SeeSR

状态：CVPR2024

单位：香港理工大学

文章链接：https://arxiv.org/abs/2311.16518

Github 链接：https://github.com/cswry/SeeSR

目录

Abstract

由于其具有最强的生成先验知识，预训练的文本生成图片（文生图）扩散模型，逐渐被更广泛地用于解决真实图像超分问题。但是，由于低分辨率图像（LR）的图像结构被破坏，其语义信息也变得含糊不清。因此，还原出的高分辨率图像（HR）可能会出现语义错误，从而影响超分表现。为了解决这一问题，我们推出了一种基于语义的错十，来更好地保证超分过程中的语义一致性。首先，我们训练一个能够适应高度退化图像的语义提取器。这个模块能够从较为模糊的图像中获取语义信息。语义信息包括 soft 和 hard 两种，hard 用于直接注入文生图模型的 prompt，而 soft 则提供额外的信息约束。这些语义信息可以帮助模型输出细节且语义信息正确的结果。此外，在预测过程中，我们也将 LR 图像融合到初始噪声中，来缓和扩散模型生成过多随机细节的问题。

Introduction

传统方案介绍

传统方案预设简单的退化核，但会因为过于专注一致性而使得输出图像过于平滑（缺少细节）。

GAN 模型采用对抗训练的方式，牺牲了一部分一致性，而提升了图像质量。

但由于合成训练集和现实图像中的差异，以上的模型在实际图片测试中表现不佳。为了解决这一问题，RISR 被提出，也有许多 GAN 模型表现良好，如 BSRGAN，Real-ESRGAN。但由于其对抗训练的不稳定性，结果仍不够完善。LDL 可以通过细节的检验防止效果不佳的增生物，但不能生成额外细节。

近来，DDPM 逐渐战胜了 GAN，而其内部丰富的先验知识，也被诸多工作尝试用于 RISR 任务（后文简称超分），如 StableSR，PASD，DiffBIR 等。

但这些方案都存在问题，StableSR，PASD 都进依赖 LR 为控制条件，而忽视了预训练 T2I 模型的文本信息输入。而 PASD，尝试使用现存的高维模型（某种大模型？）来提取图像中的语义文本，但其在处理复杂场景和过度模糊图像时遇到问题。

本文方案介绍

在本文中，我们总结了如何提取语义文本，才能更好的利用预训练模型内的先验知识，总结出了两条规则：

1. 文本最好能覆盖图像中全部的物体，来帮助文生图模型更好的理解图像。

2. 文本的提取需要是抗模糊的，以防止出现错误的语义指导，因此，不适合直接从 LR 中提取语义。

基于这些规则，我们提出了 SeeSR 模型，其训练分为两个阶段，第一阶段通过微调使得文本信息提取器能够适应 LR 图像。第二阶段，文本信息和 LR 共同作用在预训练模型上，生成高质量的输出。此外，在推理阶段，LR 图像还会和初始噪声进行复合，减轻扩散模型生成随机细节的倾向。

我们丰富的实验证明了 SeeSR 的优秀效果。

Related Work

基于 GAN 模型的超分

1. 第一个 SRCNN 用深度学习来做 ISR

2. 一系列模型研究更好的结构

3. 问题进步到 RISR，代表模型 BSRGAN，Real-ESRGAN

4. 额外的不自然物体，解决模型有 LDL，DeSRA，但也使得难以添加自然的细节。

DDPM（扩散模型）

扩散模型拥有强大的预训练模型存在，所以广泛运用于下游任务之中。

基于扩散模型的超分

早期的扩散超分（文中略过）

1. 一些重要的同类模型：StableSR，DiffBIR（两阶段生成），但这两个模型都忽略了预训练扩散模型所具有的文本控制路径。

2. PASD 则尝试了引入文本信息（通过 ResNet，Yolo，BLIP），但其难以处理复杂的场景和过于模糊的图像，而本文的目标就是探究如何高效的提取并利用文本信息。

Methodology

动机和结构总览

Motivation

我们尝试研究了三种文本信息提取方案：分类，描述，标签，并采用三种之前工作中的模型来进行这三种信息的提取。

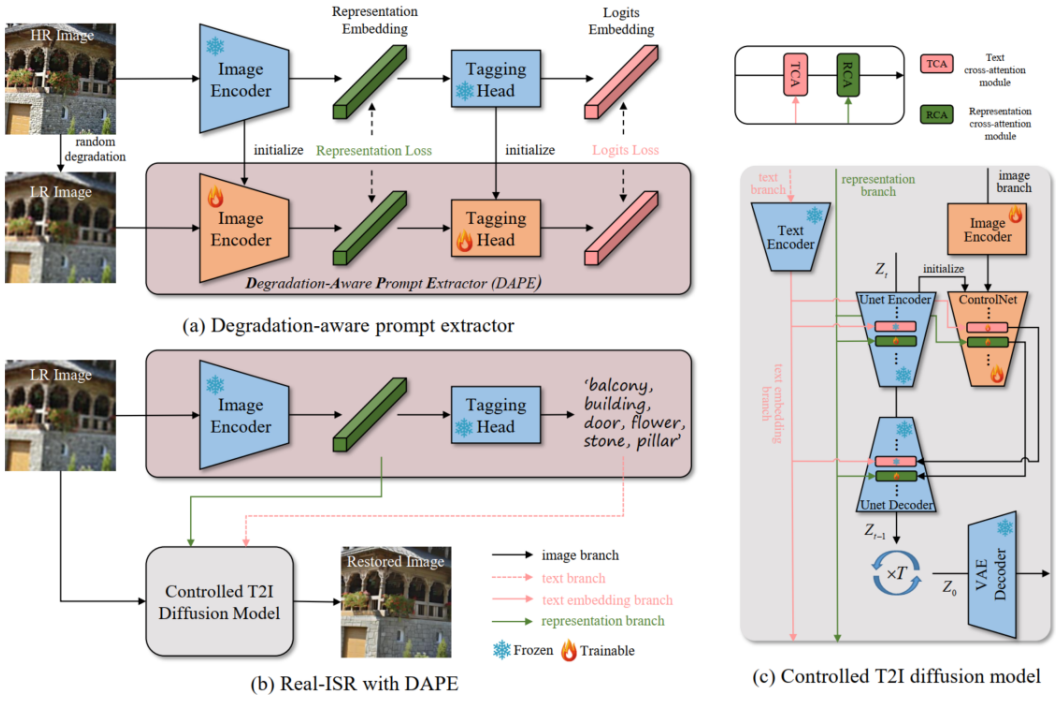
分类提取法只为每幅图片提供一个类别标签。但这种方案几乎难以给出足够的信息，在实验中发现其与不提供 prompt 的对照组几乎难以分辨。

描述提取法为图片提供一段话，提供了丰富的信息。但其缺点在于，额外的介词副词会干扰预训练模型，且更容易在 LR 图片上识别出错误的语义。

而标签类的文本则表现更好，即便其文本内不含位置信息，但预训练模型中本身具有的语义分割能力，就足够理解并排布这些物体。但其仍然会在 LR 上进行错误识别。

可以看出，只要解决标签类文本提取的错误识别问题，就可以得到较好的文本提取方案。

结构总览



本文的结构图如上图所示。图中(a)部分（DAPE）描述了如何微调文本提取器，使其能够在 LR 图像上提取出和 HR 图像接近的结果。在完成微调后，这一文本提取器被放入(b)中，进行文本提取，通过固定的文本影响路线，进入到扩散模型之中，而(c)展示了信息具体是如何注入的，且(b)中还需要对 ControlNet 进行适应性训练。

适应模糊的文本提取器

本文中的文本信息提取器（DAPE）模型，是通过微调一个预训练 tag 提取模型（RAM）得到。这一结构存在量阶段，第一阶段为经过 Encoder 的特征图（soft），第二阶段为特征图编码成的文字（hard）。

训练时的 loss 由 hard 和 soft 两个部分各自求 loss 组成，且 soft 信息用于补全 hard 信息（hard 信息设定了一个数量限制，只标注重要的 tag）。

训练 SeeSR 模型

由于 controlnet 的成功，本文也采用其结构作为加入控制的方式。图中的半个 u-net 直接由预训练模型中拷贝得到。attention 结构则采用了 PASD 中的，来引入语义指导，即图中的 RCA 模块用于引入 soft 信息，TCA 模块则用于引入 hard 信息。这些模块被一并拷贝到了 controlnet 中。（其中预训练模型中本来就具有 TCA 模块）

controlnet 路径则用于引入 LR 的图像信息，LR 经过 image encoder 进入 controlnet，最后再注入 U-net 之中造成影响。

对 SeeSR 模型的训练较为简单，loss 即为 diffusion 训练使用的 loss。为了减少训练开销，只训练 ControlNet 和 U-net 中的 attention 层。

推理噪声加入 LR

预训练的模型，在训练时并不会真的将图像转变为纯噪声。而大多数预训练模型，又会以纯噪声开始推理，就造成了训练和推理时的差异。这一差距在超分工作中表现为，可能会使得光滑的表面被增添过多的噪点。

为了解决这一问题，我们将 LR 的信息加入到了初始的噪声中，且这一方案适用于大多数的超分模型。

Experiments

依照之前的工作，实验专注于 4 倍缩放。

实验设置

数据集

训练集：DIV2K，DIV8K，Flickr2K，OST，FFHQ（前一万张图像）

退化方案：采用 Real-ESRGAN 的退化方案来生成 LR 图像。

测试集：

1. 随机裁剪 3K 张 DIV2K 图像（512×512），并使用同样的退化方案得到 LR

2. 采用了两个真实图像数据集 RealSR 和 DRealSR，并将其中心裁剪（似乎是某种退化方案）出 128×128 的图像。

3. 采用了一个新建的真实世界数据集，RealLR200，38 张来自早前文章，47 张来自 DiffBIR，50 张来自 DiffBIR，50 张来自 VideoLQ，65 张由我们自行收集。

实验细节

一些具体的实验细节

训练使用 8 块 V100 完成

评估指标

PSNR，SSIM 为图像相似度指标，LPIPS，DISTS 为有参考的图像评估指标，FID 用于衡量初始图像和生成图像的分布差距，NIQE，MANIQA，MUSIQ，CLIPIQA 为无参考图像评估指标。

对比模型

基于 GAN 的模型：BSRGAN，Real-ESRGAN，LDL，FeMaSR，DASR

基于 Diffusion 的模型：LDM，StableSR，ResShift，PASD，DiffBIR

与 Sota 模型的比较

定量比较

由于基于扩散模型的超分方案能够额外生成细节，所以其 PSNR，SSIM 指标都不够高，比不过基于 GAN 的超分方案，这也包括 LPIPS/DISTS 这种有参考的图像评估指标，而 FID，及无参考的评估指标 CLIPIQA，MUSIQ 和 MANIQA 则 DM 有显著优点。本文中的 SeeSR 也在这几个指标上表现优异。

定性比较

通过对展示的图片进行分析，而得出 SeeSR 效果较之之前的工作更佳的结论。

用户实验

遵循 SR3 中提出的愚人率测试方案，开展了用户实验。人造图像与 GT 进行对比，而真实图像则与多个超分结果进行 对比。

得到了 36.6%的愚人率，以及 56.2%的对比选择率（这部分实验有问题）。

语义保存测试

通过使用 COCO 数据集中的测试集，以及一些语义信息检测指标，证明了 SeeSR 在语义信息正确率上有相当好的表现。

消融实验

LRE（也就是 LR 注入到噪声的方案）效果实验。

DAPE 和 RAM（微调前的预训练模型）对比实验，DAPE 在 LR 图像上的表现更优。

DAPE 及其 Hard，soft 信息使用情况的消融实验。

复杂度分析

SeeSR 具有 2283.7M 大的可训练参数，推理（128 to 512）耗时（V100 显卡）7.24s

更多的视觉比较

更多的视觉对比展示。

（这篇文章的实验部分格式很标准，可以作为之后实验部分的模板）

Conclusion

一些简要总结。