

**论文标题:** Adding Conditional Control to Text-to-Image Diffusion Models

**论文作者:** Lvmin Zhang, Anyi Rao, and Maneesh Agrawala Stanford University

**论文出处:** International Conference on Computer Vision (ICCV)

**论文年份:** 2023

**作者单位:** 斯坦福大学

**关键词:** 条件控制, 扩散模型, 模型重用

**论文概述:** 本文提出了一种条件控制神经网络架构 ControlNet, 旨在为大型、预训练的文本到图像扩散模型添加空间条件控制。该架构以准备就绪的大型扩散模型为基础, 通过重用深入且稳健的预训练编码层实现学习多样化的条件控制。该架构通过引入“零卷积”(零初始化的卷积层)连接来确保微调过程中不受有害噪声影响。为了验证有效性, 团队针对各种条件控制, 例如边缘、深度、分割、人体姿势等进行测试, 实验结果表明提出的架构可以使得扩散模型适用于各种条件控制并保证训练鲁棒性。

**研究问题:** 通过让用户提供额外的图像来直接指定他们所需的图像组成来实现更细粒度的空间控制。

**研究意义:** 扩散模型可以根据输入条件进行定向精准地生成目标图像, 为 AIGC 发展注入动力, 带来图像制作产业的进一步变革

**早期方法:**

### 1. 控制文本到图像模型的方法:

**空间掩码:** 这是一种在生成图像时用于控制图像特定区域的技术。通过在模型中加入空间掩码, 可以指定哪些区域应该被保留或修改, 从而实现更精细的图像控制。

**图像编辑指令:** 这种方法允许用户通过指令直接对生成的图像进行编辑, 比如修改颜色、添加元素等。

**个性化微调:** 通过微调模型, 可以使其适应特定用户或特定任务的需求, 从而生成更加个性化的图像。

### 2. 解决文本到图像生成问题的方法:

**无训练技术:** 对于一些简单的问题, 如生成图像的不同变体或图像修复, 可以通过无需额外训练的技术来解决。例如, 通过限制去噪扩散过程或编辑注意力层激活, 可以在不重新训练整个模型的情况下实现这些功能。

### 3. 端到端学习和数据驱动解决方案:

对于更复杂的问题, 如从深度信息生成图像、从姿态信息生成图像等, 通常需要采用端到端的学习方法和数据驱动的解决方案。这是因为这些问题涉及到多个复杂因素之间的相互作用, 需要模型能够从大量数据中学习到这些因素的映射关系。

**相关工作及优缺点:**

#### 1. GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models

**优点:** 相比无引导策略, 生成的图像更有照片一样的感觉, 更符合人类感官

**缺点:** 使用了成本昂贵的 CLIP 排序

#### 2. High-Resolution Image Synthesis with Latent Diffusion Models

**优点:** 首次允许在降低复杂性和保留细节之间达到近乎最佳点, 从而大大提高视觉保真度。在各种任务上极具竞争力的性能, 包括文本到图像合成、无条件图像生成和超分辨率, 同时与基于像素的 DM 相比, 显著降低了计算要求。

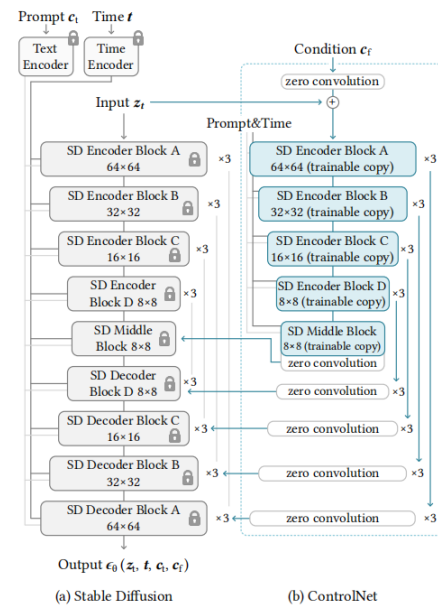
**缺点:** 生成的图像仍难以达到人类视觉的美感

#### 3. Learning Transferable Visual Models From Natural Language Supervision

优点：模型可以很好地实现无训练迁移，且能表征学习、认知学习等 30 多种任务上与之前的有监督模型进行 PK

缺点：但是生成的图像质量一般

文章核心思路：



文章的神经网络结构基于稳定扩散模型进行改进，通过引入零卷积块和控制条件实现模型指导，结构如上图所示

论文数据集：

常用的网上公开图像数据集

实验设置：

1. 基本实验：文章用具有稳定扩散的控制网测试了各种反射条件，包括 Canny Edge, Depth Map, Normal Map, M-LSD lines, HED soft edge, ADE20K segmentation, Openpose 和 user sketches.
2. 消融研究：文章设计了三组消融实验，其中一组是用高斯权值初始化的标准卷积层，第二组是用标准卷积层替换零卷积，第三组是用一个卷积层替换每个块的可训练副本，
3. 效果评估：文章通过提供了 4 个提示设置来测试真实世界用户可能的行为，评测不同设置下的模型效果。这四种设置分别是(1)没有提示；(2)没有充分覆盖对象的提示不足，例如，本文的默认提示“高质量、详细、专业的图像”；(3)冲突的提示改变了条件图像的语义；(4)描述必要内容语义的完美提示，如“好房子”。
4. 对比实验：文章将工作与另外几个相关工作进行了对比，包括 PITI, LDM, Sketch-Guided Diffusion 和 Taming Transformers。实验表明，相比其他网络，控制网可以更稳健地处理不同的条件反射图像，并达到清晰和干净的结果。