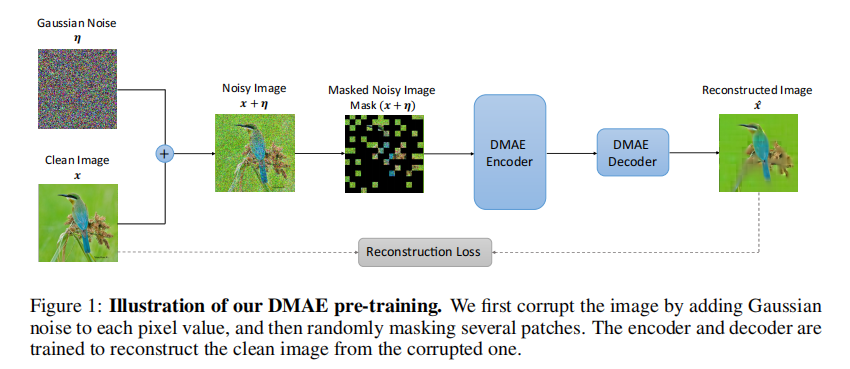
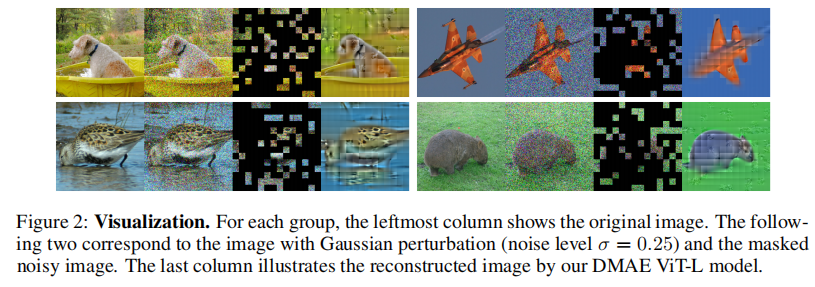
1. DMAE：通过在每个像素值上添加高斯噪声来破坏图像，然后随机屏蔽几个补丁。编码器和解码器被训练从损坏的图像中重建干净的图像。

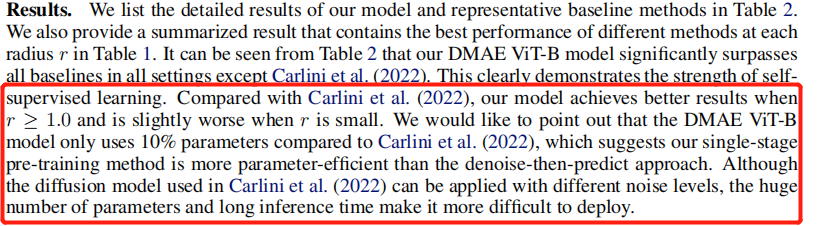


采用像素级均方误差作为损失函数。与MAE略有不同的是，在所有补丁上计算损失，因为模型也可以学习未屏蔽位置的purification。在训练前的过程中，编码器和解码器从头开始进行联合优化，在学习下游任务时去掉解码器。

为了重建原始图像，编码器和解码器必须从未被掩蔽的补丁中学习语义，并同时去除噪声。为了强制执行编码器（而不是解码器）来学习健壮的语义特征，我们在He等人（2022年）之后，通过设置一个较小的隐藏维数和深度的值来控制解码器的容量。



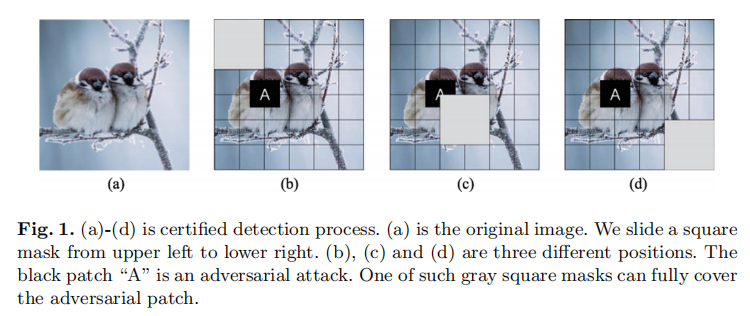
**重建后的效果一般**



与Carlini等人（2022）相比，我们的模型在r≥为1.0时获得了更好的结果，而在r较小时则略差。我们想指出的是，与Carlini等人（2022年）相比，DMAE ViT-B模型只使用了10%的参数，这表明我们的单阶段预训练方法比去噪后预测方法的参数效率更高。虽然Carlini等人（2022）中使用的扩散模型可以应用于不同的噪声水平，但大量的参数和较长的推理时间使其部署更加困难。

2. ViP：提供了第一个开发针对双补丁攻击的认证检测的研究，其中，允许攻击者反向操纵两个不同区域内的像素。

依次滑动



手机屏幕截图

中度可信度描述已自动生成

图片包含 日程表

描述已自动生成

3. Certified Defences Against Adversarial Patch Attacks on Semantic Segmentation

我们提出了分割平滑，这是第一个（据我们所知）认证的恢复或认证的基于检测的防御，针对语义分割模型的对抗性补丁攻击（第4节）。

去除平滑可以做认证的检测和恢复与任何可能的分割模型，而不需要微调或任何其他自适应。

5个GPU

图形用户界面, 网站

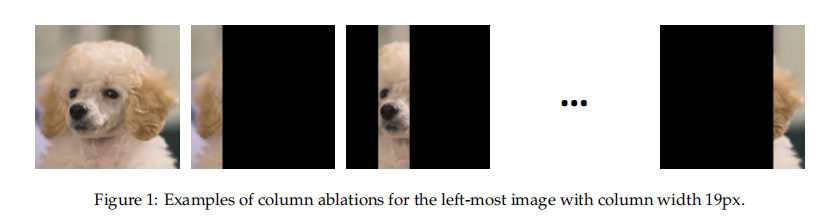
描述已自动生成

日历

描述已自动生成

4. Certified Patch Robustness via Smoothed Vision Transformers

感觉是依次平滑



图形用户界面

描述已自动生成