

---

# StyleCLIPDraw: Coupling Content and Style in Text-to-Drawing Synthesis

---

**Peter Schaldenbrand**  
Robotics Institute  
Carnegie Mellon University  
Pittsburgh, PA 15213  
pschalde@andrew.cmu.edu

**Zhixuan Liu**  
School of Data Science  
The Chinese University  
of Hong Kong  
Shenzhen, China  
zhixuanliu@cuhk.edu.cn

**Jean Oh**  
Robotics Institute  
Carnegie Mellon University  
Pittsburgh, PA 15213  
jeanoh@nrec.ri.cmu.edu

## Abstract

Generating images that fit a given text description using machine learning has improved greatly with the release of technologies such as the CLIP[4] image-text encoder model; however, current methods lack artistic control of the style of image to be generated. We introduce StyleCLIPDraw which adds a style loss to the CLIPDraw[1] text-to-drawing synthesis model to allow artistic control of the synthesized drawings in addition to control of the content via text. Whereas performing decoupled style transfer on a generated image only affects the texture, our proposed coupled approach is able to capture a style in both texture and shape, suggesting that the style of the drawing is coupled with the drawing process itself. More results and our code are available at <https://github.com/pschaldenbrand/StyleCLIPDraw>

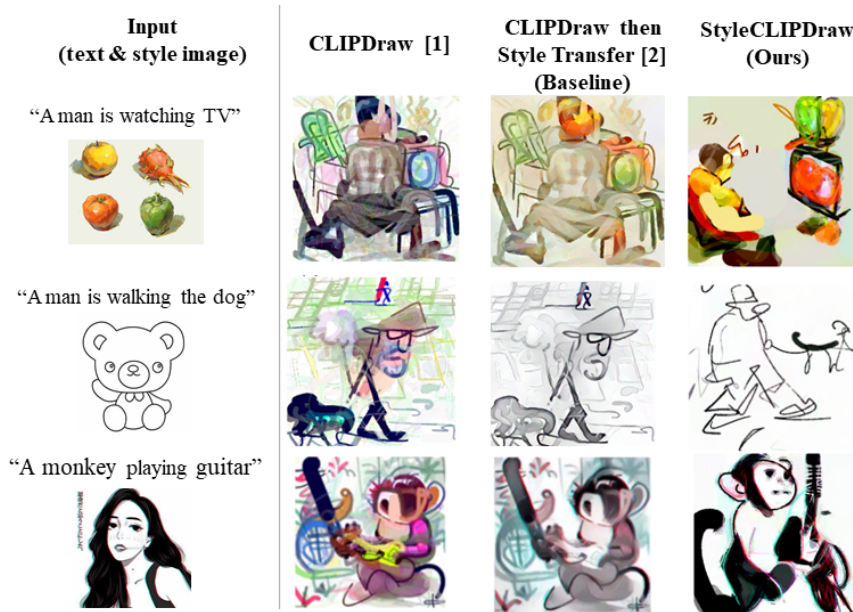


Figure 1: Comparing styled text-to-drawing results. The baseline is formed by using CLIPDraw [1] to convert the input text into an image, then performing style transfer [2]. Our model, StyleCLIPDraw, couples style and content by generating the drawing using both the text and style simultaneously.

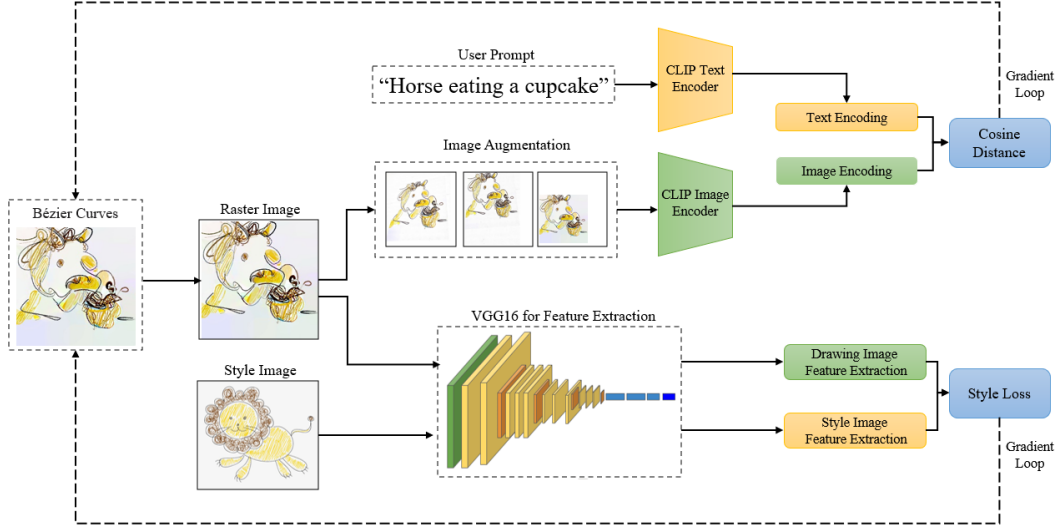


Figure 2: The StyleCLIPDraw model architecture. The Bézier curve representation is converted into a raster image and is then used to compute two losses: One loss for aligning the content of the image with the user’s text prompt and the other for aligning style.

**Introduction** Text-to-drawing synthesis models provide a method for anyone to generate artistic visuals by describing the contents in natural language. While drawing by hand or using software to render images can require specialized skills and training that not everyone may have the means for, text-to-drawing synthesis can serve as a method for individuals to generate visuals by proxy of a skill that they do have, language. Models such as CLIPDraw [1] or Dall-E [5] can create images based on text inputs, but these existing approaches lack additional control of the generation process such as style. We introduce an additional input to the CLIPDraw text-to-drawing synthesis model to alter the style of the generated visuals.

**Method** Unlike most other image generation models, CLIPDraw produces drawings consisting of a series of Bézier curves defined by a list of coordinates, a color, and an opacity. The drawing begins as randomized Bézier curves on a canvas and is optimized to fit the given style and text. The StyleCLIPDraw model architecture is shown in Fig. 2. The brush strokes are rendered into a raster image via differentiable model [3]. There are two losses for StyleCLIPDraw that correspond to each input. The text input and the augmented raster drawing are fed the the CLIP model and the difference in embeddings are compared using cosine distance to compute a loss that encourages the drawing to fit the text input. The image is augmented to avoid finding shallow solutions to optimizing through the CLIP model. The raster image and the style image are fed through early layers of the VGG-16 [6] model (per the STROTSS style-transfer algorithm [2]) and the difference in extracted features form the loss that encourages the drawings to fit the style of the style image.

**Results** Because CLIP[4] is trained on such a large, diverse dataset, StyleCLIPDraw works on almost arbitrarily general text prompts. While the StyleCLIPDraw results shown in Fig. 1 lack some realism with respect to fitting the text prompt exactly, all images show strong evidence of the given style and have a resemblance to the text which describes them. The content of the StyleCLIPDraw drawings is no less recognizable than that of the CLIPDraw drawings, indicating that adding the style loss did not adversely affect the optimization for the text description loss.

**Conclusions** In the second example of Fig. 1, the baseline results in a gray drawing with shading, but the given style image was a black, line drawing. StyleCLIPDraw was able to synthesize a proper line drawing that fit the content from the input text description. Performing style transfer on the CLIPDraw drawings changes the colors and textures of the drawings but does not change the shape of the content which is required for drawings in a certain style. Simultaneously optimizing style and content in the drawing synthesis process is able to properly affect the shape of the image to fit the given style, indicating that the style of the drawing is coupled with the drawing process itself.

## Ethical Considerations

StyleCLIPDraw relies heavily on the feedback from the CLIP[4] image-text encoding model. CLIP was trained on 400 million image-text pairs scraped from the internet, and this dataset is not made publicly available. As pointed out in the original CLIPDraw paper[1], the biases in this data will be reflected in the generated images from the model. The biases of the CLIP model have been investigated[4], and it is important to recognize the presence of them when utilizing StyleCLIPDraw.

## References

- [1] Kevin Frans, LB Soros, and Olaf Witkowski. Clipdraw: Exploring text-to-drawing synthesis through language-image encoders. *arXiv preprint arXiv:2106.14843*, 2021.
- [2] Nicholas Kolkin, Jason Salavon, and Gregory Shakhnarovich. Style transfer by relaxed optimal transport and self-similarity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10051–10060, 2019.
- [3] Tzu-Mao Li, Michal Lukáč, Michaël Gharbi, and Jonathan Ragan-Kelley. Differentiable vector graphics rasterization for editing and learning. *ACM Transactions on Graphics (TOG)*, 39(6):1–15, 2020.
- [4] 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. Learning transferable visual models from natural language supervision. *CoRR*, abs/2103.00020, 2021.
- [5] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. *arXiv preprint arXiv:2102.12092*, 2021.
- [6] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.