DIVE: 用于判别任务的条件扩散模型的反转

Yinqi Li, Hong Chang, *Member, IEEE,* Ruibing Hou, *Member, IEEE,*
Shiguang Shan, *Fellow, IEEE,* and Xilin Chen, *Fellow, IEEE*

***Abstract*—扩散模型在图像和视频生成等各种生成任务中表现出显著的进步。本文研究了利用预训练的扩散模型来执行判别任务的问题。具体而言，我们通过“反转”一个预训练的布局到图像的扩散模型，将预训练的冻结生成扩散模型的判别能力从分类任务 [1], [2] 扩展到更复杂的目标检测任务。为此，分别提出了一种基于梯度的离散优化方法来替换繁重的预测枚举过程，以及一种先验分布模型来更准确地利用贝叶斯定理。实验证明，该方法在 COCO 数据集上的表现与基本的判别目标检测基线相当。此外，我们的方法可以在不牺牲准确性的情况下，大大加快之前基于扩散的方法 [1], [2] 进行分类的速度。代码和模型可在 https://github.com/LiYinqi/DIVE 获得。**

***Index Terms*—Diffusion model, generative modeling, discriminative Task, object detection, visual recognition.**

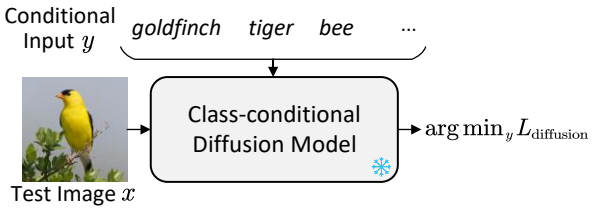
I. 引言

最近，生成模型如扩散模型 [3]–[8]、自回归模型 [9]–[12] 和生成对抗网络（GAN） [13]–[15]，由于它们能够合成照片般逼真的图像，在研究界获得了越来越多的关注。这种卓越的生成能力表明这些方法能够准确地建模数据分布并学习有效的图像表示。

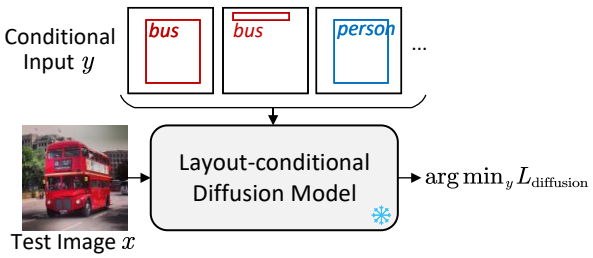
利用预训练的生成模型，一些工作 [16]–[22] 合成了一组训练样本，然后使用这些合成的训练集来训练相应的模型以进行判别任务。其他一些工作 [23]–[29] 则使用训练好的生成模型来提取特征，并在此基础上学习判别头。值得注意的是，这两类方法仍然依赖并训练判别模型以执行判别任务。而在这项工作中，我们研究纯粹的预训练生成模型是否可以完成判别任务。也就是说，我们专注于直接将预训练的生成模型转变为“判别模型”，而不微调模型参数或训练额外的判别头。我们相信，这样的研究可以更本质地揭示预训练生成模型的判别能力。

这种生成到判别的传统转换方法是通过贝叶斯定理来实现的，该方法通过计算 $\arg\max_y p(x|y)p(y)$ 来完成判别任务 $\arg\max_y p(y|x)$，其中 y 是任务标签，x 是输入图像 [30]。基于这一方法，最近，[1], [2] 将预训练的类（和文本）条件扩散模型，一类基于似然的生成模型，转换为生成分类器。对于在均匀先验 $p(y)$ 假设下的 C 分类任务，他们的解决方案是查询模型 C 次以获得每个类别的似然，然后选择最高的一种，如 Figure 1 (a) 中所示。在本文中，我们探讨如何将这种反向方法应用于更复杂的判别任务，例如对象检测，如 Figure 1 (b) 中所示。我们认为这一思路是有趣的，因为它为复杂的对象检测任务提供了一种纯粹的基于生成的解决方案，帮助研究人员从判别能力的角度理解预训练的扩散模型，并附带地为扩散模型提供了一种新的评价指标。同时，由于以下挑战，以往基于扩散的分类方法 [1], [2] 不能直接扩展：(1) 检测数据集的先验标签分布，即对象布局，不是均匀的，因此不能忽略；(2) 更为关键的是，不可能为测试图像枚举所有潜在的检测标签，因为对象及其位置的组合十分庞大，几乎是无数的。

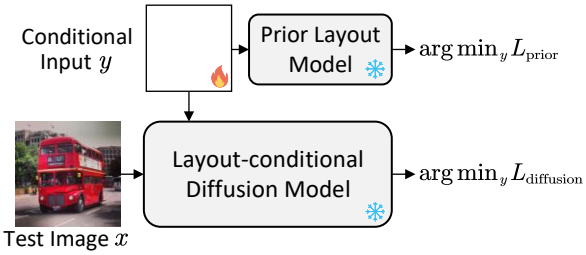
为了解决第一个问题，我们训练一个扩散模型来建模先



(a) Enumeration-based Diffuser Classifier [1], [2].



(b) Directly extending (a) to the object detection task.



(c) Our proposed optimization-based prior-integrated framework.

Fig. 1: 重新利用预训练的条件扩散模型用于判别任务而无需调整模型参数。$L_{\text{diffusion}}$ 和 L_{prior} 表示扩散模型 [4] 的训练损失，即所建模数据分布 $p(x|y)$ 和 $p(y)$ 的下界。

Y. Li, H. Chang (corresponding author), S. Shan and X. Chen are with Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, 100190, China, and University of Chinese Academy of Sciences, Beijing 100049, China.

R. Hou is with Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, 100190, China.

E-mail: yinqi.li@vipl.ict.ac.cn, { changhong, houruibing, sgshan, xlchen } @ict.ac.cn

验 $p(y)$ 。结合预训练的条件扩散模型 $p(x|y)$ ，我们通过后验最大化获得辨别预测 $p(y|x)$ 。对于第二个问题，我们提出不再像之前的工作那样测试每个潜在标签 y ，而是将 y 视为待优化的结果，通过最大化后验目标来找到它。具体来说，我们冻结训练好的扩散模型的参数，同时允许条件输入成为可学习的嵌入，并使用梯度方法来优化它。整体框架在 Figure 1 (c) 中有所展示，并在 Section III 中详细描述。我们称这种方法为扩散模型反转（DIVE），因为它将一个 y -to- x 模型反转为 x -to- y 。

我们主要在标准目标检测任务的背景下介绍和评估我们的方法。COCO 上的实验 [31] 显示，使用我们的方法逆转扩散模型在与基本判别检测器如 Faster R-CNN [32] 的对比中取得了具有竞争力的结果。此外，我们将基于优化的方法应用于图像分类任务。与之前基于枚举的方法 [1] 相比，我们的方法显著提高了速度，同时保持了准确性。虽然我们的方法与判别模型之间仍存在性能和效率差距，但我们希望这项工作能够帮助研究人员进一步认识预训练生成模型的判别能力，并可能作为复杂判别任务的一种新颖解决方案。

我们的贡献可以总结如下：

- 我们将反转扩散模型用于判别任务的范式从分类扩展到目标检测。据我们所知，这是首次将图像生成模型反转用于目标检测任务的研究工作。
- 我们提出学习模块，使反演范式在理论上更精确，在实现上更可行，以应对更复杂的检测任务。
- 我们通过实验证明，我们的方法不仅能达到与基本判别检测器相媲美的结果，并且比基于枚举的生成分类器运行得更快。

II. 相关工作

扩散模型是一种生成模型，通过在前向过程中逐渐向数据中添加噪声并通过马尔可夫链学习逆转此过程来逼近数据分布，从而生成新数据 [3], [4] 。扩散模型由于其合成高保真内容的能力，在各种生成任务中获得了极大的关注，包括图像生成 [6]–[8], [33], [34] 、图像编辑 [35]–[37] 和视频生成 [38]–[40] 等。

由于内容生成能力的优越性，最近使用生成模型来辅助或执行区分性任务的趋势有所增加。典型的工作可以分类为：

生成图像以训练辨别模型 [16]–[22], [41]–[47] ；

(2) 从预训练生成模型中提取表示，然后通过可学习的判别头部进行处理 [23]–[29], [48]–[51] ；

(3) 使用架构修改和额外的训练 [52]–[59] 根据图像生成任务标签；

(4) 利用来自模型的注意力图 [60] 完成与分割相关的任务 [19], [61]–[64] ，这些任务主要基于文本到图像扩散模型 [7], [8] 。

在这些工作中，(4) 仅适用于特定架构，(1) 至 (3) 需要额外训练，并且其中一些甚至包含判别模型。总之，它们没有从根本上揭示预训练图像生成模型的判别能力。

与所有这些工作不同的是，一些最新方法使用贝叶斯法则直接将图像生成扩散模型转换为执行辨别任务。[1], [2] 通过转换类别 [65] （或文本 [7], [8] ）的条件扩散模型来构建标准（或零样本）图像分类器，[66]–[68] 专注于构建鲁棒的扩散分类器。另一方面，[69], [70] 将这种范式扩展到图像-文本匹配任务，并利用其评估文本到图像生成模型。在这项工作中，我们进一步探索是否可以在模型反演范式下完成其他更复杂的辨别任务，比如对象检测，即对于测试图像没有一组预定义可能的类别 [1], [2] 或文本 [69], [70] 标签。

A. 生成模型逆转

生成模型反演通常指的是寻找给定图像的初始噪声表示，通过这个表示可以使用生成模型重建图像 [71], [72] 。这通常用于通过操纵反演的噪声表示来编辑真实图像 [71]–[74] 。对于 GANs，反演可以通过基于优化的方法实现，这种方法通过最小化重建损失来优化初始噪声 [75]–[77] ；基于编码器的方法，通过学习一个将图像映射到潜在噪声的编码器实现 [78]–[81] ；或结合这两种方法 [71], [82] 。对于扩散模型，反演可以通过在 DDIM 中提出的广义确定性前向扩散过程来完成 [83] 。随后的一些改进工作包括更精确更快速的反演 [84], [85] 以及扩展到文本条件模型 [86], [87] 。

虽然这些工作将图像反转到噪声空间以实现精确重建，本工作则是将扩散模型反转到条件输入空间以生成分类预测。最近，[88]–[90] 将扩散模型反转到条件输入空间，但他们的目标是将给定图像的概念 [88] 或完整内容 [89], [90] 反转为文本，并使用反转的文本生成新图像，即仍朝向重建目标，这与我们进行判别任务的目标不同。

III. 方法

在本节中，我们首先回顾条件扩散模型的背景，然后介绍我们将这些模型重新用于在目标检测背景下执行辨别任务的方法。

A. 条件扩散模型基础

1) 扩散模型 : 扩散模型 [3]–[5] 是基于似然的生成模型，经过训练以逆转扩散过程。正向过程被定义为一个马尔可夫链结构，逐步在干净图像 x_0 上添加噪声，持续若干时间步长 $t = 1 \cdots T$ [3] 。在逆向过程中，模型学习预测添加到图像上的噪声量 ϵ ，即 $\epsilon_\theta(x_t, t)$ ，或对于条件扩散模型为 $\epsilon_\theta(x_t, t, v_\theta(y))$ ，其中 x_t 是带噪图像，y 是条件输入例如类别标签，$v_\theta(\cdot)$ 将 y 映射到嵌入。具体来说，模型可以通过优化条件对数似然的变分界限进行训练 [4] ：

$$\max_\theta \log p_\theta(x|y) \approx \min_\theta \mathbb{E}_{\epsilon \sim N(0,1),t}\left[||\epsilon - \epsilon_\theta(x_t, t, v_\theta(y))||_2^2\right], \tag{1}$$

，其中 x, y 来自训练数据（我们使用 x 代表 x_0 以简化）并且 t 从 $\{1 \cdots T\}$ 中均匀采样。

条件分支将输入 y 编码到去噪网络中，实现可控的图像合成，例如类别到图像、文本到图像和布局到图像的生成。类别标签的编码可以很容易地通过一个可学习的嵌入层来实现，该层将整数类别索引映射到嵌入向量，就像部分工作所做的一样 [6], [7], [14], [65], [91] 。

潜在扩散模型(LDM)[7] 能够处理各种条件输入，包括图像布局如检测标签。基于这一能力，我们在本文中提出了一种模型反演方法。如 Figure 2 所示(左)，类似于类标签的编码过程，每个图像中的边界框 $(m_{min}, n_{min}, m_{max}, n_{max})$ 与相应的类别 c 首先通过可学习的词汇映射到嵌入向量中。这些框和类别的嵌入形成一个序列。由于不同图像通常包含不同数量的对象，因此序列使用 [none] 嵌入进行填充以实现固定长度。然后，这个序列由一个 transformer 编码器 [60] 处理，并通过交叉注意力层 [60] 整合到主要的去噪网络中。词汇和 transformer 编码器作为整体扩散模型的组成部分进行训练。

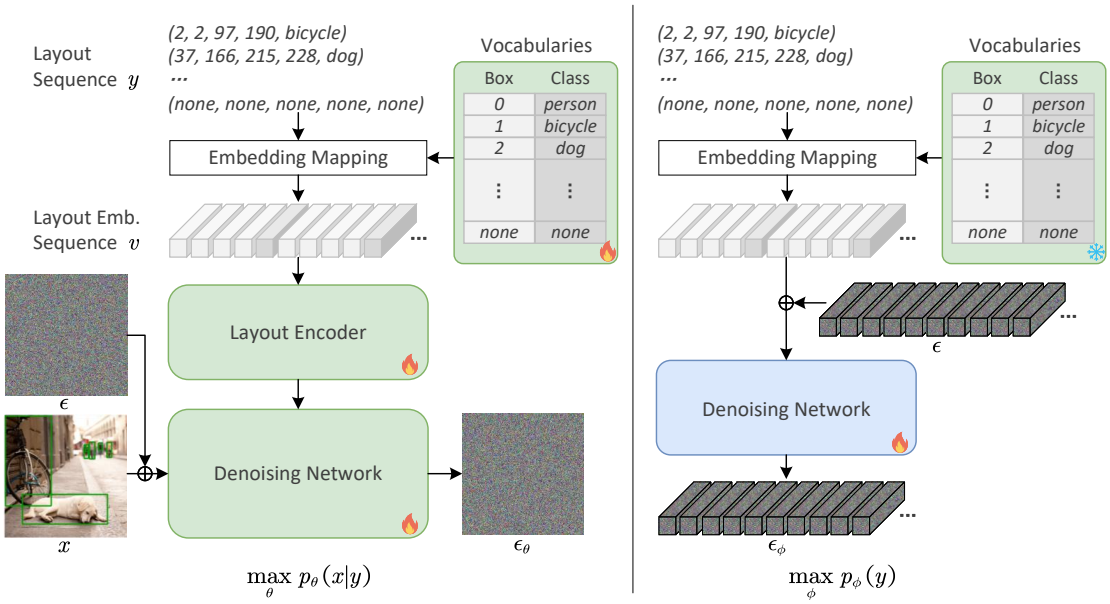


Fig. 2: 版面条件图像生成模型（左）和先验版面模型（右）的训练框架图示。图像 x 中的边界框仅用于可视化。

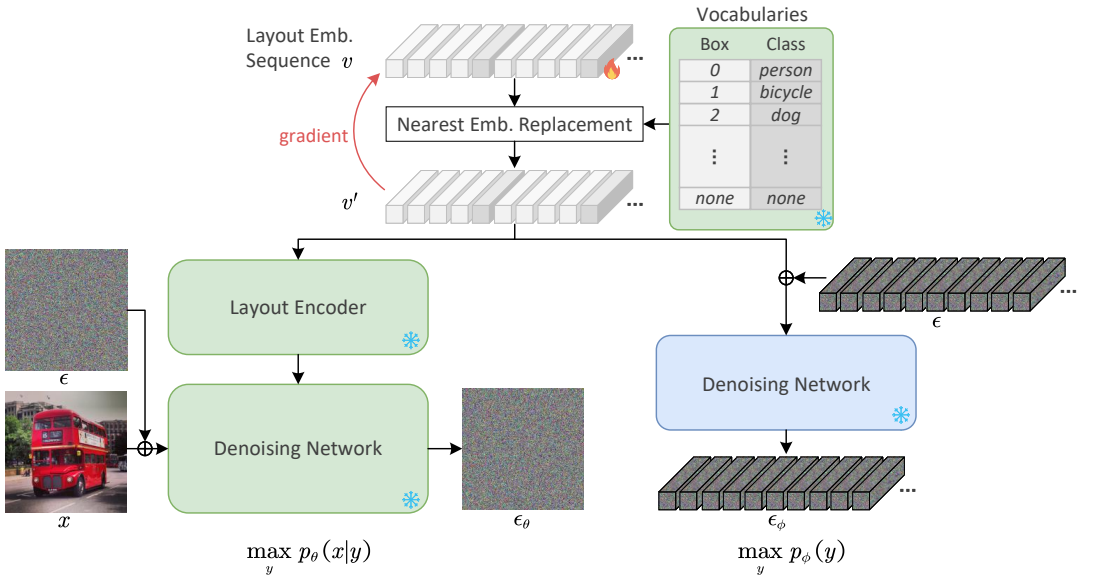


Fig. 3: 使用训练好的布局条件图像生成模型和先验布局模型进行物体检测。

B. 用于判别任务的条件扩散模型反演

给定预训练的条件扩散模型 $p_\theta(x|y)$，继 [1], [2] 之后，我们使用它们通过贝叶斯规则执行判别任务：

$$\arg\max_y p_\theta(y|x) = \arg\max_y p_\theta(x|y)p(y). \tag{2}$$

一个理论问题出现了：条件扩散模型只捕捉了条件似然 $p_\theta(x|y)$，而没有考虑先验 $p(y)$。在训练于平衡分类数据集如 ImageNet-1k [92] 上的类别条件扩散模型中，当通过计算每个类别的后验 [1] 将模型转换为“扩散分类器”时，忽略公式 (2) 中的 $p(y)$ 是合理的，因为先验是均匀分布的（即，对每个可能的类别来说都是相同的）。不幸的是，对于目标检测任务而言，这样的均匀先验分布根本不成立。例如，稀有布局（如人的上方有一辆自行车）或者具有不寻常宽高比的事物，应该具有相对较低的先验概率。因此，在使用公式 (2) 获得扩散检测器时，$p(y)$ 不能被忽略。

为了解决上述问题，我们建议对训练数据集中标签的底层先验分布进行建模。理想情况下，学习到的模型可以为任何检测布局 y 提供一个合理的 $p(y)$ 值。我们构建了一个扩散模型 $p_\phi(y)$ 来建模先验标签分布。与条件图像生成模型不同，这个先验模型是在标签空间中运行的标签生成模型。具体到目标检测，输入是噪声布局嵌入的序列，模型训练用于预测加入的噪声序列：如 Figure 2 （右）所示。先验模型的嵌入映射过程使用布局条件模型中冻结的预训练词汇，这里的去噪网络架构是一个 transformer [60]，与布局条件模型中的布局编码器相似。

通过学习到的先验知识和预训练的条件模型，从理论上讲，我们可以根据方程 (2) 精确计算贝叶斯后验概率的最大值，从而创建一个目标检测器。另一个挑战是，对于给定的测试图像，有无数潜在的检测标签。一方面，物体的数量是不确定的。此外，即使对于单个物体，随着其尺寸的增加，图像上可能的边界框数量呈二次增长。因此，需要一种非枚举的方法。

1) 基于优化的扩散模型反演： 我们并不是尝试无数种可能的预测然后选择具有最高后验值的那个，而是建议通

过最大化后验目标直接优化预测 y 。通过将方程 (1) 和 (??) 代入方程 (2) ，我们得到

$$y^* = \arg\min_y \{\mathbb{E}_{\epsilon\sim N(0,1),t}[||\epsilon - \epsilon_\theta(x_t, t, v_\theta(y))||_2^2] + \mathbb{E}_{\epsilon\sim N(0,1),t}[||\epsilon - \epsilon_\phi(v_\theta(y)_t, t)||_2^2]\}, \quad (3)$$

，其中我们只是简单地添加了两个扩散模型的损失。

一个实际的问题是嵌入映射（词汇查找）操作 $v_\theta(\cdot)$ 不可微。因此，我们转向嵌入空间来部署我们的可学习参数 v ，并通过使用相应词汇中的最近嵌入 $NN_{\text{value}}(v)$ 来替换这些可学习参数 v ，从而将它们限制在 $v_\theta(\cdot)$ 的离散输出空间中。此外，在反向传播过程中，我们将梯度从替换 $NN_{\text{value}}(v)$ 复制到原始的可学习参数 v 。这种离散优化技巧使得梯度可以流回到我们的可学习参数，类似于离散自编码器的训练 [9]–[12] 。

形式上，方程 (3) 实现为：

$$v^* = \arg\min_v \{\mathbb{E}_{\epsilon\sim N(0,1),t}[||\epsilon - \epsilon_\theta(x_t, t, v')||_2^2] + \mathbb{E}_{\epsilon\sim N(0,1),t}[||\epsilon - \epsilon_\phi(v'_t, t)||_2^2]\}, \quad (4)$$

$$v' := \text{sg}\left[NN_{\text{value}}(v) - v\right] + v, \quad (5)$$

，其中 sg 是停止梯度运算符，NN 表示最近邻，默认使用余弦距离作为度量标准。整个框架在 Figure 3 中进行了说明。

优化后，最终的预测 y^* 可以通过在预训练词汇中找到 v^* 最近嵌入的索引来获得：

$$y^* = NN_{\text{index}}(v^*). \quad (6)$$

当使用方程 (4) 在每一步学习 v 时，噪声和时间步总是从相应的分布中采样，类似于扩散模型的训练过程。我们对 v 进行若干次优化步骤的更新。为了获得 v^* ，我们不能直接比较每次更新的损失值，因为它们的噪声和时间步值不同。因此，我们预先保存了一组固定的噪声和时间步，并使用它们定期评估优化过程中的损失值。这些值作为选择 v^* 的监控器。Section IV 中报告的所有结果都是以这种方式获得的。

我们在初步实验中经验发现，用冻结词汇中的 [none] 嵌入来初始化可学习的嵌入 v ，在分类和检测任务上效果都很好。因此，我们保留了这一设计不变。在目标检测任务中，从优化后的序列获取检测对象时，我们去掉了不包含数值的对象和非法对象（即那些 $m_{max} \le m_{min}$ 或 $n_{max} \le n_{min}$ 的对象）。Algorithm 1 总结了对象检测任务优化过程的伪代码。

Algorithm 1: DIVE 算法用于目标检测

Input: pretrained layout-to-image model $\epsilon_\theta(x_t, t, v)$ with embedding vocabularies $v_\theta : y \to v$, pretrained prior layout model $\epsilon_\phi(v_t, t)$, a set of timesteps T used for inversion (same between the two models), test image x , maximum optimization steps K , monitoring frequency

Output: prediction y of the given image x

```
预先保存两组固定噪声以进行评估
E_forθ = {ε_forθ ~ N(0,1)}^range(T)
E_forφ = {ε_forφ ~ N(0,1)}^range(T)
# 优化循环
y = [none, none, ··· , none] ; v = v_θ(y)
monitor_values = [] , corresponding_embs = []
for k in range (K) do
    词汇内离散优化
    v = stop_gradient (NN_value(v) − v) + v
    t ~ T , ε_forθ ~ N(0,1) , ε_forφ ~ N(0,1)
    x_t = addn(x, t, ε_forθ) , v_t = addn(v, t, ε_forφ)
    L = ||ε_forθ − ε_θ(x_t, t, v)||_2^2 + ||ε_forφ − ε_φ(v_t, t)||_2^2
    Update v to minimize L
    定期评估优化的 v
    if k % monitoring frequency == 0 then
        monitor_values.append(eval(v))
        corresponding_embs.append(v)
Function eval( v):
    v = NN_value(v) , drop none and illegal boxes
    Calculate val_losses over {T} like line10-12
      but uses saved fixed noises in E_forθ, E_forφ
    return averaged val_losses
# 得到最终结果
v* = corresponding_embs[argmin(monitor_values)]
return y = NN_index(v*) , drop none and illegal
  boxes
```

IV. 实验

我们在本节中进行实验，以从以下三个方面评估所提出的扩散模型反演（DIVE）方法的有效性：

- 仅通过使用扩散模型进行对象检测（Section IV-A ）。
- 加速之前的扩散分类器方法 [1] 应用于图像分类任务（Section IV-B ）。
- 为条件扩散模型提供一个独立的评估指标（ Section IV-C ）。

A. 用于目标检测的 DIVE

在本小节中，我们旨在回答以下问题，以评估我们的方法在标准目标检测任务中的表现：

- 我们的扩散模型反演方法与其他基于扩散模型的方法和纯判别方法相比如何？
- 所提出的两个组件，即先验建模和离散优化，如何影响最终结果？

1) 设置：

a) 数据集和评估协议： 我们使用目标检测数据集 COCO 2017 [31] 进行实验，该数据集包含 118k 个训练图像和 5k 个验证图像。尽管我们基于优化的方法已经使检测可实现，但由于生成方法的计算成本依然很高，我们在一个包含 500 张图像的子集上评估 DIVE。我们报告平均精度 (AP)，这是一个在多个交并比 (IoU) 阈值上进行平均的指标，被认为是在 COCO 上评估性能时最重要的单一指标。我们还根据 [31] 中的定义，提供了其他指标（单一阈值 0.5 和 0.75 的 AP 以及小、中、大尺寸物体的 AP），以彻底分析我们的方法。

由于在 COCO 目标检测数据集上没有公开可用的预训练布局到图像扩散模型，我们使用 LDM 的官方代码重新

TABLE I: 优化步骤对 DIVE 的影响。在 Nvidia 3090 GPU 上评估速度。我们观察到，如果切换到更先进的 A100 GPU，速度可以提高 ∼ 1.8 倍，如 ?? 所总结。

Optimization step	400	1000	2000	4000
Time per image	18min	45min	1.5h	3h
AP	5.7	6.4	7.1	7.2

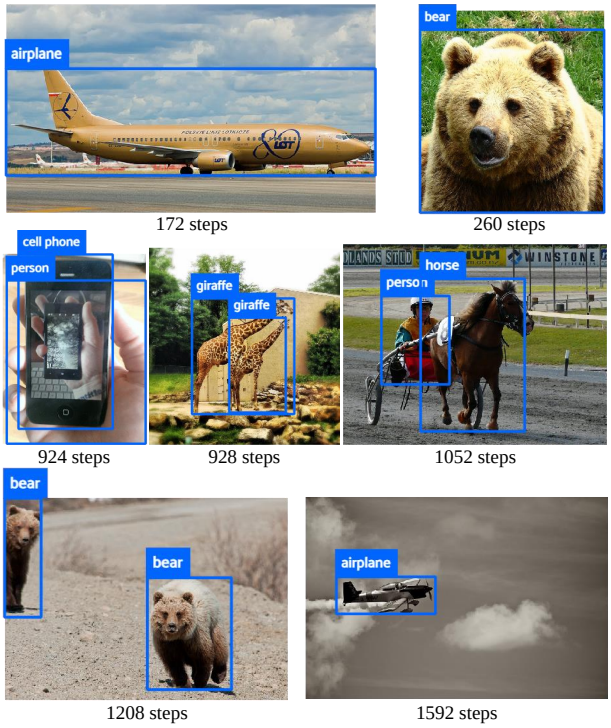


Fig. 4: 将 DIVE 检测结果可视化，并在下方显示相应的收敛优化步骤，这些结果来源于 ?? 中介绍的监控器。为了简单起见，所有图像的最大优化步骤设置为一个固定的数字 (2000)。测试集的平均收敛步骤大约为 960。输入到网络的图像分辨率为 256×256，这里以它们原始的纵横比显示，以便于更好的可视化。

训练了一个模型。去噪网络采用 UNet 架构，布局编码器是一个 transformer。按照其实现，我们在具有 8 倍下采样因子的 VQGAN 潜空间中训练模型（名为 LDM-8）。模型在图像分辨率为 256×256 的情况下进行训练，输入布局序列具有固定的 100 个对象填充长度。对象坐标保持整数精度，这意味着在边界框词汇表中存储有 $256 + 1$ 个嵌入。我们为嵌入的布局（框和类别）序列添加了序列长度可学习的位置编码。训练期间，图像中的对象在不同迭代中随机打乱，因为它们不应该有固定顺序，即对象在序列中位置相等。这种模型，其他设计不变，具有 363M 参数。我们在 200k 次迭代（即 220 个 epoch）上训练模型，使用 8 个 Nvidia A100 GPU 大约需要 5 天时间。

英寸

b) 先验布局扩散模型： 先前的布局扩散模型与布局到图像模型中的布局编码器具有相似的架构。两者都是 transformers，我们简单地继承了其实现，形成了一个 58M 大小的模型。区别在于它们的功能——在先前模型中，transformer 作为去噪网络，而在布局到图像模型中，它则作为条件编码分支。我们在 50 万次迭代（∼ 270 个 epochs）中训练了先前模型，大约需要 3 天时间，并使用 4 个 Nvidia 3090 的 GPU。

英寸

c) 潜水细节： 在使用训练过的 LDM 优化我们的嵌入时，我们继承了原始 LDM 代码中的超参数选择，除非另有说明。根据 [1]，我们使用均匀分布的时间步来提高学习和监控效率。均等间隔为 5，因此我们有 200 个不同的时间步 $\{2, 7, \cdots, 997\}$。监控器在这些不同的时间步循环后激活。由于 GPU 内存限制，我们进行批量大小为 50 的实验，由 5 步的梯度累积组成。学习率设置为 0.01，使用 AdamW 优化器 [100]。我们使用单个 Nvidia 3090 GPU 对每个图像进行逆转。关于 DIVE 的优化步骤的计算成本将在下一段中讨论。

d) DIVE 优化步骤和计算成本： 我们测试了一组优化步骤以评估其影响。Table I 显示了定量结果，其中所有图像在每次试验中使用相同的优化步骤。我们没有观察到步骤超过 2000 时的显著 AP 增益。我们注意到计算成本相对较高于常见的判别检测器，因为 DIVE 需要通过整个网络进行多次反向传播来优化结果。虽然检测任务的计算成本目前不是这项工作的主要关注点，它是为了使检测任务可通过预训练冻结的图像生成模型实现并检查其判别能力，DIVE 的高成本推动我们探索是否存在潜在的减少成本而不损害检测性能的可能性。

为了达到这个目的，我们检查了一些图像的收敛步骤，并在 Figure 4 中展示了它们与检测可视化结果一起。可以看到，对于一些具有更大和更少物体的“简单”测试图像，优化过程收敛得更快。而对于“困难”的图像，DIVE 需要更多的步骤才能收敛。基于这一观察，我们可以得出结论，使用自适应优化步长策略是一种在保持检测性能的同时减少总体计算成本的可行方法。例如，如果监控值（损失）在若干步内没有减少，我们可以使用提前停止策略来终止优化过程。我们进一步检查了测试集的平均收敛步骤，大约为 960，这表明提前停止方法总体上可以达到速度提升的 ∼ 2× 倍。目前，为了简单起见，以下实验中报告的 DIVE 结果均使用固定的 2000 步优化来处理所有图像。

英寸

2) 与判别检测器的比较：

a) 基线： 我们首先将我们纯生成的方法与生成-判别混合方法进行比较，这些方法也利用了预训练的图像生成扩散模型。Synthetic Data 使用从布局到图像模型生成的合成训练数据训练判别检测模型。Diffusion Feature 使用预训练的扩散模型来提取特征，作为判别对象检测器的骨干。这些基线本质上是判别方法，我们使用 Faster R-CNN [32] 作为它们的后端。对于 Diffusion Feature 方法，除了 LDM-8，我们还使用文本到图像模型 Stable Diffusion [7] (SD v1.5) 作为特征提取模块来进行测试，它是在比 COCO 数据集大得多的网络规模数据集上训练的，按照 [27] 的方法。

我们进一步与广泛认可的纯判别对象检测器 Faster R-CNN [32] 和 DETR [101] 进行比较，以及一种较新的方法 DiffusionDet [54]（DiffDet S @ N_{eval}），该方法在基于判别图像特征的提议框空间中部署扩散过程，其中 S 和 N_{eval} 是本文 [54] 中引入的两个超参数。我们为这些检测器采用了各种骨干网络（ResNet-50 和 ResNet-101 [102]）和优化的架构（特征金字塔网络（FPN） [103] 和膨胀的最后阶段（DC5） [104]）。我们使用 `detectron2` [105] 来实现 Faster R-CNN 和 DiffDet 基线，并使用官方代码实现 DETR 基线。Faster R-CNN 和 DiffDet 模型使用 3 × 计划（∼ 37 个 epoch）进行训练，这是 `detectron2` 中的典型设置。由于基于 Transformer 的架构通常需要更长的训练计划，DETR 模型训练超过 200 个 epoch。这些判别检测器在分辨率为 256×256 的条件下训练，使用随机裁剪和水平翻转增强，并从头开始训练，与布局到图像扩散模型的设置一致。虽然由于架构和训练成本的差异，

TABLE II: 在 COCO 验证集上的目标检测。我们与同样使用图像生成扩散模型的生成-判别混合方法进行了比较。我们还与先进的纯判别方法进行了比较。T 代表每张图像的处理时间，其是在 Nvidia 3090 GPU 上评估的。

Method	Backbone	T	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Generative + discriminative methods:								
Synthetic Data	R50	90ms	4.7	8.7	4.5	0.0	3.1	9.8
Diffusion Feature	LDM-8	47ms	6.9	13.4	5.9	0.0	3.5	15.8
Diffusion Feature	SD	41ms	7.9	15.7	7.3	0.0	1.2	18.9
Generative methods:								
DIVE (ours)	LDM-8	1.5h	7.1	11.0	7.1	0.0	1.2	16.1
Discriminative methods:								
Faster R-CNN	R50	90ms	6.8	12.7	5.7	0.3	5.0	13.5
Faster R-CNN	R101-FPN	29ms	9.6	17.3	9.5	2.2	9.0	18.3
DETR	R50	38ms	9.3	16.9	8.7	1.3	6.4	27.6
DETR	R101-DC5	47ms	14.7	24.9	13.9	2.6	11.6	25.9
DiffDet (1 @ 300)	R101-FPN	37ms	14.6	24.4	14.6	2.9	13.8	26.9
DiffDet (4 @ 500)	R101-FPN	41ms	15.2	25.7	14.8	3.4	14.4	27.8

很难进行真正公平的比较，但我们的比较旨在彻底评估反向扩散模型的优势和劣势。

英寸

b) 结果： Table II 显示了对比结果 [1] 。与使用相同预训练扩散模型的其他生成-判别-混合方法相比，DIVE 优于合成数据，并具有与扩散特征（LDM-8）相当的竞争力。这些方法的相应可视化结果显示在 Figure 5 的左侧。配备了 SD 主干的扩散特征基线比我们的 DIVE 和其他基于 LDM-8 的方法表现更好。这是因为文本到图像的 SD 在比 LDM-8 的 COCO 更大的数据集上进行了预训练，将更丰富的先验知识融入到网络中。

与纯粹的判别检测器相比，DIVE 在与基础判别检测器 Faster R-CNN R50 的比较中达到了竞争力的结果，但落后于使用更深骨干网或现代架构的更强检测器。值得注意的是，据我们所知，我们是首次展示使用冻结的预训练图像生成模型可以成功解决具有挑战性的目标检测任务，并与判别方法取得竞争性表现。

此外，我们通过与 Faster R-CNN R50 进行比较，对 DIVE 进行了更深入的分析。Faster R-CNN R50 在整体性能上表现出色。一个有趣的现象是，尽管在 AP_{50} 上落后于 Faster R-CNN R50，但 DIVE 在 AP_{75} 上表现更佳，并在总体 AP 上略胜一筹。这一现象表明，DIVE 在正确预测真阳性边界框时具有更高的精确度，同时未检测到一些对象。这些未检测到的对象通常是小型和中型对象，这在较差的 AP_S 和 AP_M 值中有所体现。不巧的是，句法数据和扩散特征基线也遭遇了这一问题。我们推测，这可能是因为 LDM 和 SD 在潜在空间中训练，该空间具有一个 8 的降采样因素。在压缩建模空间中，较小的对象被调整为更小的尺寸。

英寸

3) 消融研究： 在这一小节中，我们研究了我们方法中提出组件的影响。除了对先验模型进行消融外，我们还研究了如果不使用离散优化技巧的影响，即使用公式 (4) 作为目标，而不使用来源于公式 (5) 的词汇内离散优化技巧（在这种情况下，v' 是 v ）。我们在 100 张图像的较小子集上进行了消融研究。可以从 **??** 看出，定量上，这两个组件对于整体表现都很重要。Figure 5 中的定性结果则更多地揭示了原因，下面我们进行分析。

[1]判别检测器的重现结果相对较低，是因为我们使用了较小的 256×256 图像尺寸，这是针对生成模型做公平比较时常用的方法。

从 Figure 5 可以看出，如果我们不将先验模型整合到反演过程中，反演后的序列（预测）将包含许多冗余物体和非法框。合成图像说明了原因。这些冗余物体不会影响条件模型的生成过程（例如，前两行），这是一个有趣的现象，类似于文本到图像扩散模型攻击文献中的发现 [106]–[108] 。然而，这些冗余会对检测性能产生负面影响，导致许多误报。只有先验模型告诉优化过程哪种布局类型更常见时，我们才能获得更好的检测结果（DIVE 默认）。

在嵌入空间的连续优化情况下（即，不使用方程 (5) ），优化过程只需要重构（学习内容的）给定图像。这个情况可以类比于 Textual Inversion [88] ，它学习如何将给定图像的新词注入预训练的文本到图像模型中，而不考虑新词的语义。对于我们的情况，如果我们试图解码以查看这些“新词”是什么（通过在词汇表中寻找最近的邻居），如 Figure 5 最右侧所示，有时候，我们可以得到正确的类别（例如，第一行的“person”），但在许多其他情况下，类别难以解释（例如，最后一行的“cow”），即使在先验模型的帮助下，边框也很难完成。这是因为重构目的主导了优化（Figure 5 中的最佳重构性能），而单词的可解释性则未得到关注。

英寸

a) 解耦类别和边界框： 鉴于训练的先验模型可能会过拟合训练数据分布，此消融将序列中的类别和边界框解耦。具体而言，我们分别训练两个先验分布模型，一个用于类别，另一个用于边框，并在逆变期间将它们结合使用。然而，请注意，此策略可能不是对先验布局分布进行建模的自然且适当的方式，因为它分离了类别和边框之间的关系。例如，它不能区分“一个人在自行车上”和“自行车上的一个人”，因为它们具有相同的边框布局。尽管如此，与不使用任何先验模型相比，它仍然可以学习哪些边框是合法的。**??** 的最后一行展示了结果。结果介于默认（未解耦）方法和“无先验模型”方法之间，这与上述假设一致。

B. 用于图像分类的 DIVE

在本小节中，我们将我们的基于优化的方法与之前的基于枚举的方法 Diffusion Classifier [1] 在图像分类任务中的准确性和速度进行比较。

继 [1] 中的设定之后，我们使用预训练的扩散 Transformer（DiT） [65] 作为需要反转的条件扩散模型，它是在 ImageNet-1k [92] 上训练的类别条件模型。与 [1] 中的设定相同，为了公平比较，我们使用 DiT-XL/2 在分辨率 256 × 256 下进行，并在由 2000 张图像（每类 2 张图像）组成的子集上进行评估。鉴于 ImageNet-1k 的标签分布是均匀的，在这种情况下无需先验建模模型。此处 DIVE 的不同时间步数设置为 250，以与扩散分类器保持一致。我们对 embeddings 进行 200 步的优化，批量大小为 25，在单个 A100 GPU 上反转一张图像大约需要 80 秒。

1) 结果： Table III 显示了比较结果。我们与 Diffusion Classifier [1] 、判别式分类器 [102], [109] [110]–[112] 和生成-判别式混合方法 Synthetic Data 和 Diffusion Feature 进行比较。Synthetic Data 使用 DiT-XL/2 生成的训练数据集训练一个标准的 ResNet-50。Diffusion Feature 根据 [1] 的基线设计，使用预训练的 DiT-XL/2 网络作为特征提取器，并在这些特征基础上使用 ImageNet 真实数据集训练一个修改版的 ResNet-18。从 Table III 可以看出，DIVE 几乎达到了与 Diffusion Classifier 相同的准确度。该结果验证了所提出的基于优化的方法在纯生成的图像分类方面的有效性，代表了枚举型方法的理想替代。

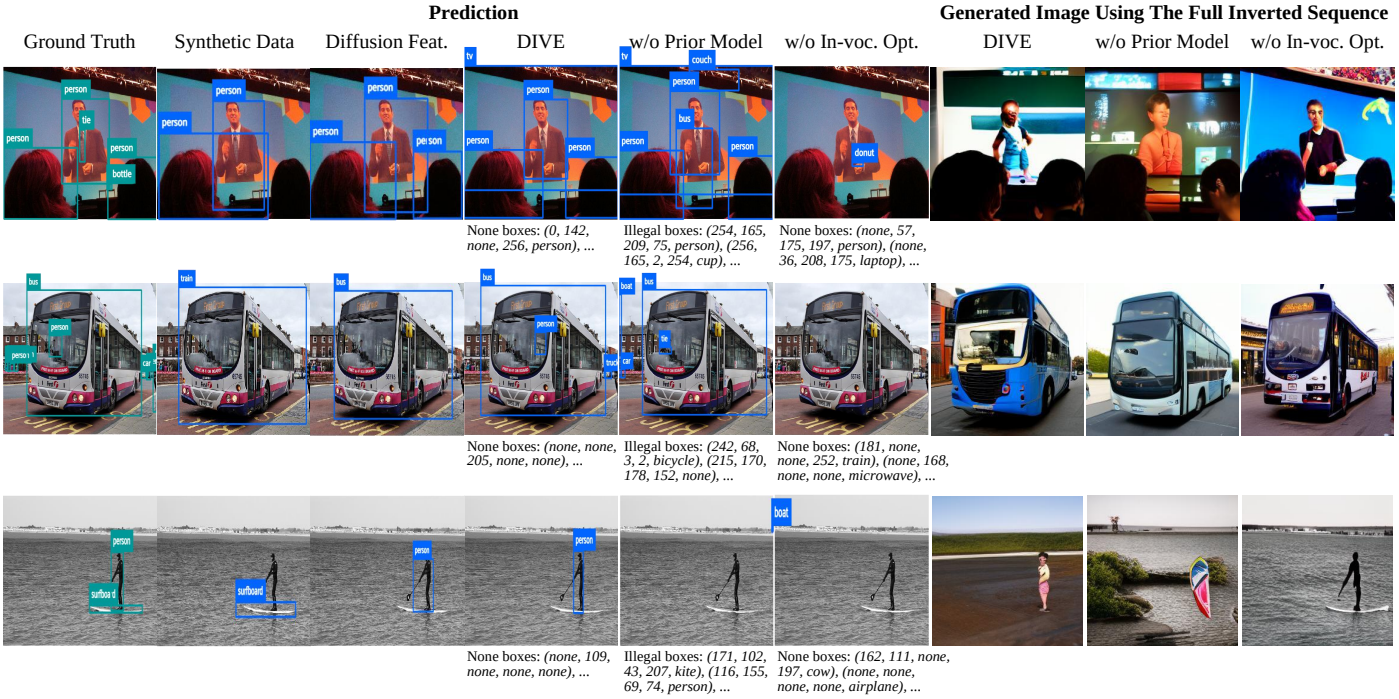


Fig. 5: 物体检测结果的可视化。除了将 DIVE 与使用相同预训练扩散模型的其他生成基线进行比较外，我们还展示了先验模型和词汇内离散优化方法的影响。对于这些消融实验，我们在底部展示了一些额外的丢弃对象（不含值的和非法的框）在反向序列中的可视化，以更清楚地展示不同方法的行为。在右侧，我们展示了通过将完整反向序列输入到预训练的布局到图像模型所生成的图像。放大以获得更好的可视化效果。

TABLE III: ImageNet-1k 上的图像分类。速度在 A100 GPU 上进行测试。Diffusion Classifier 的准确率来自 [1]，其速度使用官方代码进行测试。

Method	Time per image	Accuracy
Generative + discriminative methods:		
Synthetic Data	9 ms	67.1
Diffusion Feature	31 ms	69.2
Generative methods:		
Diffusion Classifier	$\sim$ 1100 s	77.3
DIVE (ours)	80 s	77.2
Discriminative methods:		
AlexNet	8 ms	57.5
ResNet-18	8 ms	70.6
ResNet-50	9 ms	77.6
ResNeXt101	15 ms	79.7
ViT-L/32	12 ms	78.0
ViT-B/16	9 ms	81.5
Swin-B	21 ms	83.7

值得注意的是，我们的基于优化的方法在保持准确性的同时，实现了对枚举方法的 $\sim 14\times$ 倍加速。这是因为枚举方法需要尝试测试数据集的所有可能预测，这会导致显著的计算成本。与将测试图像直接映射到预测标签的判别模型相比，由于探索可能标签的成本，生成方法在效率上仍然落后。进一步加速这些生成方法的潜在方向包括减少探索步骤或预训练模型的参数。

为了确定通过使用提前停止策略来停止 DIVE 优化过程所能实现的最大加速比，我们在此检查测试图像的收敛步骤，就像在之前的检测部分 IV-A1d 中所做的那样。我们发现，这里的测试集的平均收敛步骤大约是 106（最大值是 200），这表明减少探索步骤总体上最多可以实现 $\sim 2\times$ 的加速。

我们在这里再次对词汇内优化方法进行消融，以验证其有效性。没有这个优化组件，我们的准确率为 60.3，表明其在基于优化的方法中的重要性。

读者可能注意到我们在前面的检测和分类实验中使用了不同的 GPU。这只是为了更有效地调度我们有限的资源而没有其他原因。具体来说，我们在进行正式实验之前测试了使用不同 GPU 进行不同任务的速度。正如 Table IV 总结的那样，我们观察到从 3090 切换到 A100 在分类任务上实现了比检测任务更大的加速比，导致了之前的选择。在这里，我们进行了重复实验以评估受 GPU 切换影响的任务性能。

Table IV 展示了结果，其中检测实验是在消融研究的实验规模下进行的（参见 **??** ）。可以看到，检测 AP 和分类准确率略有波动。

C. 用于条件扩散模型评估的 DIVE

根据 [69], [70]，我们使用我们提出的 DIVE 从判别能力的角度评估不同的条件扩散模型。虽然使用 DIVE 作为生成模型的一般评估指标可能存在争议 - 不适用于像 GAN 这样的非似然基础模型 - 但它具有不依赖于外部模型的优点，例如计算 [114] 的常用的 ImageNet 预训练的 InceptionV3 [113]。然而，本小节的主要目的是观察 DIVE 的判别能力是否与评估生成模型的其他广泛使用的指标一致。

英寸我们评估了前面提到的两种条件扩散模型。对于类别条件模型，我们比较了 DiT-XL/2 [65] 与 LDM-4 [7]，前者基于前一个小节中使用的 transformer 架构 [60]，而后者基于卷积 UNet 架构 [99]。对于布局条件模型，我们比较了在 Section IV-A 中使用的模型 (LDM-8) 与我们自己训练的另一个较小的 LDM (LDM-8-S)。对于其他评价指标的比较，我们使用了 Fréchet Inception 距离（FID）

TABLE IV: GPU 对不同任务的影响。

Task	GPU	Time per image	AP / Acc
Detection	3090	90 min	10.3
	A100	50 min	10.6
Classification	3090	240 s	76.6
	A100	80 s	77.2

TABLE V: 比较具有不同评估指标的条件扩散模型。分别使用准确率和 AP 作为 DIVE 指标。表 (a) 中的其他数字来自 [65] 。Prec. 代表精度。

(a) Comparing class-conditional models

Model	# Params	FID ↓	IS ↑	Prec. ↑	Recall ↑	DIVE ↑
LDM-4	400M	3.60	247.67	0.87	0.48	62.7
DiT-XL/2	675M	2.27	278.24	0.83	0.57	77.2

(b) Comparing layout-conditional models

Model	# Params	FID ↓	Prec. ↑	Recall ↑	DIVE ↑
LDM-8-S	65M	24.90	0.45	0.62	4.5
LDM-8	363M	18.36	0.53	0.63	7.1

[114] 、Inception 分数 (IS) [115] 、以及精度/召回 [116] 。

1) 结果： 结果显示在 Table V 中。可以观察到，我们的判别指标结果与大多数其他指标一致，例如 FID，它评估真实图像与生成图像之间的分布相似性。在 Table V (a) 中，DIVE 与 Precision 之间的差异是个例外。这表明一个具有更高生成多样性 (高 Recall) 而非更好图像真实性 (高 Precision) 的模型可能具有更强的判别能力 (高 DIVE)。

V. 结论

在本文中，我们提出了一种基于优化的方法，用于反转条件扩散模型，以在贝叶斯规则框架下生成判别性标签。我们通过实验证明，所提出的基于优化的方法在图像分类任务中比以前的枚举法快得多，并使得仅使用预训练生成模型来完成更复杂的目标检测任务成为可能。我们希望本文的结论能够帮助识别预训练图像生成模型的内部判别能力。未来的研究可以进一步加速所提出的基于优化的方法，或者将其扩展到更复杂的密集任务，例如语义分割。然而，需要注意的是，简单地将所提出的方法扩展到密集任务将会创建一个较大的优化空间，并且单个条件像素对最终目标的贡献相对较小，这可能会带来问题，需要解决。添加一些正则化项可能会提供一个可行的解决方案。

VI.

致谢 本工作由中国国家自然科学基金 (NSFC) 资助：62376259，62306301，以及国家博士后创新人才支持计划 (编号 BX20220310)。

References

[1] Alexander C Li, Mihir Prabhudesai, Shivam Duggal, Ellis Brown, and Deepak Pathak. Your diffusion model is secretly a zero-shot classifier. In *IEEE ICCV*, pages 2206–2217, 2023.
[2] Kevin Clark and Priyank Jaini. Text-to-image diffusion models are zero-shot classifiers. *NeurIPS*, 36, 2023.
[3] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*, pages 2256–2265, 2015.
[4] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 33:6840–6851, 2020.
[5] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *ICLR*, 2021.
[6] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat GANs on image synthesis. *NeurIPS*, 34:8780–8794, 2021.
[7] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE CVPR*, pages 10684–10695, 2022.
[8] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. *NeurIPS*, 35:36479–36494, 2022.
[9] Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. *NeurIPS*, 30, 2017.
[10] Ali Razavi, Aaron van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with VQ-VAE-2. *NeurIPS*, 32, 2019.
[11] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *IEEE CVPR*, pages 12873–12883, 2021.
[12] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *ICML*, pages 8821–8831, 2021.
[13] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *NeurIPS*, 27, 2014.
[14] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *ICLR*, 2019.
[15] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *IEEE CVPR*, pages 4401–4410, 2019.
[16] Ruifei He, Shuyang Sun, Xin Yu, Chuhui Xue, Wenqing Zhang, Philip Torr, Song Bai, and Xiaojuan Qi. Is synthetic data from generative models ready for image recognition? In *ICLR*, 2023.
[17] Mert Bülent Sarıyıldız, Karteek Alahari, Diane Larlus, and Yannis Kalantidis. Fake it till you make it: Learning transferable representations from synthetic imagenet clones. In *IEEE CVPR*, pages 8011–8021, 2023.
[18] Shekoofeh Azizi, Simon Kornblith, Chitwan Saharia, Mohammad Norouzi, and David J. Fleet. Synthetic data from diffusion models improves imagenet classification. *TMLR*, 2023.
[19] Weijia Wu, Yuzhong Zhao, Mike Zheng Shou, Hong Zhou, and Chunhua Shen. DiffuMask: Synthesizing images with pixel-level annotations for semantic segmentation using diffusion models. In *IEEE ICCV*, pages 1206–1217, 2023.
[20] Quang Nguyen, Truong Vu, Anh Tran, and Khoi Nguyen. Dataset diffusion: Diffusion-based synthetic data generation for pixel-level semantic segmentation. *NeurIPS*, 36, 2023.
[21] Kai Chen, Enze Xie, Zhe Chen, Yibo Wang, Lanqing Hong, Zhenguo Li, and Dit-Yan Yeung. GeoDiffusion: Text-prompted geometric control for object detection data generation. In *ICLR*, 2024.
[22] Haoyang Fang, Boran Han, Shuai Zhang, Su Zhou, Cuixiong Hu, and Wen-Ming Ye. Data augmentation for object detection via controllable diffusion models. In *IEEE WACV*, pages 1257–1266, 2024.
[23] Dmitry Baranchuk, Andrey Voynov, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. Label-efficient semantic segmentation with diffusion models. In *ICLR*, 2022.
[24] Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. Open-vocabulary panoptic segmentation with text-to-image diffusion models. In *IEEE CVPR*, pages 2955–2966, 2023.
[25] Xingyi Yang and Xinchao Wang. Diffusion model as representation learner. In *IEEE ICCV*, pages 18938–18949, 2023.
[26] Daiqing Li, Huan Ling, Amlan Kar, David Acuna, Seung Wook Kim, Karsten Kreis, Antonio Torralba, and Sanja Fidler.

DreamTeacher: Pretraining image backbones with deep generative models. In *IEEE ICCV*, pages 16698–16708, 2023.

[27] Wenliang Zhao, Yongming Rao, Zuyan Liu, Benlin Liu, Jie Zhou, and Jiwen Lu. Unleashing text-to-image diffusion models for visual perception. In *IEEE ICCV*, pages 5729–5739, 2023.

[28] Neehar Kondapaneni, Markus Marks, Manuel Knott, Rogerio Guimaraes, and Pietro Perona. Text-image alignment for diffusion-based perception. In *IEEE CVPR*, pages 13883–13893, 2024.

[29] Guangkai Xu, Yongtao Ge, Mingyu Liu, Chengxiang Fan, Kangyang Xie, Zhiyue Zhao, Hao Chen, and Chunhua Shen. Diffusion models trained with large data are transferable visual models. *arXiv preprint arXiv: 2403.06090*, 2024.

[30] Andrew Ng and Michael Jordan. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. *NeurIPS*, 14, 2001.

[31] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, pages 740–755, 2014.

[32] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. *NeurIPS*, 28, 2015.

[33] Yuqi Jiang, Qiankun Liu, Dongdong Chen, Lu Yuan, and Ying Fu. AnimeDiff: Customized image generation of anime characters using diffusion model. *IEEE TMM*, pages 1–13, 2024.

[34] Yifei Xu, Xiaolong Xu, Honghao Gao, and Fu Xiao. SGDM: An adaptive style-guided diffusion model for personalized text to image generation. *IEEE TMM*, 26:9804–9813, 2024.

[35] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. SDEdit: Guided image synthesis and editing with stochastic differential equations. In *ICLR*, 2022.

[36] Shidong Cao, Wenhao Chai, Shengyu Hao, Yanting Zhang, Hangyue Chen, and Gaoang Wang. DiffFashion: Reference-based fashion design with structure-aware transfer by diffusion models. *IEEE TMM*, 26:3962–3975, 2024.

[37] Cong Zhang, Wenxia Yang, Xin Li, and Huan Han. MMGInpainting: Multi-modality guided image inpainting based on diffusion models. *IEEE TMM*, 26:8811–8823, 2024.

[38] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *NeurIPS*, 35:8633–8646, 2022.

[39] Daizong Liu, Jiahao Zhu, Xiang Fang, Zeyu Xiong, Huan Wang, Renfu Li, and Pan Zhou. Conditional video diffusion network for fine-grained temporal sentence grounding. *IEEE TMM*, 26:5461–5476, 2024.

[40] Minglu Zhao, Wenmin Wang, Tongbao Chen, Rui Zhang, and Ruochen Li. TA2V: Text-audio guided video generation. *IEEE TMM*, 26:7250–7264, 2024.

[41] Minheng Ni, Zitong Huang, Kailai Feng, and Wangmeng Zuo. ImaginaryNet: Learning object detectors without real images and annotations. In *ICLR*, 2023.

[42] Zebin You, Yong Zhong, Fan Bao, Jiacheng Sun, Chongxuan Li, and Jun Zhu. Diffusion models and semi-supervised learners benefit mutually with few labels. *NeurIPS*, 36, 2023.

[43] Lisa Dunlap, Alyssa Umino, Han Zhang, Jiezhi Yang, Joseph E Gonzalez, and Trevor Darrell. Diversify your vision datasets with automatic diffusion-based augmentation. *NeurIPS*, 36, 2023.

[44] Lihe Yang, Xiaogang Xu, Bingyi Kang, Yinghuan Shi, and Hengshuang Zhao. FreeMask: Synthetic images with dense annotations make stronger segmentation models. *NeurIPS*, 36, 2023.

[45] Laurynas Karazija, Iro Laina, Andrea Vedaldi, and Christian Rupprecht. Diffusion models for open-vocabulary segmentation. In *ECCV*, 2024.

[46] Jiahao Xie, Wei Li, Xiangtai Li, Ziwei Liu, Yew Soon Ong, and Chen Change Loy. Mosaicfusion: Diffusion models as data augmenters for large vocabulary instance segmentation. *arXiv preprint arXiv:2309.13042*, 2023.

[47] Saksham Suri, Fanyi Xiao, Animesh Sinha, Sean Culatana, Raghuraman Krishnamoorthi, Chenchen Zhu, and Abhinav Shrivastava. Gen2Det: Generate to detect. In *IEEE CVPRW*, 2024.

[48] Ziyi Li, Qinye Zhou, Xiaoyun Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Open-vocabulary object segmentation with diffusion models. In *IEEE ICCV*, pages 7667–7676, 2023.

[49] Soumik Mukhopadhyay, Matthew Gwilliam, Yosuke Yamaguchi, Vatsal Agarwal, Namitha Padmanabhan, Archana Swaminathan, Tianyi Zhou, and Abhinav Shrivastava. Do text-free diffusion models learn discriminative visual representations? In *ECCV*, 2024.

[50] Shiyin Dong, Mingrui Zhu, Kun Cheng, Nannan Wang, and Xinbo Gao. Bridging generative and discriminative models for unified visual perception with diffusion priors. In *IJCAI*, pages 740–748, 2024.

[51] Suraj Patni, Aradhye Agarwal, and Chetan Arora. ECoDepth: Effective conditioning of diffusion models for monocular depth estimation. In *IEEE CVPR*, pages 28285–28295, 2024.

[52] Tomer Amit, Tal Shaharbany, Eliya Nachmani, and Lior Wolf. SegDiff: Image segmentation with diffusion probabilistic models. *arXiv preprint arXiv:2112.00390*, 2021.

[53] Ting Chen, Lala Li, Saurabh Saxena, Geoffrey Hinton, and David J Fleet. A generalist framework for panoptic segmentation of images and videos. In *IEEE ICCV*, pages 909–919, 2023.

[54] Shoufa Chen, Peize Sun, Yibing Song, and Ping Luo. DiffusionDet: Diffusion model for object detection. In *IEEE ICCV*, pages 19830–19843, 2023.

[55] Yiqun Duan, Xianda Guo, and Zheng Zhu. DiffusionDepth: Diffusion denoising approach for monocular depth estimation. *arXiv preprint arXiv:2303.05021*, 2023.

[56] Yuanfeng Ji, Zhe Chen, Enze Xie, Lanqing Hong, Xihui Liu, Zhaoqiang Liu, Tong Lu, Zhenguo Li, and Ping Luo. DDP: Diffusion model for dense visual prediction. In *IEEE ICCV*, pages 21741–21752, 2023.

[57] Hsin-Ying Lee, Hung-Yu Tseng, Hsin-Ying Lee, and Ming-Hsuan Yang. Exploiting diffusion prior for generalizable dense prediction. In *IEEE CVPR*, pages 7861–7871, 2024.

[58] Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *IEEE CVPR*, pages 9492–9502, 2024.

[59] Xianyun Wang, Linhong Wang, Zhenchen Yang, Jiacong Zhou, Yuchen Zheng, Feng Chen, Richang Hong, Jun Yu, and Fan Yang. DSIS-DPR: Structured instance segmentation and diffusion prior refinement for dental anatomy learning. *IEEE TMM*, 2024.

[60] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 30, 2017.

[61] Junjiao Tian, Lavisha Aggarwal, Andrea Colaco, Zsolt Kira, and Mar Gonzalez-Franco. Diffuse attend and segment: Unsupervised zero-shot segmentation using stable diffusion. In *IEEE CVPR*, pages 3554–3563, 2024.

[62] Koutilya Pnvr, Bharat Singh, Pallabi Ghosh, Behjat Siddiquie, and David Jacobs. LD-ZNet: A latent diffusion approach for text-based image segmentation. In *IEEE ICCV*, pages 4157–4168, 2023.

[63] Jinglong Wang, Xiawei Li, Jing Zhang, Qingyuan Xu, Qin Zhou, Qian Yu, Lu Sheng, and Dong Xu. Diffusion model is secretly a training-free open vocabulary semantic segmenter. *arXiv preprint arXiv:2309.02773*, 2023.

[64] Changming Xiao, Qi Yang, Feng Zhou, and Changshui Zhang. From text to mask: Localizing entities using the attention of text-to-image diffusion models. *Neurocomputing*, 610:128437, 2024.

[65] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *IEEE ICCV*, pages 4195–4205, 2023.

[66] Roland S Zimmermann, Lukas Schott, Yang Song, Benjamin Adric Dunn, and David A Klindt. Score-based generative classifiers. In *NeurIPS Workshop on Deep Generative Models and Downstream Applications*, 2021.

[67] Huanran Chen, Yinpeng Dong, Zhengyi Wang, Xiao Yang, Chengqi Duan, Hang Su, and Jun Zhu. Robust classification via a single diffusion model. In *ICML*, 2024.

[68] Huanran Chen, Yinpeng Dong, Shitong Shao, Zhongkai Hao, Xiao Yang, Hang Su, and Jun Zhu. Your diffusion model is secretly a certifiably robust classifier. *arXiv preprint arXiv: 2402.02316*, 2024.

[69] Benno Krojer, Elinor Poole-Dayan, Vikram Voleti, Christopher Pal, and Siva Reddy. Are diffusion models vision-and-language reasoners? *NeurIPS*, 36, 2023.

[70] Sai Saketh Rambhatla and Ishan Misra. SelfEval: Leveraging the discriminative nature of generative models for evaluation. *arXiv preprint arXiv: 2311.10708*, 2023.

[71] Jun-Yan Zhu, Philipp Krähenbühl, Eli Shechtman, and Alexei A Efros. Generative visual manipulation on the natural image manifold. In *ECCV*, pages 597–613, 2016.

[72] Weihao Xia, Yulun Zhang, Yujiu Yang, Jing-Hao Xue, Bolei Zhou, and Ming-Hsuan Yang. GAN inversion: A survey. *IEEE TPAMI*, 45(3):3121–3138, 2022.

[73] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *IEEE CVPR*, pages 1921–1930, 2023.

[74] Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. MasaCtrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In *IEEE ICCV*, pages 22560–22570, 2023.

[75] Antonia Creswell and Anil Anthony Bharath. Inverting the generator of a generative adversarial network. *IEEE TNNLS*, 30(7):1967–1974, 2018.

[76] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2StyleGAN: How to embed images into the StyleGAN latent space? In *IEEE ICCV*, pages 4432–4441, 2019.

[77] Peihao Zhu, Rameen Abdal, Yipeng Qin, John Femiani, and Peter Wonka. Improved StyleGAN embedding: Where are the good latents? *arXiv preprint arXiv:2012.09036*, 2020.

[78] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. Encoding in style: A StyleGAN encoder for image-to-image translation. In *IEEE CVPR*, pages 2287–2296, 2021.

[79] Omer Tov, Yuval Alaluf, Yotam Nitzan, Or Patashnik, and Daniel Cohen-Or. Designing an encoder for StyleGAN image manipulation. *ACM TOG*, 40(4):1–14, 2021.

[80] Yuval Alaluf, Omer Tov, Ron Mokady, Rinon Gal, and Amit Bermano. HyperStyle: StyleGAN inversion with hypernetworks for real image editing. In *IEEE CVPR*, pages 18511–18521, 2022.

[81] Jialu Huang, Jing Liao, and Sam Kwong. Unsupervised image-to-image translation via pre-trained StyleGAN2 network. *IEEE TMM*, 24:1435–1448, 2022.

[82] Jiapeng Zhu, Yujun Shen, Deli Zhao, and Bolei Zhou. In-domain GAN inversion for real image editing. In *ECCV*, pages 592–608, 2020.

[83] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2021.

[84] Bram Wallace, Akash Gokul, and Nikhil Naik. EDICT: Exact diffusion inversion via coupled transformations. In *IEEE CVPR*, pages 22532–22541, 2023.

[85] Guoqiang Zhang and W Bastiaan Kleijn. Exact diffusion inversion via bi-directional integration approximation. *arXiv preprint arXiv:2307.10829*, 2023.

[86] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *IEEE CVPR*, pages 6038–6047, 2023.

[87] Daiki Miyake, Akihiro Iohara, Yu Saito, and Toshiyuki Tanaka. Negative-prompt inversion: Fast image inversion for editing with text-guided diffusion models. *arXiv preprint arXiv:2305.16807*, 2023.

[88] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *ICLR*, 2023.

[89] Chen Wei, Chenxi Liu, Siyuan Qiao, Zhishuai Zhang, Alan Yuille, and Jiahui Yu. De-diffusion makes text a strong cross-modal interface. In *IEEE CVPR*, pages 13492–13503, 2024.

[90] Shweta Mahajan, Tanzila Rahman, Kwang Moo Yi, and Leonid Sigal. Prompting hard or hardly prompting: Prompt inversion for text-to-image diffusion models. In *IEEE CVPR*, pages 6808–6817, 2024.

[91] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

[92] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *IJCV*, 115:211–252, 2015.

[93] Wan-Cyuan Fan, Yen-Chun Chen, DongDong Chen, Yu Cheng, Lu Yuan, and Yu-Chiang Frank Wang. Frido: Feature pyramid diffusion for complex scene image synthesis. In *AAAI*, volume 37, pages 579–587, 2023.

[94] Jiaxin Cheng, Xiao Liang, Xingjian Shi, Tong He, Tianjun Xiao, and Mu Li. LayoutDiffuse: Adapting foundational diffusion models for layout-to-image generation. *arXiv preprint arXiv: 2302.08908*, 2023.

[95] Guangcong Zheng, Xianpan Zhou, Xuewei Li, Zhongang Qi, Ying Shan, and Xi Li. LayoutDiffusion: Controllable diffusion model for layout-to-image generation. In *IEEE CVPR*, pages 22490–22499, 2023.

[96] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. COCO-Stuff: Thing and stuff classes in context. In *IEEE CVPR*, pages 1209–1218, 2018.

[97] Zhengyuan Yang, Jianfeng Wang, Zhe Gan, Linjie Li, Kevin Lin, Chenfei Wu, Nan Duan, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. ReCo: Region-controlled text-to-image generation. In *IEEE CVPR*, pages 14246–14255, 2023.

[98] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. GLIGEN: Open-set grounded text-to-image generation. In *IEEE CVPR*, pages 22511–22521, 2023.

[99] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pages 234–241, 2015.

[100] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019.

[101] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, pages 213–229, 2020.

[102] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE CVPR*, pages 770–778, 2016.

[103] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *IEEE CVPR*, pages 2117–2125, 2017.

[104] Yi Li, Haozhi Qi, Jifeng Dai, Xiangyang Ji, and Yichen Wei. Fully convolutional instance-aware semantic segmentation. In *IEEE CVPR*, pages 2359–2367, 2017.

[105] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019.

[106] Raphaël Millière. Adversarial attacks on image generation with made-up words. *arXiv preprint arXiv:2208.04135*, 2022.

[107] Haomin Zhuang, Yihua Zhang, and Sijia Liu. A pilot study of query-free adversarial attack against stable diffusion. In *IEEE CVPRW*, pages 2385–2392, 2023.

[108] Hongcheng Gao, Hao Zhang, Yinpeng Dong, and Zhijie Deng. Evaluating the robustness of text-to-image diffusion models against real-world attacks. *arXiv preprint arXiv:2306.13103*, 2023.

[109] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.

[110] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet classification with deep convolutional neural networks. *NeurIPS*, 25, 2012.

[111] Saining Xie, Ross B. Girshick, Piotr Dollár, Z. Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *IEEE CVPR*, 2016.

[112] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and B. Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *IEEE ICCV*, 2021.

[113] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *IEEE CVPR*, pages 2818–2826, 2016.

[114] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. *NeurIPS*, 30, 2017.
[115] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training GANs. *NeurIPS*, 29, 2016.
[116] Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. *NeurIPS*, 32, 2019.