

FINE-GRAINED TEXT TO IMAGE SYNTHESIS SUPPLEMENTARY MATERIALS

Xu Ouyang, Ying Chen, Kaiyue Zhu, Gady Agam

Illinois Institute of Technology
Computer Science
USA

1. APPENDIX

1.1. Training of the network

In this section, we show the structure of our proposed FG-RAT GAN with auxiliary classifier and contrastive learning as in Figure 1.

1.2. Evaluation metrics

The Inception Score [2] can measure a synthetic image quality by computing the expected Kullback Leibler divergence (KL divergence) between the marginal class distribution and conditional label distribution:

$$IS = \exp(\mathbb{E}_x KL(p(y|x)||p(y))) \quad (1)$$

where $p(y|x)$ is the conditional label distribution of features extracted from the middle layers of the pretrained Inception-v3 model for generated images, and $p(y)$ is the marginal class distribution. IS gives a score that tells us if each image made by the model is clear and distinct, and if the model can make a wide range of different images. We want models that make a mix of clear images, so a higher IS is better.

The Frechet Inception Distance [1] that is given by:

$$d^2(F, G) = |\mu_x - \mu_y|^2 + \text{tr}|\Sigma_x + \Sigma_y - 2(\Sigma_x \Sigma_y)^{1/2}| \quad (2)$$

where F, G are two distributions of features extracted from the middle layers of a pretrained Inception-v3 model for generated and real images. The parameters $\mu_x, \mu_y, \Sigma_x, \Sigma_y$ are the mean vectors and covariance matrices of F and G . While IS checks image clarity and variety, FID checks if they look real. We want our model’s images to look like real photos, so a lower FID is better.

1.3. Comparison results

Figure 2 only shows synthesized images generated by RAT GAN and our proposed FG-RAT GAN on the Oxford-102 flower dataset since LAFITE did not train or test on this dataset and VQ-Diffusion did not post their pretrained model on this dataset.

We investigate the effects of different strategies we added to the basic RAT GAN model for text to image synthesis to demonstrate their significance on both the CUB-200-2011 bird and Oxford-102 flower datasets. We train three different models: A proposed FG-RAT GAN with auxiliary classifier, a proposed FG-RAT GAN with contrastive learning, and a proposed FG-RAT GAN with combination of auxiliary classifier and contrastive learning. The results are summarized in Table 1.

We compare with the DALLE-2 and Stable Diffusion which are the most popular models for text to image synthesis task. Since neither DALLE-2 nor Stable Diffusion did not train on the CUB-200-2011 bird dataset and Oxford-102 flower dataset, we only show the visualized results in Figure 3 and in Figure 4.

Figure 3 and Figure 4 show synthesized images generated by DALLE-2, Stable Diffusion, and our proposed FG-RAT GAN on the bird and flower dataset. There are six samples which belong to two different classes in each dataset. As we can see, our proposed FG-RAT GAN can generate fine-grained images which highly correspond to the given captions. Additionally, each synthesized image is more similar to other synthesized images in the same class. Thus, we demonstrate that our proposed FG-RAT GAN can reach better visualized results compared with DALLE-2 and Stable Diffusion.

2. REFERENCES

- [1] D. C. Dowson and B. V. Landau. The Fréchet distance between multivariate normal distributions. *Journal of Multivariate Analysis*, 12(3):450–455, September 1982. 1
- [2] Tim Salimans, Ian J. Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016*, pages 2226–2234, 2016. 1

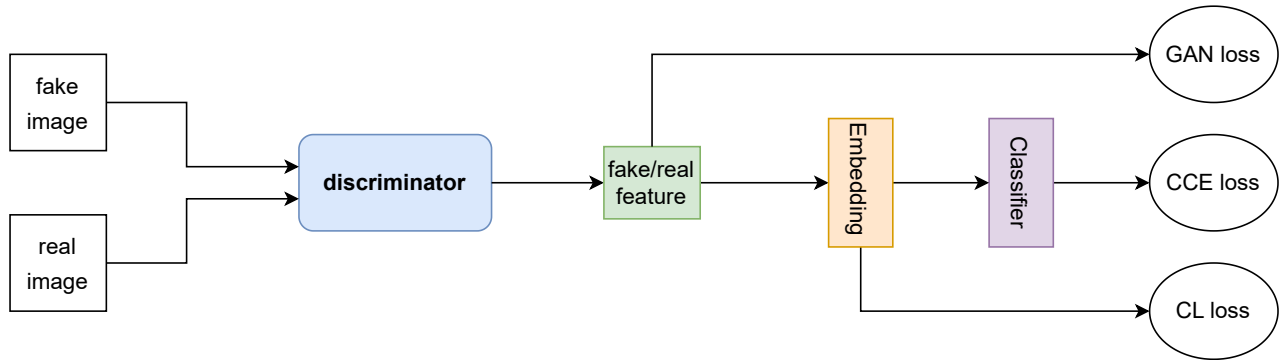


Fig. 1. The structure of the discriminator with auxiliary classifier and contrastive learning. The original output of the discriminator is still used to compute the GAN loss, and meanwhile followed by one fully connected layer to decrease the feature dimension. Next, the fully connected layer is followed by one embedding layer for contrastive learning. Then, the embedding layer is followed by a classifier for image classification.

Model	CUB bird dataset		Oxford flower dataset	
	IS \uparrow	FID \downarrow	IS \uparrow	FID \downarrow
RAT GAN	4.83	12.12	3.62	12.90
RAT GAN + classifier (our)	5.08	9.90	3.45	9.55
RAT GAN + contrastive learning (our)	4.84	9.10	3.66	10.63
FG-RAT GAN (our)	4.99	8.66	3.45	9.14

Table 1. Comparison of RAT GAN, proposed FG-RAT GAN with auxiliary classifier, proposed FG-RAT GAN with contrastive learning, and proposed FG-RAT GAN with combination of auxiliary classifier and contrastive learning on the CUB-200-2011 bird and Oxford-102 flower datasets. Each row presents a different model. The first column is the name of each model. The second and third columns show the IS and FID scores for the CUB bird dataset. The fourth and fifth columns show the IS and FID scores for the Oxford flower dataset. As can be observed, in CUB bird dataset, the proposed FG-RAT GAN with classifier reaches the highest IS score and the proposed FG-RAT GAN with classifier and contrastive learning reaches the lowest FID score. In the Oxford flower dataset, the proposed FG-RAT GAN with contrastive learning reaches the highest IS and the proposed FG-RAT GAN with classifier and contrastive learning reaches the lowest FID.

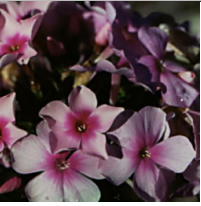
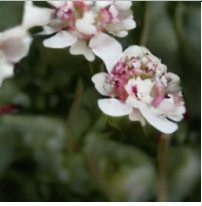



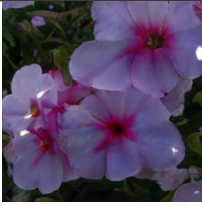


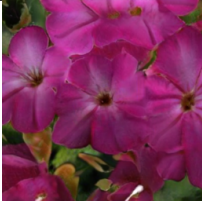




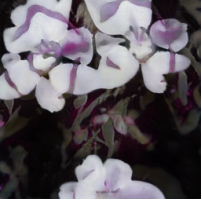


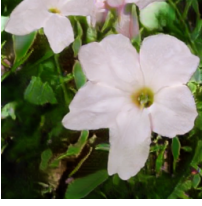

Class	Caption	Target	RAT GAN	Our FG-RAT GAN
Class 032 <i>image_05587.png</i>	the petals of the flowers are various shades of pink and have five individual petals.			
Class 032 <i>image_05602.png</i>	a large group of light pink flowers with dark pink centers.			
Class 032 <i>image_05604.png</i>	these flowers are mostly pink but some of them have white parts located closer to their stamens.			
Class 049 <i>image_06209.png</i>	this flower has thin white petals as its main feature.			
Class 049 <i>image_06216.png</i>	the petals on this flower are white with yellow stamen.			
Class 049 <i>image_06224.png</i>	the flower has petals of a white color with a many yellow stamen.			

Fig. 2. Examples of generated images using RAT GAN and the proposed FG-RAT GAN with classifier and contrastive learning trained on the Oxford flower dataset. Each row represents a different sample (image size = 256x256). The first column is the sample detail including class and specific image name. The second column is the caption. The third column is the corresponding target image. The fourth column is the image generated by RAT GAN. The fifth column is the image generated by our proposed FG-RAT GAN. As we can see, our proposed FG-RAT GAN can generate more realistic images where each image is similar to other images within the same class. For example, the 5th row generates a flower with white petals and yellow stamen as in the description, the 6th row generates a flower with white petals and yellow stamen as in the description, and both samples are similar to each other given they belong to the same class.

Class	Caption	Target	DALLE-2	Stable Diffusion	Our FG-RAT GAN
Class 001 Black Footed Albatross 0001_796111.png	the entire body is dark brown with a white band encircling where the bill meets the head.				
Class 001 Black Footed Albatross 0002_55.png	this bird has wings that are brown and has a big bill.				
Class 001 Black Footed Albatross 0005_796090.png	this bird has large feet and a broad wingspan with all grey coloration.				
Class 014 Indigo Bunting 0001_12469.png	this bird has a short, pointed blue beak, it also has a blue tarsus and blue feet.				
Class 014 Indigo Bunting 0047_12966.png	a small colorful bird with teal feathers covering its body, with green speckles on its vent and abdomen.				
Class 014 Indigo Bunting 0059_11596.png	a small purple bird, with black primaries, and a thick bill.				

Fig. 3. Examples of generated images using DALLE-2, Stable Diffusion, and the proposed FG-RAT GAN trained on the CUB bird dataset. Each row represents a different sample (image size = 256x256). The first column is the sample detail including class and specific image name. The second column is the caption. The third column is the corresponding target image. The fourth column is a generated image from DALLE-2. The fifth column is a generated image from Stable Diffusion. The sixth column is a generated image from our proposed FG-RAT GAN. As we can see, our proposed FG-RAT GAN can generate more realistic images where each image is similar to other images within the same class. For example, in the 1st row the proposed FG-RAT GAN generates a bird with dark brown body and white band encircling near the bill as specified in the caption, in the 3rd row it generates a bird with all gray body as specified in the caption, and both examples are similar to each other given that they belong to the same class.

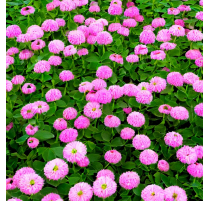
Class	Caption	Target	DALLE-2	Stable Diffusion	Our FG-RAT GAN
Class 032 <i>image_05587.png</i>	the petals of the flowers are various shades of pink and have five individual petals.				
Class 032 <i>image_05602.png</i>	a large group of light pink flowers with dark pink centers.				
Class 032 <i>image_05604.png</i>	these flowers are mostly pink but some of them have white parts located closer to their stamens.				
Class 049 <i>image_06209.png</i>	this flower has thin white petals as its main feature.				
Class 049 <i>image_06216.png</i>	the petals on this flower are white with yellow stamen.				
Class 049 <i>image_06224.png</i>	the flower has petals of a white color with a many yellow stamen.				

Fig. 4. Examples of generated images using RAT GAN and the proposed FG-RAT GAN with classifier and contrastive learning trained on the Oxford flower dataset. Each row represents a different sample (image size = 256x256). The first column is the sample detail including class and specific image name. The second column is the caption. The fourth column is a generated image from DALLE-2. The fifth column is a generated image form Stable Diffusion. The sixth column is a generated image from our proposed FG-RAT GAN. As we can see, our proposed FG-RAT GAN can generate more realistic images where each image is similar to other images within the same class. For example, the 5th row generates a flower with white petals and yellow stamen as in the description, the 6th row generates a flower with white petals and yellow stamen as in the description, and both samples are similar to each other given they belong to the same class.