

This is the weekly report by Wentao from 2022/08/23 to 2022/09/05

topics

- Diffusion model
- VAE
- Gesture Generation
- stable diffusion
- skin lesion classification

Reading list

- **Diffusion model**

[1] A DIFFUSION GENERATED METHOD FOR ORTHOGONAL MATRIX-VALUED FIELDS

[2] The iterative convolution–thresholding method (ICTM) for image segmentation

[3] DIFFUSION GENERATED METHODS FOR DENOISING TARGET-VALUED IMAGES

[4] High-Resolution Image Synthesis with Latent Diffusion Models

[5] Hierarchical Text-Conditional Image Generation with CLIP Latents

- **Gesture generation**

[1] Towards Automatic Speech to Sign Language Generation

[2] ANONYSIGN: Novel Human Appearance Synthesis for Sign

[3] Speech Drives Templates: Co-Speech Gesture Synthesis with Learned Templates

[4] Speech Gesture Generation from the Trimodal Context of Text, Audio, and Speaker Identity

[5] Text2video: text-driven talking-head video synthesis with personalized phoneme - pose dictionary

[6] SEEG: Semantic Energized Co-speech Gesture Generation

[7] SignPose: Sign Language Animation Through 3D Pose Lifting

[8] Gesture2Vec: Clustering Gestures using Representation Learning Methods for Co-speech Gesture Generation

[9] Evaluation of text-to-gesture generation model using convolutional neural network

[10] Everybody Sign Now: Translating Spoken Language to Photo Realistic Sign Language Video

[11] Progressive Transformers for End-to-End Sign Language Production

[12] Modeling Intensification for Sign Language Generation: A Computational Approach

[13] Can Everybody Sign Now? Exploring Sign Language Video Generation from 2D Poses

[14] Analyzing Input and Output Representations for Speech-Driven Gesture Generation

[15] Learning Hierarchical Cross-Modal Association for Co-Speech Gesture Generation

- **skin lesion classification**

details

VAE

Diffusion model

说到扩散模型，一般的文章都会提到能量模型（Energy-based Models）、得分匹配（Score Matching）、朗之万方程（Langevin Equation）等等，简单来说，是通过得分匹配等技术来训练能量模型，然后通过郎之万方程来执行从能量模型的采样。

从理论上来讲，这是一套很成熟的方案，原则上可以实现任何连续型对象（语音、图像等）的生成和采样。但从实践角度来看，能量函数的训练是一件很艰难的事情，尤其是数据维度比较大（比如高分辨率图像）时，很难训练出完备能量函数来；另一方面，通过郎之万方程从能量模型的采样也有很大的不确定性，得到的往往是带有噪声的采样结果。所以很长时间以来，这种传统路径的扩散模型只是在比较低分辨率的图像上做实验。

如今生成扩散模型的大火，则是始于2020年所提出的[DDPM \(Denoising Diffusion Probabilistic Model\)](#)，虽然也用了“扩散模型”这个名字，但事实上除了采样过程**的形式有一定的相似之外，DDPM与传统基于郎之万方程采样的扩散模型可以说完全不一样，这完全是一个新的起点、新的篇章。

准确来说，DDPM叫“渐变模型”更为准确一些，扩散模型这一名字反而容易造成理解上的误解，传统扩散模型的能量模型、得分匹配、郎之万方程等概念，其实跟DDPM及其后续变体都没什么关系。有意思的是，DDPM的数学框架其实在ICML2015的论文[《Deep Unsupervised Learning using Nonequilibrium Thermodynamics》](#)就已经完成了，但DDPM是首次将它在高分辨率图像生成上调试出来了，从而引导出了后面的火热。由此可见，一个模型的诞生和流行，往往还需要时间和机遇。

拆楼建楼：

实际上，相比于传统的扩散模型，DDPM实际上更像VAE而不是扩散模型。很多文章在介绍DDPM时，上来就引入转移分布，接着就是变分推断，一堆数学记号下来，先吓跑了一群人，再加之人们对传统扩散模型的固有印象，所以就形成了“需要很高深的数学知识”的错觉。事实上，DDPM也可以有一种很“大白话”的理解，它并不比有着“造假-鉴别”通俗类比的GAN更难。

首先，我们想要做一个像GAN那样的生成模型，它实际上是将一个随机噪声 z 转换成一个数据样本 x 的过程：

如果我们把一个噪声看成一个没有修建的地基，那么我们增加高斯噪声的过程就等于是在这个地基上添砖加瓦，最后我们变换完成，得到一个完整的样本数据。

设

$$x = x_0 \rightarrow x_1 \rightarrow x_2 \rightarrow \dots \rightarrow x_{T-1} \rightarrow x_T = z$$

x_0 为建好的高楼大厦（数据样本）， x_T 为拆好的砖瓦水泥（随机噪声），假设“拆楼”需要 T 步，整个过程可以表示为建高楼大厦的难度在于，从原材料 x_T 到最终高楼大厦 x_0 的跨度过大，很难理解 x_T 是怎么一下子变成 x_0 的。但是，当我们有了“拆楼”的中间过程 x_1, x_2, \dots, x_{T-1} 后，我们知道 $x_{T-1} \rightarrow x_T$ 代表着拆楼的一步，那么反过来 $x_T \rightarrow x_{T-1}$ 不就是建楼的一步？如果我们能学会两者之间的变换关系 $x_{T-1} = \mu(x_T)$ ，那么从 x_T 出发，反复地执行 $x_{T-1} = \mu(x_T)$ 、 $x_{T-2} = \mu(x_{T-1})$ 、...，最终不就能造出高楼大厦 x_0 出来？

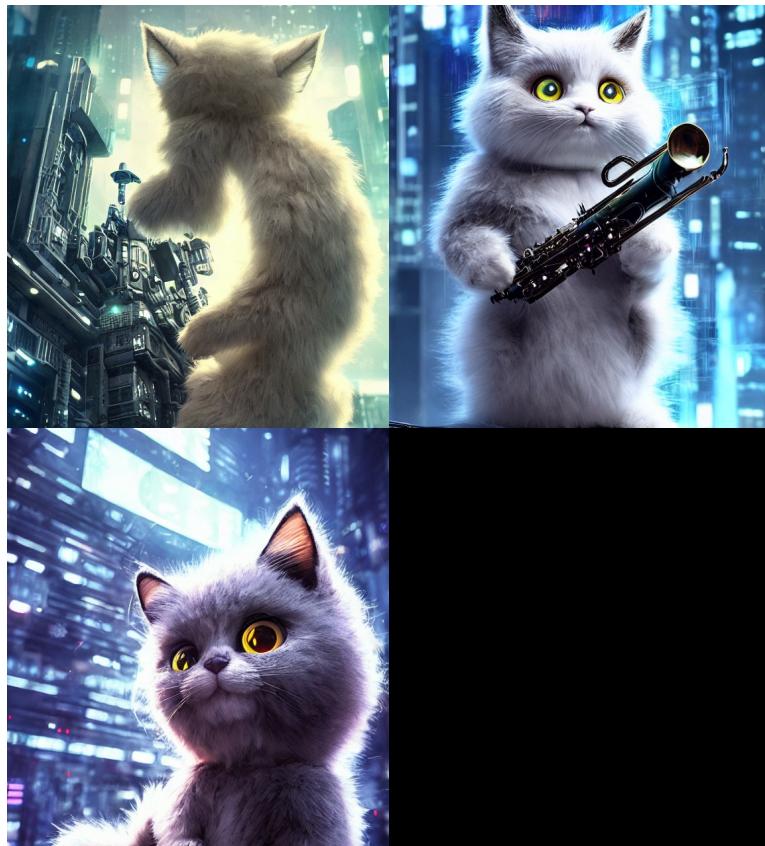
Stable Diffusion

首先，从名字[Stable Diffusion](#)就可以看出，这个主要采用的扩散模型（Diffusion Model）。

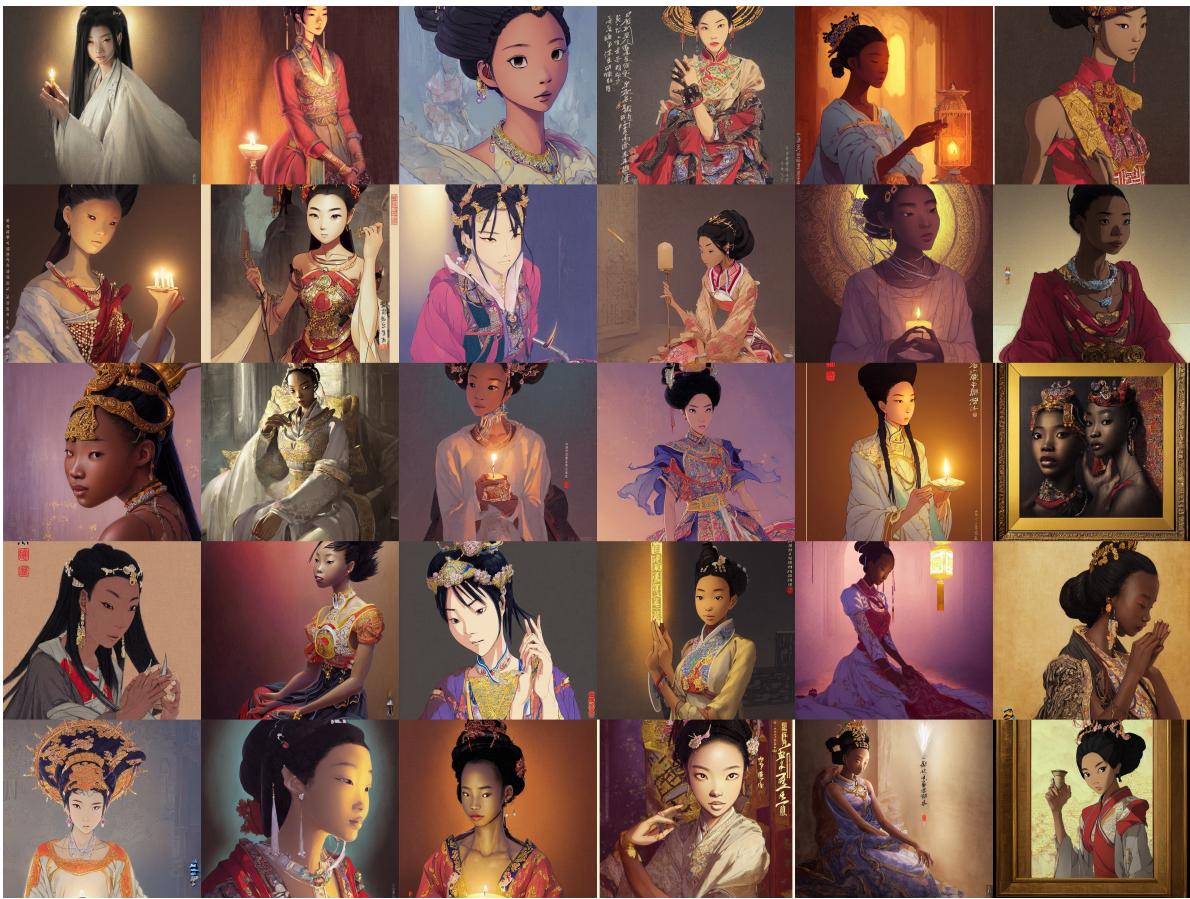
简单来说，扩散模型就是去噪自编码器的连续应用，逐步生成图像的过程。stable diffusion延续了ddpm, dalle, glides, dalle2等一系列工作，首先，使用编码器将图像 x 压缩为较低维的潜在空间表示 z （ x ）它与时间步长 t 一起，以简单连接和交叉两种方式，注入到潜在空间表示中去。随后在 z （ x ）基础上进行扩散与去噪。换言之，就是模型并不直接在图像上进行计算，从而减少了训练时间、效果更好。值得一提的是，Stable Diffusion的上下文机制非常灵活， y 不光可以是图像标签，就是蒙版图像、场景分割、空间布局，也能够相应完成。其中上下文（Context） y ，即输入的文本提示，用来指导 x 的去噪。

stable diffusion是由慕尼黑大学机器视觉与学习研究小组和Runway的研究人员，基于CVPR2022的一篇论文《High-Resolution Image Synthesis with Latent Diffusion Models》，并与其他社区团队合作开发的一款开源模型。核心数据集是LAION-5B的一个子集，它是专为基于CLIP的新模型而创建。同时，它也是首个在4000个A100 Ezra-1 AI超大集群上进行训练的文本转图像模型。

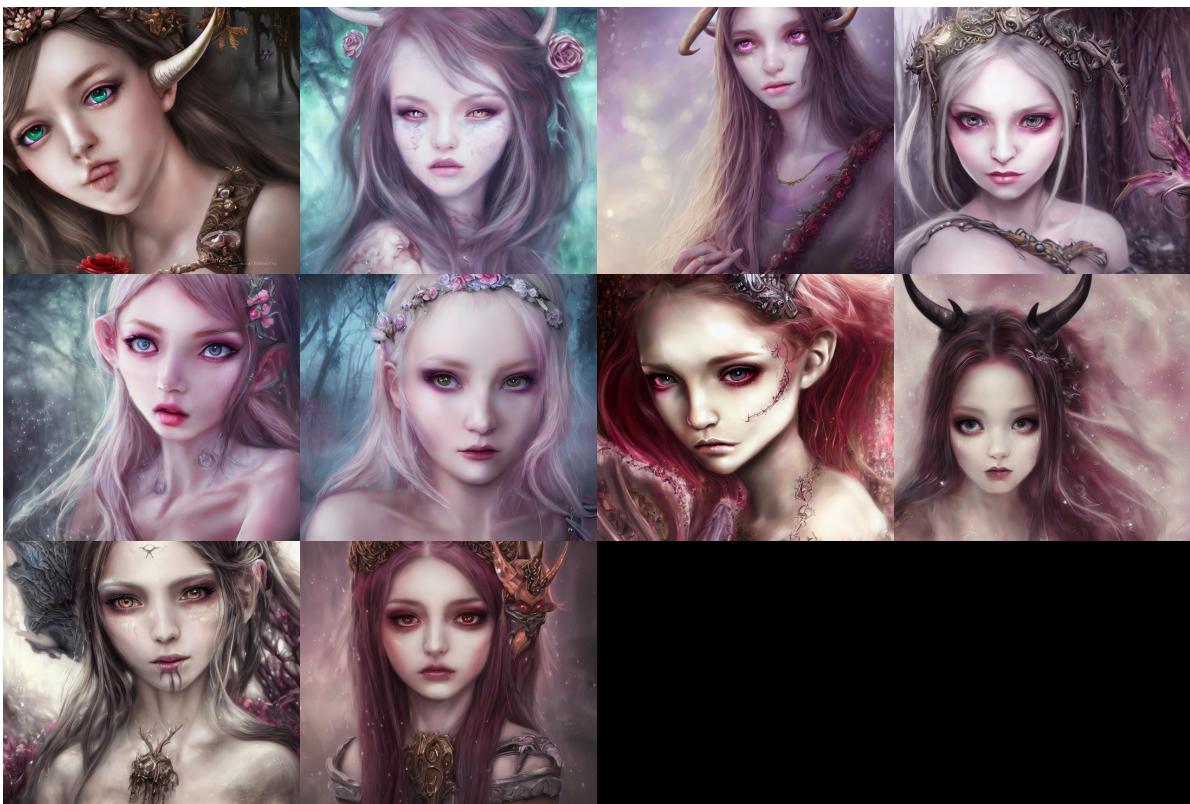
目前我成功在本地部署项目，并且部署了网页GUI，尝试了使用文本（领域内一般叫prompt）进行生成。在本地3070ti显卡上，大约15s可以生成一张图片，以下是一些生成的图片和相应的prompt：



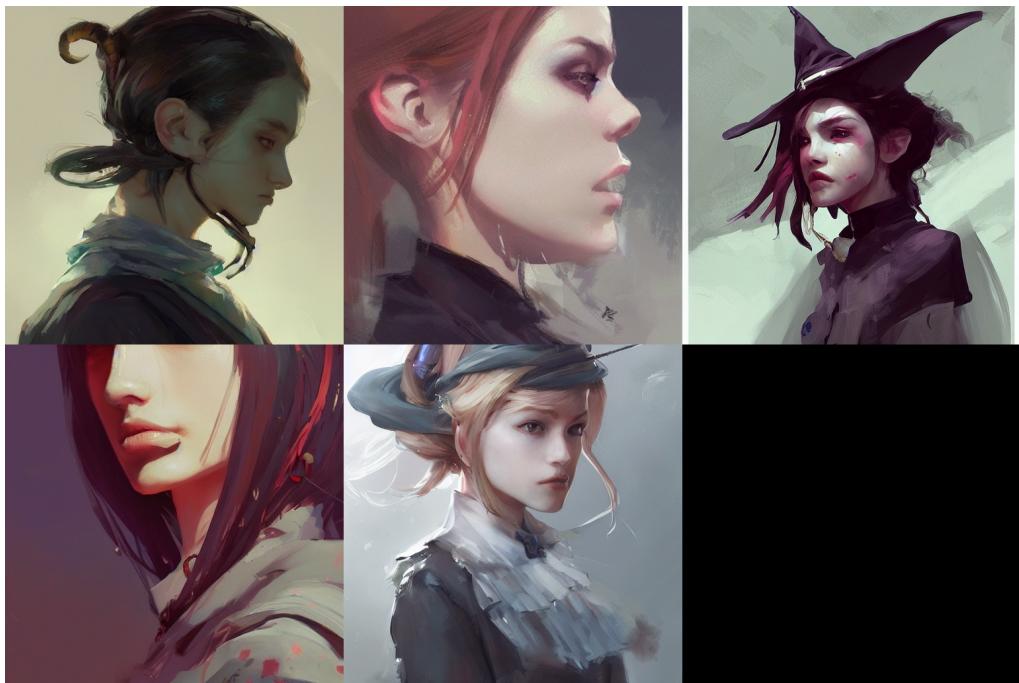
(high_quality_3_d_render_very_cute_fluffy_cyborg!!_cat_plays_trumpet,_cyberpunk_highly_detailed,_unreal_engine_cinematic_smooth)



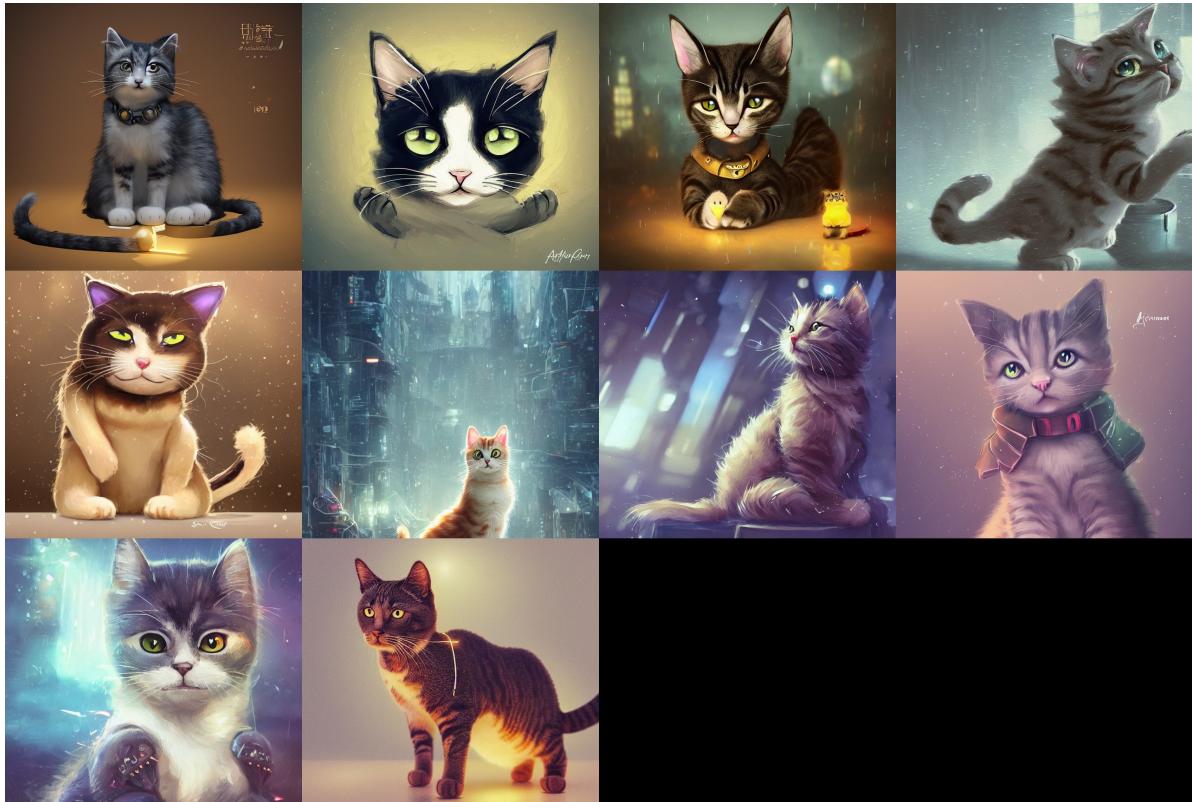
(a_portrait_of_a_chinese -
_african_princess,_candle_light,_finely_detailed_features,_perfect_art,_at_an_ancient_city,_gapmoe
_yand)



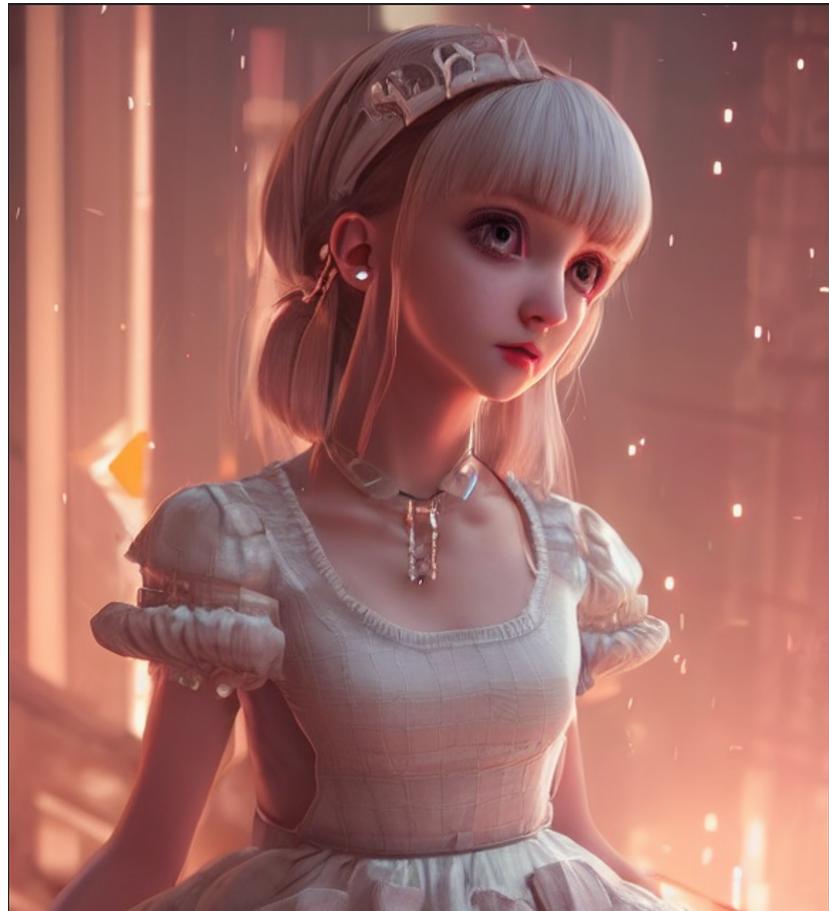
(realistic_beautiful_gorgeous_natural_cute,_fantasy,_elegant,_lovely,_princess_girl,_art_drawn_ful
l_hd,_4_k,_highest_quality)



(side_portrait_of_a_rugged_girl_witch_wearing_magic_school_uniform,_cinematic,_elaborate,_elegant,_masterpiece,_illustration,_dig)



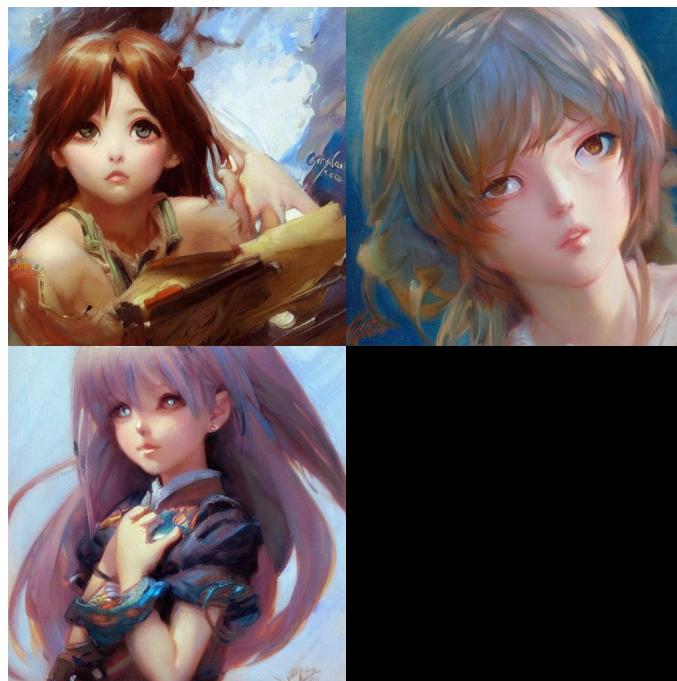
(cat_theme_logo,_cat_theme_banner,_cat_design,_art_photography_style,_trending_on_artstation,_warm_light,_lovely_and_cute,_fantas)

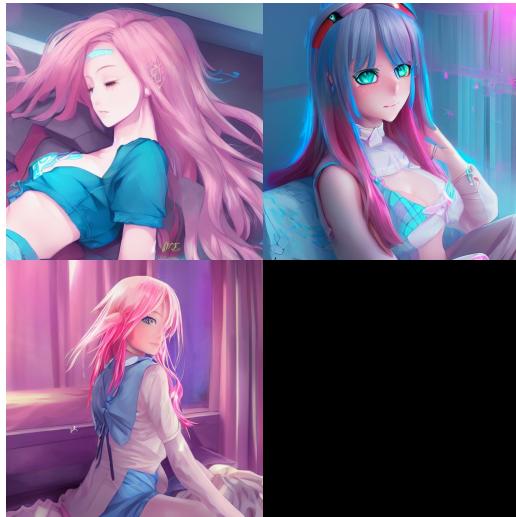


(high_quality_3_d_render_very_cute_princess_girl!!_princess,intricate_cotton_dress,_girl_plays_ga
mes,_cyberpunk,_unreal_engine)

现在应用的diffusion generation model都有一个很重要的方法就是 text guided generation (也就是用文本去引导diffusion model的扩散方向) , 领域内一般把这个文本叫做 (prompt)

也有from image to image的应用 (masked or crop)





总而言之，想要生成效果好，最重要的就是有一个适当的prompt和一个大的尺寸（我的电脑最多只能运行576x640）

Stable Diffusion Text-to-Image Unified Stable Diffusion Image-to-Image Unified

realistic beautiful gorgeous natural cute fantasy elegant lovely princess girl art drawn full hd 4 k highest quality in artstyle by professional artists wlop, taejune kim, yan gis

Height: 512 Width: 512
Classifier Free Guidance Scale (how strongly the image should follow the prompt): 7.5
Seed (blank to randomize):
Batch count (how many batches of images to generate): 2
Batch size (how many images are in a batch; memory-hungry): 1

Sampling Steps: 50
Sampling method (k_lms is default k-diffusion sampler): k_lms
Simple Advanced
Submit on enter? (no means multiline): Yes No

Seed: 124151765 Image # and click Copy to copy to img2img: 1

其中prompt的选取是最讲究的，一般可以分成五类（类型、内容，风格，作者，符号），我这里把内容和风格部分又做了一些细分：

- 图片风格 (anime、pen and ink、unreal engine cinematic smooth, A digital illustration、Full shot sketch、Lineart only、film、portrait)
- 图片内容 (cat/girl)
- 图片关系 (动作, 位置关系, doing something, on, in等)
- 图片类型(drone of footage、a cover of)
- 视角 (side view, low angle、medium shot、closeup headshot、full shot、side portrait)

- 维度 (2d/3d render)
- 图片质量 (intricate、high quality uhd8k, sharp focus highly detailed、post - processing high resolution、hdr)
- 符号 (问号, 感叹号? !)
- 光线 (moody light, *matte painting*、volumetric lighting)
- 色彩 (*fantasy vivid colors*、yellow color scheme)
- 画家描述 (greg rutkowski and thomas Kinkade)
- 绘画平台 (Trending on artstation)
- 年代 (renaissance background、neoclassical、ultrarealistic、traditional chinese painting)
- 服饰 (intricate cotton dress、lolitafashion)
- 皮肤 (porcelain skin)
- 半身 (half body, full body)
- 场景 (wuxia、war)
- 国家 (Chinese、african)
- 形容词 (gorgeous)

其中风格，类型，质量类的prompt会对生成的图片有巨大的影响。

我还做了一些实验，主要探索以下问题：

Q1:语序问题是否影响生成？

A:会影响生成，暂时不确定前后的影响哪个更大



(girl and boy)



(boy and girl)

Q2:单个词语和句子的区别?

A:单个单词之间似乎彼此不会对内容产生太大的影响（耦合程度低），在一个句子内，即使说明了内容间的关系，还是不可避免地会互相影响



(girl; boy使用分号隔开)



(girl and boy, 其中boy的部分似乎会受到girl的影响)

Q3：某些符号是否会对生成有影响？

A：问号和感叹号会对表情和场景有一定的影响



(girl ? ? boy, 其中两个问号似乎给图片产生了一个疑问的神情)



(girl ! ! ? ? boy, 在两个问号的基础上加了两个感叹号让神情变得更加惊疑)

- 后续会继续探索动画生成的部分

Gesture Generation

对姿态生成论文的调研报告汇总在项目的文档内（包括15篇论文的简介，参考价值和一个总结）

reference

[1]苏剑林. (Jun. 13, 2022). 《生成扩散模型漫谈（一）：DDPM = 拆楼 + 建楼》 [Blog post]. Retrieved from <https://kexue.fm/archives/9119>

[2]《预训练语言模型的前世今生》 [Blog post] <https://www.cnblogs.com/nickchen121/p/16470569.html>

[3]DALL·E 2【论文精读】https://www.bilibili.com/video/BV17r4y1u77B?spm_id_from=333.337.search-card.all.click&vd_source=44aab6c8d4a7cdc7288eec8d0f60b618

[4]变分自编码器 VAE 鲁鹏 https://www.bilibili.com/video/BV1Zq4y1h7Tu?spm_id_from=333.337.search-card.all.click&vd_source=44aab6c8d4a7cdc7288eec8d0f60b618