DRCT: Diffusion Reconstruction Contrastive Training towards Universal Detection of Diffusion Generated Images

摘要

扩散模型在视觉内容生成方面取得了重大进展，但也对生成的图像检测提出了越来越高的要求。现有的检测方法已经取得了相当大的进步，但是当检测由未知的扩散模型生成的图像时，它们通常会遭受精度的显著下降。在本文中，我们试图从困难样本分类的角度解决生成图像检测器的可推广性。基本思想是，如果分类器可以区分与真实图像非常相似的生成图像，那么它也可以有效地检测不太相似的样本，甚至可能是由不同扩散模型产生的样本。基于这一思想，我们提出了扩散重建对比学习 (DRCT)，这是一个通用框架，以增强现有检测器的泛化性。DRCT 通过高质量的扩散重建生成困难样本，并采用对比训练来指导扩散伪影的学习。此外，我们还建立了一个百万规模的数据集 DRCT-2M，包括 16 种类型的扩散模型，用于评估检测方法的普遍性。广泛的实验结果表明，用 DRCT 增强的检测器在交叉集测试中实现了超过 10% 的准确性改进。

关键词：扩散模型；生成图像检测；扩散重建；对比学习

1 引言

近年来，基于去噪扩散模型的图像生成技术发展迅速，新的生成模型不断出现。这些技术为数字创作、商业广告、新闻发布和社交娱乐等应用提供了高效的内容编辑和生成工具。但也存在恶意误用的风险，如编造假新闻、误导舆论、干扰政治选举、侵犯版权等。因此，迫切需要开发用于检测生成图像的技术，以维护可信的网络空间环境。

Hugging Face 和 CIVITAI 等 AI 模型共享平台提供了一系列复杂的基于扩散的图像生成模型及其变体，方便用户通过简单的文本交互生成多样化的图像内容。图像生成模型的广泛多样性和可用性对检测方法的通用性提出了相当大的挑战。它要求生成的图像检测器不仅能够识别由已知的生成模型产生的图像，而且能够识别尚未参与检测器训练的新开发的模型产生的图像。

在这项工作中，论文从困难样本分类的角度解决了生成图像检测的普遍性。其核心思想是，如果分类器能够区分难以检测的生成图像和真实图像，那么它也很可能在识别更容易的样本方面具有良好的泛化能力。这启发了论文作者，让分类器专注于从困难样本中学习，以达到更好的泛化能力。在生成图像检测的任务中，如果能够构建大量的困难样本来训练检测器，结合合理的引导，检测器的泛化性有望得到一定程度的提高。

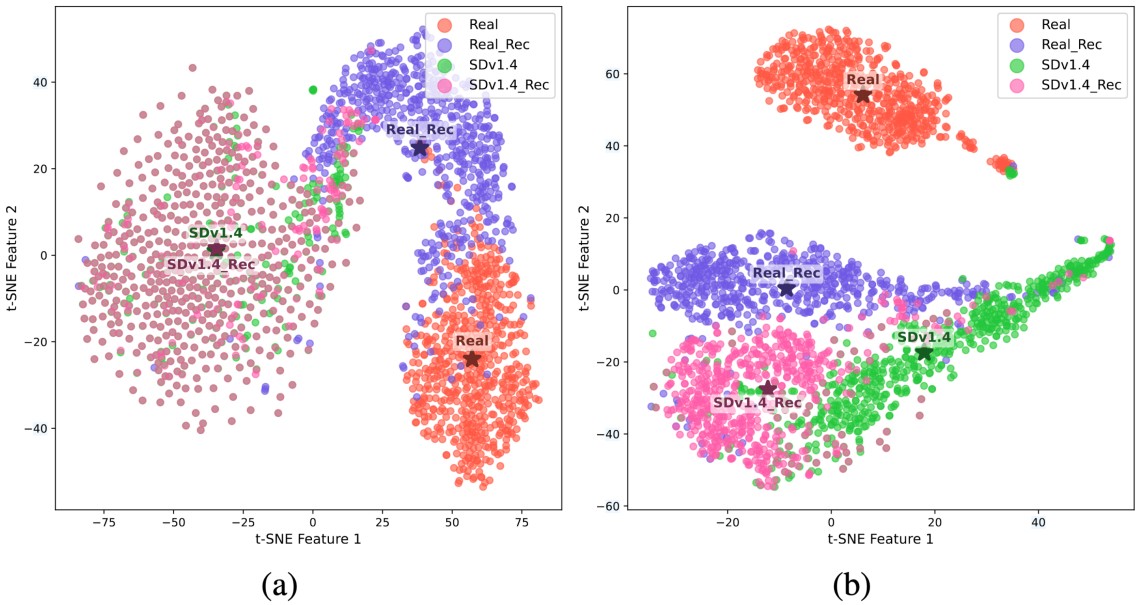

图 1. 具有聚类中心的 t-SNE 嵌入的可视化. (a) 是在使用 DRCT 之前，(b) 是在使用 DRCT 之后

为了进一步证明困难样本如何增强泛化性,原文从预训练检测器 Convnext-base [14](Conv-B）中分类层之前的最后一个特征层中提取真实图像、真实重建图像、使用 Stable Diffusion 1.4 版生成的图像以及生成然后重建的图像的特征。然后使用 t-SNE 将这四组特征投影到 2D 空间中，如图1（a）所示。可以看出，真实图像的特征点分布在真实重建图像的特征点附近，表明它们在区分两者方面的相对困难，从而真实重建图像可以作为困难样本。通过用真实重建样本和生成然后重建的样本微调预训练的检测器，检测器在区分真实图像和困难样本 (真实重建样本) 方面学习更好的辨别能力。因此，检测器可能能够识别生成的样本及其重建的对应物，如图1（b）所示。

原论文的主要贡献有以下四个部分：

1. 原文作者发现，在真实图像上的扩散重建可以很好地保留图像的视觉内容，同时在得到的图像上留下扩散模型的内在指纹。这些重建图像可以作为检测器学习真实图像和生成图像之间的细微差异的信息困难样本，为增强检测器的泛化能力提供了有效的方法。

2. 基于上述观察和对比学习,原文提出了一种新的训练框架,称为扩散重建对比训练(DRCT)。所提出的框架可以显著提高扩散生成图像检测器的检测精度和泛化能力。

3. 原文创建了 DRCT-2M 基准数据集，其中包括 200 万张大小从 256 到 1024 的高质量生成图像，涵盖了 16 种典型的 stable diffusion 模型。DRCT-2M 数据集还包括名为 DRCT-2M-Wild 的 136k 生成图像，这些图像是从 8 个真实世界的生成平台手动收集的。DRCT-2M 数据集为生成图像检测任务中的性能评估和比较提供了全面的基准。

4. 原文进行了广泛的实验，以验证和评估提出的 DRCT 框架在各种设置以及现实世界场景下的有效性、鲁棒性和可推广性。结果表明，配备 DRCT 的探测器在检测精度方面取得了显著的提高。

2 相关工作

2.1 基于扩散模型的图像生成

生成对抗网络（GANs）[6] [20] 和变分自动编码器（VAE）[12] [29] 一直是图像生成领域的先行者。然而，它们在控制生成的图像内容方面的局限性为新的范例铺平了道路。由 Ho 等人 [10] 介绍去噪扩散概率模型（DDPMs）在生成可与 GANs 生成的图像相媲美的高质量图像方面显示出前景，从而标志着基于扩散模型的图像生成技术发展的一个重要里程碑。随后的工作重点是增强结构 [3] [24] 和采样效率 [18] [15] 的扩散模型。Rombach 等人提出的潜在扩散模型（LDM）[23] 是流行的开源 Stable Diffusion（SD）技术的基础，一直是进一步研究的催化剂。SD 的显著扩展包括用于改进图像生成控制的 ControlNet [34]、用于高分辨率图像的 SDXL [21] 以及用于加速采样的 LCM-RoLA [16] 和 SD-turbo [25]。

2.2 生成图像检测

在过去的几年里，生成图像的检测主要集中在基于 GAN 的图像上 [11] [5] [28]。Wang 等人 [30] 证明了在 JPEG 压缩和模糊的 ProGAN [11] 图像上训练的简单 CNN 分类器可以推广到其他基于 GAN 的生成图像。然而，Corvi 等人 [2] 发现，仅在基于 GAN 的图像上训练的分类器在推广到基于扩散的生成图像时面临困难。对于基于扩散的生成图像的检测，Sha 等人 [26] 利用 CLIP [22] 作为主干网络的多模态融合技术，并发现使用 BLIP 生成的 [13] 字幕作为文本模型的输入具有鲁棒性。Ojha 等人 [19] 采用大型预训练 CLIP 模型作为具有最近邻分类器的特征提取器，实现了有希望的泛化。Wu 等人 [32] 为他们的检测器采用了基于 CLIP 的文本-图像对比学习方法。Wang 等人 [31] 远离基于 CLIP 的方法提出了扩散重建误差（DIRE）方法来检测基于扩散的生成图像，利用了扩散模型无法准确重建真实图像的洞察力。同样，Ma 等人 [17] 开发了基于扩散的生成图像检测的逐步误差（SeDID）方法，利用中间阶跃噪声特征进行分类。[33] 引入了一种通过交叉注意力增强的双流检测网络，集成了手工制作的 SRM 特征和 RGB 深度特征。Zhong 等人 [35] 专注于富纹理区域和差纹理区域之间的像素间相关性的对比，从而产生了结合 SRM 滤波器和 CNN 分类器的双流模型。对于更精细的检测，Guarnera 等人 [7] 和 Guo 等人 [8] 提出了能够区分不同基于 GAN 和基于扩散的生成图像，以及局部伪造。Bird 等人 [1] 利用简单的 CNN 分类器进行二元分类，而 Epstein 等人 [4] 采用在线学习方法，对新发布的技术生成的图像进行增量训练。他们还强调了数据增强技术在提高检测 Stable Diffusion 修复图像中的像素级分割性能方面的功效。

3 本文方法

如前所述，扩散重建对比训练（DRCT）的思想是通过在适当的指导下用困难样本训练检测器来增强检测器的泛化性。本节的结构如下。我们首先介绍 DRCT 的总体框架，然后描述其技术细节，包括扩散重建和对比训练。

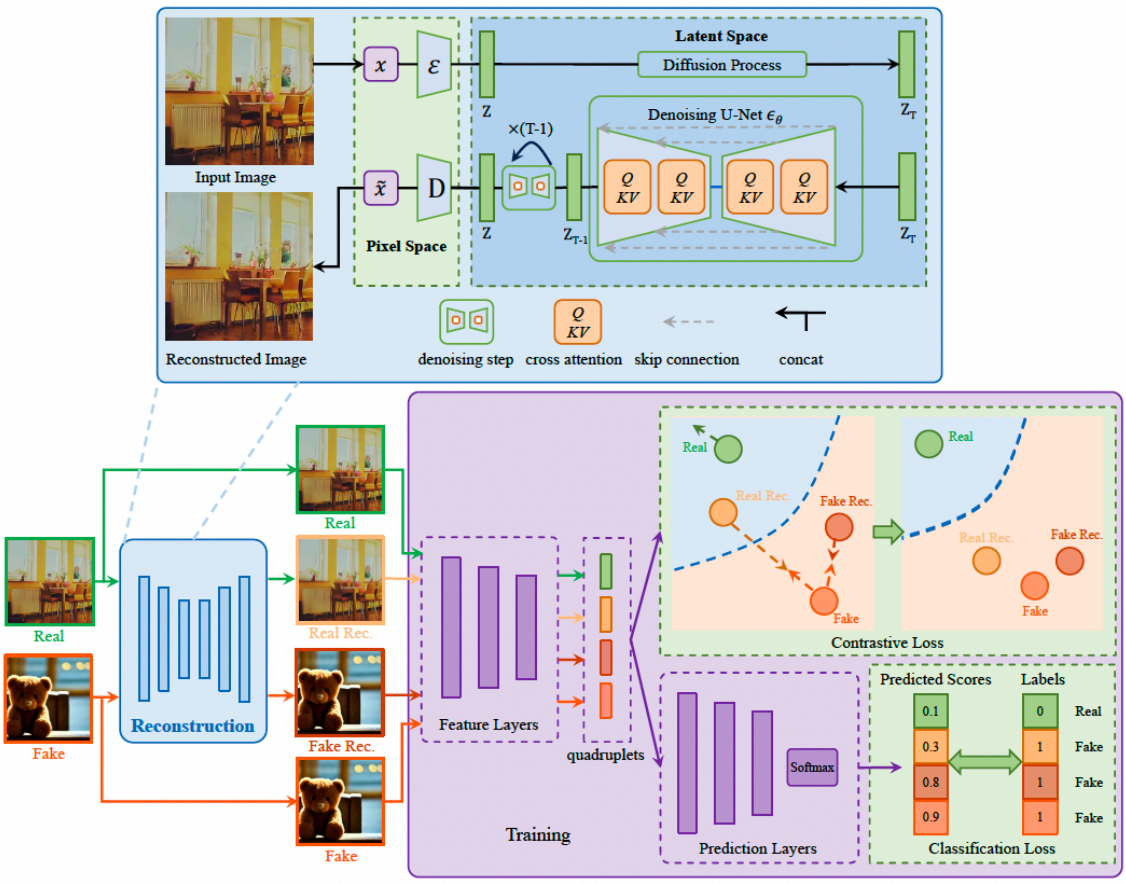

图 2. 扩散重建对比训练 (DRCT) 框架的工作流程。

3.1 DRCT 框架

图2展示了 DRCT 的框架。DRCT 框架由两个主要阶段组成：重建阶段和训练阶段。在重建阶段，通过使用一个或多个选择的基于扩散的生成模型重建真实图像和生成图像两者来产生大量图像样本，然后准备用于分类器的训练。在训练阶段，四种类型的样本，包括真实图像、真实重建图像、假图像（即生成图像）和假重建图像，用于计算对比度损失和分类损失。这两个损失函数指导分类器学习更好的特征表示，并将真实图像识别为真实图像，将其他三种类型的图像识别为假图像。

3.2 扩散重建

Stable Diffusion 框架内的重建依赖于迭代去噪图像的条件扩散过程。这一过程在 Stable Diffusion（SD）模型中得到了阐述 [23]，该模型利用一个潜在变量来调节扩散过程的表现力。扩散前向过程可以表示为递增地向图像添加噪声，并且去噪过程通过迭代地降低噪声以恢复原始图像来逆转这一点。前向扩散过程可以表示为:

$$x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon, \tag{1}$$

其中 $\epsilon \sim \mathcal{N}(0, I)$, $t = 0, \ldots, T$. 此处,x_0 表示原始数据,x_t 表示 t 个扩散步骤后的噪声数据, 并且 $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$.

DDIM 的逆过程 [27] 是确定性的，可以表示为:

$$x_{t-1} = \sqrt{\alpha_{t-1}} \frac{x_t - \sqrt{1-\alpha_t}\epsilon_\theta(x_t, t)}{\sqrt{\bar{\alpha}_t}} + \sqrt{1-\alpha_{t-1}}\epsilon_t, \tag{2}$$

其中 $\alpha_{t-1} = \frac{\bar{\alpha}_{t-1}}{\bar{\alpha}_t}$, $t = T, \ldots, 1$, $\epsilon_\theta(x_t, t)$ 是由 θ 参数化的去噪神经网络预测的噪声, $\epsilon_t \sim \mathcal{N}(0, I)$ 是高斯噪声.

3.3 对比训练

原论文在框架中采用了基于边际的对比损失 [9]，使正对更接近，同时通过边际将负样本对分开。这种方法简化了损失计算，并在数学上表示为:

$$\begin{aligned}\mathcal{L}_{\text{contrastive}} = \frac{1}{N}\sum_{i}^{N}[Y \cdot D_w(i)^2 \\ + (1-Y) \cdot \max(0, m - D_w(i))^2]\end{aligned} \tag{3}$$

其中，N 是样本对的总数，Y 是每个样本对的二值标签（如果两个样本具有相同的分类标签，即真实或虚假，那么 $Y = 1$，否则 $Y = 0$ ）。$D_w(i)$ 是每对样本之间的欧几里得距离。$m > 0$ 是负样本对的边界值，在原文的实验中，m 的默认值为 1.0。

要最小化的总体目标是对比损失和二值分类交叉熵损失的组合，由平衡参数 λ 加权:

$$\mathcal{L}_{\text{total}} = \lambda\mathcal{L}_{\text{contrastive}} + (1-\lambda)\mathcal{L}_{\text{cross-entropy}} \tag{4}$$

其中交叉熵损失定义为:

$$\mathcal{L}_{\text{cross-entropy}} = -\sum_{i=1}^{N}[y_i \log(p_i) + (1-y_i)\log(1-p_i)], \tag{5}$$

其中 y_i 是第 i 个样本的真实标签，p_i 是第 i 个样本属于正类的预测概率。参数 $\lambda \in [0, 1)$ 调节损失之间的权衡，在实验中，λ 的默认值为 0.3。

4 复现细节

4.1 与已有开源代码对比

本次论文复现的代码基于论文中给出的代码仓库进行实验的复现，仓库网址为https://github.com/beibuwandeluori/DRCT。不同的是，本次复现加入了对 DRCT 模型的关于对抗样本的初步实践与防御探索。在评估流程中整合了 FGSM（快速梯度符号法）攻击，使得模型能够在测试阶段同时面对原始测试数据和经过 FGSM 生成的对抗样本。通过修改评估函数 eval_model，新增 eval_model_with_fgsm，在其中嵌入 FGSM 攻击函数 fgsm_attack。该攻击函数严格按照 FGSM 原理，计算输入图像对于模型的梯度，依据梯度符号生成对抗扰动，并添加到原始图像上，同时确保生成的对抗样本像素值处于合法范围。

```python
# 使用FGSM生成对抗样本
def fgsm_attack(model, images, labels, eps=0.03):
    images = images.clone().detach().requires_grad_(True)
    outputs = model(images)
    loss = nn.CrossEntropyLoss()(outputs, labels)
    model.zero_grad()
    loss.backward()
    # 获取梯度的符号
    gradient_sign = images.grad.data.sign()
    # 生成对抗扰动并添加到原始图像上
    perturbed_images = images + eps * gradient_sign
    # 对图像进行裁剪，确保像素值在合法范围内
    perturbed_images = torch.clamp(perturbed_images, 0, 1)
    return perturbed_images.detach()
```

除了开源代码中原有的准确率、AUC 等评估指标，新增了如平均精度均值（mAP）、误报率（FNR）等指标，以更全面地衡量模型在复杂场景下的性能表现。在评估函数中调用 sklearn.metrics 库的相关函数，如 recall_score、precision_score、f1_score 等，精确计算新增指标，为模型性能分析提供更丰富的数据支持。

```python
# 在评估模型时，同时考虑原始数据和FGSM攻击生成的对抗样本进行评估
def eval_model_with_fgsm(model, epoch, eval_loader, is_save=True, is_tta=False,
    threshold=0.5, save_txt=None,fgsm_eps=0.03):
    #···省略部分代码
    # 生成并处理FGSM对抗样本
        with torch.set_grad_enabled(True):
            perturbed_img = fgsm_attack(model, img, label, eps=fgsm_eps)
        #···省略部分代码

    # 处理FGSM对抗样本的评估指标计算
    outputs_fgsm = torch.cat(outputs_fgsm, dim=0).cpu().numpy()
    labels_fgsm = torch.cat(labels_fgsm, dim=0).cpu().numpy()
    labels_fgsm[labels_fgsm > 0] = 1
    auc_fgsm = roc_auc_score(labels_fgsm, outputs_fgsm)
    recall_fgsm = recall_score(labels_fgsm, outputs_fgsm > threshold)
    precision_fgsm = precision_score(labels_fgsm, outputs_fgsm > threshold)
    binary_acc_fgsm = accuracy_score(labels_fgsm, outputs_fgsm > threshold)
    f1_fgsm = f1_score(labels_fgsm, outputs_fgsm > threshold)
    fnr_fgsm = calculate_fnr(labels_fgsm, outputs_fgsm > threshold)

    #···省略其余代码
```

4.2 实验环境搭建

本实验基于 Linux 系统，采用 Python3.8.20 版本，借助 PyTorch2.1.2 深度学习框架实现 DRCT 相关代码，配合 torchvision、sklearn、matplotlib 等库分别用于图像数据处理、模型评估、可视化操作。硬件配置包括 8 张 NVIDIA A100 80GB PCIe，利用 GPU 并行计算加速任务。通过 Anaconda 虚拟环境隔离依赖冲突，为实验构建稳定、高效的运行环境。数据集

选取该论文贡献的一份数据集，DRCT-2M/SDv1.4 数据集进行训练，基准模型选用 Conv-B，在 DRCT 框架下进行训练。

5 实验结果分析

Method	DR	SD Variants						Turbo Variants		LCM Variants		ControlNet Variants			DR Variants			Avg.
		LDM	SDv1.4	SDv1.5	SDv2	SDXL	SDXL-Refiner	SD-Turbo	SDXL-Turbo	LCM-SDv1.5	LCM-SDXL	SDv1-Ctrl	SDv2-Ctrl	SDXL-Ctrl	SDv1-DR	SDv2-DR	SDXL-DR	
CNNSpot	-	99.87	99.91	99.90	97.55	66.25	86.55	86.15	72.42	98.26	61.72	97.96	85.89	82.84	60.93	51.41	50.28	81.12
F3Net	-	99.85	99.78	99.79	88.66	55.85	87.37	68.29	63.66	97.39	54.98	97.98	72.39	81.99	65.42	50.39	50.27	77.13
CLIP/RN50	-	99.00	99.99	99.96	94.61	62.08	91.43	83.57	64.40	98.97	57.43	99.74	80.69	82.03	65.83	50.67	50.47	80.05
GramNet	-	99.40	99.01	98.84	95.30	62.63	80.68	71.19	69.32	93.05	57.02	89.97	75.55	82.68	51.23	50.01	50.08	76.62
De-fake	-	92.1	99.53	99.51	89.65	64.02	69.24	92.00	93.93	99.13	70.89	58.98	62.34	66.66	50.12	50.16	50.00	75.52
Conv-B	-	**99.97**	**100.0**	**99.97**	95.84	64.44	82.00	80.82	60.75	99.27	62.33	99.80	83.40	73.28	61.65	51.79	50.41	79.11
UnivFD	-	98.30	96.22	96.33	93.83	91.01	93.91	86.38	85.92	90.44	88.99	90.41	81.06	89.06	51.96	51.03	50.46	83.46
DIRE	SDv1	98.19	99.94	99.96	68.16	53.84	71.93	58.87	54.35	**99.78**	59.73	99.65	64.20	59.13	51.99	50.04	49.97	71.23
DIRE	SDv2	54.62	75.89	76.04	**99.87**	59.90	93.08	**99.77**	57.55	87.29	72.53	67.85	99.69	64.40	49.96	52.48	49.92	72.55
DRCT/Conv-B (ours)	SDv1	99.91	99.90	99.90	96.32	83.87	85.63	91.88	70.04	99.66	78.76	**99.90**	95.01	81.21	**99.90**	95.40	75.39	90.79
DRCT/Conv-B (ours)	SDv2	99.66	98.56	98.48	99.85	96.10	**98.68**	99.59	83.30	98.45	93.78	96.68	**99.85**	**97.66**	93.91	**99.87**	**90.39**	**96.55**
DRCT/UnivFD (ours)	SDv1	96.74	96.26	96.33	94.89	**96.24**	93.46	93.43	92.94	91.17	95.01	95.60	92.68	91.95	94.10	69.55	57.43	90.49
DRCT/UnivFD (ours)	SDv2	94.45	94.35	94.24	95.05	95.61	95.38	94.81	**94.48**	91.66	**95.54**	93.86	93.48	93.54	84.34	83.20	67.61	91.35

图 3. DRCT-2M 上 DRCT 和其他生成图像检测器的准确度（ACC，%）比较

图3报告了 DRCT-2M 数据集的检测精度比较。大多数方法在由与 SDv1.4 相关的扩散模型生成的图像上表现出极高的 ACC，如 LDM、SDv1.5、LCM-SDv1.5 和 SD-Ctrl。然而，当检测未知的和实质性改变的扩散模型（如 SDv2、SDXL、SDXL-Refiner、SDXL-Turbo、LCM-SDXL 和 SDXL-Ctrl）时，这些方法的 ACC 显著下降。特别是对于 SDXL 生成的图像，除了 UnivFD 之外，现有方法的 ACC 仅在 53%-67% 之间。用于检测 SDv1-DR、SDv2-DR 和 SDXL-DR 的真实重建图像的现有方法的 ACC 下降到 50%-65%。

相比之下，论文提出的 DRCT 框架在所有类型的 SD 生成图像中实现了卓越的 ACC。实验还比较了不同扩散模型（即 SDv1 或 SDv2）用于重建时的检测 ACCs。如图所示，仅使用 SDv1 进行 DRCT 重建，与基线检测器 Conv-B 相比，平均检测 ACC 已经从 79.11% 提高到 90.79%。对于 UnivFD，平均 ACC 从 83.46% 提高到 90.49%。当使用 SDv2 进行重建时，与基线检测器 Conv-B 相比，平均检测 ACC 可进一步提高至 96.55%。这表明更好的重建模型有助于在 DRCT-2M 数据集上实现更好的检测性能。

表 1. 论文原文、本文复现以及在 FGSM 对抗样本下的模型准确率（ACC，%）比较

Method	DR	SD Variants						Turbo Variants		LCM Variants		ControlNet Variants			DR Variants			Avg.
		LDM	SDv1.4	SDv1.5	SDv2	SDXL	SDXL-Refiner	SD-Turbo	SDXL-Turbo	LCM-SDv1.5	LCM-SDXL	SDv1-Ctrl	SDv2-Ctrl	SDXL-Ctrl	SDv1-DR	SDv2-DR	SDXL-DR	
DRCT/Conv-B (paper)	SDv1	99.91	99.90	99.90	96.32	83.87	85.63	91.88	70.04	99.66	78.76	**99.90**	95.01	81.21	**99.90**	95.40	75.39	90.79
DRCT/Conv-B (mine)	SDv1	99.96	99.96	99.95	97.62	82.03	80.41	91.74	68.47	99.70	77.16	**99.96**	94.13	81.27	**99.96**	91.62	66.51	89.40
DRCT/Conv-B (FGSM)	SDv1	60.77	57.06	57.28	49.93	45.56	46.72	46.04	45.22	55.38	45.78	55.02	47.10	46.36	53.31	48.28	48.76	50.51

表1报告了在 DRCT-2M 数据集下，原文、复现以及面对对抗样本下的检测精度比较。在大多数方法下，复现得到的数据和原文中给出的数据基本符合。但是在面对 FGSM 攻击的时候，模型的准确度收到了较大的影响，平均准确度从 89.40% 下降到了 50.51%。可以看出，开源代码仅专注于在正常数据分布下对扩散生成图像的检测，未考虑对抗样本的存在。而改进后的代码引入 FGSM 攻击，主动探究模型在面对对抗攻击时的脆弱性，通过生成对抗样本并在评估过程中使用，能够直接观察模型性能的下降情况，为后续的防御策略制定提供依据。

为了进一步验证 DRCT 框架的有效性，实验还按照与 GenImage 相同的实验方案进行了比较。所有检测方法都在 GenImage 的 SDv1.4 子集上进行训练。具体来说，对于 DIRE 和 DRCT 的训练，重建模型也是 SDv1。图4中的比较结果显示，所有比较的方法在 SDv1.4、

Method	Midjourney	SDv1.4	SDv1.5	ADM	GLIDE	Wukong	VQDM	BigGAN	Avg.
CNNSpot	84.92	99.88	99.76	53.48	53.80	99.68	55.50	49.93	74.62
F3Net	77.85	98.99	99.08	51.20	54.87	97.92	58.99	49.21	73.51
CLIP/RN50	83.30	99.97	99.89	54.55	57.37	99.52	57.90	50.00	75.31
GramNet	73.68	98.85	98.79	51.52	55.38	95.38	55.15	49.41	72.27
De-fake	79.88	98.65	98.62	71.57	78.05	98.42	78.31	74.37	84.73
Conv-B	83.55	**99.99**	**99.92**	51.75	56.27	99.92	58.41	50.00	74.98
UnivFD	91.46	96.41	96.14	58.07	73.40	94.53	67.83	57.72	79.45
DIRE	50.40	**99.99**	**99.92**	52.32	67.23	**99.98**	50.10	49.99	71.24
DRCT/Conv-B (ours)	**94.63**	99.88	99.82	61.78	65.92	99.91	74.88	58.81	82.08
DRCT/UnivFD (ours)	91.50	95.01	94.41	**79.42**	**89.18**	94.67	**90.03**	**81.67**	**89.49**

图 4. GenImage 上 DRCT 和其他生成图像检测器的准确度（ACC，%）比较

SDv1.5 和 Wukong 子集上实现了非常高的检测精度。然而，在其他子集（如 Midjourney、ADM、GLIDE、VQDM 和 BigGAN）中可以观察到 ACC 的明显下降，特别是在非基于扩散的生成方法 BigGAN 上。在结合提出的 DRCT 框架后，基准检测器 Conv-B 和 UnivFD 分别实现了 7.1% 和 10.04% 的平均 ACC 改善。这验证了 DRCT 框架在增强所涉及的检测器的可推广性方面的稳定有效性。

表 2. 论文原文、本文复现模型准确率（ACC，%）比较

Method	Midjourney	SDv1.4	SDv1.5	VQDM	BigGAN	Avg.
DRCT/Conv-B (paper)	94.63	99.88	99.82	74.88	58.81	85.60
DRCT/Conv-B (mine)	97.13	99.49	99.06	68.75	54.54	83.79

表2报告了在 GenImage 数据集下，原文、复现的检测精度比较。在大多数方法下，复现得到的数据和原文中给出的数据基本符合。

6 总结与展望

原论文提出了一个通用的框架，扩散重建对比训练，以增强现有方法检测基于扩散的生成图像的泛化性。使用 DRCT 框架，基准检测器可以实现显著的改进，表明 DRCT 框架的有效性。此外，还构建了大规模、高质量的图像数据集 DRCT-2M，用于检测器的训练以及有效性、泛化性、鲁棒性等的评估。原论文提出未来的工作包括改进 DRCT，以跟上扩散模型的发展。探索 DRCT 增强检测器学习的特征的可解释性可以提供对真实图像和生成图像之间根本差异的见解。

而本人在阅读和复现的过程中，理解原文给出的代码后，尝试在其中尝试整合了 FGSM 攻击，使得模型能够在测试阶段同时面对原始测试数据和经过 FGSM 生成的对抗样本，并且收集对应的数据进行比较。本文在实践对抗样本攻击检测模型的同时，探索 DRCT 模型对对抗样本的受影响程度。未来可以继续尝试在模型中添加检测或者防御对抗样本的功能，进一步提升 DRCT 模型的性能，拓展模型的功能边界。

参考文献

[1] Jordan J Bird and Ahmad Lotfi. Cifake: Image classification and explainable identification of ai-generated synthetic images. *IEEE Access*, 2024.

[2] Riccardo Corvi, Davide Cozzolino, Giada Zingarini, Giovanni Poggi, Koki Nagano, and Luisa Verdoliva. On the detection of synthetic images generated by diffusion models. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[3] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

[4] David C Epstein, Ishan Jain, Oliver Wang, and Richard Zhang. Online detection of ai-generated images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 382–392, 2023.

[5] Joel Frank, Thorsten Eisenhofer, Lea Schönherr, Asja Fischer, Dorothea Kolossa, and Thorsten Holz. Leveraging frequency analysis for deep fake image recognition. In *International conference on machine learning*, pages 3247–3258. PMLR, 2020.

[6] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

[7] Luca Guarnera, Oliver Giudice, and Sebastiano Battiato. Level up the deepfake detection: a method to effectively discriminate images generated by gan architectures and diffusion models. In *Intelligent Systems Conference*, pages 615–625. Springer, 2024.

[8] Xiao Guo, Xiaohong Liu, Zhiyuan Ren, Steven Grosz, Iacopo Masi, and Xiaoming Liu. Hierarchical fine-grained image forgery detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3155–3165, 2023.

[9] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06)*, volume 2, pages 1735–1742. IEEE, 2006.

[10] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[11] Tero Karras. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

[12] Diederik P Kingma. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[13] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022.

[14] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986, 2022.

[15] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *Advances in Neural Information Processing Systems*, 35:5775–5787, 2022.

[16] Simian Luo, Yiqin Tan, Suraj Patil, Daniel Gu, Patrick von Platen, Apolinário Passos, Longbo Huang, Jian Li, and Hang Zhao. Lcm-lora: A universal stable-diffusion acceleration module. *arXiv preprint arXiv:2311.05556*, 2023.

[17] Ruipeng Ma, Jinhao Duan, Fei Kong, Xiaoshuang Shi, and Kaidi Xu. Exposing the fake: Effective diffusion-generated images detection. *arXiv preprint arXiv:2307.06272*, 2023.

[18] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International conference on machine learning*, pages 8162–8171. PMLR, 2021.

[19] Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. Towards universal fake image detectors that generalize across generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24480–24489, 2023.

[20] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2337–2346, 2019.

[21] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

[22] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[23] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[24] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022.

[25] Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. In *European Conference on Computer Vision*, pages 87–103. Springer, 2025.

[26] Zeyang Sha, Zheng Li, Ning Yu, and Yang Zhang. De-fake: Detection and attribution of fake images generated by text-to-image generation models. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, pages 3418–3432, 2023.

[27] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

[28] Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, and Yunchao Wei. Learning on gradients: Generalized artifacts representation for gan-generated images detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12105–12114, 2023.

[29] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

[30] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros. Cnn-generated images are surprisingly easy to spot... for now. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8695–8704, 2020.

[31] Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, Hezhen Hu, Hong Chen, and Houqiang Li. Dire for diffusion-generated image detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22445–22455, 2023.

[32] Haiwei Wu, Jiantao Zhou, and Shile Zhang. Generalizable synthetic image detection via language-guided contrastive learning. *arXiv preprint arXiv:2305.13800*, 2023.

[33] Ziyi Xi, Wenmin Huang, Kangkang Wei, Weiqi Luo, and Peijia Zheng. Ai-generated image detection using a cross-attention enhanced dual-stream network. In *2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 1463–1470. IEEE, 2023.

[34] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023.

[35] Nan Zhong, Yiran Xu, Zhenxing Qian, and Xinpeng Zhang. Rich and poor texture contrast: A simple yet effective approach for ai-generated image detection. *arXiv preprint arXiv:2311.12397*, 2023.