CycleGAN with Feature-Consistency Loss

Hao Zeng Xinyuan Lian Xiuzhen Zhang

Abstract

Image-to-image translation is a class of vision and graphics problems and there exist many works to solve it. CycleGAN is a such framework to achieve the image-to-image translation with cycle consistency loss. We propose a novel framework based on CycleGAN and update the loss function by designing the feature-consistency loss. Considering feature-consistency loss can incentivize the framework to keep more information in the training process. We evaluate our framework and find the translation preforms better than CycleGAN. Besides, the size of datasets is an important factor in training process. In this report, we also propose some new ways to train the model on imbalanced size of datasets.

1. Introduction

Many problems in image processing [5], computer graphics [3], and computer vision [4] can be posed as "translating" an input image into a corresponding output image. In analogy to automatic language translation, we define automatic image-to-image translation [8, 7, 12] as the problem of translating one possible representation of a scene into another, given sufficient training data. In recent years, many works have been proposed to solve the image-to-image translation problem and it mainly contains conditional and unconditional models (e.g. [9, 12, 10, 1, 7, 11]).

Training a model for image-to-image translation typically requires a large datasets of paired examples. These datasets can be difficult and expansive to obtain, and in some cases impossible, such as photographs of paintings by long dead artists. To solve the absence of paired examples, cycleGAN introduced in [12] used cycle-consistency adversarial networks to transfer the style of an image to that of a reference. CycleGAN can be a technique for training unsupervised image translation models via the GAN architecture using unpaired collections of images from two different domains.

With large enough capacity, a network can map the same set of input images to any random permutation of images in the target domain, where any of the learned mappings can induce an output distribution that matches the target distribution. Thus, CycleGAN gives the cycle consistency loss to reduce the space of possible mapping functions. However, cycle consistency loss in on the pixel level and it only achieves the color mapping in the training process. To improve this case, we focus on both pixel level and feature level. Considering the feature can help the framework keep more information during the training process and the result can be shown more realistic. We compare our new framework with CycleGAN with the same datasets. Through the experiment, we can find that the feature consistency loss will give better results.

Besides conditional GANs, unconditional GANs have shown remarkable success in generative realistic, high quality samples when trained on calss specific datasets. However, capturing the distribution of highly diverse datasets with multiple object classes is still considered a major challenge and often requires conditioning the generation on another signal or training the model for a specific task. SinGAN proposed in [9] is an unconditional generative model that can be learned from a single natural image. Their model is trained to capture the internal distribution of patches within the image, and is then able to generate high quality, diverse samples that carry the same visual content as the image. The utility of SinGAN is in a wide range of image manipulation tasks.

Motivated by the above two works, we want to combine SinGAN with the structure of CycleGAN in this project. The main focus of our work is to generate different styles of images with known unpaired images. Although our goal is similar to CycleGAN, we only need fewer unpaired images to achieve the same function, which means our model needs less data to reduce the expanses of obtaining datasets.

In a word, our work in this project can be shown as following.

- Propose a new framework based on CycleGAN with feature-consistency loss.
- Image-to-image translation with imbalanced size of datasets.

2. Related Work

Generative adversarial networks have achieved impressive results in image generation, image editing and representation learning. Recent methods adopt the same

idea for conditional image generation applications, such text2image, image inpainting, and future prediction, as well as to other domains like videos and 3D data. The key to GAN's success is the idea of an adversarial loss that forces the generated images to be, in principle, indistinguishable from real photos. We adopt an adversarial loss to learn the mapping such that the translated images cannot be distinguished from images in the target domain.

CycleGAN focuses on unpaired image-to-image translation. It does not rely on any task-specific, predefined similarity function between the input and output, nor do the authors assume that the input and output have to lie in the same low-dimensional embedding space. This makes the method a general-purpose solution for many vision and graphics tasks. CycleGAN introduces a loss to push two generators G and F to be consistent with each other. However, cycle consistency loss in CycleGAN is only on the pix level which means it only achieves the color mapping in the training process. Different from the cycle consistency loss, we introduce a new loss, feature-consistency loss. Feature-consistency loss can help the framework keep more information on features.

Image-generating machine learning models are typically trained with loss functions based on distance in the image space which often leads to over-smoothed results. The paper [2] proposed a class of loss functions, which they called deep perceptual similarity metrics that mitigate this problem. Instead of computing distance in the image space, the authors compute distances between image features extracted by deep neural networks. This metric better reflects perceptually similarity of images and thus leads to better results. Motivated by these improved methods [12, 10, 2], we propose the feature-consistency loss to integrate their advantages on image-to-image traslation.

SinGAN [9] is to learn an unconditional generative model that captures the internal statistics of a single training image. It is conceptually similar to the conventional GAN setting, except that the training samples are patches of a single image, rather than whole image samples from a database. With SinGAN, we can deal with more general natural images and go beyond texture generation. The demand of datasets is always a problem in training process and it also exists in the unpaired image-to-image translation setting. We try to use SinGAN to generate more images with fewer data and apply the data into CycleGAN in order to achieve the same function with CycleGAN in such imbalanced size of datasets.

3. Problem Formulation

Our goal is to learn mapping functions between two images X and Y given training samples $\{x_i\}_{i=1}^{N}$ where $x_i \in X$ and $\{y_j\}_{j=1}^{M}$ where $y_i \in Y$. We denote the data distribution as $x \sim p_{data}(x)$ and $y \sim p_{data}(y)$. Similar to CycleGAN, our model includes two mapping $G : X \rightarrow Y$ and $F : Y \rightarrow X$. In addition, we introduce two adversarial discriminators D_X and D_Y, where D_X aims to distinguish between images $\{x\}$ and translated images $\{F(y)\}$. In the same way, D_Y aims to discriminate between $\{y\}$ and $\{G(x)\}$. Our objective contains two types of terms: adversarial losses for matching the distribution of generated images to the data distribution in the target domain and feature-consistency losses to prevent the learned mappings G and F from contradicting each other.

4. Method

Our objective contains two types of terms: adversarial losses for matching the distribution of generated images to the data distribution in target domain and feature-consistency losses to reduce the space of possible mapping functions such that the learned mappings G and F won't contradict each other.

4.1. Adversarial Loss

The adversarial loss $\mathcal{L}_{\text{GAN}}$ penalizes for the distance between the distribution of patches and the distribution of patches in generated samples. We use the WGAN-GP [6], which we found to increase training stability, where the final discrimination score is the average over the patch discrimination map. The adversarial loss can be expressed as:

$$\mathcal{L}_{\text{GAN}}(G, D_Y, X, Y) = \mathbb{E}_{y \sim p_{data}(y)}[\log D_Y(y)] + \mathbb{E}_{x \sim p_{data}(x)}[\log(1 - D_Y(G(x)))].$$

4.2. Feature-Consistency Loss

With large enough capacity, a network can map the same set of input images to any random permutation of images in the target domain, where any of the learned mappings can induce an output distribution that matches the target distribution. Thus, adversarial losses alone cannot guarantee that the learned function can map an individual input x to a desired output y. To further reduce the space of possible mapping functions, cycleGAN proposes the notion of cycle consistency loss.

Cycle consistency loss can encourage generators to output the images which are similar with inputs in structure. However, cycle consistency loss is enforced at pixel level, which means it is too strong and will cause undesired results in some situations. Instead of relying on CycleGAN to reconstruct the input image pixels, we argue that it should reconstruct the general structures. Inspired by the DeePSiM [2] which uses a CNN feature level distance combined with pixel level distance to better reflect perceptually similarity of images such that it can lead to better results, We introduce an L_1 loss on CNN features extracted by the corresponding discriminator which has learned good features

on the domain we are interested in. The feature-consistency loss can be expressed as:

$$\begin{aligned}&\mathcal{L}_{fc}(G, F, D_X, X, \theta_t)\\&= \mathbf{E}_{x\sim p_{data}(x)}[\theta_t||f_{D_x}(F(G(x))) - f_{D_X}(x)||_1\\&+ (1-\theta_t)||F(G(x)) - x||_1],\end{aligned}$$

where f_D is the extracted feature from the last layer of the corresponding discriminator D, and $\theta_t \in [0, 1]$ is the ration between the discriminator feature level and pixel level loss. In particular, θ should be low because the features are not so good at the beginning of the training, and then linearly increase to a high value close to 1. The reason why θ_t should not be equal to 1 is that we also need pixel level consistency to prevent excessive hallucination and undesired objects in the images.

4.3. Full Objective

At epoch t, our full objective is:

$$\begin{aligned}\mathcal{L}(G, F, D_X, D_Y, t) = &\mathcal{L}_{\text{GAN}}(G, D_Y, X, Y)\\&+ \mathcal{L}_{\text{GAN}}(F, D_X, Y, X)\\&+ \lambda_t\mathcal{L}_{fc}(G, F, D_X, X, \theta_t)\\&+ \lambda_t\mathcal{L}_{fc}(F, G, D_Y, Y, \theta_t),\end{aligned}$$

where λ_t controls the relative importance of the two objectives and will linearly decrease to a small value in training process such that we can get better generators. Notice that λ_t can not decrease to zero since we need to keep the function of cycle consistency to prevent the generators being unconstrained and mode collapse. We aim to solve:

$$G^*, F^* = \arg\min_{G,F}\max_{D_X,D_Y}\mathcal{L}(G, F, D_X, D_Y, t).$$

5. Experiments

In this section, we first compare our improved CycleGAN with feature consistency loss against the origin CycleGAN on unpaired datasets. We then study the parameter setting for our methods. Finally, we demonstrate the performance of CycleGAN in imbalance size of datasets, and try to prevent mode collapse phenomenon by training similar image from SinGAN.

5.1. Datasets

We use the datasets horse2zebra and monet2photo from original CycleGAN[12]. Due to the limit of time, we only choose 100 horse and 100 zebra images as a subset of horse2zebra, which mainly contain one horse or zebra in each image. For monet2photo, we do the same thing for two style images.

5.2. Training details

We use the same technique as CycleGAN,for example, we replace the negative log like- lihood objective by a least-squares loss for stable and higher quality results. For generator, we use 9 residual blocks for 256×256 images. For the discriminator networks we use 70×70 PatchGANs, which aim to classify whether 70×70 overlapping image patches are real or fake.

For CycleGAN, we set $\lambda = 10$, use the Adam solver with a batch size of 1. All networks were trained from scratch with a learning rate of 0.0002. We keep the same learning rate for the first 100 epochs and linearly decay the rate to zero over the next 100 epochs.

For feature-cycle consistency loss, we keep the learning rate 0.0002 for the first 100 epochs and linearly decay the rate to zero over the next 100 epochs. we reduce λ_t from 10 to 3 linearly during the iterations and increase θ_t from 0 to 0.9 if there is no special mention.

5.3. The performance of feature consistency loss

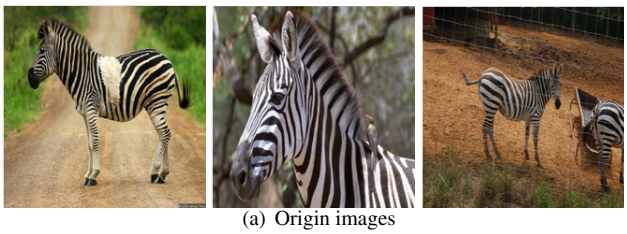
In this subsection, we compare our methods with origin CycleGAN. The experiments were carried out in horse2zebra sub-dataset (we choose 100 horse and 100 zebra for trainning as mentioned above). The results can be see in Figure.1 and Figure.2. We then found feature consistency loss help to capture the details of horse better than origin cycle consistency loss. It means that this method can learn more key features.

However, there are some failure cases (e.g. Figure.3). This method is also difficult to prevent the question caused by color similarity.

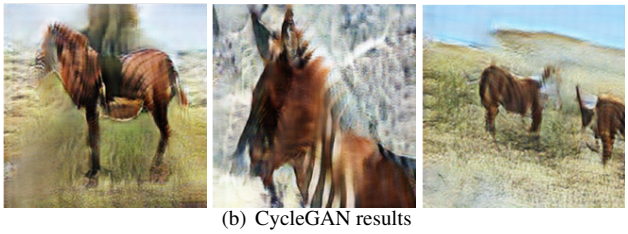
5.4. Parameter tuning results

As mentioned above, cycle consistency loss is a kind of regularization method. Large λ_t cause model to learn similarity at feature level. However, if $\lambda_t = 0$ generators will become unconstrained, and suffer from mode collapse. For the ratio parameter θ, $\theta \to 1$ may cause image lose the information of background. So we keep it a large value but less than 1.

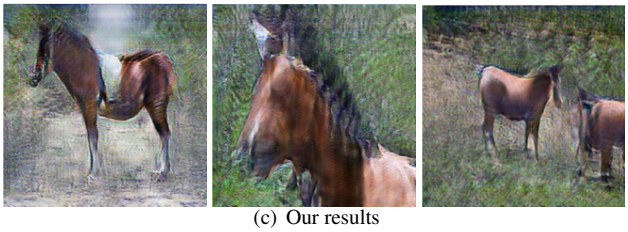
Let T be the total epochs, in order to study how the parameters affect the result, we took $\lambda_T \in \{1, 3, 5\}$ and $\theta_T \in \{0.5, 0.7, 0.9\}$. For all experiments, λ start to decrease linearly from 10 and θ increase linearly from 0. It can be seen in 4, the better parameters for horse2zebra subdataset are $\lambda_T = 3$ and $\theta_T = 0.9$. The results also present the larger θ_t help to capture the feature of horse, and importance of suitable λ choices. A smaller λ will allow the model to bring more dramatic changes to the input images, so θ should be decreased in order to preserve more pixel level details.



(a) Origin images

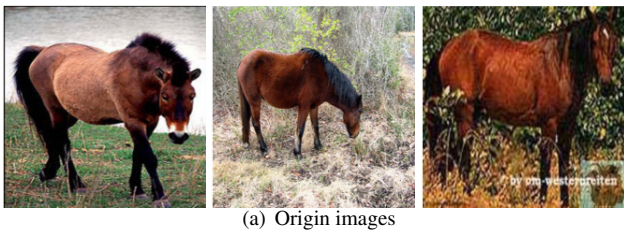


(b) CycleGAN results

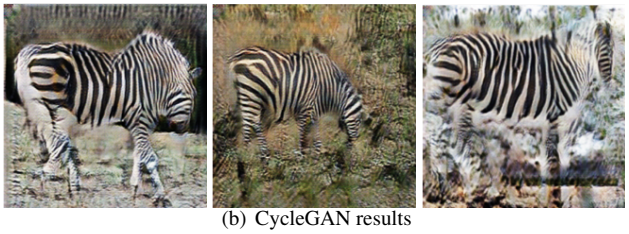


(c) Our results

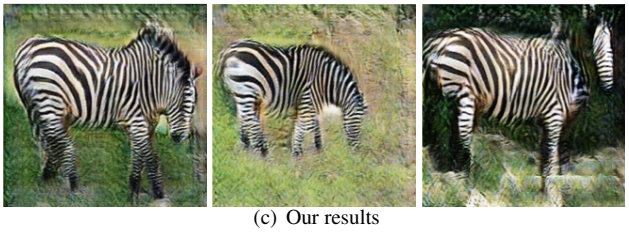
Figure 1. The results of horse2zebra sub-dataset. From top to bottom (a to c): input zebra images, Origin CycleGAN results, our results. The feature consistency loss help model to preserve more detail of the image.



(a) Origin images



(b) CycleGAN results



(c) Our results

Figure 2. The results of horse2zebra sub-dataset. From top to bottom (a to c): input zebra images, Origin CycleGAN results, our results. The feature consistency loss help model to preserve more detail of the image.

5.5. A study for imbalance datasets size

At most examples based on GAN, the training set size depends on tasks and domains. The more complicated task/domain is, the more images we need. For CycleGAN, if there is too much difference between the two domains, then it need more image to train. To solve this problem, we try use SinGAN [9] to expand data domain. One can be seen as a method to generate similar natural images, so it may also a method to expand dataset. Based on this, for two different mode datasets, we try to reduce one dataset size and generated some similar images by SinGAN.

In this experiment, we use a subset of monet2photo for 100 Monet's paints and 100 photo. We stable photo dataset at size 100, and decrease the paints data from 100 to 10, see Figure.5. We then generate new similar paints image from the 10 images (each generate 10 examples by inputting noise to SinGAN). The experiments show that the SinGAN may not as an effective method

5.6. A study for single-to-single translation

Based on the difference of generating from one image in SinGAN and image translation on unpaired dataset in CycleGAN. A very intuitive idea is that we get random images generated by SinGAN which has been trained on input images to augment the dataset. Motiveted by this, we train two SinGANs on two unpaired images and then get 50 random images generated from each input images in SinGAN respectively. Then we train CycleGAN on the augmented dataset and the results are shown in Figure.6

(a) Failure result

Figure 3. A failure case. Left: input image. Right: translation image. While meadow has the similar color with horse, then it may be changed into zebra.

Another idea is that we combine these two models by integrate CycleGAN into every scale of SinGAN and expecct that the noise randomness will be helpful in solving small dataset problem. So we proposed a model shown in Figure.7.

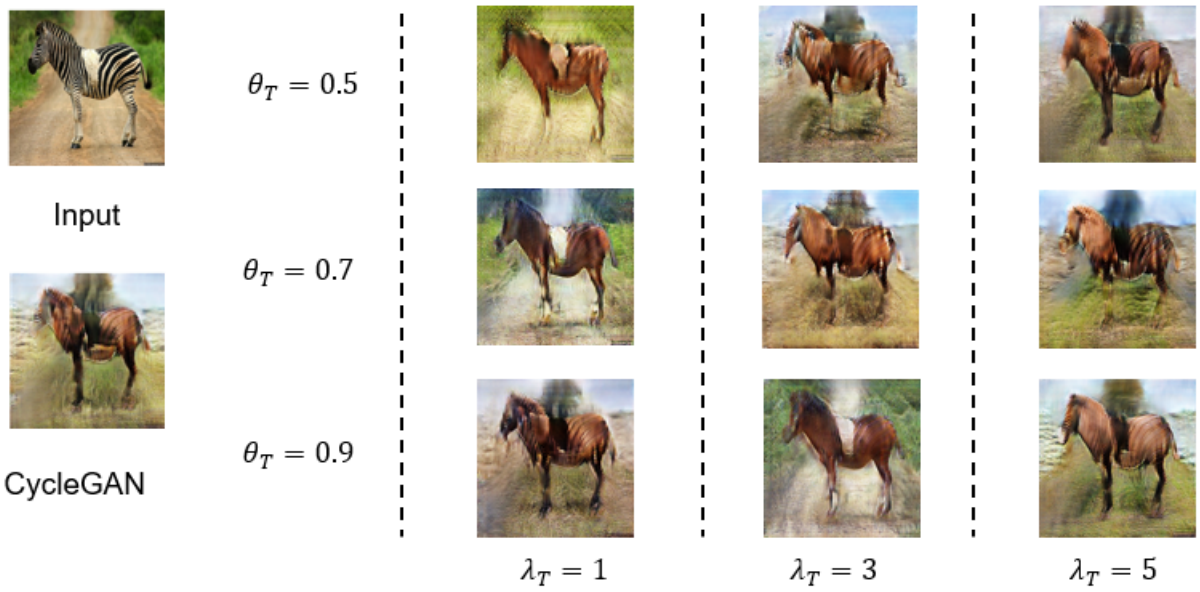
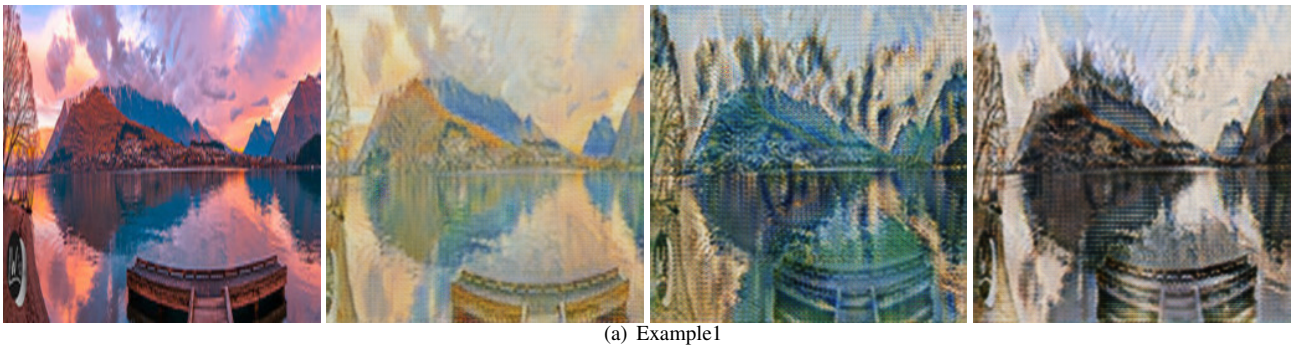
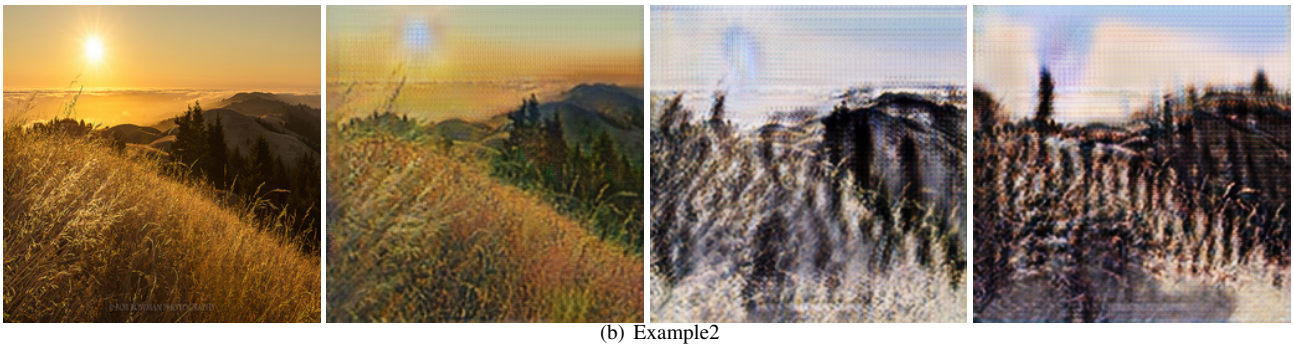


Figure 4. Parameter tuning results. The results from different θ_T and λ_T combination. The input image and original CycleGAN result are on the leftmost. Each line has same θ_T and each column has same λ_T. It illustrates larger θ_t help to capture the feature of horse and samll λ will drive image to change more dramatically.



(a) Example1



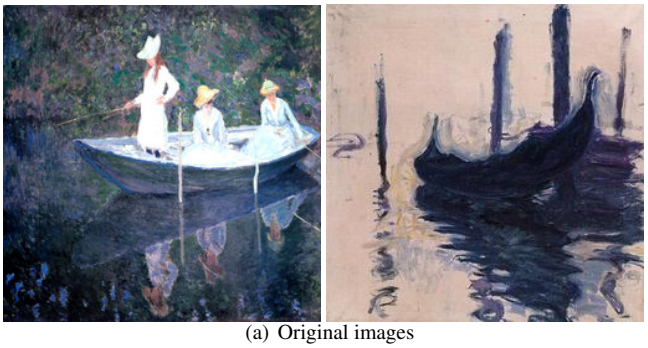
(b) Example2

Figure 5. The results of monet2photo sub-dataset. From left to right, they are input images, images from the model trained by 100-100 dataset, images from the model trained by 10-100 dataset and images from the model trained by 10-100 dataset, where each image of the small size dataset is expanded by SinGAN.

We train this model on a horse image and a zebra image from the horse2zebra dataset. However, the result shows that we cannot prevent the model collapse and we get the same images with the original input images.

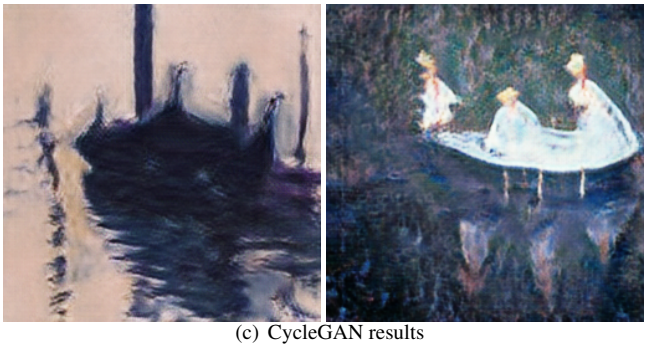
6. Conclusion

In this project, we propose a novel framework based on CycleGAN to achieve image-to-image translation and intro-



(a) Original images

(b) SinGAN results



(c) CycleGAN results

Figure 6. The results of monet single-to-single dataset. From top to bottom (a to c): input monet images, random results generated from SinGAN, cycleGAN results.

duce the L_{fc} loss based on the distance in feature spaces and decay the weight of cycle consistency loss as the training processes. We show that this method improve obviously in some cases. So, we think that this model is considerable for specific image-to-image translation tasks. However, this model is sensitive to the sensitive to parameter changes, and may cause mode collapse if the setting is not good.

For the research of imbalance datasets, we have tried some basic methods on solving the bad performance when CycleGAN is trained by the imbalanced datasets. However, the results are not as good as expected.

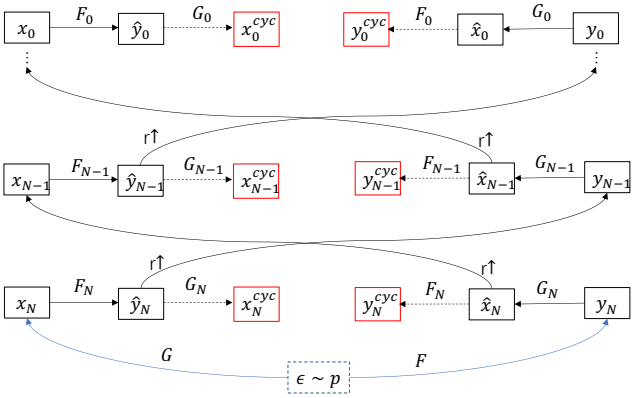


Figure 7. MixGAN structure: There are two SinGAN with the same scales. The same scale in thse two SinGAN can be considered as a CycleGAN and their outputs will be upsampled as the input to the next scale of different SinGAN.

References

[1] A. Almahairi, S. Rajeswar, A. Sordoni, P. Bachman, and A. Courville. Augmented cyclegan: Learning many-to-many mappings from unpaired data. *arXiv preprint arXiv:1802.10151*, 2018.

[2] A. Dosovitskiy and T. Brox. Generating images with perceptual similarity metrics based on deep networks. In *Advances in neural information processing systems*, pages 658–666, 2016.

[3] J. D. Foley, A. Van Dam, S. K. Feiner, J. F. Hughes, and R. L. Phillips. *Introduction to computer graphics*, volume 55. Addison-Wesley Reading, 1994.

[4] D. A. Forsyth and J. Ponce. *Computer vision: a modern approach*. Prentice Hall Professional Technical Reference, 2002.

[5] R. C. Gonzales and R. E. Woods. *Digital image processing*, volume 2. Prentice hall New Jersey, 2002.

[6] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville. Improved training of wasserstein gans. In *Advances in neural information processing systems*, pages 5767–5777, 2017.

[7] P. Isola, J.-Y. Zhu, T. Zhou, and A. 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.

[8] M.-Y. Liu, T. Breuel, and J. Kautz. Unsupervised image-to-image translation networks. In *Advances in neural information processing systems*, pages 700–708, 2017.

[9] T. R. Shaham, T. Dekel, and T. Michaeli. Singan: Learning a generative model from a single natural im-

age. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4570–4580, 2019.

[10] T. Wang and Y. Lin. Cyclegan with better cycles.

[11] Z. Yi, H. Zhang, P. Tan, and M. Gong. Dualgan: Unsupervised dual learning for image-to-image translation. In *Proceedings of the IEEE international conference on computer vision*, pages 2849–2857, 2017.

[12] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.