UltraPixel

状态：arxiv 文章（24.7.4）

单位：港科大（广州）/华为诺亚方舟实验室

文章链接： https://arxiv.org/abs/2407.02158

Github 链接： UltraPixel Gallery (jingjingrenabc.github.io)

目录

摘要 1

背景 2

贡献 2

实现 3

实验 4

不足 5

摘要

大尺寸图像的生成一直以来都面临着较大的挑战。而本文推出的 UltraPixel 方案可以进行任意尺度的高质量图像生成（从 1K 到 6K）。主要实现方式为先生成低分辨率的图像用以作为约束，然后生成高分辨率且富有细节的图像。此外，本文的方案由于是外加了小模块进行训练，所以对于原有模型的参数增加只有 3%，因此也不会带来更大的存储和计算开销。最后，本文通过实验证明了本文方案与当前最佳方案之间的对比情况。

文中几次提到了传统的：生成 LR 图像+对该图像进行超分=生成 HR 图像的流程。实际上，本文的思想也大致如此，只不过 LR 图像不再被作为超分的输入，而是另一个生成过程的控制条件。

LR　　BSRGAN

Ours

因此表现出的结果便是，相较于常规的超分方案，其得到的图像会更为合理且富含细节，如上图所示。当然，这实际上是通过放弃了与原图的相似性才得到的，上图中可以看出，本文方案得到的结果与 LR 差距实际上较大。但考虑到图像生成的语境，这种不相似并不会有什么问题。

背景

各种不同的文生图模型都有令人惊艳的表现。但这些模型大都只能处理 1024x1024 尺寸的图像。随着显示技术的发展，人们对于超高清图像的需求越发旺盛，即便有部分模型针对高分辨率图像进行训练，其生成效果仍然不足以令人满意。

1024×1024　2048×2048　3072×3072　4096×4096

StableCascade

简单来说，就是现有的生成模型在生成不同尺寸图像时，尝试使用同一套先验知识，这就会导致内容的出错，如上图中，图像尺寸翻倍后，车辆的上部就出现了两个顶棚。

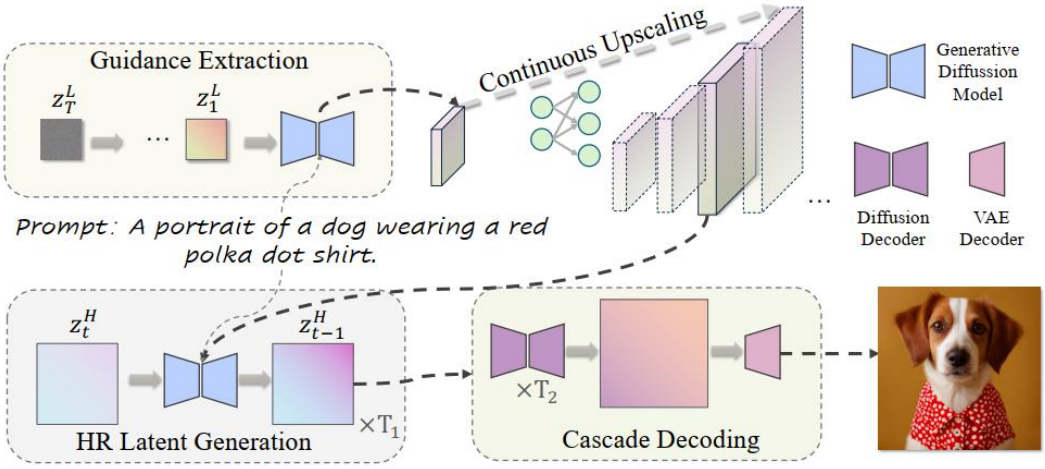
将图像分布可视化，则是不同分辨率的生成输出，分布差异较大（如上图左）。而理想中的分布应当不会因为生成尺寸而出现过大区别（如上图右）

贡献

1：本文推出了 UltraPixel，一个可以进行高质量任意尺度图像生成的模型。其能直接适应不同尺度的超分图像，且不会额外加入过多的参数。

2：通过和现有的文生图方案进行对比，得出了本文模型在多尺度下都能有更佳效果的结论。

实现



本文整体流程图如上图所示，输入的语义信息先被注入到一个标准的 1024x1024 文生图模型中，得到一张 LR 图像（在隐空间中，所以大小为 24x24）。然后这张 LR 图像被训练的神经网络缩放至目标大小，再和语义信息一起注入到 HR 生成模型中进行高清图像的生成。得到结果后进入到 Decoder 中，得到最后的图像输出。

Figure 4: Illustration of continuous upscaling by implicit neural representation.

其中，LR 图像并非直接作为条件输入，而是经过神经网络放大后再作为条件。这里的神经网络可以理解为隐空间上的简单超分网络，保持图像的内容整体不变，但大小变化。

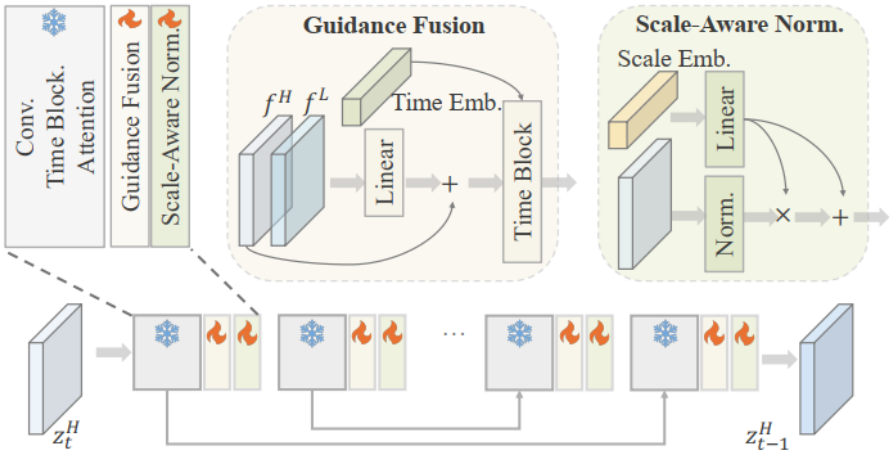
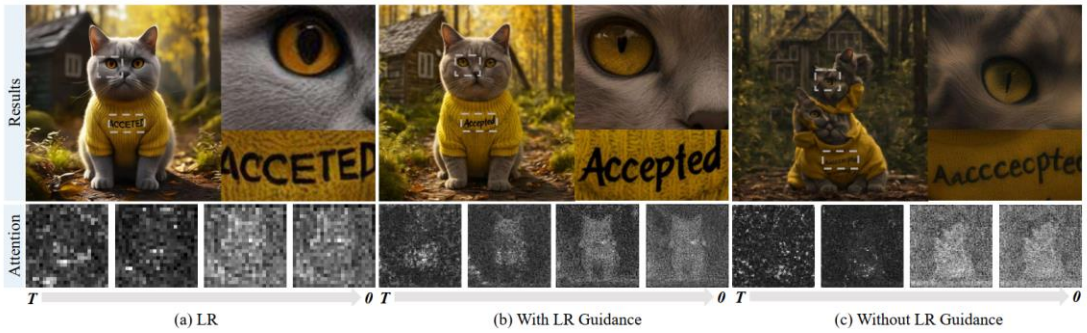


Figure 5: Architecture details of generative diffusion model.

而本文的 HR 超分模块，则同样也是使用预训练的扩散模型微调而来。可训练部分为额外加入的两个控制层，一个层注入 LR 信息，一个层注入尺寸信息（不同尺寸图像自然需要不同的引导）。也正因如此，实质上 HR 生成模块和基础的 LR 生成模块公用大部分参数，文章才会提出只增加了额外 3%的参数。

实验



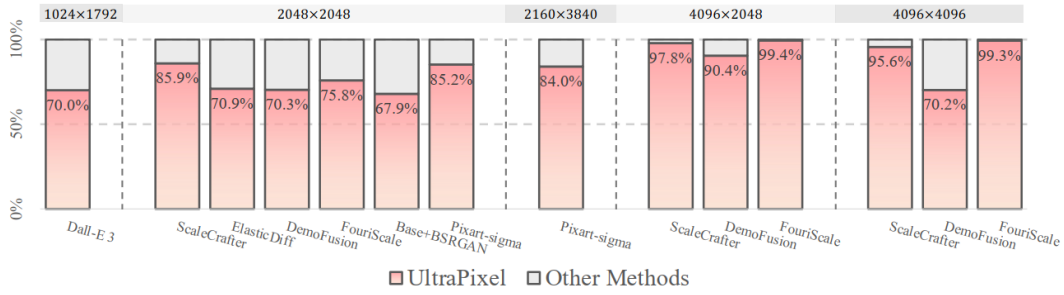
上图很好的展现了本文的特点：LR 图像是直接用预训练模型生成的，质量优秀，构图合理，就是分辨率不够高。而本文的模型以这一图像为指导，生成了一个结构一致，但细节更佳丰富的图像。而如果缺乏 LR 图像的指导，那么得到的图像即便细节尚可（字母都清晰可见）但整体结构会十分错乱。

也就是说，目前大尺寸图像的生成缺乏结构信息，而 LR 作为条件，提供了结构信息，从而解决了这一问题。

Table 1: Quantitative comparison with other methods. Our UltraPixel achieves state-of-the-art performance on all metrics across different resolutions.

Resolution(H × W)	Method	FID_P ↓	FID↓	IS_P ↑	IS↑	CLIP↑	Latency(sec.) ↓
1024 × 1792	DALL·E 3	88.44	86.16	16.43	18.30	29.66	-
	Ours	**60.5**	**63.53**	**17.84**	**26.89**	**35.34**	8
2048 × 2048	ScaleCrafter [14]	64.75	73.79	**15.41**	22.53	31.79	45
	ElasticDiffusion [13]	77.19	65.37	11.12	21.97	32.95	295
	DemoFusion [11]	54.86	63.97	13.38	28.07	32.98	97
	FouriScale [20]	68.79	86.71	7.70	18.08	30.70	74
	Base + BSRGAN [45]	48.52	64.00	13.67	29.87	33.53	11+6
	Pixart-Σ [3]	54.35	63.96	14.87	27.13	31.18	57
	Ours	**44.74**	**62.50**	14.95	**30.52**	**35.43**	**15**
2160 × 3840	Pixart-Σ [3]	49.86	63.87	10.89	25.35	30.86	111
	Ours	**46.06**	**62.41**	**11.91**	**25.65**	**34.98**	**31**
4096 × 2048	ScaleCrafter [14]	101.58	120.71	9.04	12.15	23.71	190
	DemoFusion [11]	51.16	75.28	10.81	21.83	29.95	325
	FouriScale [20]	128.03	137.16	3.82	10.41	21.98	197
	Ours	**42.60**	**64.69**	**11.76**	**25.36**	**34.59**	**33**
4096 × 4096	ScaleCrafter [14]	74.02	98.11	9.07	14.53	31.79	580
	DemoFusion [11]	47.40	**61.11**	9.99	26.40	33.14	728
	FouriScale [20]	72.23	105.12	8.12	14.81	27.73	573
	Ours	**44.59**	62.12	**10.27**	**27.69**	**35.18**	**78**

定量实验方面，以五种尺寸和多种不同的方案进行对比，得出了本文各项指标都较优秀的结果。



定性实验方面则可以看出，本文方案大多数时候都显著胜于对照方案。

不足

本文自行提出的不足在于，训练集中的数据质量和数量不足，导致的效果不佳，尤其是在面对复杂场景时。

个人认为，实际上目前超分的方案也同样是在一个扩散模型中加入 LR 作为控制条件，所以本文的方案实质上与之前并没有很大的区别，或许是基模型的优化（今年二月的 Stable Cascade，超过 SD-XL，而超分方案许多还停留在 SD2.1）导致的效果提升。