A Fast Text-Driven Approach for Generating Artistic Content

Marian Lupaşcu[†], Ryan Murdock[†], Ionuț Mironică, Yijun Li
Adobe Research
{lupascu,rmurdock,mironica,yijli}@adobe.com

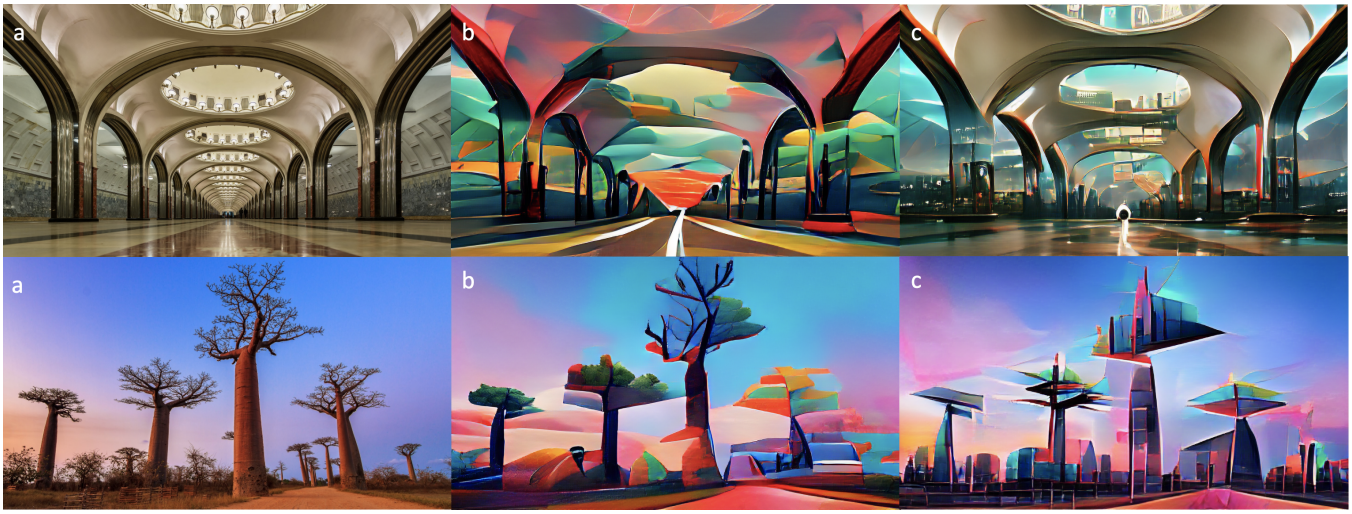

Figure 1: Results of our proposed method, where (a) represents the content image, (b) is the content image stylized with the text "A bold color landscape depicting the road everyone lives on in the style of romanticism paintings | Low poly" and (c) represents the content image stylized with the text "City of the future | Geometric art".

CCS CONCEPTS

• **Computing methodologies** → **Learning latent representations**; **Mixture models**.

KEYWORDS

Generative art, image generation, optimization, style transfer, stylization with text, synthetic art

ACM Reference Format:
Marian Lupaşcu[†], Ryan Murdock[†], Ionuț Mironică, Yijun Li. 2022. A Fast Text-Driven Approach for Generating Artistic Content. In *Proceedings of SIGGRAPH '22 Posters.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/8888888.7777777

1 INTRODUCTION

Finding a source of inspiration for visual art creation can be a difficult task. Even with the online search engine, it is still labor-intensive because manually crawling the web is time-consuming. Therefore, a large number of artistic prototypes must be explored before an artist has a better feel of what their future creations look like structurally but also artistically. In this work, the image can be stylized according to the user's requirements with a text prompt, a style image, or a combination of style parameters (Figure 1). The only limitations are the user's imagination and the models involved in the generative process.

Despite the considerable progress of generative models, there is not a common solution that allows stylization with image and text. The Conditional-driven GAN [Reed et al. 2016] first extracted visual details from the text and then synthesized the output image conditioned on these previously extracted features. The Attention-driven GAN [Xu et al. 2018] introduced a similarity module between text features and image features. The DALLE [Ramesh et al. 2021] and StyleCLIP [Patashnik et al. 2021] are the state-of-the-art approaches of generating and manipulating images via text. However, the DALLE does not support the stylization with a style example and StyleCLIP is mainly limited in generating images from a single domain: faces, cars, etc.

2 APPROACH

Our generative network is guided by the CLIP [Radford et al. 2021] to generate images that closely match a description using gradient descent through backpropagation. We use the VQGAN [Esser et al. 2021] as the generative network, but it can be replaced with other generative models such as BigGAN [Brock et al. 2019] and diffusion models [Dhariwal and Nichol 2021].

The whole pipeline of our proposed method is shown in Figure 2. The first step in the stylization process is represented by the projection of the stylization images and text in the CLIP latent space. Also at this step, the learnable tensor from the latent space

[†]Equal contribution, ordered alphabetically.

SIGGRAPH '22 Posters, August 08-11, 2022, Vancouver, Canada

ACM ISBN 978-1-4503-1234-5/22/07.
https://doi.org/10.1145/8888888.7777777

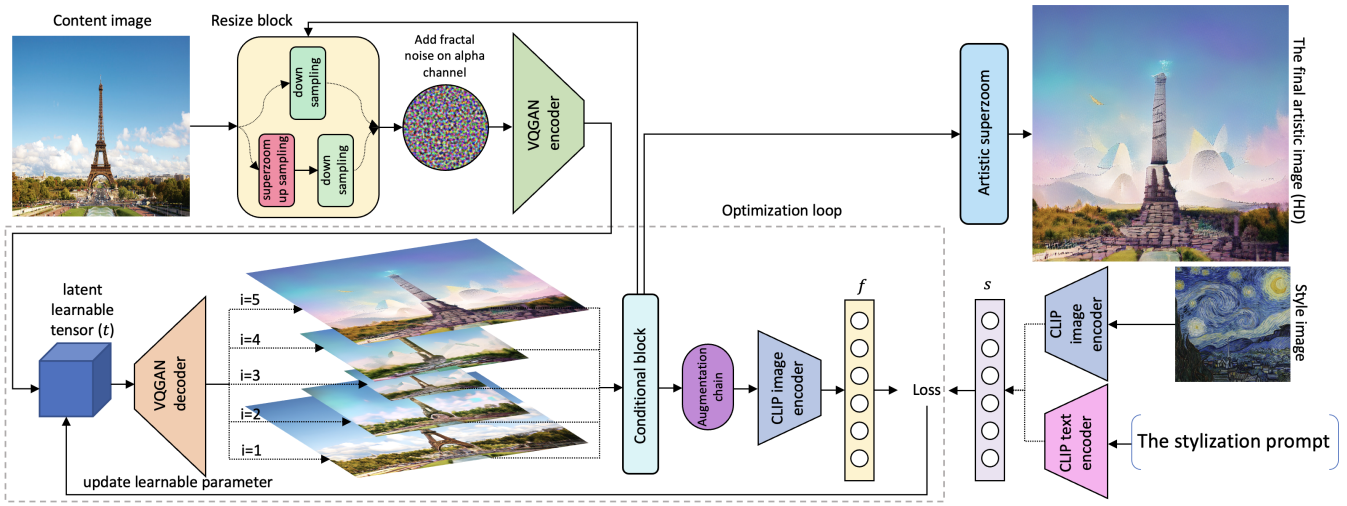

Figure 2: Overview of the proposed stylization pipeline with CLIP and a generative network (VQGAN). The optimization loop is framed in the dotted rectangle. The vector f represents the projection in the CLIP space of the image from the current iteration and the vector s represents the projection in the CLIP space of a style parameter (image or text).

of the generative model is defined. In our case where the generative model is VQGAN, this tensor t is the encoding of the content image, $t = Encoder(O)$ where $Encoder(\cdot)$ is the VQGAN encoder and O is the content image (for the case without content images, t is randomly sampled in VQGAN latent space). This step is performed only once regardless of the number of iterations.

A stylization step represents a change to this learnable tensor, which is then decoded by the generative model into an RGB image. This modification is performed as follows: in the first phase, the learnable tensor is passed through the generative model, obtaining an RGB image. In the second phase, this image is projected in the CLIP latent space. The third phase involves calculating the loss function between the projection of the learnable tensor and the projection of the style parameters in CLIP latent space. In the fourth phase, the gradient with respect to the learnable tensor t can be computed using standard back-propagation. At the end, the learnable parameter is updated by decreasing the partial derivatives (taken from the gradient) of the learnable parameter on each component. The loss function is:

$$\mathcal{L} = \frac{2}{|\mathcal{S}|} \sum_{s \in \mathcal{S}} \arcsin\left(\frac{1}{2} \|\widehat{CLIP}(\widehat{Decoder}(t)) - \widehat{CLIP}(s)\|_2\right)^2, \quad (1)$$

where $\widehat{\cdot}$ represents the operation of normalizing a vector, $\mathcal{S}$ represents the set of stylization parameters (images and texts), $CLIP(\cdot)$ is the function that encodes an image or a text in the CLIP latent space, t is the learnable tensor and $Decoder(\cdot)$ is the function that project the learnable tensor in the space of RGB images.

To improve the speed and quality of the results, we use hierarchical scaling to different resolutions, fractal noise combined with an augmentation chain and artistic super-resolution. Hierarchical scaling to different resolutions allows fast styling of intermediate images at lower resolutions than the initial one. Noise fractal added to the intermediate results in combination with an augmentation chain (random resize for each dimension independently, random crop, random perspective, random horizontal flip and random Gaussian noise) removes regular surfaces from the content image, which in turn eliminates the problem of vanishing gradients. Artistic super-resolution increase the resolution of the generated images and create specific paintings effects (e.g., brush strokes).

The average run-time of generating an artistic output of size 2048×2048 starting from a content image or from a random point in the VQGAN latent space is about 4.9s (depending on the stylization parameters) on a single GeForce RTX-3090 GPU.

3 CONCLUSION

In this work, we propose a complete framework that generates visual art. Unlike previous stylization methods that are not flexible with style parameters (i.e., they allow stylization with only one style image, a single stylization text or stylization of a content image from a certain domain), our method has no such restriction. In addition, we implement an improved version that can generate a wide range of results with varying degrees of detail, style and structure, with a boost in generation speed. To further enhance the results, we insert an artistic super-resolution module in the generative pipeline. This module will bring additional details such as patterns specific to painters, slight brush marks, and so on.

REFERENCES

Andrew Brock, Jeff Donahue, and Karen Simonyan. 2019. Large Scale GAN Training for High Fidelity Natural Image Synthesis. *ArXiv* abs/1809.11096 (2019).
Prafulla Dhariwal and Alexander Nichol. 2021. Diffusion models beat gans on image synthesis. *NeurIPS* (2021).
Patrick Esser, Robin Rombach, and Bjorn Ommer. 2021. Taming Transformers for High-Resolution Image Synthesis. In *CVPR*.
Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. 2021. StyleCLIP: Text-Driven Manipulation of StyleGAN Imagery. In *ICCV*.
Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *ICML*.
Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-Shot Text-to-Image Gen.. In *ICML*.
Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. 2016. Generative Adversarial Text to Image Synthesis. In *ICML*.
Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. 2018. AttnGAN: Fine-Grained Text to Image Generation With Attentional Generative Adversarial Networks. In *CVPR*.