Manipulating Generative Models using Sketches and/or Interactive User Inputs

AIGo 3rd

Team members: Jian Kim, Chanhyeok Choi, Jaewoo Heo / Professor: JaeJun Yoo / TA: Dayeong Baek

Abstracts

This research began with the idea of finding a way to easily adjust the generative model even for those who do not know the detailed network of GAN. GAN could only be handled by experts who knew in detail how models were trained and produced images. In particular, advanced engineering techniques were needed to induce GAN to generate slightly different forms of images. This was a major obstacle for the general public to use GAN for creative activities.

We developed the model which can simultaneously reflect the user's sketch and text. So user can manipulate GAN model with various shape of images which user wanted. in that way, we devised a multimodal model that allows the user to train the model from various angles by giving additional text containing the meaning he wants to request to GAN as well as sketch. Therefore, we tried to solve the problem of diversity in GAN through text and sketch.

Introduction & Research Purpose

Sketch your own GAN :

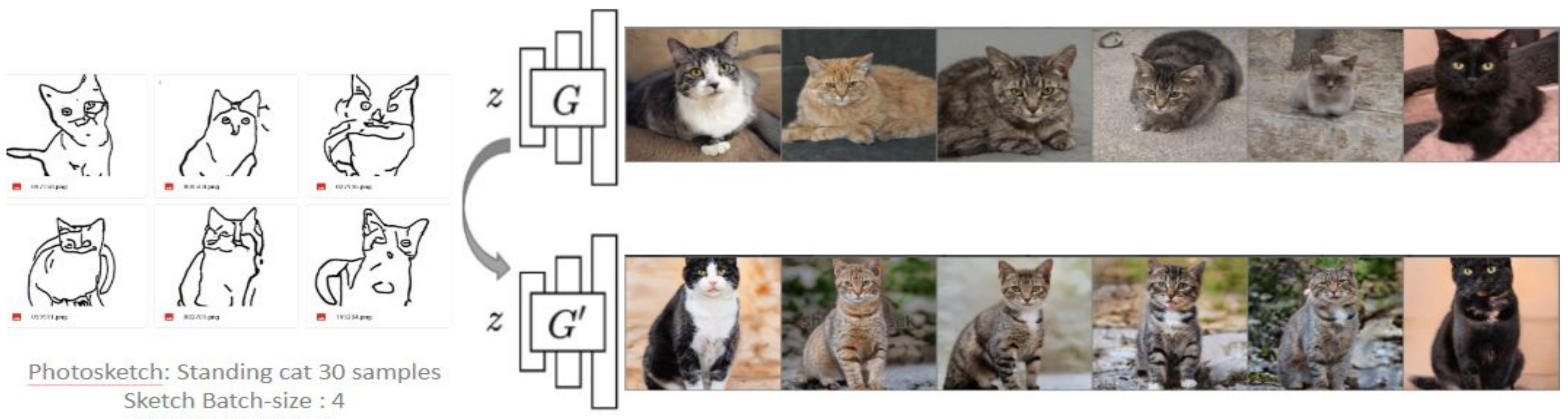
- Description: our baseline model. **When the user draws a sketch in the desired form, the GAN is trained to generate an image in the form of a sketch.**
- limitation : **can't generate more diverse texture and reflect local information,** such as a standing pink cat or a standing open mouth cat. So, we propose a multimodal model that includes both sketch and text in order to generate diverse images.

CLIP :

- Description : text-image embedding layers, suggested by openAI

Research Purpose :

- Apply **CLIP** to **Sketch your own GAN**, Develop the model which can adjust both the text contents and sketch which users suggested.



Photosketch: Standing cat 30 samples
Sketch Batch-size : 4
Iteration : 40,000

Research Methods

We used Cross-Domain Adversarial Learning and Image space Regularization proposed by Sketch your own GAN together.

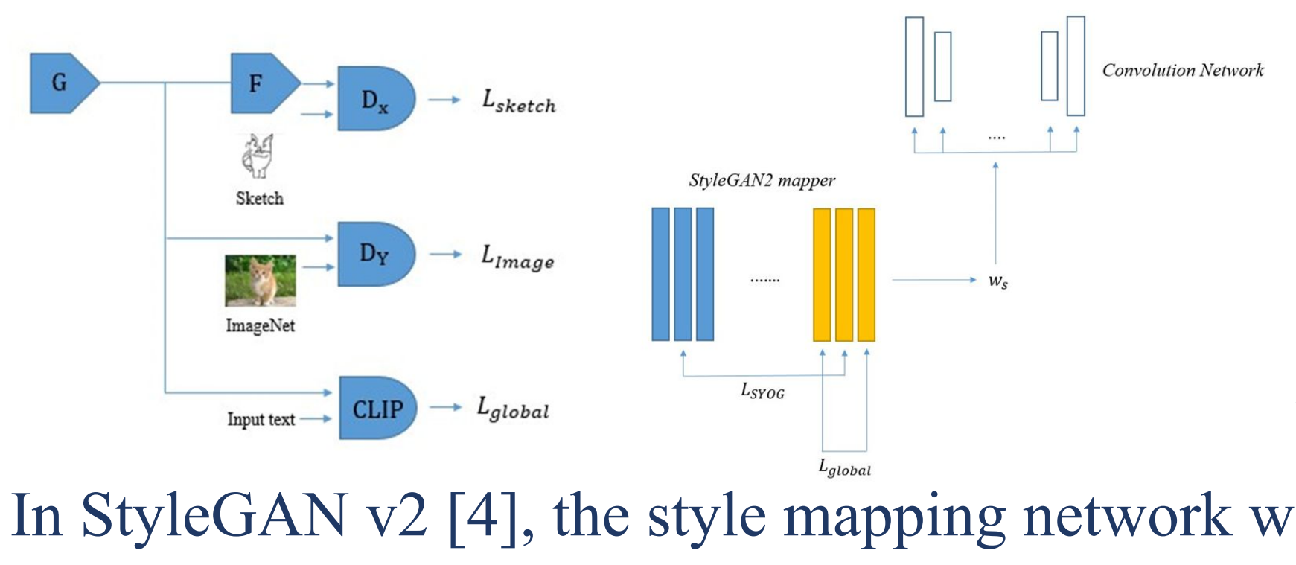
Due to the loss function on the right, a high-quality image that follows the shape of the sketch suggested by the user is created.

$$L_{sketch} = E_{y\sim p_{data}(y)} log(D_y(y)) + E_{z\sim p(z)} log(1 - D_y(F(G(z)))) \quad (1)$$

$$L_{image} = E_{x\sim p_{data}(x)} log(D_x(x)) + E_{z\sim p(z)} log(1 - D_x(G(z))) \quad (2)$$

$$L_{SYOG} = L_{sketch} + \boldsymbol{\beta} L_{image}$$

1. Text-Image CLIP Domain Optimization



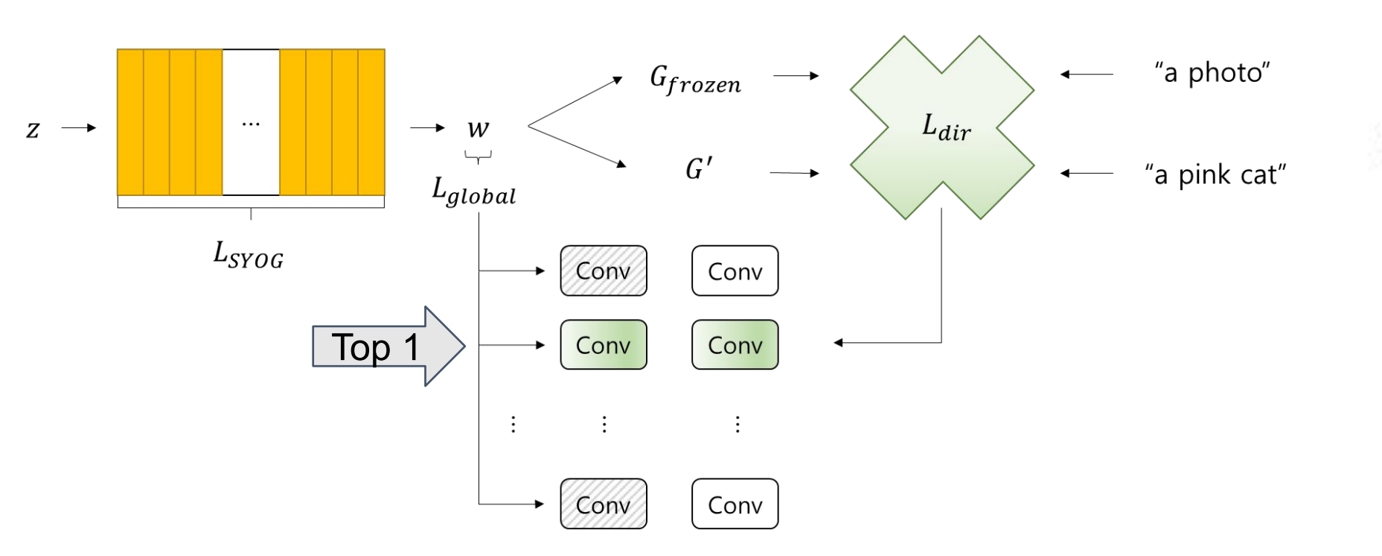
$$L_{global} = D_{CLIP}(G(w), t_{target})$$

$$L = L_{SYOG} + L_{global}$$

In order to reflect the text suggested by the user to the image, L_global was configured using CLIP.

In StyleGAN v2 [4], the style mapping network was learned with L_{SYOG}, and at the same time, only some layers were learned with L_{global}.

2. Text-Image CLIP Space Alignment



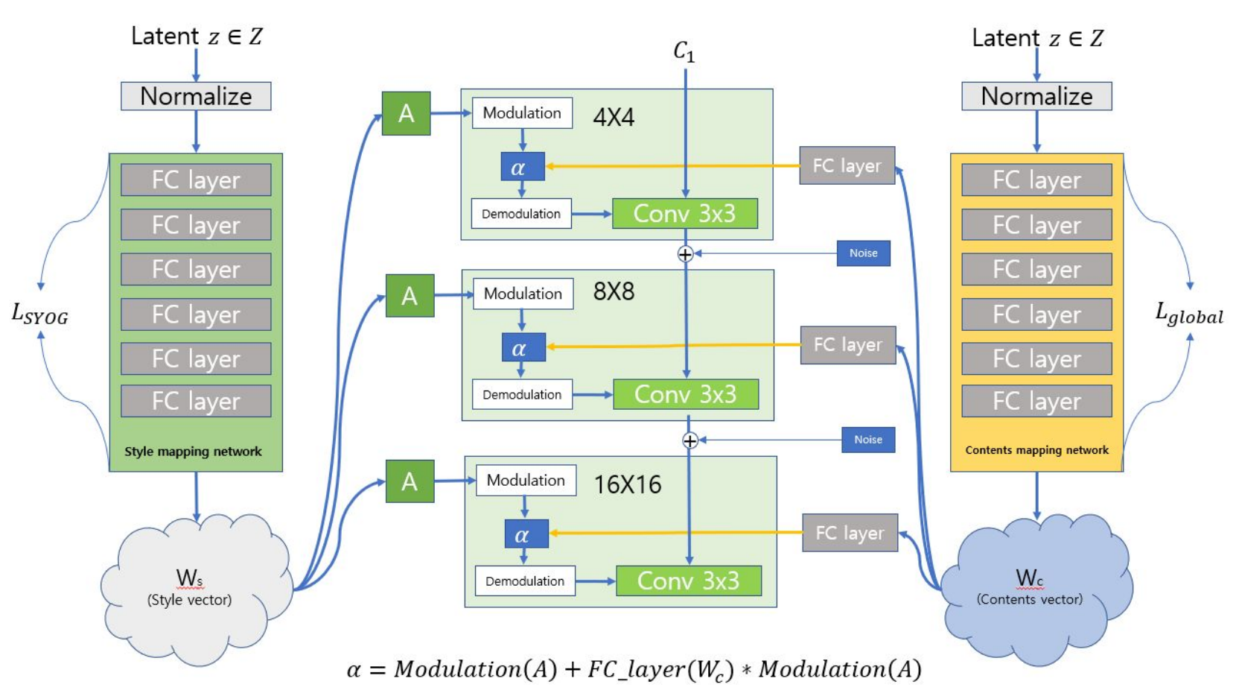
$$\Delta T = E_T(t_{target}) - E_T(t_{source})$$

$$\Delta I = E_I(G_{train}(w)) - E_I(G_{frozen}(w))$$

$$L_{direction} = 1 - \frac{\Delta I \cdot \Delta T}{|\Delta I||\Delta T|}$$

In method 1, the influence of CLIP was very large, and an inappropriate image was created. To solve this problem, we referred to the top-k algorithm and CLIP directional loss used in StyleGAN-NADA [7]. A style mapping network was learned with L_{SYOG} and.
(The top-k algorithm is a method of selecting the k-th layer layer in which the most change occurs.)

3. Separate Content and Style Codes



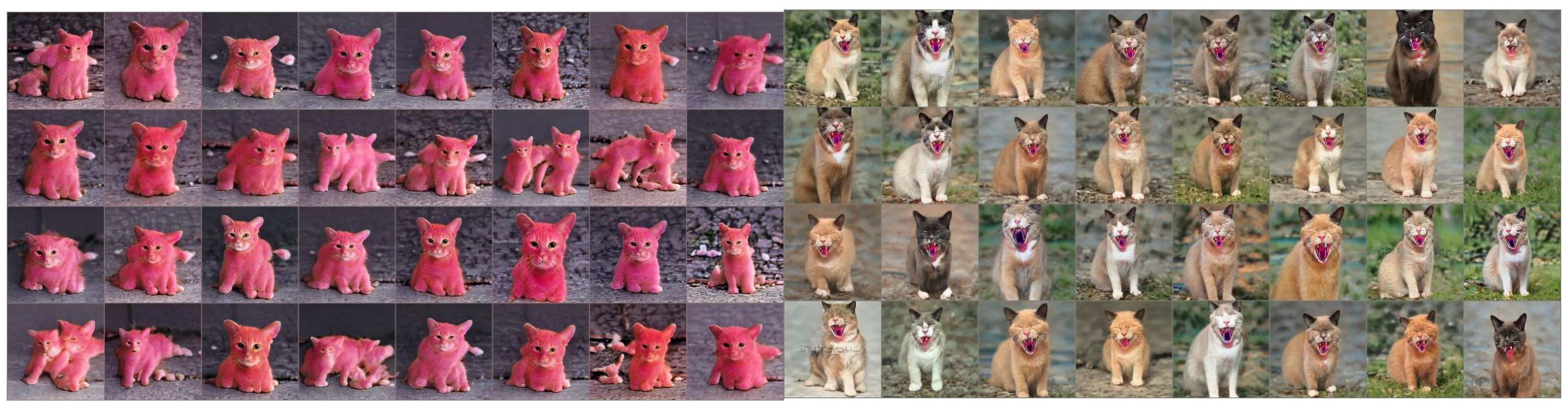
$$\alpha = Modulation(A) + FC\_layer(W_c) * Modulation(A)$$

As a result of the two experiments, the quality of the image increased, but we could see that the collision between L_{image} and $L_{direction}$ lowered the quality of the image. In order to solve the conflict of loss that appeared in methods 2, we tried to change the structure of the generator. A content mapping network was added to create a structure that separately learns L_{SYOG} and L_{global}. This is a diagonalGAN [9] using StyleGAN v1 applied to StyleGAN v2, and the content mapping network was trained with clip global loss.

Research Results

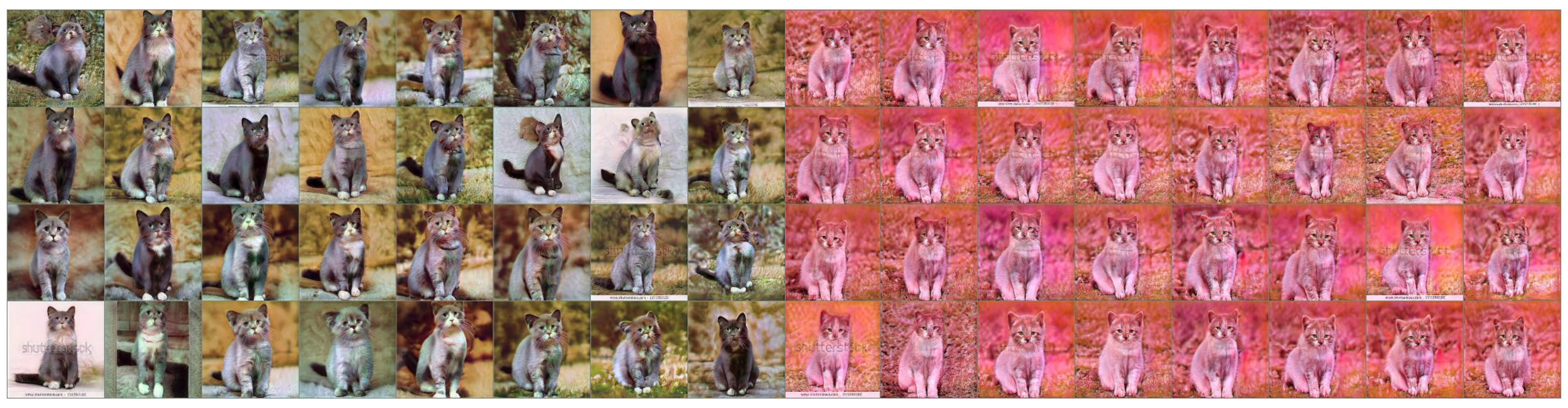
As sketch inputs, we used 30 standing cat sketches from PhotoSketch provided by Sketch your own GAN. "StyleGAN v2" was used as the pretrained image generation model, and "Vit-B/32" was used as the pretrained CLIP model. The iteration term is fixed at 40,000. The final images are 32 photos of 1024 x 1024 resolution taken using fixed random noise.

1-1. Result of applying Global CLIP Loss to the last four layers of style mapping network



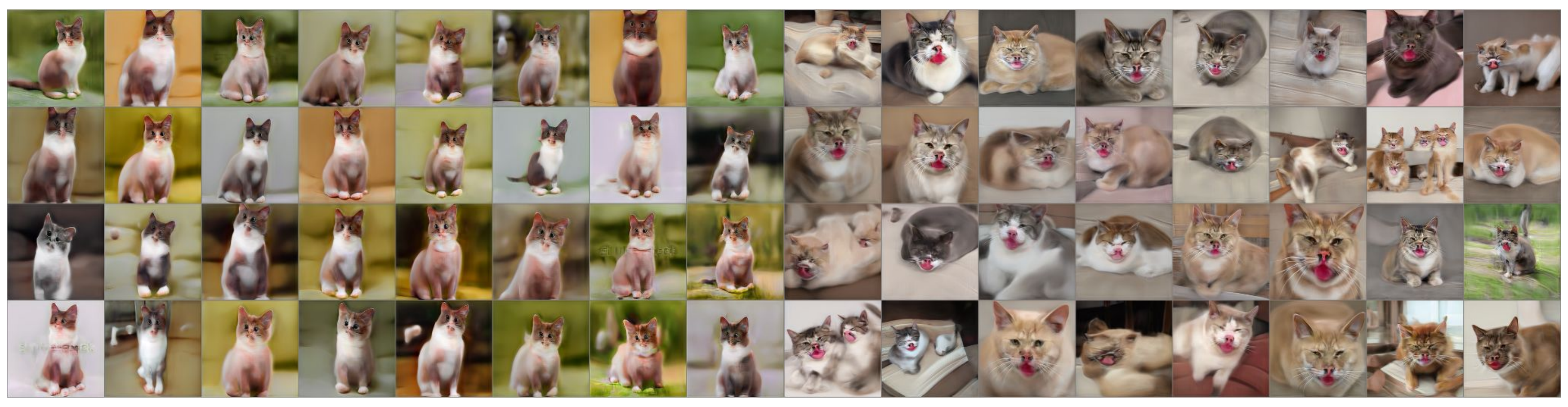
Left image was trained with "a photo of pink cat" text, and right image was trained with "a photo of open mouth cat" text. Although it followed the pink cat and open mouth cat, we could see their shape collapsing.

2-1. Result of using StyleGAN-NADA method



Both images were trained with "a photo of pink cat" text. $L_{direction}$ learned once when L_{SYOG} is learned 100 times. The image on the right is the learning result after removing L_{image}. Through experiments, it was confirmed that learning could not be performed well because of the conflict between L_{image} and CLIP.

3-1. Result of using Diagonal GAN method



Left image was trained with "a photo of pink cat" text, and right image was trained with "a photo of open mouth cat" text. However, the most satisfactory results were obtained among the experimental results obtained without adjusting the learning process (hyperparameters, iteration control, etc.), and provided clues for improvement in the future. The image on the right is the result after removing L_{SYOG}, and you can see that CLIP is well displayed as contents.

Conclusions

Significance of this research :

1. **It is the first attempt in the field of GAN manipulation using both sketch and text.**
   - We experimentally show that manipulation is possible just by training certain layers with CLIP and Sketch without retraining the entire network, and a generated image reflect both contents of text and sketch.
2. **CLIP itself has found its limitation.**
   - Training the model with CLIP simply focuses on minimizing loss. Changes that are not text-related are added, resulting in lower loss. This means that the model is not learned in the desired direction, and the generated image collapses at a certain moment during learning.

Future Work :

1. **Easing of competition between CLIP Loss and Image Loss.**
2. **To manipulate the model in real time and reduce costs for securing the learning dataset of the original model.**

References

1. Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial networks. arXiv preprint arXiv:1406.2661, 2014.
2. Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan: How to embed images into the stylegan latent space? In Proceedings of the IEEE international conference on computer vision, pages 4432–4441, 2019.
3. Sheng-Yu Wang, David Bau, and Jun-Yan Zhu. Sketch Your Own GAN. In IEEE International Conference on Computer Vision (ICCV). 2021.
4. Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020.
5. Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. arXiv preprint arXiv:2103.17249, 2021.
6. Gihyun Kwon and Jong-Chul Ye. Clipstyler: Image style transfer with a single text condition. ArXiv, abs/2112.00374, 2021.
7. R. Gal, O. Patashnik, H. Maron, G. Chechik, and D. Cohen-Or, "StyleGAN-NADA: CLIP-guided domain adaptation of image generators," arXiv preprint arXiv:2108.00946, 2021.
8. Rameen Abdal, Peihao Zhu, John Femiani, Niloy J Mitra, and Peter Wonka. Clip2stylegan: Unsupervised extraction of stylegan edit directions. arXiv preprint arXiv:2112.05219, 2021.
9. Gihyun Kwon and Jong Chul Ye. Diagonal attention and style-based gan for content-style disentanglement in image generation and translation. In Proc. Int'l Conf. Computer Vision, 2021.
10. Gihyun Kwon and Jong-Chul Ye. One-shot adaptation of gan in just one clip. ArXiv, abs/2203.09301, 2022.