

图像迁移模型 PnP 的复现与基于 SAM 的改进

李一凡 敬鼎豪

2024 年 1 月 20 日

1 论文介绍

我们选取的论文工作是 CVPR2023 的一篇论文，名称为 Plug-and-Play Diffusion Features for Text-Driven Image-to-Image Translation。这篇论文聚焦的主要任务是把给定图像基于文字 prompt 进行迁移，得到对应的生成图片。可以理解为一种从图片到图片的翻译工作，在保留原图的基本布局和结构特征的基础上，在风格和场景上去尽量贴合文字的描述，从而得到我们的生成图片。

1.1 背景

随着文生图领域中扩散模型（diffusion model）的提出，出现了很多 text-to-image 的生成模型。这些生成模型有强大的生成能力，但由于文字的描述信息有限，生成模型在生成结果上可能具有不确定性和多样性，在这个基础上，此篇论文想要为用户提供更多在生成图片上的控制力，于是就把任务范式从单一的文字生成图片变为图片到图片的翻译，给的文字描述只会改变风格和场景，不会产生大的结构性变动。这样用户就能按照自己想要的风格去做图像的迁移。

1.2 方法

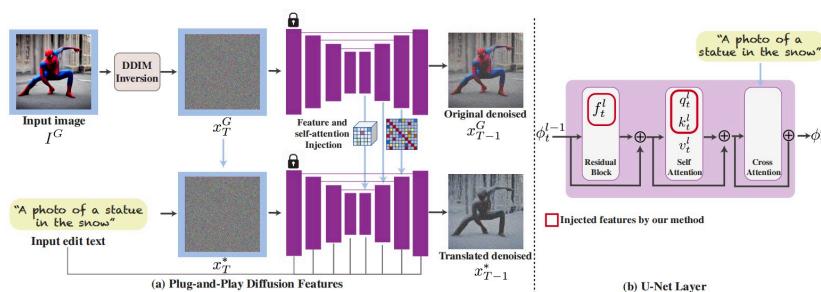


图 1: 论文中的 pipeline

此篇论文利用了 pretrained 的 stable diffusion model 来作为架构的主要组成部分，不需要进一步的微调或者训练。如图 1 所示，整个架构分为两个过程，首先是提取 feature 的过程（图中上面部分的 stream），给定我们的 input image，我们用 DDIM inversion 的过程来得到噪声图，接着用 DDIM 来还原我们的噪声图，在这个还原过程中，经过 UNet 的每一层 layer，都能得到我们的相应 feature，这个 feature 我们需要在对应的时间步注入到下面的生成过程中，从而保留原图的结构特征。下面的生成过程中，同样的噪声图，也是经过 DDIM 逐步去噪还原，但是在 UNet 的相应步中，除了有上面 stream 所注入的原图 feature，更有文字 prompt

的特征，在两方面的作用下得到我们的生成图。图片中的右半部分是 UNet 的细节图，有三个 block 的参与，在 residual block 中注入我们原图解噪过程对应时间步的 feature f_t^l ，然后经过 self attention（自交叉注意力机制）来强化结构性特征的自相似性，最后再通过交叉注意力机制引入和文字 prompt 的交互。

1.3 结果

论文通过实验对比了 UNet 中不同 layer 注入的 feature 对于生成图片的结果影响，选取了 layer 4 作为主要的提取特征层。最终从论文呈现的结果上来看，论文采用的实验方法，在最优的参数选择下（例如 layer 4），能取得优于其他 baselines 的结果。具体表现为两方面：一是相比于一些方法能保留更多的结构性信息，生成图片在结构上的特征不会因为文字描述而发生大的变化。二是相比于一些方法，原图的外观等信息没有太影响图片的生成，生成图片更贴合文字的描述。总的来说就是在两个输入间的平衡做的比现有方法都要好。

2 改进原因

原论文中虽然对于图片的翻译有比较好的效果，但是对于用户来说，常常只想要保留主体的结构信息，我们并不希望生成图片的背景信息也和原图片的背景信息具有相同的结构特征，而是更希望生成图片的背景信息去更贴合于文字的描述场景。我们可以看下面的两个例子：



图 2: 样例 1

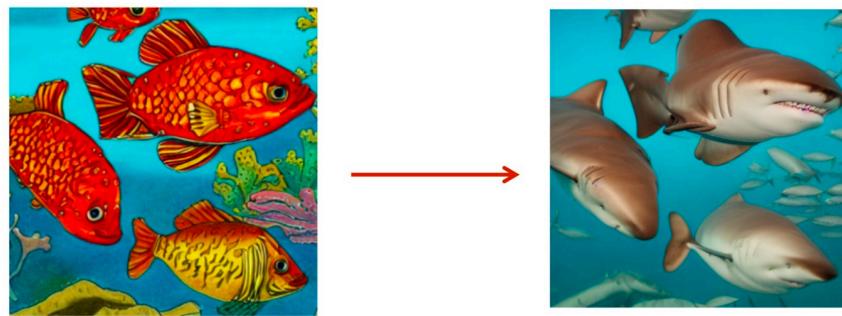


图 3: 样例 2

如图2所示，样例 1 中，我们的 target prompt 是 an iron robot in the snow，从生成结果上来看，主体从人到 robot 的迁移是比较理想的，但是在背景上，原图中的草垛变成了生成图片上的黄色光晕，这个是不太符合我们的 snow 的场景的，所以我们想要减弱这种原图背景

对于生成图片的影响。

如图3所示，样例 2 也具有相似的问题，从图片的转换过程中我们可以看到，用户应该是想把左图中的金鱼都转化成鲨鱼，但是在转化过程中，左边图背景的水草也变成了一群鲨鱼堆。某种程度上这是有点偏离实现目标的。

我们在此基础想要对原论文中的方法进一步改进，在原图片转化到生成图片的过程中，去减弱原图片背景干扰信息对于生成图片的影响。使得生成图片在主体物体上的结构信息与原图保持一致，但是在背景方面弱化与原图的关联，从而生成更合理、更贴合文字描述的图片。

3 改进方案及实验结果

3.1 动机

如前所述，原论文根据输入图像的 DDIM Inversion 结果再生成过程中提取出的特征及自注意力图对文本引导的图像生成过程进行正则化，因而可能存在以下几个方面导致图像迁移的效果由于背景冗余信息而变差：

1. DDIM Inversion 过程没有考虑目标文本而掺杂了过多的背景冗余信息，而原论文的生成起始噪声正是这一经过反向计算的伪噪声。在短时间内通过数学推导的方式提出更适配于图像迁移任务的 DDIM Inversion 算法有一定困难，因此我们尝试了将文本引导生成过程的起始噪声由 inversion 结果替换为随机高斯噪声，以此探测初始伪噪声与特征注入环节相比，是否会提供更多信息。详见 4.2 节。
2. 原论文使用较为直接的特征及自注意力图的注入方式，即直接将某一特定时间步以后的去噪网络的特征和自注意力图替换成携带有原图结构信息的特征和自注意力图。我们认为这样稍显粗暴的替换方式可能并不适配于 Stable Diffusion 的去噪网络和无训练、无微调的实验设置，如此替代操作可能会损害文本引导的生成过程。因此我们尝试了线性插值、高斯平滑化技巧等特征插入方式，以期自适应地进行特征的注入与引导。详见 4.3 节。
3. 我们观察到原论文方法最影响视觉体验的问题在于背景信息的冗余注入，而一个直观的想法是将背景与待迁移内容的主体分割开来。直觉上，在文本引导的生成过程中，Stable Diffusion 会更加关注主体的生成，因为背景拥有的高频信息远少于前景主体，而背景的生成会因为有文本引导的存在而变得容易。因此我们尝试了（1）将背景分割后删去，只留前景主体作为输入图像进行前述特征提取及文本引导的再生成的过程；（2）将背景弱化一定程度，将弱化背景图作为输入作上述操作。详见 4.4 节。

3.2 方案一：以高斯噪声作为再生成过程起始噪声

使用 DDIM Inversion 的噪声作为生成的初始噪声固然可以提供充分的原图信息，但也许会存在冗余的与目标文本不相关的信息被嵌入到该隐变量中，打破图像迁移任务的 fidelity 和 reality 的 trade-off。理想的研究思路是深入思考 DDIM 采样的机制并广泛调研近年来关于 DDIM Inversion 的研究，基于 SD 可以提供文本引导的特性设计方法筛选这部分冗余信息。遗憾的是时间紧迫，亦没有精力设计全新的方法。因此，我们设计直接以高斯噪声作为再生成过程起始噪声，如下图4所示。

该改进方案的结果如下图5所示。虽然图像摆脱了冗余信息的束缚，可以在背景根据目标文本自由地生成一些多样的内容，然而视觉效果却因为这种自由程度过大而变得更糟。我们再次基础上尝试注入更多时间步的特征，结果如下图6所示。不幸的是，结果并没有明显改观。究其原因，我们认为是后续进行特征注入的过程会极大程度损失 SD 本身的生成性能。除此之外，该方案十分依赖于随机种子的选取，模型生成结果的质量的方差较大。

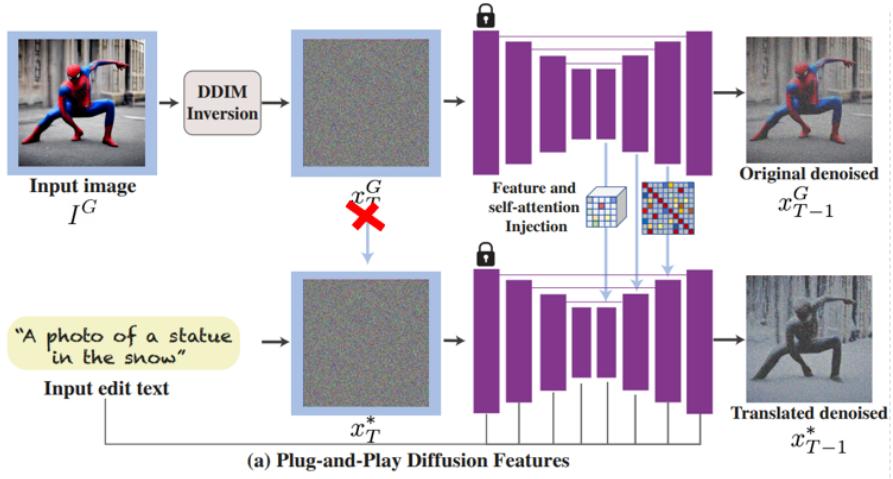


图 4: 以高斯噪声作为再生成过程起始噪声



图 5: 方案一结果，目标文本为”a photo of an iron robot in the snow”

3.3 方案二：基于线性插值的特征注入方法

如前方案一所述，直接将起始噪声替换为高斯噪声会导致生成过程不稳定的问题，我们将问题核心定位在后续的特征注入上，认为原论文中的直接替换会极大地损害 SD 的生成能力。因此，我们尝试使用线性加权的方式将生成过程中 SD 的特征及自注意力图和 Inversion 过程提取出来的特征及自注意力图进行结合，结果如下图7所示，结果貌似变得更差了。

因此综合改进方案一和二，我们得出了初步结论：经 DDIM Inversion 得到的初始噪声携带着至关重要的信息，简单地改为高斯噪声会造成后续生成过程难以弥补的损害。虽然是一次失败的尝试，但是也让我们认识到初始隐变量的生成潜力和底蕴，为后续开发新的 Inversion 算法提供可能。

3.4 方案三：基于 SAM 的背景弱化技术在图像迁移任务上的应用

如前动机所述，现存的背景冗余信息问题可以通过先将输入图的背景进行弱化或删除后得到的处理后图像进行文本引导的图像迁移。幸运的是，SAM 可以根据用户指定的 point prompt 很好地解决这一背景分割问题。我们尝试了两种背景弱化的方式：

(1) 删去不必要背景

(2) 将背景乘上 0.5 以作弱化，

经弱化背景后的图像分别如下图8, 9所示。



图 6: 方案一 + “注入更多时间步特征” 结果



图 7: 方案二结果

经过实验发现，如果直接将背景删去，即背景区域填充以 0 像素值，那么以此生成出的图像也没有背景，如图10，导致文本引导无处施加，迁移任务化作泡影。而采用背景按照 0.5 比例弱化的方式 2 会得到背景解耦效果较好的结果，但整体图像偏暗，如图11。

因此，我们在进行生成后，将生成结果再乘上之前相应的背景弱化因子 0.5 的倒数，即为 2，达到整体提亮的目的，结果如图12所示，这一结果相较于图3而言修正了本不应出现的黄色光晕，视觉效果更好。同时我们也认识到，以 DDIM Inversion 得到的隐变量作为初始去噪输入的 SD 框架甚至可以保持亮度特征，这也印证了之前方案一失败的原因：Inversion 结果包含很多图像的隐式信息，因而在其上的解耦比图像上的解耦要复杂得多。而一旦可以在隐变量空间上的解耦，那么将大大提高我们对于生成模型的控制能力。

4 更多实验结果与讨论

我们在本节展示了更多的实验结果以供参考。其中图17展示的鲨鱼样例在原论文的方法下结果中掺杂过多冗余的背景信息，而在基于 SAM 的背景弱化方法下，这些背景被柔地地转变为了水草，体现了我们方法的有效性。其他结果也可以体现我们的方法不仅加强了图像迁移过程中的可控性，还不会损害原有模型的生成能力。然而值得注意的是，我们的方法可能过多地强调了前景主体的重要性，因而在后续生成时若目标文本与原始图像的差距较大，则可能造成不真实的结果。其次，由于 DDIM Inversion 的结果会保留图像的明暗程度，这也导致生成的结果较暗，需要进一步提亮才可以得到比较好的视觉效果，而这一步会提高时间成本和方法复杂性。除此之外，我们的方法需要用户提供目标文本之外，还需要显式地指出前景主体的 prompt point，这可能会降低用户体验。总的来说，这是一次收获颇丰的复现实验与改进尝试，我们不仅积累了基于 Stable Diffusion 和 Inversion 进行图像迁移的思路的认识，还尝试了使用其他技术，如 SAM，进行原方法的改进。希望未来可以在隐变量空间中得到对 diffusion 模型的解释和控制，发掘扩散模型真正的魅力。



图 8: 删去不必要的背景



图 9: 将背景乘上 0.5 以作弱化



图 10: 删去背景得到的迁移效果



图 11: 背景按照 0.5 比例弱化得到的迁移效果

5 小组分工

李一凡: 复现实验、提出 SAM 改进思路、部分 SAM 改进实验、报告中改进方案与实验结果及讨论部分撰写

敬鼎豪: 协助复现实验、提出部分改进思路 (非 SAM)、复现 SAM 工作并参与改进实验, 报告中论文介绍和改进原因部分撰写



图 12: 图 11 的提亮结果

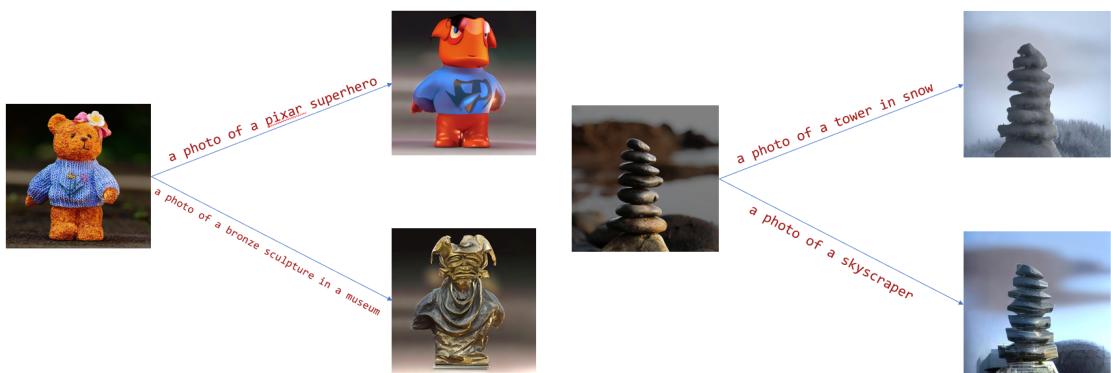


图 13: 更多结果 1

图 14: 更多结果 2

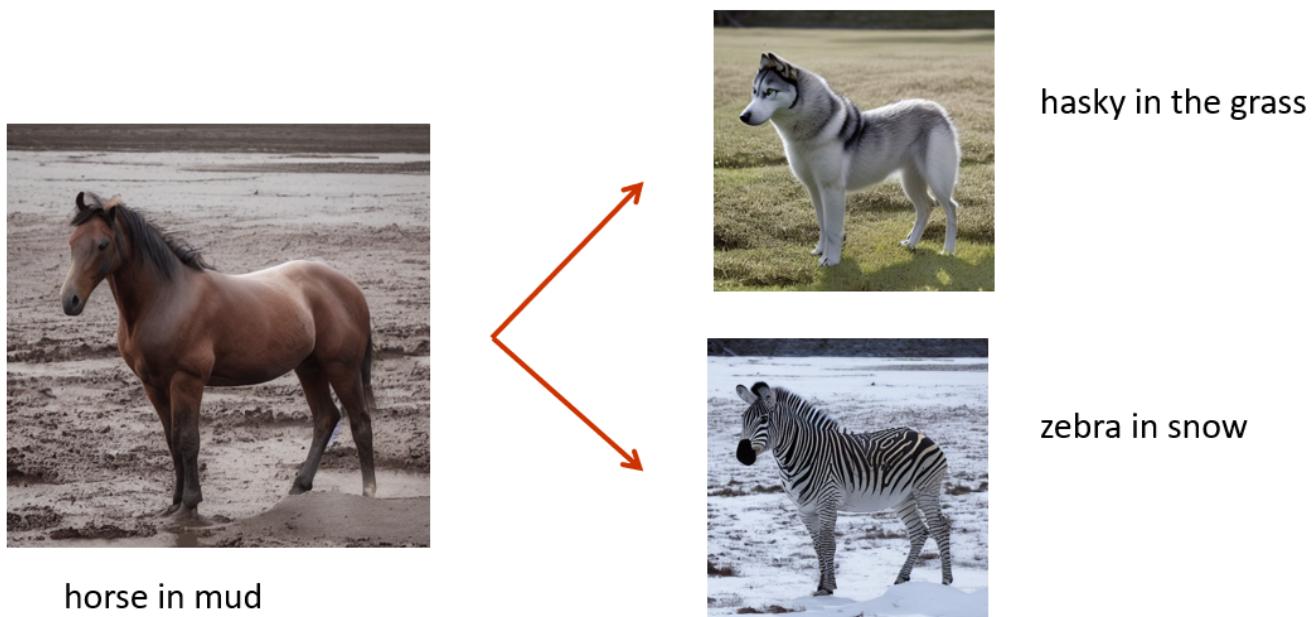


图 15: 更多结果 3

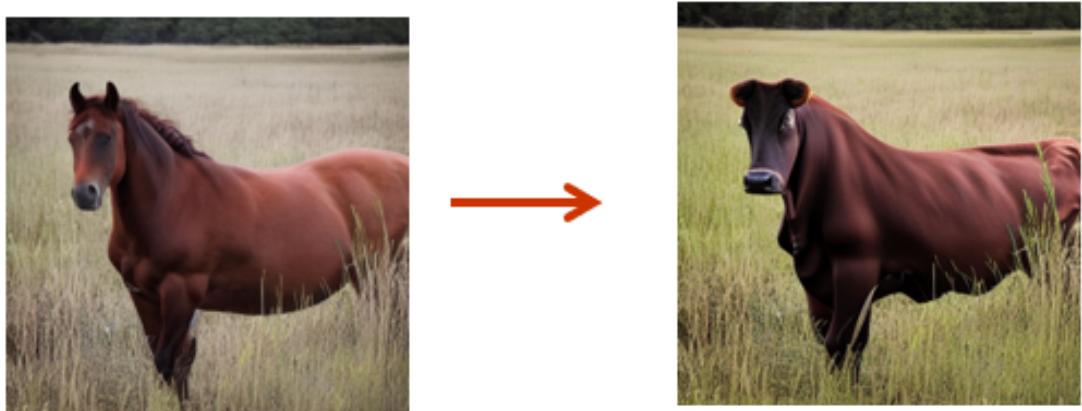


图 16: 更多结果 4



图 17: 更多结果 5