arXiv:2302.04578v2 [cs.CV] 6 Jun 2023

对抗样本有益：利用对抗样本防止扩散模型模仿绘画作品

Chumeng Liang [*1 2] **Xiaoyu Wu** [*1 2] **Yang Hua** [3] **Jiaru Zhang** [1] **Yiming Xue** [4 2] **Tao Song** [1] **Zhengui Xue** [1] **Ruhui Ma** [1] **Haibing Guan** [1]

摘要

近年来，扩散模型（DMs）推动了艺术人工智能的蓬勃发展，但也引发了新的版权担忧：侵权者利用未经授权的绘画作品训练扩散模型，从而生成风格相似的新绘画作品。为应对这些新出现的侵犯版权问题，本文首次探索并提出利用扩散模型的对抗样本来保护人类创作的艺术作品。具体而言，我们首先构建了一个理论框架，用于定义和评估扩散模型的对抗样本。随后，基于该框架，我们设计了一种名为 AdvDM 的新算法，通过优化从扩散模型逆向过程中采样的不同潜变量，利用蒙特卡洛估计生成扩散模型的对抗样本。大量实验表明，生成的对抗样本能够有效阻碍扩散模型提取其特征。因此，我们的方法可成为人类艺术家保护版权、抵御使用基于扩散模型的艺术人工智能应用的侵权者的有力工具。我们的方法代码已发布在 GitHub 上：

h
t
t
ps://github.com/mist-project/mist.git。



图*1*。分类模型和扩散模型中对抗样本工作流程的对比。扩散模型中的对抗样本通过诱导分布外特征，阻止扩散模型提取图像特征作为条件。图中所示的特征提取方法是扩散模型中的文本反演（Gal 等人，2022年），该方法曾在多个案例中引发版权担忧（MT，2022年；Deck，2022年）。

1. 引言

近年来，深度扩散模型（Sohl-Dickstein 等人，2015；Ho 等人，2020年）在计算机视觉领域蓬勃发展。凭借坚实的理论基础（Song 等人，2020b;a；Bao 等人，2022年）以及高度实用的技术（Gal 等人，2022年；Lu 等人，2022年），扩散模型已被证明在生成任务中卓有成效，包括图像合成（Ruiz 等人，2022年）、视频合成（Yang 等人，2022年）、图像编辑（Kawar 等人，2022），以及文本到 3D 合成（Poole 等人，2022）。在这些方法中，潜在扩散模型（Rombach 等人，2022）在艺术创作方面展现出强大能力，引发了艺术人工智能的商业化热潮。

尽管扩散模型在商业化方面取得了成功，但这些模型助长某些侵犯版权行为，已成为公众关注的问题。例如，文本反演（Gal 等人，2022年），这是一项基于潜在扩散模型、在大多数艺术人工智能应用中实现的新功能，只需几个样本便能模仿人类创作的绘画作品的艺术风格。已有案例表明，侵权者可以轻易地从网上获取艺术家创作的绘画作品

[*]共同贡献 [1]中国上海交通大学 [2]cheer4creativity.ai [3]英国贝尔法斯特女王大学 [4]美国纽约大学。通讯作者：Tao Song < songt333@sjtu.edu.cn>。

《 40th 国际机器学习会议，美国夏威夷州檀香山。PMLR 202, 2023年。版权归作者所有，2023年。

并借助文本反演（MT，2022年；Deck，2022年）非法将其用于训练模型。尽管艺术家有权声明禁止将其艺术作品用于训练艺术人工智能模型，但目前尚无技术能够防止或追踪这种非法使用，导致犯罪成本更低，也更难收集证据。此外，艺术家缺乏资源，难以对侵权者提起法律诉讼（Vincent，2022年）（关于艺术人工智能伦理问题的更多讨论，请参见附录A ）。因此，艺术界正在呼吁采用现成技术，保护绘画作品的版权免受艺术人工智能侵害（Vincent，2022年）。

这一保护思路受到图像分类领域对抗样本的启发（Goodfellow 等人，2014b；Madry 等人，2018年；Carlini 和 Wagner，2017年），即在图像中添加一些经过专门设计的微小扰动，使其在 AI 艺术应用中对扩散模型无法识别。这里的无法识别是指扩散模型无法将图像识别为普通图像，因此无法提取图像特征或模仿艺术风格。我们将经过扰动的图像视为针对扩散模型的对抗样本。将绘画作品转换为对抗样本，同时不损失图像语义，便可弥补艺术品版权保护技术的不足。

然而，为扩散模型生成对抗样本并非易事。与分类模型不同，扩散模型通过根据输入图像生成新图像来利用输入，而非对输入进行端到端推理。因此，对抗样本必须阻止扩散模型在某些可识别条件下提取其特征（例如风格、内容、等等）。此外，扩散模型的训练目标是通过变分界间接优化的，因此不适用于对抗样本的优化。基于这些原因，现有研究仅利用扩散模型来提高分类器的鲁棒性（Nie 等人，2022年），而针对扩散模型的对抗样本构造仍是一片空白。

本文构建了一个理论框架，用于定义和评估扩散模型的对抗样本。具体而言，对抗样本旨在保护其自身特征，使其不会在扩散模型的推理流程中被提取（见图1）。该流程由两个阶段组成：（1）条件阶段，从输入图像中提取特征作为条件；（2）生成阶段，根据这些条件生成图像。在图1所示的案例中，条件阶段由文本反演（Gal 等人，2022年）实现。我们的对抗样本通过误导条件阶段中的特征提取过程，产生分布外条件。为此，我们将对抗样本定义为具有以下优化目标：最小化图像被扩散模型识别为真实图像的概率。我们通过向图像添加微小扰动来优化该目标。随后，我们根据图1所示的流程，对对抗样本进行评估。优秀的对抗样本会导致条件生成图像的质量较差，我们可以据此评估对抗样本的质量。

在所提出的框架下，我们提出了一种为扩散模型生成对抗样本的算法。我们采用蒙特卡洛方法，在扩散模型的背景下估算根据我们的定义所给出的目标函数。我们还在真实的版权侵权场景中评估了我们的对抗样本（MT，2022年；Deck，2022年）。大量实验表明，我们的对抗样本能够有效阻止商业应用所使用的潜在扩散模型提取图像特征并模仿其风格或内容。

我们的贡献可概括为以下几个方面。

- 我们构建了一个新颖的框架，用于定义和评估扩散模型的对抗样本。据我们所知，我们是首个系统研究这一主题的团队。
- 在上述框架下，我们提出了一种端到端算法 AdvDM，用于为扩散模型生成对抗样本。
- 我们在多个数据集上进行了大量实验，涵盖单类别数据集和艺术风格数据集，以验证我们的方法能够有效保护图像，防止扩散模型对其进行学习、模仿和复制。

2. 背景

2.1. 生成建模与扩散模型

生成模型从数据中学习 $x \sim q(x)$ 并具有一个分布 $p_\theta(x)$，从中可以采样生成的数据。基于潜变量的生成模型已被证明在生成任务中十分有效，包括变分自编码器（Kingma 与 Welling，2014年；Razavi 等人，2019年）和生成对抗网络（Goodfellow 等人，2014a；Brock 等人，2018年）。这些模型通过低维空间中的潜变量 z 来拟合数据，并对联合分布 $p_\theta(x,\ z)$进行建模。

训练生成模型的一个直观思路是针对真实数据 $x \sim q(x)$ 最大化$p_\theta(x)$。然而，直接优化 $p_\theta(x)$ 较为困难，因此需要进行转换。此时，可选择变分界（Higgins 等人，2017年；Gregor 等人，2016年）$-\log p_\theta(x) \leq -\log \frac{p_\theta(x,\ z)}{q(z|x),}$ 并对其进行优化。

生成模型中的一个重要范式是条件

生成建模（Mirza & Osindero，2014年）。通常，条件生成模型使用不同形式的条件来进行图像生成，包括类别（Mirza &Osindero，2014年）、基础图像（Zhu 等人，2017年）、特征（Karras 等人，2019年；2020年），以及自然语言中的条件提示（Rombach 等人，2022年）。将条件记为 c，他们用参数 p_θ（$x|c$）对其进行建模，并支持从该分布中采样图像 x。

作为定义最先进水平的生成模型，扩散模型（Sohl-Dickstein 等人，2015；Ho 等人，2020年）通过马尔可夫链构建一系列潜变量 $x_{1:T}$。q（$x_{1:T}|x_0$）随后使用反向马尔可夫链 p_θ（$x_{0:T}$）将潜变量还原为数据 x_0。 p_θ（x）通过 p_θ的变分界进行优化，

$$-\log p_\theta(x) \leq -\log \frac{p_\theta(x_{0:T})}{q(x_{1:T}|x_0)} := \mathcal{L}_{DM}. \tag{1}$$

直观而言，扩散模型通过学习利用去噪反向过程 p_θ（$x_{0:T}$）从噪声中还原图像来生成图像。近期，研究人员将去噪反向过程应用于潜在空间（Rombach 等人，2022年），取得了突破性进展。这种潜在扩散模型（LDM）在艺术作品生成的图像质量和采样效率方面均达到了最先进水平，已成为人工智能艺术应用中的主流模型。

2.2. 对抗样本

令 p_{data}(y， x) 表示数据x 与标签 y之间的联合分布。参数为 θ 的分类模型旨在利用 p_θ($y|x$) 估计 p_{data}($y|x$)。这两个分布之间的差异可用 KL 散度量化。分类模型的优化目标可表述为

$$\arg\min_\theta KL(p_{data}(y|x)||p_\theta(y|x)). \tag{2}$$

在分类建模中，人们会采用多种损失函数项作为实现这一目标的替代优化目标。最大对数似然是一种广泛使用的损失函数，已证明它等价于式（2）Shlens，2014年中的目标（）。为简化记号，我们用 $\mathcal{L}_\theta(x$， $y)$ 表示损失项，并在优化过程中将其最小化。

分类模型中使用的神经网络容易受到对抗样本的影响：给定输入 x 及其标签 y，可以找到一个新的输入 x'，使其不会被分类为 y (Goodfellow 等人，2014b；Carlini 和 Wagner，2017）。分类模型的对抗样本 x' 可表述为

$$\begin{aligned} x' := \arg\max_{x'} \quad & \mathcal{L}_\theta(x'), \\ s.t. \quad & ||x - x'|| \leq \epsilon. \end{aligned} \tag{3}$$

许多研究探讨了与生成模型相关的对抗样本。它们主要关注由生成模型生成、并能误导分类模型的对抗样本（Kos 等人，2018年）。关于基于流的生成模型中的对抗攻击（Pope 等人，2020年）的讨论，旨在寻找与真实图像相似、但在基于流的模型中似然分数较低的样本。然而，其理论分析仅考虑基于流的模型中的情况，并采用了数据 x 服从正态分布这一强假设。此外，它并未给出适用于生成模型的对抗样本通用表述。

3. 扩散模型的对抗样本

本节将讨论如何为扩散模型（DMs）生成并评估对抗样本。首先，我们构建目标函数，使样本是真实图像且由模型采样得到的概率最小化。随后，我们提出 AdvDM，这是一种用于近似生成扩散模型对抗样本的算法。最后，我们将讨论这些对抗样本的具体评估方法。遵循扩散模型的记号（Ho 等人，2020年），本节将图像 x 表示为 x_0，并将潜变量 z 表示为 $x_{1:T}$。

3.1. 扩散模型的对抗样本

我们研究这样一种对抗样本：扩散模型无法将其识别为真实图像，但它们在视觉上与真实图像相似。对于扩散模型 θ，对抗样本相对于生成分布p_θ(x)而言属于分布外样本。生成此类对抗样本的一种思路是，通过添加一个或多个尺度受到严格限制的扰动δ，使 p_θ($x+\delta$) 最小化。对扰动尺度的限制可确保扰动在人眼看来不可见，且不会损害图像语义。基于这一思路，我们定义由参数 θ表征的扩散模型的对抗样本。

定义 3.1 (扩散模型的对抗样本)。 给定一个由参数 θ 表征的扩散模型，以及真实数据的分布 q(x)，对抗样本 x' 定义为 $x' = x + \delta$，其中 $x \sim q$(x)，而 δ 由以下公式给出：

$$\begin{aligned} \delta := \arg\min_\delta & p_\theta(x+\delta), \\ \text{where } & x \sim q(x), \|\delta\| \leq \epsilon, \\ & \epsilon \text{ is a constant and usually small.} \end{aligned} \tag{4}$$

然而，在扩散模型中，p_θ 无法实际计算。借助潜变量 p_θ（x），我们可以通过蒙特卡洛估计 p_θ（$x+\delta$）。为此，我们展开

关于潜变量 $x_{1:T}$ 的 $p_\theta(x)$，

$$p_\theta(x) = \int p_\theta(x_{0:T})\mathrm{d}x_{1:T}. \tag{5}$$

我们将对抗样本记为 $x+\delta$，其为 x'，其中 $\|\delta\| \leq \epsilon$。公式（5）表明，可以通过蒙特卡洛最小化 $p_\theta(x)$：通过对 $p_\theta(x_{0:T})$ 进行最小化，并采用 $x_{1:T}$的不同采样过程，我们可以近似地最小化 $p_\theta(x')$。令 $u(x'_{1:T})$ 表示 $x'_{1:T}$ 的分布。我们可以将优化目标改写为以下形式：

$$\begin{aligned} \delta &:= \arg\min_{\delta} \mathbb{E}_{x'_{1:T}\sim u(x'_{1:T})} p_\theta(x'_{0:T}), \\ &\text{where } x \sim q(x), x' = x + \delta. \end{aligned} \tag{6}$$

扩散模型的一个优势在于，后验分布 $q(x'_{1:T}|x'_0)$ 是一个参数固定且完全独立于 x'_0 的高斯分布。因此，可以使用 $q(x'_{1:T}|x'_0)$ 对 $p_\theta(x'_{0:T})$ 进行正则化。我们通常使用 $p_\theta(x'_{0:T})$ 的负对数项。由此即可推导出目标函数的最终形式。

$$\begin{aligned} &\min_{\delta} \mathbb{E}_{x'_{1:T}\sim u(x'_{1:T})} p_\theta(x'_{0:T}) \\ =&\max_{\delta} \mathbb{E}_{x'_{1:T}\sim u(x_{1:T})} - \log p_\theta(x'_{0:T}) \\ =&\max_{\delta} \mathbb{E}_{x'_{1:T}\sim u(x'_{1:T})} - \log \frac{p_\theta(x'_{0:T})}{q(x'_{1:T}|x'_0)} \\ =&\max_{\delta} \mathbb{E}_{x'_{1:T}\sim u(x'_{1:T})} \mathcal{L}_{DM}(x', \theta). \end{aligned} \tag{7}$$

直观地说，公式（7）通过最大化用于训练扩散模型的损失来生成对抗样本 x'，其中采用从 $u(x'_{1:T})$ 中采样的不同潜变量。

3.2. AdvDM：通过蒙特卡洛生成对抗样本

本小节中，我们提出了 AdvDM，这是一种用于生成扩散模型的对抗样本的算法。受分类任务中现有对抗攻击方法的启发（Goodfellow 等人，2014b；Madry 等人，2018；Carlini 和 Wagner，2017），我们利用优化目标的梯度。不同之处在于，我们无法解析地计算目标函数 $\mathbb{E}_{x_{1:T}\sim u(x_{1:T})} \mathcal{L}_{DM}(\theta)$ 的梯度，因为它是一个期望的梯度。如第 3.1 节所述，我们通过蒙特卡洛计算期望梯度来估计它。每次迭代中，我们采样一个 $x'_{1:T} \sim u(x'_{1:T})$，并据此计算 $\mathcal{L}_{DM}(\theta)$ 的梯度。随后，我们使用该梯度执行一步梯度上升。该估计总结于公式（8）：

$$\nabla_{x_0} \mathbb{E}_{x_{1:T}\sim u(x_{1:T})} \mathcal{L}_{DM}(\theta) \approx \mathbb{E}_{x_{1:T}\sim u(x_{1:T})} \nabla_{x_0} \mathcal{L}_{DM}(\theta). \tag{8}$$

我们遵循现有的对抗攻击方法（Good- fellow 等人，2014b；Madry 等人，2018年），并应用符号函数来约束估计梯度的尺度。令 $x_0^{(i)}$ 表示优化过程中第 i 步的对抗样本。第 $(i+1)$ 步的对抗样本通过步长为 α 的符号梯度上升生成，

$$x_0^{(i+1)} = x_0^{(i)} + \alpha \mathrm{sgn}(\nabla_{x_0^{(i)}} \mathcal{L}_{DM}(\theta)|_{x_{1:T}^{(i)}\sim u(x_{1:T}^{(i)})}), \tag{9}$$

其中，sgn 表示符号函数。

直观而言，AdvDM 会对不同的潜变量进行采样，并针对每次采样得到的不同$u(x_{1:T})$，在扩散模型的损失上迭代执行一步梯度上升。在实践中，我们令 $q(x_{1:T}|x_0)$为后验分布，因为实验结果表明，这样能够带来良好的性能。我们在算法 1中总结了 AdvDM。实现细节见附录 D.2。

算法 1 AdvDM：扩散模型的对抗样本

输入：数据 x_0、参数 θ、蒙特卡洛次数 N、步长 α
输出：对抗样本 x'_0
初始化 $x_0^{(0)} \leftarrow x_0$。
对于 $i = 1$**到** N，**执行**
　采样 $x_{1:T}^{(i)} \sim q(x_{1:T}^{(i)}|x_0^{(i)})$
　$\delta^{(i)} \leftarrow \alpha \mathrm{sgn}(\nabla_{x_0^{(i)}} \mathcal{L}_{DM}(\theta)|_{x_{1:T}^{(i)}})$
　$x_0^{(i)} \leftarrow x_0^{(i-1)} + \delta^{(i)}$
结束循环
$x'_0 \leftarrow x_0^{(N)}$

3.3. 对抗样本质量评估

扩散模型 θ 通过从以下内容采样得到的图像质量进行评估 $p_\theta(x)$ (Goodfellow 等人，2014a；Ho 等人，2020年）。这种采样称为扩散模型的 推理 。与分类模型不同，扩散模型不会直接将图像作为输入，而是通过从图像中提取特征并以这些特征为条件生成图像来利用图像。我们的评估主要聚焦于这种发生过侵犯版权情况的条件推断。对于无条件推理，模型会采样噪声并生成图像。这个过程没有输入图像，也不会引发版权方面的担忧，因此不纳入我们的评估。

延续对抗样本领域的现有研究（Goodfellow 等人，2014b；Dai 等人，2018年；Jia 和 Liang，2017年），我们通过衡量对抗样本对图像生成性能的损害程度，评估其在扩散模型推理过程中的效果。如图1所示，推理过程分为两个阶段。在条件阶段，扩散模型从输入图像中提取特征。在生成阶段，模型利用这些特征作为条件来生成新图像。我们将

条件记为 c，将特征提取过程记为 $p_\theta(c|x)$。实际上，c可以是概括图像语义的自然语言提示（Gal 等人，2022年；Ruiz 等人，2022年），也可以是与图像相关的潜变量（Rombach 等人，2022年）。

扩散模型 θ 随后可以生成一幅图像 x_g，其条件 c_g 采样自 $p_\theta(c|x)$。我们将此过程建模为 $p_\theta(x|c_g)$，其中 $x_g \sim p_\theta(x|c_g)$。注意，$p_\theta(x|c_g) = p_\theta(x)^{p_\theta(c_g|x)\ p_\theta(c_g)}$。我们假设 c_g 与图像样本 x之间存在依赖关系。

假设 3.2（c_g **与** x**之间的依赖关系**）。c_g 是一个采样自 $p_\theta(c|x)$ 的条件。我们有 $p_\theta(c_g|x) / p_\theta(c_g) \geq 1$。

假设 3.2在大多数情况下是合理的，因为 θ 是一个经过训练的扩散模型，在语义上保证了样本与条件之间的强关系。在该假设下，$p_\theta(x|c)$ 可以作为 $p_\theta(x)$ 的上界，并作为采样的另一种分布。按照对扩散模型的常规评估方式，我们还通过将质量指标$\mathbf{D}(\cdot)$应用于从$p_\theta(x|c)$采样得到的图像，对 $p_\theta(x|c)$ 进行评估。评估过程总结于算法 2中。

算法 2评估用于扩散模型

输入： 对抗样本 x_{adv}、扩散模型 θ、样本质量指标 $\mathbf{D}(\cdot)$**输出：** 样本质量 $\mathcal{Q}$初始化数据集 $x_r \leftarrow x_{adv}$样本 $c_g \sim p_\theta(c|x_r)$ 通过采样生成图像 $x_g \sim p_\theta(x|c_g)$ $\mathcal{Q} \leftarrow \mathbf{D}(x_g, x_r)$

良好的对抗样本可以防止 $p_\theta(c|x)$ 准确地提取 c，从而导致较差的样本质量 x_g，这可以通过样本质量指标 $\mathbf{D}(\cdot)$来衡量。实践中，我们选择 Fr´echet Inception Distance (FID)(Heusel 等人，2017年) 和精确率 ($prec.$) (Kynk¨a¨anniemi 等人，2019年) 作为$\mathbf{D}(\cdot)$。

$(c|x)$ 和 $p_\theta(x|c)$ 的三种情景 将在扩散模型的对抗样本评估中予以考虑，具体如下：它们要么已经是（MT，2022年；Deck，2022年），要么可能是利用艺术人工智能侵犯版权的情景。三种情景的详细信息见附录 D.1。

1. **基于文本反演的文本到图像生成**：给定一小批图像 x，这些图像描绘了同一类别的物体，$p_\theta(c|x)$ 用一个 S^* 自然语言词语抽象图像中的物体。令条件 c_g 为 S^*。这通常通过扩散模型中嵌入的语言模型实现。$p_\theta(x|c_g)$ 随后使用扩散模型生成 x_g 以 S^*为条件的图像。该场景如图1所示。

2. **基于文本反演的风格迁移**：给定一小批图像 x，这些图像描绘了相同艺术风格的物体，$p_\theta(c|x)$ 用一个 S^* 自然语言词语抽象图像的共同艺术风格。令 c_g 为S^*。$p_\theta(x|c_g)$ 随后使用扩散模型生成以 x_g 为条件的图像 S^*。在实践中，为了更好地进行可视化，我们从另一个源图像 x_s 在步骤 t 时的潜变量$z_{s,t}$开始第 t 步的生成。该生成过程可以通过$p_\theta(x|c_g, z_{s,t})$精确表述。

3. **图像到图像合成**：给定一幅图像 x，$p_\theta(c|x)$ 在去噪步骤 t采样潜变量 z_t。令c_g 为 z_t。我们基于 z_t，从步骤 t开始生成过程 $p_\theta(x|c_g)$。

4. 实验

本节中，我们评估所提出的 AdvDM，以生成扩散模型的对抗样本。由于我们的动机是帮助保护绘画作品，防止其被人工智能艺术应用非法使用，因此我们选择潜在扩散模型（Rombach 等人，2022年）（LDM）作为骨干模型[1]，该模型是人工智能艺术应用中使用的主流模型（见附录 E）。AdvDM 在潜在扩散模型上的实现细节见附录 D.2。我们固定 l_∞范数作为生成所有对抗样本的约束。遵循现有的对抗样本研究，我们将采样步数设为 40，将每步扰动预算设为 1/255，总预算设为 8/255，并将批量大小设为 4。我们在 LSUN（Yu 等人，2015年）和 WikiArt（Nichol，2016年）类别上开展实验。所有实验均使用 8 个 NVIDIA RTX A4000 GPU。所有实验的可视化结果见附录 B。此外，我们还在更多基于潜在扩散模型的条件生成工具上评估 AdvDM，并在附录 F中展示结果。我们还讨论了其他可用于生成扩散模型对抗样本的方法，并将其与 AdvDM 进行比较。结果列于附录 C。

4.1. 基于文本反演的文本到图像生成

我们首先评估对抗样本在结合文本反演的文本到图像生成任务中的表现，正如第3.3节所述。为定量评估 AdvDM，我们从 LSUN-cat、LSUN-sheep 和 LSUN-airplane 中随机选取 1,000 张图像。所有实验设置均遵循文本反演论文（Gal 等人，2022年）中的设定。我们将图像分成每组 5 张，并针对每组优化一个条件提示 c，即 S^*（Gal 等人，2022年）中的伪词。S^* 是潜在扩散模型（LDM）中嵌入的语言模型语义空间中的一个词向量（Radford

[1]https://ommer-lab.com/files/latent-diffusion/nitro/txt2img-f8-large/model.ckpt

表 *1.* 基于文本反演的文本到图像生成

数据集	LSUN 猫类			LSUN-绵羊			LSUN-飞机		
指标	FID ↑	*prec.* ↓	*recall.*	FID ↑	*prec.* ↓	*recall.*	FID ↑	*prec.* ↓	*recall.*
无攻击	34.94	0.5643	0.1531	32.81	0.6378	0.1228	39.22	0.5016	0.2765
ADVDM	127.04	0.1708	0.061	203.5	0.0058	0.378	169.67	0.0263	0.3235

等人，2021），其预期能够捕捉 5 张图像中的对象，例如 LSUN-cat 图像中的猫。对于每组 5 张图像及S^*，我们默认将优化迭代步数设为 5,000。随后，我们使用每个伪词 S^* 进行条件生成，借助 LDM 的文本到图像功能生成 50 张图像，因此每个数据集共生成 10,000 张生成图像。默认情况下，所有图像均调整为 256×256。这一生成过程既针对干净图像进行，也针对 AdvDM 生成的对抗样本进行。

我们通过两个指标评估生成图像的样本质量：弗雷歇 Inception 距离（FID）和精确率（′），二者都用于衡量生成图像与训练图像之间的相似度。对于基于对抗样本生成的图像，较高的 FID 和较低的精确率表明这些图像无法捕捉对抗样本中的物体。作为参考，我们还在实验中计算了召回率（*prec.*）。然而，基于干净图像生成的图像与基于对抗样本生成的图像之间的召回率差异无法预测。这是因为召回率衡量的是生成图像的多样性，而非生成图像与训练图像之间的相似度；后者并非我们研究动机所关注的问题。上述指标的实现细节见附录D.1。

结果见表1。我们的对抗样本显著提高了条件生成图像的 FID，并降低了其精确率。与此同时，召回率的变化没有呈现一致趋势。这表明，我们的对抗样本能够有效保护其中的内容，防止其被提取为生成条件。

4.2. 风格迁移的定性结果

对抗样本的一项重要评估场景是使用潜在扩散模型进行风格迁移，此类场景中已发生多起侵犯版权的事件（Deck，2022年；MT，2022年）。在这些案例中，侵权者首先使用若干图像样本，通过文本反演训练一个伪词 S^*（Gal 等人，2022年），如第4.1节所述。随后，他们在图像到图像条件生成中利用 S^*，生成模仿样本图像艺术风格的图像。

为了评估我们的对抗样本抵御这种风格迁移的性能，我们沿用这一场景，比较以干净图像和对抗样本为条件生成的图像的样本质量

分别从 WikiArt 数据集中选取 10 位艺术家的 20 幅绘画作品，并为每位艺术家训练一个 S^*。其他设置与第 4.1 节中的设置相同。如图 2 所示，结果表明，以 S^* 对抗样本进行训练为条件生成的图像，其风格与输入图像显著不同。这说明 AdvDM 可有效用于防范非法风格迁移，保护版权。我们进一步开展实验，探究我们的对抗样本是否适用于商业化的人工智能艺术应用 Stable Diffusion。结果见附录 E。

4.3. 图像到图像合成的定性结果

如第3.3节所述，图像到图像生成是衡量对抗样本质量的另一种场景。我们首先将 AdvDM 应用于 Pexels 上的若干开源照片[2]，以生成对抗样本。然后，我们使用 Stable Diffusion 提供的图像到图像流程，基于这些对抗样本和干净图像生成图像。Stable Diffusion 是一种大规模商业化潜在扩散模型[3]。我们在图3中比较生成图像的质量。与基于干净图像生成的图像相比，基于对抗样本生成的图像不够逼真。

4.4. 消融研究

采样步数。 蒙特卡洛中的采样步数对估计准确性至关重要，因此从理论上说，它会对 AdvDM 生成的对抗样本产生显著影响。为探究这一超参数的作用，我们在 LSUN 飞机数据集上进行实验：随机选取 100 张图像，并按照第4.1节所述的设置生成 1,000 张图像，唯一不同之处是采样步数。采样步数从 10 变化到 1,000。结果如图 4 所示。随着采样步数增加，FID 大致呈上升趋势，而精确率则大致下降。这表明采样步数越多，对抗样本的质量越好。

[2]https://www.pexels.com/ [3] https://github.com/huggingface/diffusers

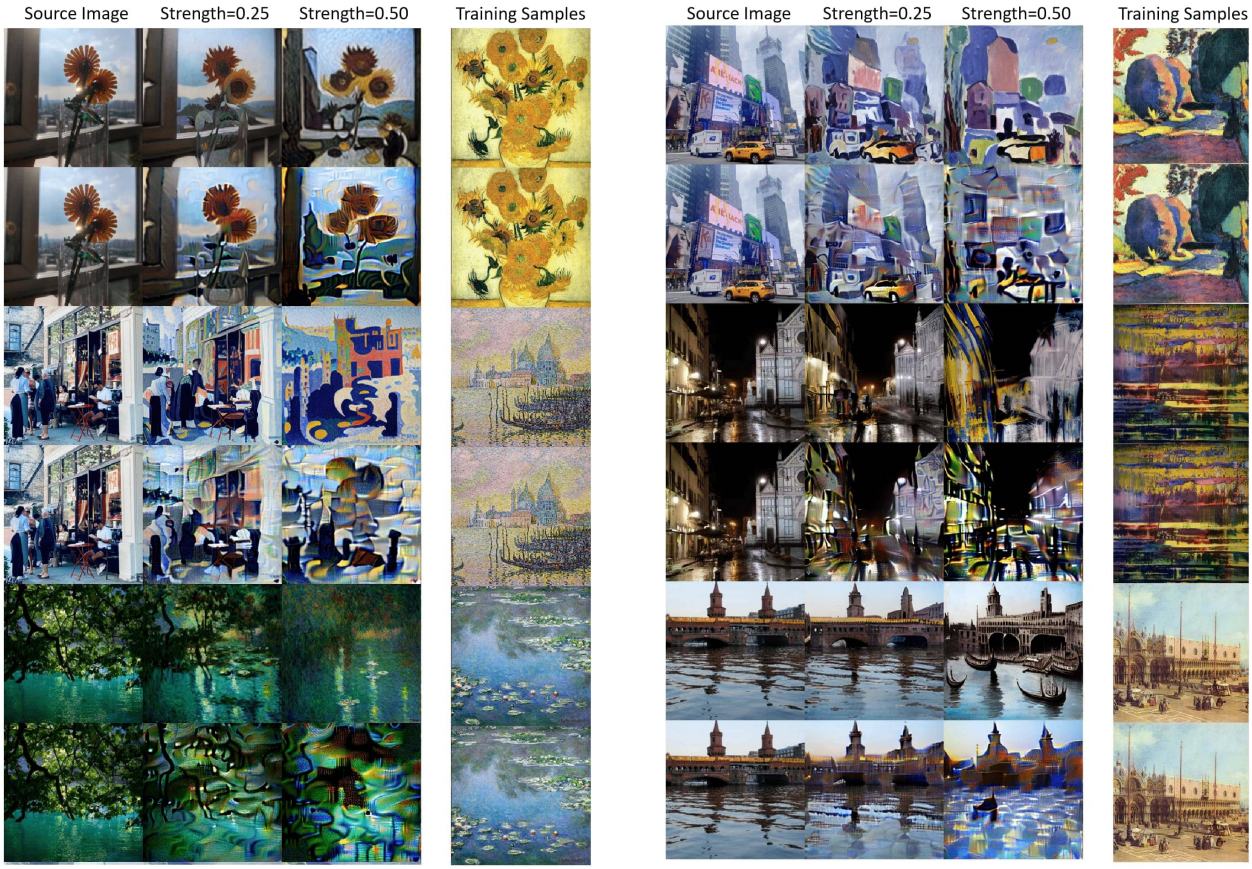


图 2. WikiArt 各类别风格迁移生成图像质量的比较（Nichol，2016年）。每组图像使用相同的源图像。我们使用文本反演（Gal 等人，2022年）从 WikiArt 的训练样本中提取风格，并将其单独列出。每组中，顶行展示基于从干净样本中提取的风格生成的图像，底行展示基于从对抗样本中提取的风格生成的图像。Strength 是一个超参数，表示源图像的风格在多大程度上被目标风格覆盖。与干净图像相比，潜在扩散模型无法从对抗样本中捕捉风格。

表2。AdvDM 在不同采样步数下的比较。推理时间是在 NVIDIA RTX A4000 GPU 上生成 1,000 张图像的对抗样本所需的平均时间。单位为秒。

采样步数	FID↑	*prec.* ↓	推理时间
10	122.9	0.05	1.803
40	186.05	0.037	6.342
1000	211.88	0.011	166.6

如表 2 所示，采样步数增加似乎会增强攻击效果，但也会导致推理时间延长。为平衡性能与推理时间之间的权衡，我们在主要实验中将默认采样步数设为 40。

扰动预算。 我们还研究了扰动预算对 AdvDM 生成的对抗样本质量的影响。除扰动预算外，我们还遵循第 4.1节中的设置。扰动预算 ϵ 在 2/255 到 32/255 之间变化。结果如下：

表3。 文本到图像生成场景中不同扰动预算下 AdvDM 的效果

LSUN 飞机	指标 FID↑	*prec.* ↓	*recall*
无攻击	54.03	0.659	0.242
ϵ=2	54.49	0.295	0.276
ϵ=4	116.79	0.09	0.342
ϵ=8	186.05	0.037	0.464
ϵ=16	217.09	0.015	0.569
ϵ=32	240.30	0.001	0.801

见表 3。我们观察到，即使扰动预算较小（4/255），AdvDM 也已能显著影响生成图像的质量。可视化结果见附录 B.1。

4.5. AdvDM 与预处理对抗防御

目前尚无研究专门探讨扩散模型的对抗防御问题。一种潜在的

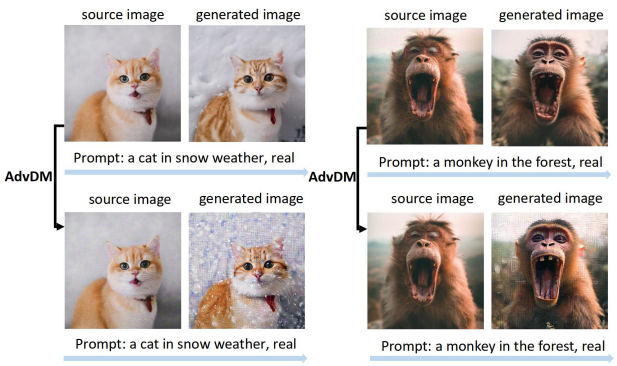
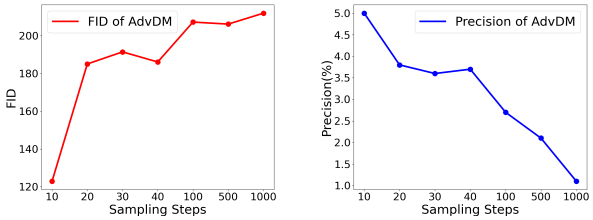


图 3。图像到图像生成中条件生成图像的比较。使用从我们的对抗样本中提取的条件时，LDM 会生成不真实的图像。



(a) FID-采样步数　(b) 精确率-采样步数

图 4。(a) AdvDM 的 FID 和采样步数。(b) AdvDM 的精确率和采样步数。

防御 AdvDM 的方法是利用预处理对抗防御，这类方法着重于消除对抗扰动。其优势在于无需重新训练生成模型或改变模型架构。鉴于此，我们对 AdvDM 生成的对抗样本进行 JPEG 压缩（Das 等人，2017年）、TVM（Guo 等人，2017年）和 SR（Mustafa 等人，2019年）处理。关于 AdvDM 的实验设置与第 4.1 节相同。

AdvDM 在基于预处理的对抗防御下的结果总结于表4中。可以观察到，JPEG 和 TVM 对 AdvDM 攻击的有效性都有限。SR 在防御方面表现更强，尤其体现在 FID 上。然而，对于精确率而言，其有效性并不显著。这表明，尽管预处理防御能够在一定程度上抵御 AdvDM，但无法完全恢复原始图像的语义信息。此外，如图5所示，基于对抗样本生成的图像与基于干净样本生成的图像之间存在显著差异。

尽管取得了上述成果，我们还采用 DiffPure（一种基于净化的先进对抗防御方法）来评估 AdvDM 的鲁棒性。实验结果见附录F.4。

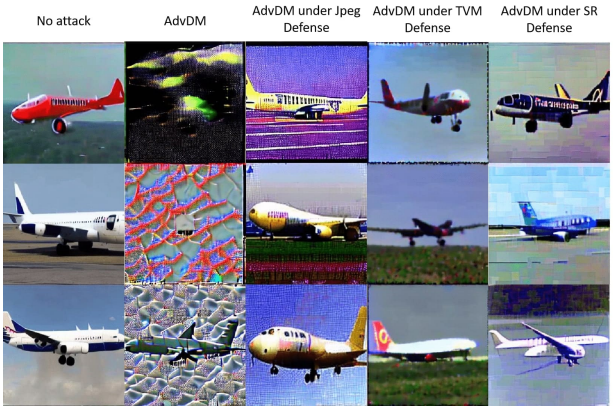


图 5。基于不同训练图像条件生成的图像可视化结果。在 AdvDM 下，所有防御都无法完美保持图像质量。

表 4。采用基于预处理的对抗防御进行基于文本反演的文本到图像生成

防御	无防御			JPEG		
指标	FID ↑	*prec.* ↓	*recall*	FID ↑	*prec.* ↓	*recall*
无攻击	39.22	0.5016	0.2765	39.19	0.5098	0.2639
AdvDM	169.67	0.0263	0.3235	61.67	0.1046	0.3208

防御	TVM			SR		
指标	FID ↑	*prec.* ↓	*recall*	FID ↑	*prec.* ↓	*recall*
无攻击	44.21	0.2513	0.1766	32.67	0.3397	0.2332
AdvDM	50.95	0.1744	0.2065	40.88	0.1673	0.2360

5. 相关工作

长期以来，对抗样本一直是不同场景中的重要研究课题，包括图像分类（Goodfellow 等人，2014b）和图（Dai 等人，2018年；Z¨ugner 等人，2018年）、文本理解（Jia 和 Liang，2017年）以及决策（Lin 等人，2017年）。我们对生成模型中对抗样本的定义，受到图像分类领域相关定义的启发（Carlini 和 Wagner，2017年）。

现有研究已经探索了针对不同生成模型的对抗样本，但尚未形成适当的框架。扩散模型被用于提升分类器的对抗鲁棒性（Nie 等人，2022年）。Kos等人研究了如何使生成模型生成会被错误分类的图像（Kos 等人，2018年）。线性流模型的对抗样本理论（Dinh 等人，2014年；2016年；Kingma &Dhariwal，2018年）已经提出，但该理论建立在数据呈正态分布这一强假设之上，而这一假设并不现实（Pope 等人，2020年）。另一项研究利用了针对分类器的替代攻击（Fetaya 等人，2019年），我们在附录 C 中将其与我们的方法进行了比较。

6. 结论

本文首次探索并提出了一个理论框架，用于定义扩散模型中的对抗样本，以保护人类创作的艺术作品。基于该框架，我们提出了一种为扩散模型生成对抗样本的算法。大量实验表明，我们的工作为抵御生成式人工智能提供了一种版权保护范式，也为人类艺术家提供了强有力的工具，使其艺术作品免遭基于扩散模型的人工智能艺术应用未经授权的使用。

致谢

本研究部分得到了中国国家自然科学基金（编号：61872234、61732010）、上海市可扩展计算与系统重点实验室以及英特尔公司（UFunding 12679）的支持。我们向来自北京大学的 Yichuan Mo 和 Qingsi Lai 致以衷心感谢，感谢他们提供了宝贵的审阅意见和反馈。

贡献：Chumeng Liang 和 Xiaoyu Wu 均为共同第一作者，并对本文作出了等量贡献。本文所研究的问题最初由 Chumeng Liang 和 Yiming Xue 提出，并由 Xiaoyu Wu 加以完善。算法由 Xiaoyu 和 Chumeng 设计。在此算法基础上，Chumeng 构建了理论框架。随后，Xiaoyu 设计并开展了实验，以评估该算法。

参考文献

Aufderheide, P. 和 Jaszi, P. 重拾合理使用：如何恢复版权中的平衡。芝加哥大学出版社，2018年。

Baio, A. 侵入式扩散：一位不情愿的插画师如何发现自己被变成了 AI 模型。https://waxy.org/2022/11/invasive-diffusion-how-one-unwilling-illustrator-found-herself-turned-into-an-AI-model/，2022年。

Bao, F.、Li, C.、Zhu, J. 和 Zhang, B. Analytic-DPM：扩散概率模型中最优反向方差的解析估计。*arXiv*预印本 *arXiv:2201.06503*，2022年。

Brock, A.、Donahue, J. 和 Simonyan, K. 用于高保真自然图像合成的大规模生成对抗网络训练。*arXiv* 预印本 *arXiv:1809.11096*，2018年。

Carlini, N. 和 Wagner, D. 《神经网络鲁棒性评估探讨》。发表于*S&P*，2017年。

Carr, S. 和 Jeffrey, N.《集体诉讼起诉书》。*Sarah An- derson* 等人诉 *Stability AI LTD.* 等人，2022年。

Dai, H.、Li, H.、Tian, T.、Huang, X.、Wang, L.、Zhu, J. 和 Song, L.《针对图结构数据的对抗攻击》。载于*ICML*，2018年。Das, N.、Shanbhogue, M.、Chen, S.-T.、Hohman, F.、Li, S.、Chen, L.、Kounavis, M. E. 和 Chau, D. H.《Shield：利用 JPEG 压缩实现深度学习的快速实用防御与疫苗接种》。载于*SIGKDD*，2018年。

Deck, A. AI 生成艺术引发日本动漫社群强烈反弹。https://restofworld.org/2022/ai-backlash-anime-artists/，2022年。

Dhariwal, P. 和 Nichol, A. 扩散模型在图像合成方面优于生成对抗网络。发表于*NeurIPS*，2021。

Dinh, L.、Krueger, D. 和 Bengio, Y. NICE：非线性独立成分估计。*arXiv* 预印本 *arXiv:1410.8516*，2014年。

Dinh, L.、Sohl-Dickstein, J. 和Bengio, S. 使用 Real NVP 进行密度估计。*arXiv* 预印本 *arXiv:1605.08803*，2016年。

Fetaya, E.、Jacobsen, J.-H.、Grathwohl, W. 和 Zemel, R. 理解条件生成模型的局限性。*arXiv*预印本 *arXiv:1906.01171*，2019年。

Fisher III, W. W. 重构合理使用原则。《哈佛法律评论》，101:1659，1987。

Franceschelli, G. 和 Musolesi, M. 生成式深度学习中的版权问题。《数据与政策》，4:e17，2022年。

Gal, R.、Alaluf, Y.、Atzmon, Y.、Patashnik, O.、Bermano, A. H.、Chechik, G. 和 Cohen-Or, D. 一张图像胜过千言万语：使用文本反演实现文本到图像生成的个性化。*arXiv* 预印本 *arXiv:2208.01618*，2022年。

Goodfellow, I. J.、Pouget-Abadie, J.、Mirza, M.、Xu, B.、Warde-Farley, D.、Ozair, S.、Courville, A. 和 Bengio, Y. 生成对抗网络。载于*NeurIPS*，2014a。

Goodfellow, I. J.、Shlens, J. 和 Szegedy, C. 解释并利用对抗样本。*arXiv* 预印本 *arXiv:1412.6572*，2014b。

Gregor, K.、Besse, F.、Jimenez Rezende, D.、Danihelka, I. 和 Wierstra, D. 迈向概念压缩。发表于*NeurIPS*，2016年。

Guo, C., Rana, M., Cisse, M., and VanDerMaaten, L. 使用输入变换抵御对抗图像。*arXiv* 预印本 *arXiv:1711.00117*，2017年。

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. 通过双时间尺度更新规则训练的生成对抗网络会收敛到局部纳什均衡。发表于*NeurIPS*，2017年。

Higgins, I.、Matthey, L.、Pal, A.、Burgess, C.、Glorot, X.、Botvinick, M.、Mohamed, S. 和 Lerchner, A. β-VAE：利用受约束变分框架学习基本视觉概念。在*ICLR*，2017年。

Ho, J.、Jain, A. 和 Abbeel, P. 去噪扩散概率模型。在*NeurIPS*，2020年。

Jia, R. 和 Liang, P. 用于评估阅读理解系统的对抗样本。*arXiv* 预印本 *arXiv:1707.07328*，2017年。

Karras, T.、Laine, S. 和 Aila, T. 面向生成对抗网络的基于风格的生成器架构。在*CVPR*，2019年。

Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J. 和 Aila, T. 分析并提升 StyleGAN 的图像质量。发表于*CVPR*，2020年。

Kawar, B., Zada, S., Lang, O., Tov, O., Chang, H., Dekel, T., Mosseri, I. 和 Irani, M. Imagic：基于文本的真实图像编辑与扩散模型。*arXiv* 预印本 *arXiv:2210.09276*，2022年。

Kingma, D. P. 和 Dhariwal, P. Glow：采用可逆1×1卷积的生成流。发表于*NeurIPS*，2018年。

Kingma, D. P. 和 Welling, M. 自动编码变分贝叶斯。发表于*ICLR*，2014年。

Kos, J., Fischer, I. 和 Song, D. 面向生成模型的对抗样本。发表于*S&P*研讨会，2018年。

Kynk¨a¨anniemi, T.、Karras, T.、Laine, S.、Lehtinen, J. 和 Aila, T. 用于评估生成模型的改进精确率与召回率指标。在*NeurIPS*，2019年。

Lin, Y.-C.、Hong, Z.-W.、Liao, Y.-H.、Shih, M.-L.、Liu, M.-Y. 和 Sun, M. 深度强化学习智能体的对抗攻击策略。*arXiv*预印本 *arXiv:1703.06748*，2017年。

Lu, C.、Zhou, Y.、Bao, F.、Chen, J.、Li, C. 和 Zhu, J. DPM-Solver：用于扩散概率模型采样、约 10 步内完成采样的快速常微分方程求解器。*arXiv* 预印本 *arXiv:2206.00927*，2022年。

Madry, A.、Makelov, A.、Schmidt, L.、Tsipras, D. 和 Vladu, A. 面向能够抵御对抗攻击的深度学习模型。在*ICLR*，2018年。

Mirza, M. 和 Osindero, S. 条件生成对抗网络。*arXiv* 预印本 *arXiv:1411.1784*，2014年。

MT, D. 人工智能艺术如何解放艺术家，而非取代他们。https://medium.com/thesequence/how-ai-art-can-free-artists-not-replace-them-a23a5cb0461e，2022年。Mustafa, A.、Khan, S. H.、Hayat, M.、Shen, J. 和 Shao, L. 将图像超分辨率作为抵御对抗性攻击的防御手段。*IEEE* 图像处理汇刊，29：1711–1724，2019年。Nichol, K. 按数字绘画，WikiArt。https://www.kaggle.com/c/painter-by-numbers，2016年。

Nie, W., Guo, B., Huang, Y., Xiao, C., Vahdat, A., and Anandkumar, A. 用扩散模型进行对抗净化。*arXiv* 预印本 *arXiv:2205.07460*，2022年。

Papernot, N., McDaniel, P., and Goodfellow, I. 机器学习中的可迁移性：从现象到使用对抗样本的黑盒攻击。*arXiv* 预印本 *arXiv:1605.07277*，2016年。

Poole, B., Jain, A., Barron, J. T., and Mildenhall, B. Dreamfusion：使用二维扩散进行文本转三维。*arXiv* 预印本 *arXiv:2209.14988*，2022年。

Pope, P., Balaji, Y., and Feizi, S. 基于流的生成模型的对抗鲁棒性。发表于*AISTATS*，2020年。

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., 等. 从自然语言监督中学习可迁移的视觉模型. 载于*ICML*, 2021.

Razavi, A., Van den Oord, A., 和 Vinyals, O. 使用 VQ-VAE-2 生成多样化高保真图像. 载于*NeurIPS*, 2019年.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., 和 Ommer, B. 使用潜在扩散模型进行高分辨率图像合成. 载于 *CVPR*, 2022年.

Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., 和 Aberman, K. DreamBooth：微调文本到图像扩散模型以实现主体驱动生成. *arXiv* 预印本 *arXiv:2208.12242*, 2022年.

Rumelhart、D. E.、Hinton、G. E. 和 Williams、R. J. 通过误差传播学习内部表征。技术报告，加利福尼亚大学圣迭戈分校拉霍亚认知科学研究所，1985年。

Schuhmann, C., Vencu, R., Beaumont, R., Kaczmarczyk, R., Mullis, C., Katta, A., Coombes, T., Jitsev, J., 和 Komatsuzaki, A. Laion-400m：由 CLIP 筛选的 4亿个图文对开放数据集。*arXiv* 预印本 *arXiv:2111.02114*，2021。

Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M. 等。Laion-5b：用于训练下一代图文模型的开放大规模数据集。*arXiv* 预印本 *arXiv:2210.08402*，2022年。

Shlens, J. 关于 Kullback-Leibler散度和似然的注释。*arXiv*预印本 *arXiv:1404.2000*，2014年。

Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., and Ganguli, S. 利用非平衡热力学进行深度无监督学习。发表于*ICML*，2015年。Song, J., Meng, C., and Ermon, S. 去噪扩散隐式模型。*arXiv* 预印本 *arXiv:2010.02502*，2020a。Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. 通过随机微分方程进行基于分数的生成式建模。*arXiv* 预印本 *arXiv:2011.13456*，2020b。Sullivan, J. E. 数字时代视觉艺术的版权：现代爱丽丝漫游奇境记。《艺术管理、法律与社会杂志》，26(1)：41–59，1996年。Vincent, J. 关于人工智能版权的可怕真相：没人知道接下来会发生什么。

https://www.theverge.com/2344685/generative-ai-copyright-infringement-legal-fair-use-training-data，2022年。Xia, X., Xu, C., and Nan, B. 用于花卉分类的 Inception-V3。发表于*ICIVC*，2017年。Yang, R., Srivastava, P., and Mandt, S. 用于视频生成的扩散概率建模。*arXiv* 预印本 *arXiv:2203.09481*，2022年。Yu, F., Seff, A., Zhang, Y., Song, S., Funkhouser, T., and Xiao, J. LSUN：利用深度学习并让人类参与其中构建大规模图像数据集。*arXiv* 预印本 *arXiv:1506.03365*，2015年。Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. 使用循环一致对抗网络进行无配对图像到图像转换。发表于*ICCV*，2017年。Z¨ugner, D., Akbarnejad, A., and G¨unnemann, S. 面向图数据的神经网络对抗攻击。发表于*SIGKDD*，2018年。

A. 伦理问题

本节将讨论基于生成式人工智能的最先进人工智能艺术应用所涉及的一些伦理问题，以及我们预期自身工作将在这些问题中发挥什么作用。

由扩散模型驱动的人工智能艺术应用大幅降低了艺术创作的门槛，重塑了艺术市场。然而，这些进步背后隐藏着尚未解决的版权问题。

未经图像所有者同意使用受版权保护的训练数据，可能构成未经授权的复制和分发，从而产生侵犯版权的法律责任。训练数据的一个主要来源是 LAION（Schuhmann 等人，2021；2022年），这是一个包含训练图像及其文本说明的大规模数据集。LAION 中的大量图像未经图像所有者同意，便从商业图像托管网站抓取而来（Carr 和 Jeffrey，2022年）。其他涉及未获许可的艺术作品的生成过程也存在同样的问题。例如，未经授权，利用小规模人工智能艺术应用的功能学习某位特定艺术家的绘画作品。

版权法保护作者复制、发行、表演和展示艺术作品的专有权利（Franceschelli & Musolesi，2022年）。这一法律框架使得在数字时代，未经版权许可使用他人的艺术作品极有可能构成侵权（Sullivan，1996）。在人工智能艺术的整个过程中，将未经授权的艺术作品从其最初发布的平台转移到人工智能数据库，以及出售或分发包含此类数据库的程序，都可能构成对原始艺术作品的复制和发行。这与人工智能艺术应用创作艺术作品的机制有关。人工智能艺术应用通过拟合训练图像，进而重组学习所得数据来生成新图像，这可以被理解为一种特殊形式的复制。对于一些具有鲜明、广为人知特征的艺术作品，例如迪士尼拥有的卡通形象，这种复制很容易被发现（Baio，2022年）。因此，在近期针对多家从 Stable Diffusion 中获利的公司的诉讼中，代表作品被用于训练这些生成式人工智能工具的艺术家的原告律师将扩散模型称为“21 世纪的拼贴工具”（Carr& Jeffrey，2022年）。

针对这些问题，为人工智能艺术应用提供的一种可能辩护是合理使用原则（Fisher III，1987）。合理使用的例子包括批评、评论、新闻报道、教学、学术研究和研究。为了科学研究目的，使用受版权保护的图像训练人工智能很可能构成合理使用。然而，生成部分则不然——

它以商业目的为导向，并为这些公司创造了数百万美元的收益。此外，人工智能艺术应用直接与艺术家竞争，成为其替代品，而这些艺术家正是其获取训练数据的来源。所有这些事实都不利于认定其属于合理使用。

版权法旨在平衡不同参与者的利益（Aufderheide 和 Jaszi，2018年），同时也关乎人类创造力的前景。一方面，研究人员为发展艺术人工智能付出了巨大努力。这类技术革新了艺术表达方式。另一方面，由于创作速度较慢、成本却高得多，艺术家正逐渐落于下风。法律需要时间来应对技术发展带来的新问题，我们期待法院就如何实现平衡给出答案。但在此之前，现实已严重失衡——科技公司通过挪用他人的知识产权，几乎不付出任何代价便赚取巨额利润；而艺术家只能眼睁睁看着自己赖以谋生的技能，因自身作品而大幅贬值。本文提出的保护方法旨在为艺术家提供一种武器，使他们能够依法保护自己在版权法下的法定权利。毕竟，人工智能应当公平地服务于所有人。

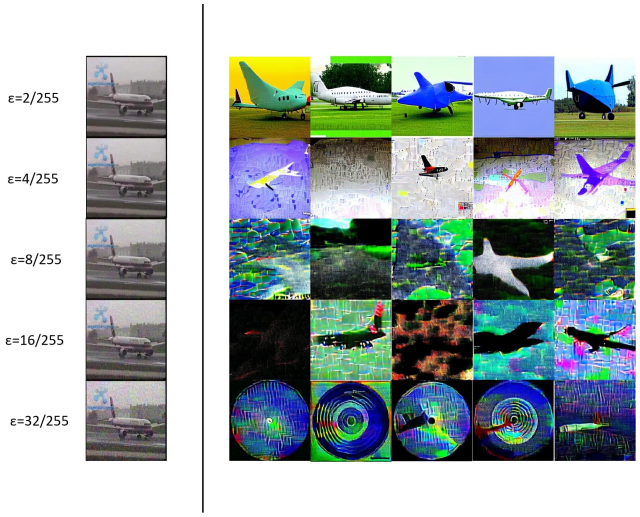


图 *6*。扰动预算消融研究的可视化。第一列：对抗样本。第二至第六列：在不同采样条件下，由对抗样本衍生出的伪词所生成的图像。

B. 更多可视化结果

B.1. 消融研究

我们在图中展示了不同扰动预算下基于对抗样本生成的图像6。扰动预算越大，飞机图像就越模糊。

B.2. 基于文本反演的文本到图像生成

我们在图中比较了对抗样本及其生成所依据的干净图像7。对抗样本与干净图像之间几乎没有人眼可见的差异。随后，我们使用文本到图像生成方法，分别以这些对抗样本和干净图像生成图像。结果如图所示8，表明我们的对抗样本会严重降低生成图像的质量。

C. 与其他潜在对抗样本的比较

目前尚无针对扩散模型对抗样本的相关研究。因此，除了 AdvDM 之外，我们还研究了其他几种针对扩散模型的潜在对抗样本，以全面了解 AdvDM。请注意，目前尚无专门用于生成扩散模型对抗样本的方法。随后，我们借鉴现有研究，探索了更多潜在方法。

在实验设置方面，这些对抗样本的步数和扰动预算均遵循第 4.4 节所述的相同设置。

C.1. 分类器上的 PGD

研究表明，一些条件生成模型甚至容易受到为分类模型生成的某些对抗样本的攻击（Fetaya 等人，2019年）。此外，神经网络之间对抗样本的可迁移性已得到广泛验证和利用（Papernot 等人，2016年；Z¨ugner 等人，2018年）。基于这一思路，我们在 InceptionV3 分类器（Xia 等人，2017年）上使用投影梯度下降（Madry 等人，2018年）（PGD）生成对抗样本。随后，我们将这些对抗样本视为可迁移至潜在扩散模型的对抗样本。该方法记为 **PGD（INCEPTIONV3）**。

C.2. 攻击嵌入层

请注意，潜在扩散模型包含一个嵌入层，可将图像投影为潜在空间中的表示。这可以视为自编码器中的编码器-解码器结构（Rumelhart 等人，1985年）。现有研究表明，编码器-解码器结构可用于生成对抗样本（Kos 等人，2018年）。受此启发，我们将 PGD（Madry 等人，2018年）应用于嵌入层。我们通过比较干净图像与对抗样本的潜在表示来计算一个新的损失项；对抗样本是通过向干净图像添加微小扰动 δ 得到的。优化目标为

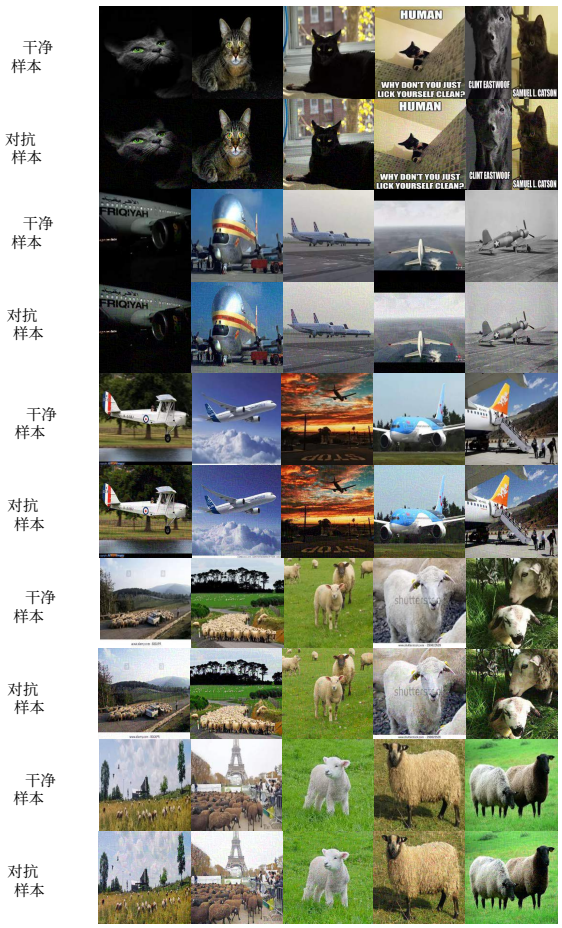


图 7. LSUN 猫类、LSUN 绵羊和 LSUN 飞机数据集中的一组干净图像和对抗图像。顶行展示干净样本，底行展示 AdvDM 下的对抗样本。

通过扰动使损失最大化。我们将此方法称为**嵌入攻击**，因为它通过针对潜在扩散模型中的嵌入层实施对抗攻击来生成对抗样本。

定义 C.1 (扩散模型的对抗样本（带嵌入攻击）)。 设 LDM 中的编码器为 $\mathcal{E}$。 x 是输入图像， δ 是在特定预算下的扰动。嵌入攻击生成的对抗样本表示为 $x' := x + \delta$，其中

$$\begin{aligned}\delta &:= \arg\max_{\delta} \mathcal{L}_{embedding}(x, \delta) \\ &= \arg\max_{\delta} \|\mathcal{E}(x) - \mathcal{E}(x+\delta)\|_2.\end{aligned} \tag{10}$$

我们将优化步骤中的对抗样本记为i。 $x^{(i)}$在实现中，我们遵循 PGD 的默认设置（Madry 等人，2018年），并随机初始化

在优化开始时，通过 $x^{(0)} = x + \epsilon z$ 初始化扰动，其中 $z \in \mathcal{N}(0, 1)$， ϵ 是攻击的扰动预算。对抗样本通过迭代式多步符号梯度上升生成，步长为 α。迭代步数设为 40。优化过程总结如下

$$x^{(i+1)} = x^{(i)} + \alpha \text{sgn}(\nabla_{x^{(i)}} \mathcal{L}_{embedding}(x, x^{(i)} - x)), \tag{11}$$

其中 sgn 表示符号函数。

C.3. PGD

另一种生成对抗样本的方法是对潜在扩散模型的损失应用 PGD。当采样步数 N 为 1 时，该方法等同于我们的方法。该方法记为 **PGD（LDM）**。

表5。在不同可能攻击下，使用对抗样本进行基于文本反演的文本到图像生成

	指标		
	FID↑	*prec.* ↓	*recall.*
无攻击	55.19	0.547	0.231
PGD（INCEPTIONV3）	56.89	0.306	0.153
嵌入攻击	175.34	**0.023**	0.352
PGD（LDM）	164.38	0.042	0.438
AdvDM	**186.05**	0.037	0.464

这些实验的结果列于表5中。从表中可以看出，在以 FID 衡量的所有基准方法中，AdvDM 取得了最佳结果。嵌入攻击也表现出相对较好的结果，尤其是在精确率方面。另一方面，针对扩散模型的 PGD 由于缺少采样过程，无法有效降低概率 p_θ，因此性能较差。分类器攻击涉及将针对分类器的攻击迁移到生成模型，但并未表现出明显效果，这表明该方法在这一设置下并不直接有效。

我们还在图 9 中展示了不同攻击下的生成效果。从可视化结果中可以观察到，在嵌入攻击下，尽管生成图像的背景中出现了噪声，但图像的语义信息并未受到很大破坏。然而，在 AdvDM 下，图像的语义信息（如形状或颜色）受到很大影响，这表明其攻击效果更强。

D. 实现细节

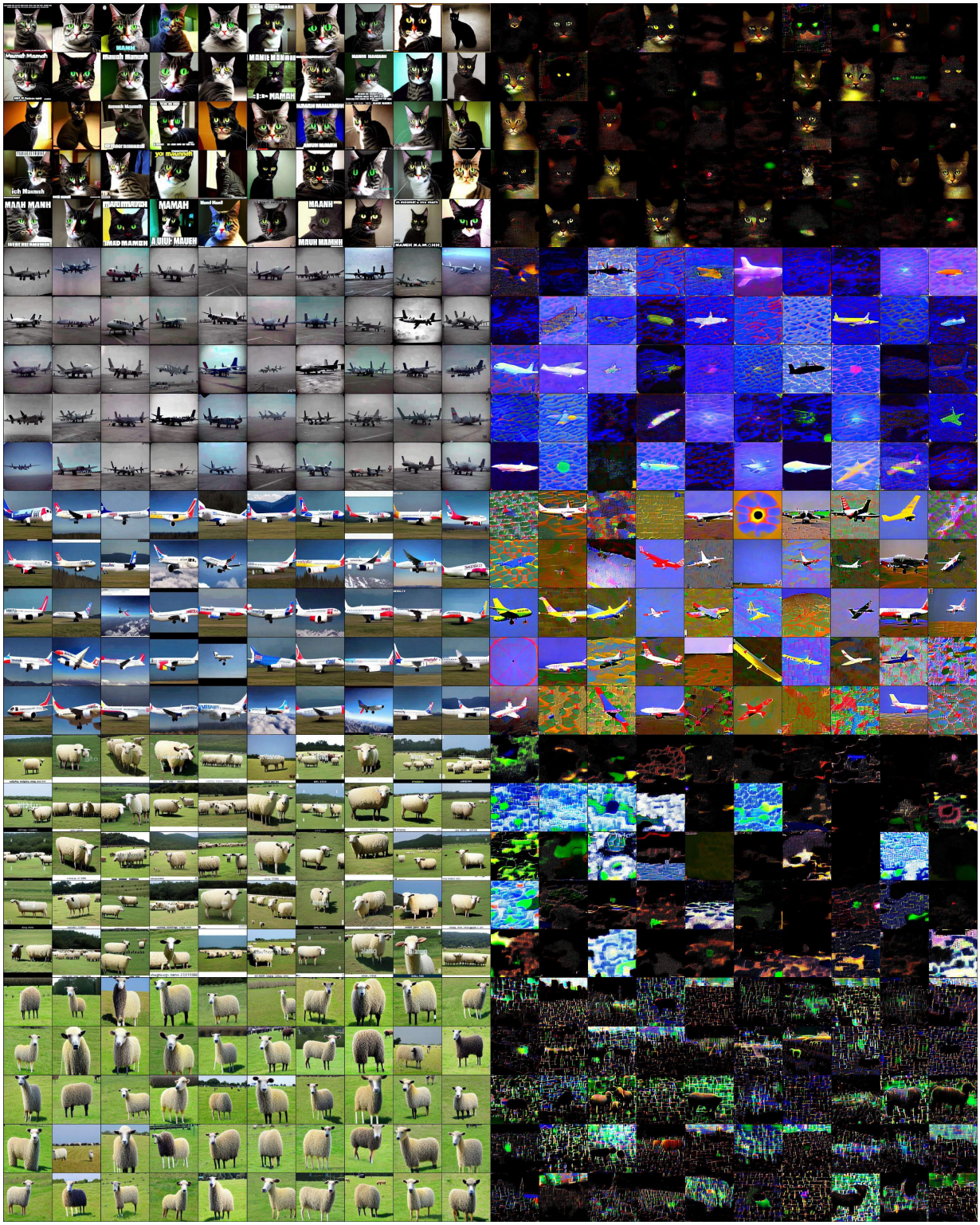


图 *8.* 左：基于文本反演对干净样本进行文本到图像生成。右：基于文本反演对对抗样本进行文本到图像生成。

算法 3 基于文本反演的文本到图像生成 输入：数据集 $\mathcal{D}$、潜在扩散模型 θ（含潜变量编码器 ε_θ 和嵌入式语言模型 $\mathcal{F}_\theta$）、优化步数 N、步长 α从 $\mathcal{D}$ 中随机选取 1,000 张图像 X将 X 分成 5 张图像一组 X: $=\{x^0, x^1, x^2, ...\}$初始化 $X_g \leftarrow \emptyset$对 X 中的 x^i 执行初始化 S^*，将其设为语义空间 $\mathcal{F}_\theta$ 中的随机词向量优化伪词 S^*: 对 $j = 1$ 到 N 执行从 x^i 中随机选取一张图像 $x\ c \leftarrow$“一张 S^* 的照片”计算 $\mathcal{L}_j = \mathbb{E}_{z\sim\varepsilon_\theta(x^i)}\ _c$、$_t||\epsilon_t - \epsilon_\theta(z, t, \mathcal{F}_\theta(c))||_2^2 S^* \leftarrow S^* - \alpha\nabla_{S^*}\mathcal{L}_j$结束循环 $c_g \leftarrow$“一张 S^* 的照片”使用 c_g 通过 θ 的文本到图像函数生成 50 张图像 x_g^i $X_g \leftarrow X_g \cup x_g^i$结束循环 计算 X_g 与 $\mathcal{D}$ 之间的 FID、$prec$ 和 $recall$

算法 4基于文本反演的风格迁移

输入：数据集 $\mathcal{D}$，潜在扩散模型 θ，带有潜在编码器 ε_θ 和嵌入式语言模型 $\mathcal{F}_\theta$，优化步数 N，步长 α 根据不同风格划分 D： $D' := \{x^0, x^1, x^2, ..., x^k\}$。初始化 $X_g \leftarrow \emptyset$，对于 x^i 中的每个 D'，初始化 S^*，使其成为 $\mathcal{F}_\theta$语义空间中的随机词向量。优化伪词 S^*: 对于 $j = 1$到 N，随机从 g^i中选取一张图像 $x\ c \leftarrow$“一幅 S^*风格的画作”计算 $\mathcal{L}_j = \mathbb{E}_{z\sim\varepsilon_\theta(x^i)}\ _{c,t}||\epsilon_t - \epsilon_\theta(z, t, \mathcal{F}_\theta(c))||_2^2 S^* \leftarrow S^* - \alpha\nabla_{S^*}\mathcal{L}_j$结束循环$c_g \leftarrow$ 包含“ S^*风格”的目标提示词$x \leftarrow$一张照片或脚本 使用 c_g, x 通过 θ的图像到图像函数生成图像 x_g^i $X_g \leftarrow X_g \cup x_g^i$结束循环 返回 X_g

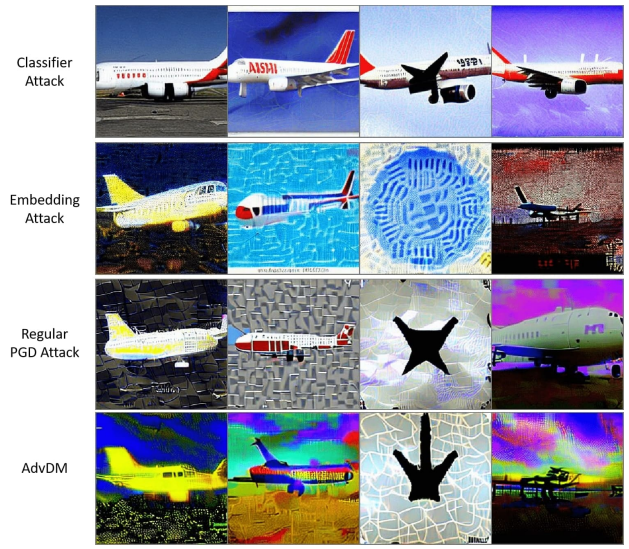


图 9. 不同攻击下生成图像的可视化结果。用于生成的伪词源自 LSUN 飞机数据集，我们可以观察到，AdvDM 能更有效地影响图像的语义信息。

算法 5图像到图像合成

输入： 源图像 X、潜在扩散模型 θ，以及潜在编码器 ε_θ 和嵌入式语言模型 $\mathcal{F}_\theta$
初始化 $X_g \leftarrow \emptyset$
对于 x，**在** X **中执行**
　$c \leftarrow$ 图像的目标提示词 x
　使用 c, x 通过图像到图像函数生成图像 θ
　$X_g \leftarrow X_g \cup x_g$
结束循环
返回 X_g

D.1. 评估详情

我们在算法 3、算法4和算法 5中分别详细说明了三种评估场景的具体流程。我们使用潜在扩散模型作者提供的预训练模型（Rombach 等人，2022年）。对于文本到图像生成和风格迁移，流程遵循文本反演论文（Gal 等人，2022年）推荐的设置。对于风格迁移，我们将强度固定为 0.5。我们还遵循该论文，将 N设为 5000。对于图像到图像生成，我们遵循（Rombach 等人，2022年）中的默认设置。

对于 FID 分数，我们使用一个开源软件包 [4]。对于精确率和召回率分数，我们使用 Dhariwal 和 Nichol（Dhariwal & Nichol，2021）提供的脚本。我们在整个类别数据集上计算这三项指标。

[4]https://github.com/w86763777/pytorch-gan-metrics

D.2. AdvDM 的实现

算法 6 在潜在扩散模型上实现 AdvDM

输入： 数据 x_0、参数 θ、去噪自编码器 ε_θ、编码器 $\mathcal{E}$，蒙特卡洛次数 N，逐步扰动预算 α，总体扰动预算 ϵ
输出： 对抗样本 x'_0
初始化 $x_0^{(0)} \leftarrow x_0$。
对于 $i = 1$ **到** N，**执行**
　样本 $x_{1:T}^{(i)} \sim q(\mathcal{E}(x_1^{(i)}) | \mathcal{E}(x_0^{(i)}))$
　样本 $t \sim U(1, T)$
　$\delta^{(i)} \leftarrow \alpha \text{sgn}(\nabla_{x_0^{(i)}} \|\mathcal{E}(x_T^{(i)}) - \varepsilon_\theta(\mathcal{E}(x_t^{(i)}), t)\|_2)$
　裁剪 $\delta^{(i)}$ s.t. $\|x_0^{(i-1)} + \delta^{(i)} - x_0^{(0)}\|_\infty \leq \epsilon$
　$x_0^{(i)} \leftarrow x_0^{(i-1)} + \delta^{(i)}$
结束循环
$x'_0 \leftarrow x_0^{(N)}$

我们在潜在扩散模型上实现 AdvDM。如算法6所示，在一系列采样步骤 $x_1^{(i)}:T$ 中，每一步都会使用一个随机时间戳 t，并将对抗扰动添加到原始图像上。

E. 针对 Stable Diffusion 的防护效果

我们的动机是保护人类艺术家创作的绘画作品，使其免遭人工智能艺术应用的模仿。值得注意的是，各种主流人工智能艺术应用 [5] [6] [7] [8] 都使用采用潜在扩散模型（LDM）架构的模型。因此，我们期望我们的对抗样本能够对这些人工智能艺术应用提供令人满意的防护效果。在此，我们开展了一项实验，以评估针对 Stable Diffusion（一款知名的人工智能艺术应用）的防护效果。值得注意的是，Stable Diffusion 使用的模型 [9] 与 LDM [10] 具有相似的架构，但规模更大、参数更多。

我们在 WikiArt 数据集（Nichol，2016年）上评估对抗样本。我们分别从三位艺术家——文森特·梵高、巴勃罗·毕加索和亨利·马蒂斯——的作品中选取了 20 幅绘画作品。然后，我们基于这些绘画作品生成对抗样本。采样步数 N 设为 100。扰动预算 ϵ 为 8/255，步长为 α 为 1/255。接着，我们在 Stable Diffusion 上使用文本反演进行风格迁移。该过程与算法 4 中所述的过程非常相似。对于伪词 S^* 的优化，优化步数为 8000，步长为 0.005。重建强度设为 0.5。我们首先在图 10、图 11 和图12 中比较用于优化 S^* 的干净绘画与对抗样本。然后，我们在图 13、图 14 和图15 中展示干净绘画和对抗样本上生成图像的结果。

[5] Text2Dream：https://deepdreamgenerator.com/#tools [6] Night Cafe：https://creator.nightcafe.studio/stable-diffusion- 图像生成器 [7] Hotpot：https://hotpot.ai/stable-diffusion [8] NovelAI：https://novelai.net/ [9] https://huggingface.co/CompVis/stable-diffusion-v-1-4- 原始版本 [10] https://ommer-lab.com/files/latent-diffusion/nitro/txt2img-f8-large/model.ckpt



图 *10.* 亨利·马蒂斯绘画作品的干净样本和对抗样本。



图 *11.* 巴勃罗·毕加索绘画作品的干净样本和对抗样本。

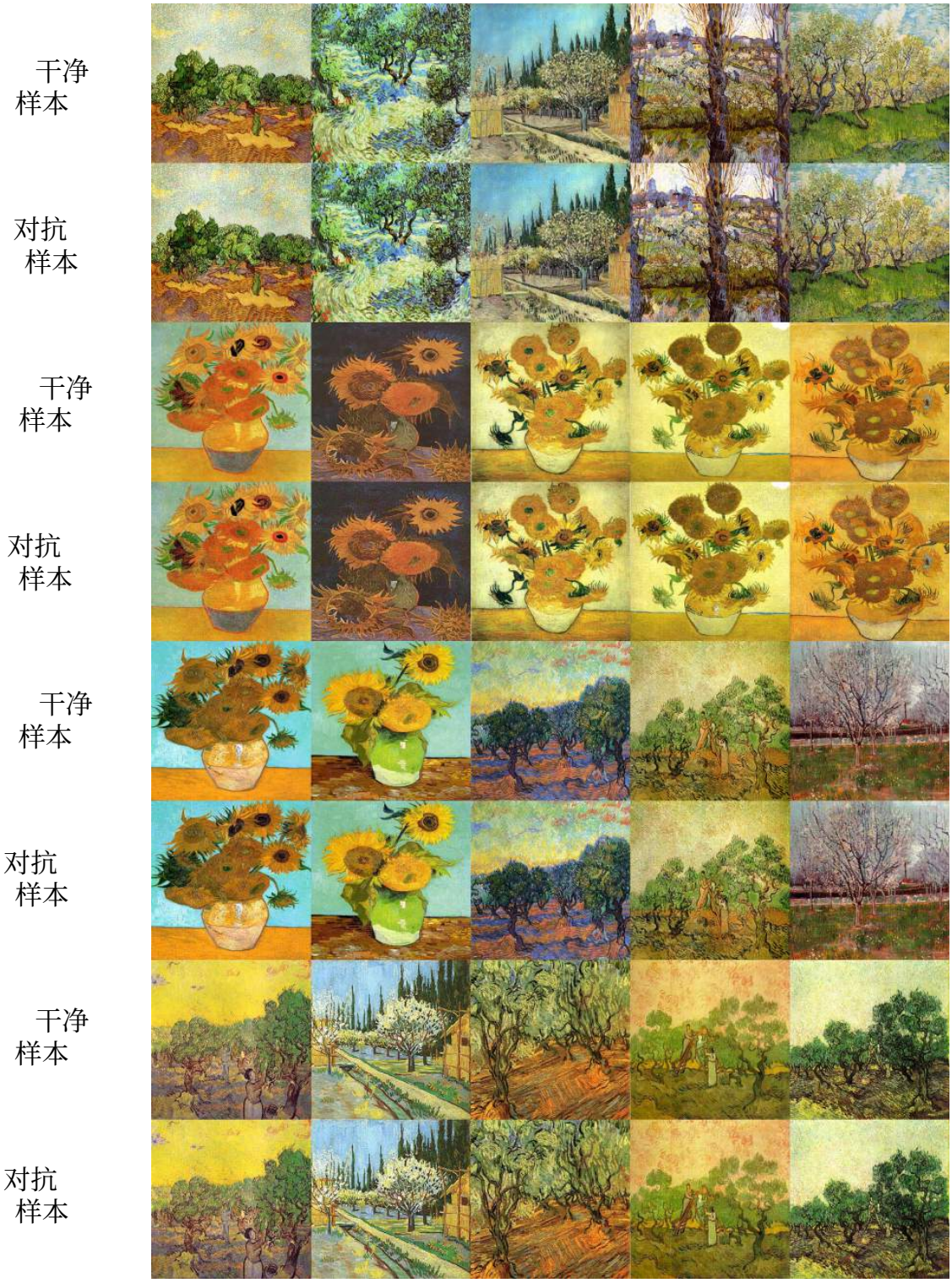


图 *12*。 文森特·梵高绘画作品的干净样本和对抗样本。



图 *13*。 **第一行：** Stable Diffusion 基于亨利·马蒂斯绘画作品的干净样本生成的图像。**第二行：** Stable Diffusion 基于亨利·马蒂斯绘画作品的对抗样本生成的图像。

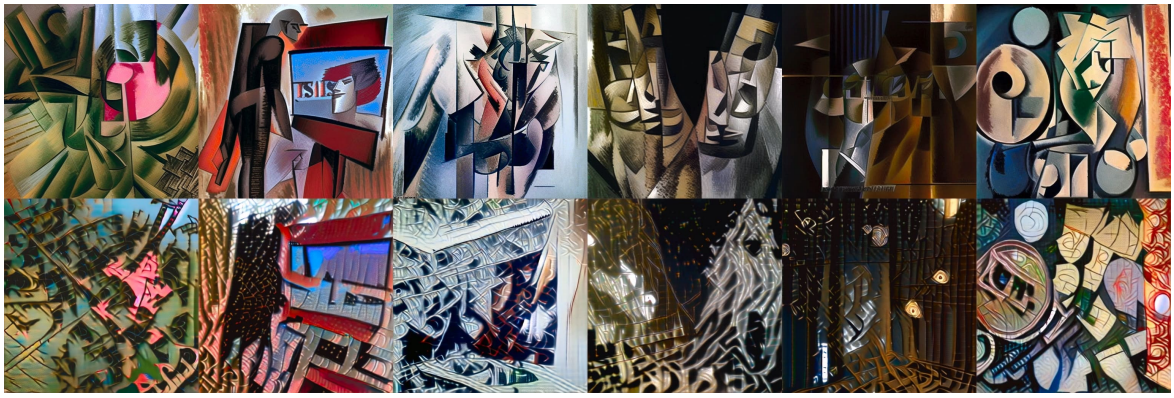


图 *14*。 **第一行：** Stable Diffusion 基于巴勃罗·毕加索绘画作品的干净样本生成的图像。**第二行：** Stable Diffusion 基于巴勃罗·毕加索绘画作品的对抗样本生成的图像。



图 *15*。**第一行：** Stable Diffusion 基于梵高绘画作品的干净样本生成的图像。**第二行：** Stable Diffusion 基于文森特·梵高绘画作品的对抗样本生成的图像。

F. 更多实验

F.1. AdvDM 在其他图像编辑任务上的表现

在本部分中，我们研究 AdvDM 在文本反演和图像到图像合成之外的图像编辑任务上的表现。我们考虑两项任务：图像修复和姿态引导合成。为展示 AdvDM 在商业艺术人工智能应用中的性能，我们选择 Stable Diffusion 1。5 [11] 作为骨干模型来生成和评估对抗样本。其他实验设置与第4节所述的设置保持一致。

图16 和图17 分别展示了图像修复和姿态引导合成的案例。可视化结果表明，当基于我们的对抗样本生成图像时，生成图像的质量很差，因而无法使用。具体而言，生成图像的内容会失去基本结构、出现奇怪的伪影，或变得过于简单。

F.2. DreamBooth 上的 AdvDM

在本部分中，我们研究 AdvDM 在 DreamBooth 上的性能。DreamBooth 是另一种可用于艺术风格迁移的主体驱动生成方法。我们使用 AdvDM 在 Stable Diffusion 1.5 上生成对抗样本，并通过使用 DreamBooth 对这些对抗样本进行风格迁移来评估其效果。我们采用 Python 库 *diffuser* [12] 的实现，将学习率设为 5×10^{-6}，步数设为 4000。其他实验设置与第4节中所述的设置保持一致。

图 18 展示了一个模仿梵高艺术风格的对比案例，使用了 20 幅绘画作品，与第4.2节中所述的设置相同。我们在两组图像上进行模仿：一组由干净绘画组成，另一组由基于这些干净绘画生成的对抗样本组成。结果表明，我们的对抗样本会在生成图像中添加混乱纹理，从而使生成图像无法使用。

F.3. AdvDM 在 scenario.gg 上的可迁移性

作为一款商业人工智能艺术应用，scenario.gg [13]除了支持 Stable Diffusion 之外，也支持艺术风格迁移，因此同样引发了人们对侵犯版权的担忧。我们开展实验，探究对抗样本能否迁移至 scenario.gg。由于 scenario.gg 由闭源扩散模型驱动，本实验旨在考察在黑盒对抗攻击设置下，AdvDM 生成的对抗样本的可迁移性。生成对抗样本时，实验设置与第 4 节所述设置保持一致。

结果如图19 和图20所示。与基于干净图像生成的图像相比，AdvDM 会在基于对抗样本生成的图像中添加混乱纹理，但效果不如白盒设置下的实验显著。不过，这些混乱纹理仍会使生成图像无法使用，从而实现我们方法的实际目标。

F.4. AdvDM 对抗防御：超分辨率和 DiffPure

AdvDM 的一个主要问题是，基于预处理的对抗防御可能会大幅削弱其效果。在本部分中，我们通过实验说明 AdvDM 在两种最先进水平的对抗防御下的有效性：超分辨率和 DiffPure。为了贴近更适合应用我们方法的真实黑盒场景，我们选择专用于艺术风格迁移的商业人工智能艺术应用 scenario.gg 作为骨干模型，以评估对抗样本的性能。如前文 F.3 所述，这恰好是一个迁移学习场景，因此能更好地验证 AdvDM 的性能。对于超分辨率，我们遵循第4.5节所述的设置。对于 DiffPure，我们使用作者提供的 DiffPure 原始实现 [14]。请注意，DiffPure 是一种基于模型的噪声净化方法，因此其效果高度取决于所用模型。在官方实现中，作者提供了三个模型，分别使用 Cifar-10、ImageNet 和 CelebA-HQ 训练，图像分辨率分别为 32×32、224×224 和 224×224。在本实验中，我们选择使用 ImageNet 训练的模型，因为它具有较高的分辨率，且该数据集的内容与我们实验中使用的绘画作品内容相对相似。所有设置均采用 DiffPure 官方实现的默认设置。

图21展示了结果。超分辨率和 DiffPure 都无法阻止我们的对抗样本在生成图像中添加混乱纹理。具体而言，基于干净样本并使用 DiffPure 生成的图像质量也很低。这是因为 DiffPure 输出图像的分辨率有限，而且在噪声净化过程中，输出图像的图像质量会下降。

[11] https://huggingface.co/runwayml/stable-diffusion-v1-5/tree/main [12] https://github.com/huggingface/diffusers/ [13] https://app.scenario.gg/

[14] https://github.com/NVlabs/DiffPure



图*16*。 **左侧：** 用于图像修复的源图像和掩码。**右上：** 基于干净图像生成的图像。**右下：** 基于对抗样本生成的图像。图像修复区域失去了一些基本结构。

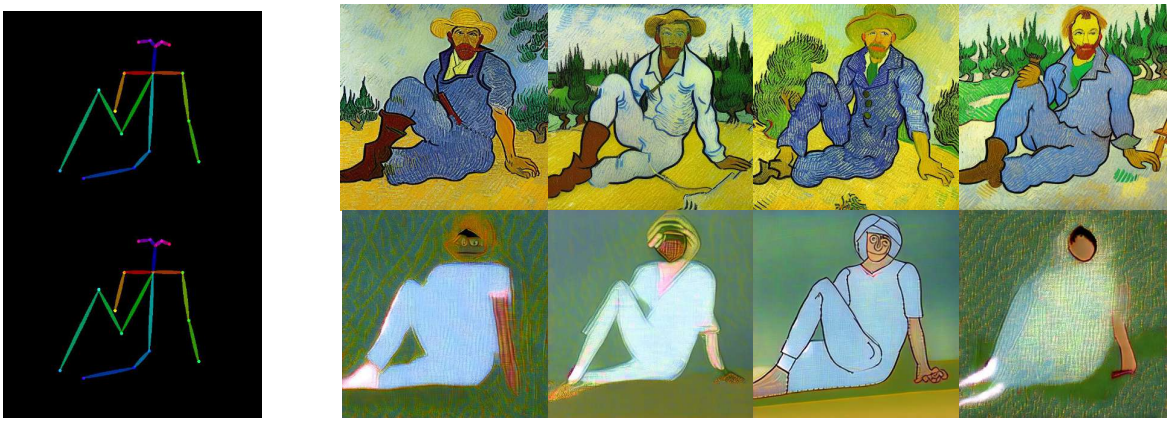


图*17*。 **左侧：** 用于生成的姿势。**右上：** 基于干净图像生成的图像。**右下：** 基于对抗样本生成的图像。基于对抗样本生成的图像失去了艺术风格的特征。

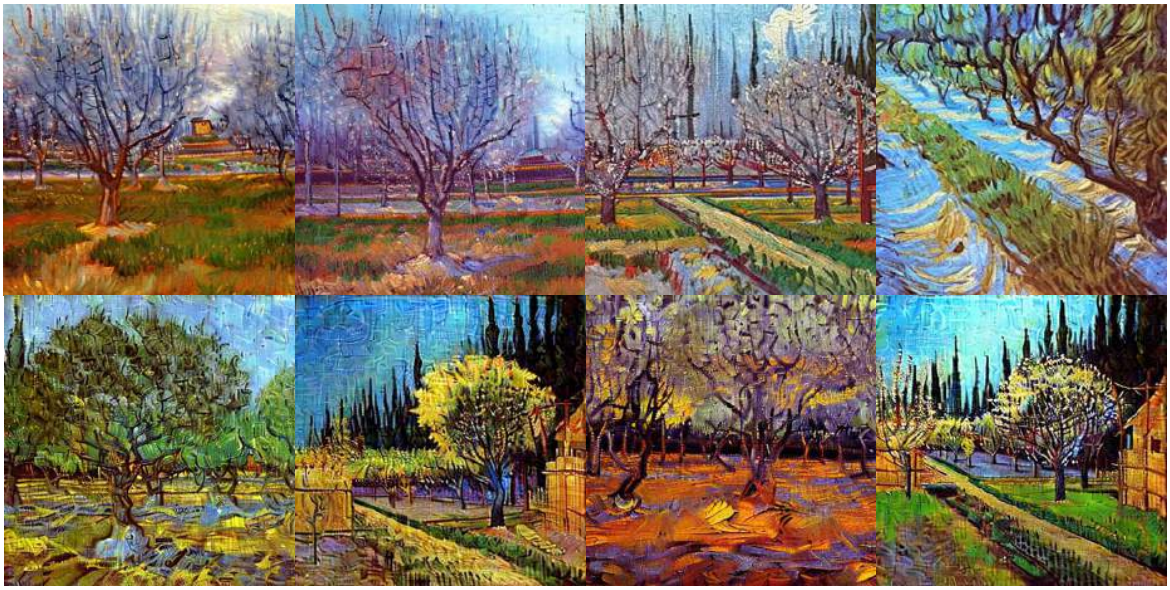


图*18*。 **上方：** 使用干净图像通过 DreamBooth 训练的 Stable Diffusion 生成的图像。**下方：** 使用带有混乱纹理的对抗样本通过 DreamBooth 训练的 Stable Diffusion 生成的图像。

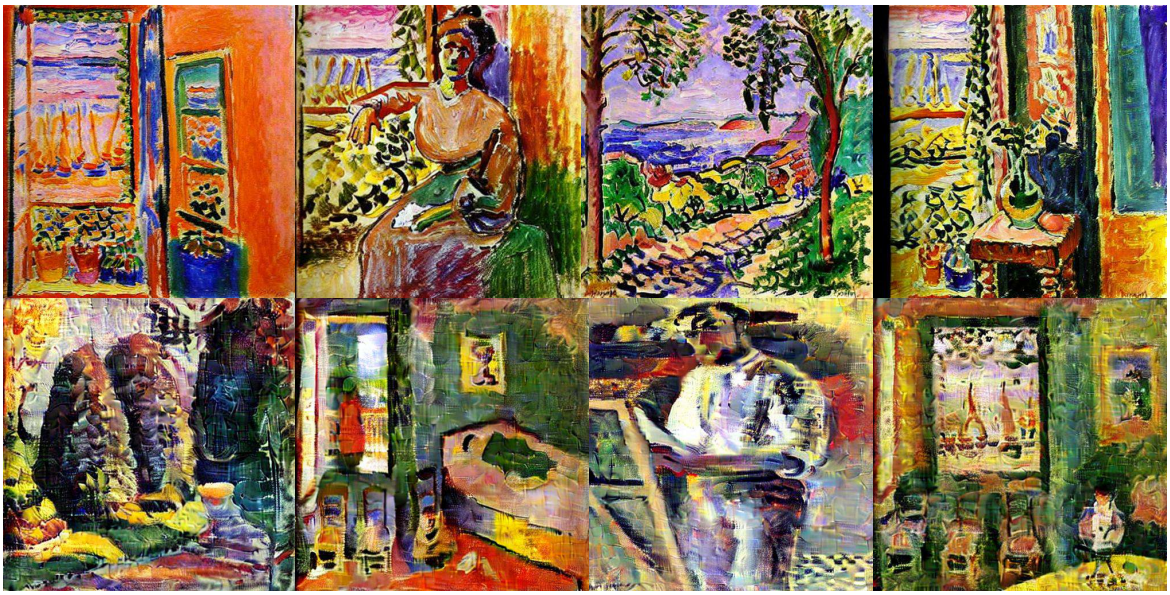


图*19*。 **第一行：** 基于亨利·马蒂斯绘画作品干净样本的生成图像，来自商业艺术生成网站 scenario.gg。**第二行：** 基于亨利·马蒂斯绘画作品对抗样本的生成图像，来自 scenario.gg。基于对抗样本生成的图像中出现了混乱纹理。

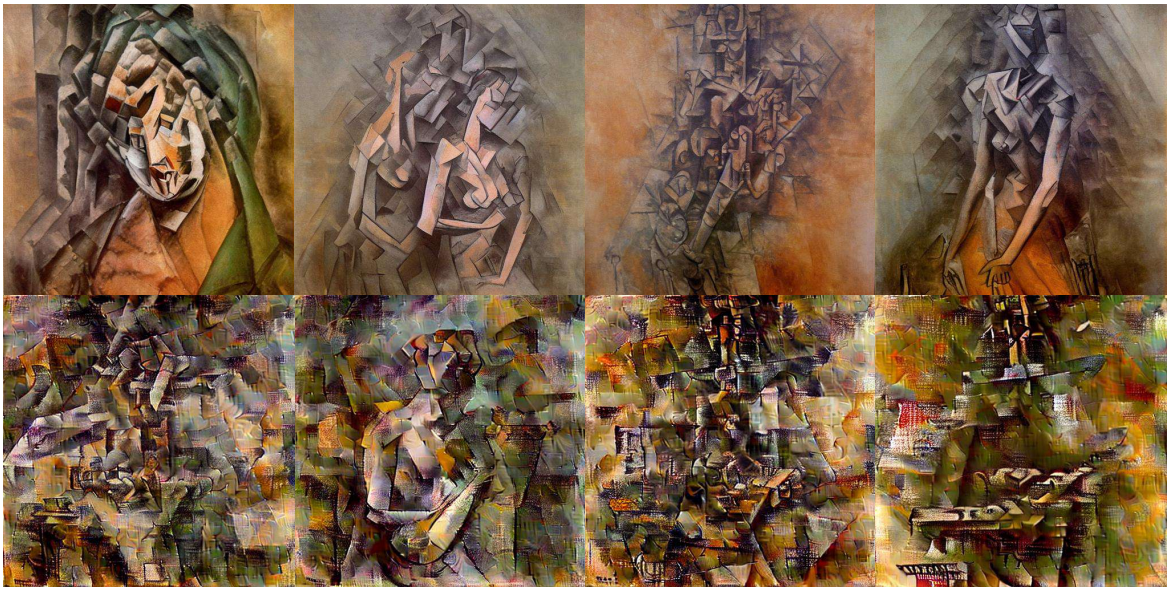


图*20*。 **第一行：** 基于巴勃罗·毕加索绘画作品干净样本的生成图像，来自商业艺术生成网站 scenario.gg。**第二行：** 基于巴勃罗·毕加索绘画作品对抗样本的生成图像，来自 scenario.gg。基于对抗样本生成的图像中出现了混乱纹理。



图*21.* scenario.gg 使用超分辨率和 DiffPure 防御处理梵高绘画作品的干净样本和对抗样本后生成的图像。**第一行：** 基于干净样本且无防御。**第二行：** 基于对抗样本且无防御。**第三行：** 基于对抗样本且使用超分辨率。**第四行：** 基于干净样本且使用超分辨率。**第五行：** 基于对抗样本且使用 DiffPure。两种防御都无法充分净化对抗扰动，因此生成图像的质量仍然较低。