# VectorFusion: Text-to-SVG by Abstracting Pixel-Based Diffusion Models

Ajay Jain  Amber Xie  Pieter Abbeel

**摘要**

扩散模型在文本到图像合成方面显示出非常优秀的结果。我们能够使用扩散模型学习并生成多样化的光栅图像。然而，设计人员经常使用图像的矢量表示形式，例如可缩放矢量图形 (SVG)，这是因为矢量图形可以缩放到任意大小。本工作证明了像素图像上训练的文本条件扩散模型可用于生成矢量图像，且无需访问带标题的 SVG 的大型数据集。通过优化可微分矢量图形光栅化器，我们的方法 VectorFusion 从预训练的扩散模型中提取抽象语义知识，生成合理的矢量图像。受最近文本转 3D 工作的启发，我们进一步使用分数蒸馏采样学习与标题一致的 SVG。为了加速生成并提高保真度，VectorFusion 还从图像样本进行初始化。实验显示出比之前的工作更高的质量。

**关键词**：扩散模型；矢量图；图像生成

## 1 引言

在图形设计和艺术领域，设计师和艺术家经常以抽象的方式表达概念，例如通过组合一些形状和线条来唤起场景的精髓。可缩放矢量图形（SVGs）提供了一种声明性格式，可以将视觉概念表达为一系列基本元素的集合。这些基本元素包括贝塞尔曲线、多边形、圆形、线条和背景颜色。SVGs是导出图形设计的事实上的标准格式，因为它们可以在用户设备上以任意高分辨率呈现，同时以紧凑的大小存储和传输，通常只有几十千字节，而且传统的栅格图像到矢量图形的转换工具存在信息损失，难以保留图像细节。然而，设计矢量图形是困难的，需要了解专业的设计工具。扩散模型在图像生成领域的成功，以及它们在新任务上的潜力，使人考虑利用强大的生成模型来创建高质量的矢量艺术作品，而不受大规模矢量图形数据集的限制。因此设计了VectorFusion。VectorFusion不仅为设计师和艺术家提供了一种新的工具来从文本描述中生成矢量图形，而且展示了扩散模型在像素图像生成之外的新应用。这种方法提供了一种新的方式来利用现有的生成模型来创造可编辑、可扩展的矢量艺术作品，同时为未来在其他领域探索扩散模型的应用提供了新的思路。

## 2 相关工作

### 2.1 矢量图像生成

可扩展向量图形 (SVG) 为表示为原语的视觉概念提供了一种声明性格式。创建 SVG 内

容的一种方法是使用序列到序列 (seq2seq) 模型来生成 SVG 脚本 [17]。这些方法严重依赖向量形式的数据集，这限制了它们的泛化能力及其合成复杂向量图形的能力。我们使用的向量合成方法不是直接学习SVG生成网络，而是针对匹配图像进行优化。

Li等人 [5]引入了一种可微光栅化器，在光栅图和矢量图之间建立了桥梁。虽然传统上在矢量图形上运行的图像生成方法需要一个基于矢量的数据集，但最近的工作已经证明了可以使用可微渲染器来克服这一限制 [12]。此外，视觉文本嵌入对比语言图像预训练模型(CLIP) [10]的最新进展使许多的方法能够成功合成矢量草图，如CLIPDraw [1]和CLIPasso [15]。本工作将可微渲染器与文本到图像扩散模型相结合以生成矢量图形，取得了较好的成果。

### 2.2 文本到图像的扩散模型

去噪扩散概率模型 (DDPM) [3]，尤其是那些以文本为条件的模型，在文本到图像合成中显示出有希望的结果。例如，无分类器指导(CFG) [4]提高了视觉质量，广泛应用于大规模文本条件扩散模型框架中，包括stable diffusion [13]、Imagen [14]。文本到图像扩散模型取得的重大进展也促进了一系列文本引导任务的发展，例如文本到3D [9]。在本工作中，我们使用stable diffusion模型为文本到SVG的生成提供监督。

### 2.3 分数蒸馏采样

自然图像建模的最新进展激发了人们对利用强大的 2D 预训练模型来恢复 3D 对象结构的重要研究兴趣。最近的工作，如DreamFusion [9]、Magic3D [6]，通过利用从2D文本到图像扩散模型派生的分数蒸馏采样(SDS)损失来进行文本到3D的生成，显示出令人印象深刻的结果。文本到SVG的发展受此启发，但得到的矢量图形的质量有限，表现出与重建的3D模型相似的过平滑现象。Wang等人 [18]将3D模型的建模扩展为随机变量，而不是像SDS那样是一个常数，并提出了变分评分蒸馏来解决文本到3D生成的过平滑问题。

## 3 本文方法

### 3.1 本文方法概述

简单的说，本工作分为两部分：第一部分从文本提示中生成光栅图像，并转化为矢量图；第二部分对生成的矢量图进行优化，提高生成质量。具体流程如图所示：

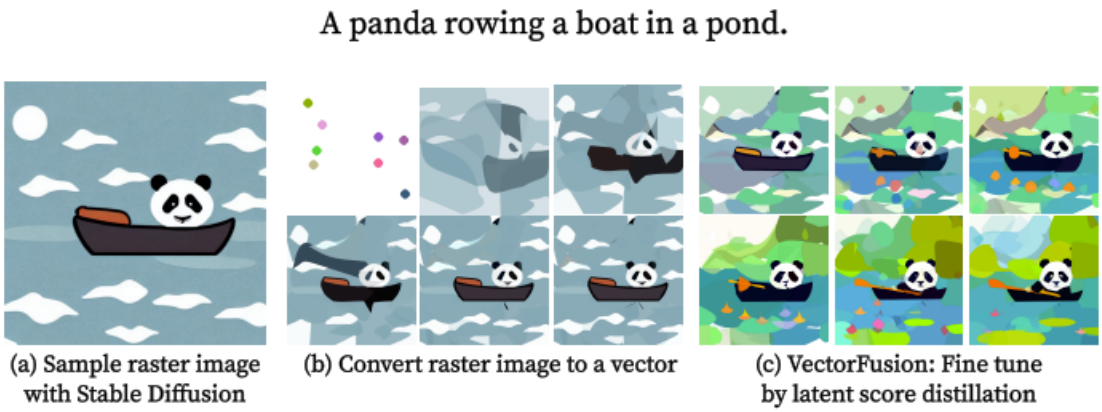


图 1. 方法示意图

### 3.2 文本到图像到矢量图的生成

我们首先构建一个两阶段管道：从稳定扩散中采样图像，然后自动对其进行矢量化。给定文本，我们使用 Runge-Kutta 解算器 [7]以 50 个采样步骤对稳定扩散中的光栅图像进行采样，引导比例 $\omega = 7.5$（Diffusers 库中的默认设置 [16]）。扩散模型生成的摄影风格和细节很难用一些恒定颜色的 SVG 路径来表达。为了鼓励使用抽象、平面矢量风格的图像生成，我们在文本后附加一个后缀："minimal flat 2d vector icon. lineal color. on a white background. trending on artstation."。由于样本可能与标题不一致，因此我们对 K 个图像进行采样，并根据 CLIP ViTB/16 [10]选择与标题最一致的样本。 CLIP 重新排序最初是由 [11]提出的。我们选择K=4。接下来，我们使用现成的逐层图像矢量化程序 (LIVE) [8]自动跟踪栅格样本，将其转换为 SVG。这会产生一组路径。上图（b）显示了分阶段优化矢量参数的过程。虽然简单，但此管道通常会创建不适合矢量化的图像。

### 3.3 基于优化的矢量图采样

之前构建的管道存在缺陷，因为样本可能不容易用一组路径表示。图 4 说明了该问题。以文本为条件，扩散模型根据分布 $p(x|y)$ 生成样本。使用 LIVE 进行矢量化可找到与该图像具有接近 L2 近似值的 SVG，而无需使用标题 y。这可能会丢失信息，并且生成的 SVG 图形可能不再与标题一致。

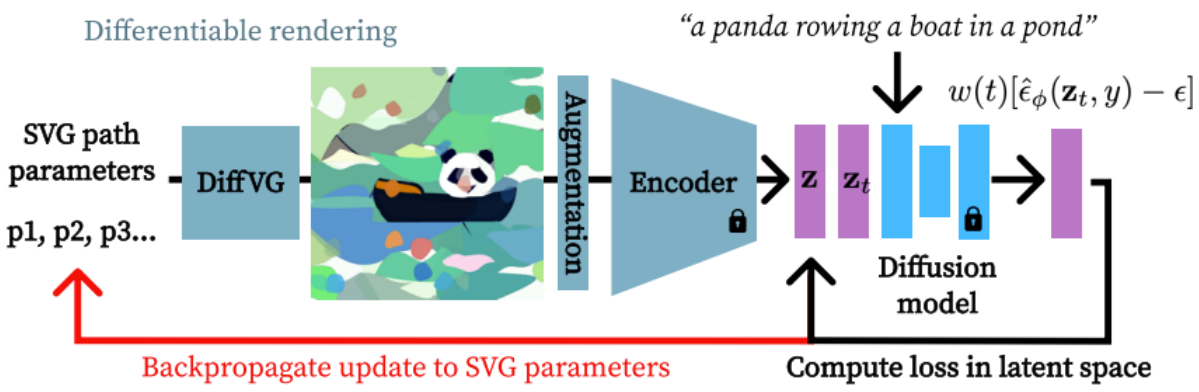


图 2. 优化过程示意

对于 VectorFusion，我们采用分数蒸馏采样来支持潜在扩散模型 (LDM)，例如开源的stable diffusion模型。我们用一组路径初始化 SVG。每次迭代，DiffVG 都会渲染 600 × 600 图像 x。与 CLIPDraw [1]一样，我们通过透视变换和随机裁剪进行增强，获得一张512×512 图像。然后，我们建议使用 LDM 编码器$E$ 计算潜在空间中的 SDS 损失。对于每次优化迭代，我们使用随机噪声对模型进行扩散–去噪过程，并优化 SDS 损失。

# 4 复现细节

## 4.1 与已有开源代码对比

本工作没有任何官方开源代码，复现过程中参考了非官方实现中的基本代码框架。这次复现过程中通过对生成的矢量表示添加额外的约束条件，来实现论文中所提到的一些特定艺术风格的矢量图生成。具体生成结果可见实验结果分析。

## 4.2 实验环境搭建

复现使用ubuntu服务器，使用TITAN XP显卡进行模型运行和矢量图优化。

## 4.3 创新点

创新点在于实现了对特定艺术风格矢量图像的生成，如草图风格和像素画风格。

# 5 实验结果分析

以下是一些生成的结果。图像和对应的文字提示如下所示。

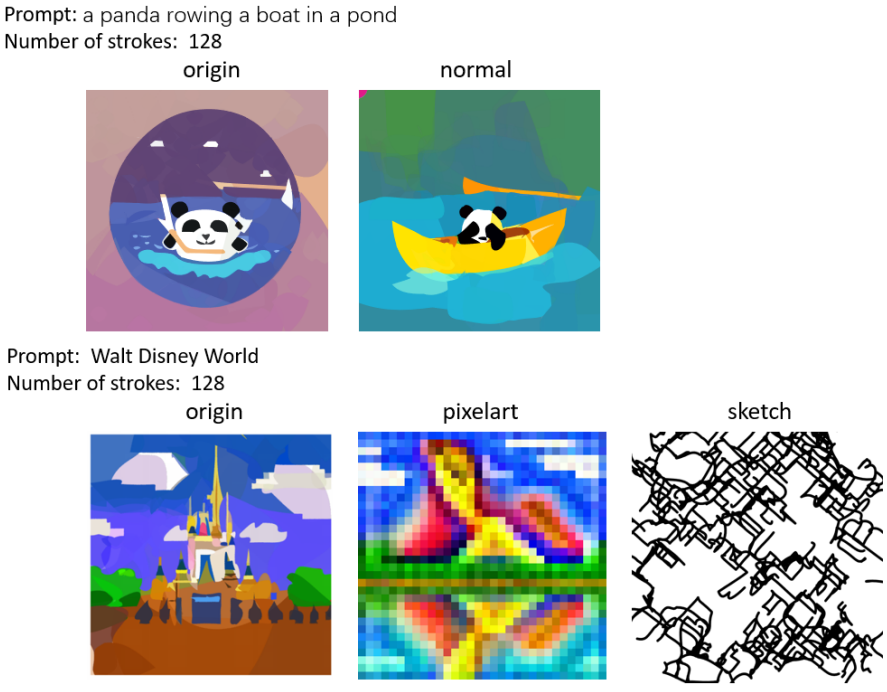


图 3. 实验结果示意。上面是一般的生成结果，下图示例是经过特定笔画约束后的生成结果。

可以看到通过对文字提示和矢量图笔画属性进行约束，我们能够生成特定风格的矢量图像。

## 6 总结与展望

本次复现中实现了Vectorfusion工作，实现了基于文本生成特定风格的矢量图。本工作的不足之处在于运行速度方面较为耗时，单图生成对于一般性能gpu需要耗时20分钟左右，结果的多样性也不足。除此之外，生成图像质量很大程度上受到生成的光栅图的质量的影响，这也意味着随着文本生成模型的进步，能够存在更多的改进空间。本工作在由光栅图到矢量图的过程中没有考虑语义信息，可以考虑在矢量化的过程中引入语义信息来减少矢量化过程中原图信息的丢失。

## 参考文献


[1] Kevin Frans, L. B. Soros, and Olaf Witkowski. Clipdraw: Exploring text-to-drawing synthesis through language-image encoders, 2021.

[2] A Geiger, P Lenz, C Stiller, and R Urtasun. Vision meets robotics: The KITTI dataset. *Int. J. Robotics Research*, 32(11):1231–1237, 2013.

[3] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020.

[4] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance, 2022.

[5] Tzu-Mao Li, Michal Lukáč, Michaël Gharbi, and Jonathan Ragan-Kelley. Differentiable vector graphics rasterization for editing and learning. *ACM Trans. Graph.*, 39(6), nov 2020.

[6] Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution text-to-3d content creation, 2023.

[7] Luping Liu, Yi Ren, Zhijie Lin, and Zhou Zhao. Pseudo numerical methods for diffusion models on manifolds. *arXiv preprint arXiv:2202.09778*, 2022.

[8] Xu Ma, Yuqian Zhou, Xingqian Xu, Bin Sun, Valerii Filev, Nikita Orlov, Yun Fu, and Humphrey Shi. Towards layer-wise image vectorization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16314–16323, 2022.

[9] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion, 2022.

[10] 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, 2021.

[11] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021.

[12] Pradyumna Reddy, Michaël Gharbi, Michal Lukáč, and Niloy J. Mitra. Im2vec: Synthesizing vector graphics without vector supervision. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7338–7347, 2021.

[13] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022.

[14] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding, 2022.

[15] Yael Vinker, Ehsan Pajouheshgar, Jessica Y. Bo, Roman Christian Bachmann, Amit Haim Bermano, Daniel Cohen-Or, Amir Zamir, and Ariel Shamir. Clipasso: Semantically-aware object sketching. *ACM Trans. Graph.*, 41(4), jul 2022.

[16] Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, and Thomas Wolf. Diffusers: State-of-the-art diffusion models. https://github.com/huggingface/diffusers, 2022.

[17] Yizhi Wang, Gu Pu, Wenhan Luo, Pengfei Wang, Yexin ans Xiong, Hongwen Kang, Zhonghao Wang, and Zhouhui Lian. Aesthetic text logo synthesis via content-aware layout inferring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[18] Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation, 2023.