the computational overhead can be ignored.

We found our model to be limited in the following aspects: 1) The proposed method requires the training image to be center-cropped and with high quality. We observed that the model failed catastrophically on several classes in the DOG dataset where the training samples have noisy backgrounds (*e.g.* a man holding the dog) 2) The natural images has more attributes that can be disentangled. For example, the pose, size and background of the bird can also be used as conditions, so that we have more control over the content of the image being generated, which will further broaden the potential applications of the image generation methods. 3) Different embedding strategies have inconsistent effects on these two datasets, which indicates a more effective and robust conditional embedding strategy is needed to explore for a better performance of the proposed architecture.

Observing the above limitations, improving the model's robustness to noise by forcing the model to focus on the foreground and enabling the model to be conditioning on more attributes are clear future directions.

6 Conclusion

In this project, we mainly explore the problem of conditional image generation with few-shot fine-grained categories. This is a challenging problem and we propose a novel solution to it. Based on unconditional image generation network, we add several effective modules including conditioning augmentation module, category discrimination module and latent reconstruction loss. We instantiate our model with an effective GAN model, SAGAN, and further evaluate our method on two standard fine-grained dataset, CUB dataset and DOG dataset. Both quantitative results and qualitative results indicate the proposed model can generate better images and ablation studies show that all individual components contribute to the full model. We will further refine each components to improve the robustness and performance of the proposed model with more challenging settings.

References

[1] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan. *arXiv preprint arXiv:1701.07875*, 2017.

[2] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[3] Jianmin Bao, Dong Chen, Fang Wen, Houqiang Li, and Gang Hua. Cvae-gan: fine-grained image generation through asymmetric training. In *Proceedings of the IEEE international conference on computer vision*, pages 2745–2754, 2017.

[4] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018.

[5] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in neural information processing systems*, pages 2172–2180, 2016.

[6] Harm De Vries, Florian Strub, Jérémie Mary, Hugo Larochelle, Olivier Pietquin, and Aaron C Courville. Modulating early visual processing by language. In *Advances in Neural Information Processing Systems*, pages 6594–6604, 2017.

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[8] Nanqing Dong and Eric Xing. Few-shot semantic segmentation with prototype learning. In *BMVC*, 2018.

[9] DC Dowson and BV Landau. The fréchet distance between multivariate normal distributions. *Journal of multivariate analysis*, 12(3):450–455, 1982.

[10] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[11] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. In *Advances in neural information processing systems*, pages 5767–5777, 2017.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[13] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

[14] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

[15] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4401–4410, 2019.

[16] Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Li Fei-Fei. Novel dataset for fine-grained image categorization. In *First Workshop on Fine-Grained Visual Categorization, IEEE Conference on Computer Vision and Pattern Recognition*, Colorado Springs, CO, June 2011.

[17] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan 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*, pages 4681–4690, 2017.

[18] Ming-Yu Liu, Xun Huang, Arun Mallya, Tero Karras, Timo Aila, Jaakko Lehtinen, and Jan Kautz. Few-shot unsupervised image-to-image translation. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[19] Volodymyr Mnih, Nicolas Heess, Alex Graves, et al. Recurrent models of visual attention. In *Advances in neural information processing systems*, pages 2204–2212, 2014.

[20] Thu Nguyen-Phuoc, Chuan Li, Lucas Theis, Christian Richardt, and Yong-Liang Yang. Hologan: Unsupervised learning of 3d representations from natural images. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7588–7597, 2019.

[21] Augustus Odena, Christopher Olah, and Jonathon Shlens. Conditional image synthesis with auxiliary classifier gans. In *International conference on machine learning*, pages 2642–2651, 2017.

[22] Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Łukasz Kaiser, Noam Shazeer, Alexander Ku, and Dustin Tran. Image transformer. *arXiv preprint arXiv:1802.05751*, 2018.

[23] Siyuan Qiao, Chenxi Liu, Wei Shen, and Alan L Yuille. Few-shot image recognition by predicting parameters from activations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7229–7238, 2018.

[24] Abhinav Sagar. Hrvgan: High resolution video generation using spatio-temporal gan. *arXiv preprint arXiv:2008.09646*, 2020.

[25] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *Advances in neural information processing systems*, pages 2234–2242, 2016.

[26] Tim Salimans, Andrej Karpathy, Xi Chen, and Diederik P Kingma. Pixelcnn++: Improving the pixelcnn with discretized logistic mixture likelihood and other modifications. *arXiv preprint arXiv:1701.05517*, 2017.

[27] David Stap, Maurits Bleeker, Sarah Ibrahimi, and Maartje ter Hoeve. Conditional image generation and manipulation for user-specified content. *arXiv preprint arXiv:2005.04909*, 2020.

[28] Qianru Sun, Yaoyao Liu, Tat-Seng Chua, and Bernt Schiele. Meta-transfer learning for few-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 403–412, 2019.

[29] Aaron Van den Oord, Nal Kalchbrenner, Lasse Espeholt, Oriol Vinyals, Alex Graves, et al. Conditional image generation with pixelcnn decoders. In *Advances in neural information processing systems*, pages 4790–4798, 2016.

[30] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. In *Advances in neural information processing systems*, pages 3630–3638, 2016.

[31] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

[32] Yaqing Wang and Quanming Yao. Few-shot learning: A survey. *arXiv preprint arXiv:1904.05046*, 2019.

[33] Yu-Xiong Wang, Ross Girshick, Martial Hebert, and Bharath Hariharan. Low-shot learning from imaginary data. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7278–7286, 2018.

[34] Xiu-Shen Wei, Jianxin Wu, and Quan Cui. Deep learning for fine-grained image analysis: A survey. *arXiv preprint arXiv:1907.03069*, 2019.

[35] Chenshen Wu, Luis Herranz, Xialei Liu, Joost van de Weijer, Bogdan Raducanu, et al. Memory replay gans: Learning to generate new categories without forgetting. In *Advances in Neural Information Processing Systems*, pages 5962–5972, 2018.

[36] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057, 2015.

[37] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. In *International Conference on Machine Learning*, pages 7354–7363. PMLR, 2019.

[38] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 5907–5915, 2017.

[39] Ruixiang Zhang, Tong Che, Zoubin Ghahramani, Yoshua Bengio, and Yangqiu Song. Metagan: An adversarial approach to few-shot learning. In *Advances in Neural Information Processing Systems*, pages 2365–2374, 2018.

[40] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman. Toward multimodal image-to-image translation. In *Advances in neural information processing systems*, pages 465–476, 2017.