arXiv:2601.00218v1 [cs.LG] 1 Jan 2026

未知感知的人工智能生成内容溯源

Ellie Thieu威斯康
星大学麦迪逊分校
thieu@wisc.edu

Jifan Zhang威斯康
星大学麦迪逊分校
jifan@cs.wisc.edu

Haoyue Bai威斯康星
大学麦迪逊分校
haoyue.bai@wisc.edu

摘要

照片级真实感生成模型的迅速发展，使得追溯合成内容的来源变得日益重要；研究重点也从二元的真假检测，转向识别生成特定图像的具体模型。我们研究如何将目标生成模型（例如 OpenAI 的 DALL·E 3）的输出与其他来源区分开来，其中包括真实图像以及由各种其他模型生成的图像。基于 CLIP 特征和一种已在先前研究中证明有效的简单线性分类器，我们建立了目标生成器归因的强基线方法，该方法仅需来自目标模型的少量标注数据，以及来自少数已知生成器的数据。然而，这种基线方法难以泛化到更具挑战性、未曾见过以及新近发布的生成器。为解决这一局限，我们提出一种约束优化方法，利用无标签野外数据。这些数据由从互联网上收集的图像构成，可能包括真实图像、未知生成器的输出，甚至目标模型自身生成的样本。该方法促使分类器将野外样本判定为非目标样本，同时明确约束其在标注数据上的性能，使其保持较高水平。实验结果表明，加入野外数据可显著提升模型对具有挑战性的未见生成器的归因性能，证明在开放世界场景中，有效利用野外无标签数据能够增强人工智能生成内容溯源。

1 引言

深度生成建模的最新进展显著提升了图像合成的质量 [1, 2, 3, 4, 5]。尽管这些进展拓展了生成模型的应用范围，但也引发了人们对滥用问题的重大担忧。旨在区分合成图像与真实图像的人工智能生成内容检测，已成为一项关键任务，并吸引了大量研究关注，催生了多种提议的方法 [6, 7, 8, 9]。然而，除了检测之外，追溯合成内容的来源同样重要，因为这能确定某幅图像具体由哪种生成模型生成。此类归因能够实现来源追踪，并通过将内容关联到负有责任的模型开发者来支持问责。例如，确定某幅图像是由 DALL·E 3 [10], 而非 Midjourney 或 Stable Diffusion [3],生成的，将直接影响生成式人工智能领域的治理、透明度和安全性。

新生成模型快速且持续地发布，对现有归因方法构成了根本性挑战。这种快速演进使得依赖监督式流程变得不切实际，因为这类流程假设在训练时已知所有可能的生成器。因此，一个关键且在很大程度上尚未解决的问题随之出现：

我们能否设计一种目标生成器归因学习框架，使其在存在未知或新发布的生成器时仍保持稳健？

一种常见方法是使用来自目标生成器的标注样本，以及少量已知非目标来源的样本来训练分类器。先前研究表明，CLIP特征 [11], 结合

预印本。审稿中。

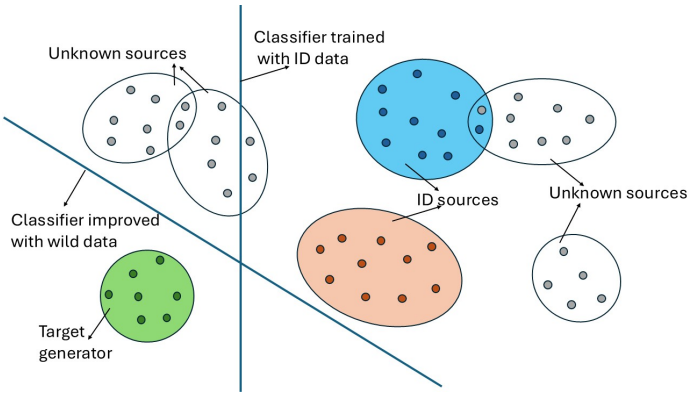

图 1：未知感知生成器归因示意图。仅使用标注数据源训练的分类器会学习到一个决策边界，将目标生成器与已知来源区分开来，但可能会将来自未知或未见生成器的样本误分类。通过在约束优化框架下纳入无标签野外数据，可以调整决策边界，更好地区分目标生成器与多种未知来源，同时保持在标注数据源上的性能。

简单的线性分类器，仅使用来自目标生成器的有限标注数据和少量已知生成器 [8]，即可实现出色的归因性能。然而，这类模型往往难以泛化到更具挑战性的新发布生成器。此外，鉴于新模型不断涌现，在实际应用中，为每一种可能的非目标生成器收集标注数据并不可行。相比之下，来自真实世界的未标注图像（例如从互联网收集的图像）数量丰富、种类多样且获取成本低。

为应对这一挑战，我们提出了一种未知感知的人工智能生成内容溯源框架，通过约束优化过程，利用无标签野外数据对给定分类器进行微调。野外数据可能包括真实图像、未知生成器的输出，甚至目标生成器生成的内容，并将其视为非目标样本，以促进泛化。我们施加约束以保持模型在原始标注数据上的性能，使其能够利用野外数据的多样性，同时维持对已知来源的准确识别。实验结果表明，我们的方法能够持续提升对 DALL·E 3 图像的归因效果，尤其是在面对 Midjourney、Firefly 和 Stable Diffusion XL 等难以区分且未见过的生成器时。结果证明，无标签野外数据能够得到有效利用，从而在开放世界设置下构建可扩展且稳健的人工智能内容归因系统。

图 1 展示了所提出框架背后的核心直觉。垂直线表示仅使用分布内数据训练的基线分类器所学习到的决策边界，而调整后的边界则展示了在所提出约束下使用无标签野外数据进行微调的效果。

我们的主要贡献总结如下：

- **一种用于未知生成器感知归因的约束微调框架**。 我们提出了一种简单且稳定的约束优化方法，利用无标签野外数据对分类器进行微调，同时明确保留其在带标签的分布内数据上的性能，从而防止模型崩溃和灾难性遗忘。

- **利用无标签野外数据提升泛化能力**。 我们提出了一种训练框架，通过施加约束，在提升对未见生成器的鲁棒性的同时，保持在可信带标签数据上的性能，从而利用无标签互联网图像，尽管这些图像具有噪声和多样性。

- **目标生成器归因的问题表述**。 我们研究一种细粒度的二元任务：判断图像是否由特定模型（如 DALL·E 3）生成。这超越了真假检测，旨在支持问责与人工智能治理。鉴于当前行业实践，相较于对众多生成器进行多分类归因，针对单一目标的归因更为实用，也更切合社会需求。

• **开放世界设置中的未知生成器感知归因**。 我们强调并通过实证研究一个挑战：当存在未知或新近发布的生成器时，如何识别目标模型的输出。这是人工智能生成内容领域中一种符合现实、却鲜有研究的设置。

2 相关工作

2.1 合成图像生成

过去十年间，图像生成建模发展迅速。早期进展由生成对抗网络（生成对抗网络）推动，使得在多个领域实现高质量图像合成成为可能 [12, 13, 14, 15, 16]。随后，研究人员探索了基于 Transformer 的架构，以进一步提升图像生成的保真度和可扩展性 [17, 18, 19]。

近年来，扩散模型已成为图像生成的主流范式，大幅推动了技术发展，并催生了 Stable Diffusion [3], DALL·E 和 DALL·E 2 [5], DALL·E 3 [10], GLIDE [20], 等广泛应用的系统及相关变体 [21, 22]。这一生态系统仍在快速演变，新的生成器不断推出，现有模型也在持续迭代优化。因此，下游检测与溯源系统必须应对不断变化的生成技术格局，其中许多技术在训练时可能尚不可用或尚不为人所知。

2.2 人工智能生成内容检测与溯源

生成模型的迅速进步推动了人工智能生成内容检测与溯源领域的大量研究。早期检测方法主要关注 GAN 生成的图像，并依赖手工设计的视觉线索。例如，先前的研究考察了眼睛和牙齿等面部伪影 [23] 或低级颜色不一致 [24]。后续方法则致力于寻找生成特定签名，包括频域伪影 [25, 26], GAN 指纹 [27],以及通过有限感受野捕获的局部伪影 [6]。

随着基于扩散模型的生成器兴起，许多面向 GAN 的检测器被证明性能大幅下降 [28]。对此，近期研究提出了针对扩散模型的检测策略 [29, 30]。例如，Synthbuster [9] 利用残差的傅里叶统计量，DE-FAKE [31]结合提示信息，而 DIRE [32] 则利用扩散模型的重建行为。其他方法则分析光照或透视等物理不一致性 [33, 34]。研究人员还提出了基于大型视觉变换器和 CLIP 表征的通用检测器 [8], , 不过这些方法通常使用来自已知生成器的标注数据进行训练。

尽管取得了这些进展，现有的大多数方法仍默认采用封闭世界设置，即训练期间已知测试时会遇到的生成器集合。正如先前研究所示 [29], , 针对特定架构或生成器训练的检测器往往无法跨模型家族泛化，这凸显了依赖模型特定伪影的局限性。

相比之下，我们研究未知感知的 人工智能生成内容溯源，这是一种更贴近现实的设置，在该设置中，测试时内容可能来自未见过或新近发布的生成器。我们的方法不假设能够获取所有可能来源的标注数据，而是利用未标注的真实场景图像混合数据，提升应对不断演进的生成模型的鲁棒性。

2.3 开放集归因

少量先前研究探讨了图像归因的开放世界设置，不过大多聚焦于生成对抗网络，或在架构层面开展工作。例如，Abady 等人 [35] 通过识别生成器架构而非特定模型来解决开放集归因问题。其他在开放世界设置下进行生成器级归因的研究 [36, 37] 主要针对基于生成对抗网络的生成器，而这类生成器通常比现代扩散模型或专有商业系统更容易处理。
为应对未见生成器，一些方法引入显式拒识或“未知”类别 [38,39]。尽管这些方法能有效避免过度自信的误分类，但它们强调的是在已知生成器之间进行正确分类，而非提升对具有挑战性的未见生成器的泛化能力。相比之下，我们的方法着重利用无标签野外数据，直接提升

对未知生成器的归因鲁棒性，无需显式拒识建模，也无需来自未见来源的带标签样本。

2.4 半监督学习

半监督学习（Semi-SL）为在分类任务中结合无标签数据与标注样本提供了通用框架 [40, 41, 42]。大多数现代半监督学习方法都基于两个核心原则：一致性正则化 和伪标签 [43, 44]。基于一致性的方法鼓励模型在数据增强下保持预测不变，而伪标签方法则为置信度高的预测所对应的无标签样本分配标签。FlexMatch [45], 、FreeMatch [46], 和 SoftMatch [47] 等扩展方法进一步通过自适应阈值和置信度校准改进了这些思路。

尽管半监督学习方法在自然图像分类基准上取得了出色的结果，但其假设在人工智能生成图像检测与溯源场景中往往不再成立。具体而言，强数据增强可能会抑制或扭曲细微的生成伪影，而这些伪影对于识别源模型至关重要。因此，在此场景中，强制增强视图之间保持一致性可能适得其反。
近期研究探索了结合大型预训练模型和视觉变换器的半监督学习 [48, 49, 50],，而 USB [51] 等基准则提供了标准化评估。然而，这些方法并未明确应对生成分布不断演变所带来的合成图像归因挑战。我们的工作还与半监督新颖性检测 [52] 及其面向分布外（OOD）检测的扩展 [53, 54, 55, 56]密切相关。这些方法旨在从已知和未知分布的混合中识别新颖样本。我们从这一研究方向汲取灵感，并将其原则应用于未知感知生成器归因问题。

3 问题设定

我们研究目标生成器归因问题：给定一幅图像$\mathbf{x}$，判断它是否由特定的目标生成模型 $\mathbb{G}_t$（例如 DALL·E 3）生成。该任务可以表述为一个二分类问题，目标是判断$\mathbf{x} \sim \mathbb{P}_{\mathbb{G}_t}$，还是来自任何非目标来源，包括其他生成模型 $\mathbb{G} \neq \mathbb{G}_t$ 或真实图像。

带标签的分布内数据。 令 $\mathcal{X} = \mathbb{R}^d$ 表示输入空间， $\mathcal{Y} = \{0, 1\}$ 表示标签空间，其中 $y = 1$ 表示图像由目标模型 $\mathbb{G}_t$生成， $y = 0$表示非目标内容。我们假设可以获得一个带标签的分布内（ID）训练数据集$\mathcal{D}_{\text{labeled}} = \{ (\mathbf{x}_i,\ y_i) \}_{i=1}^{n}$，其中每个样本$\mathbf{x}_i$ 要么来自目标生成器分布$\mathbb{P}_{\mathbb{G}_t}^{\mathcal{X}}$ 并带有标签 $y_i = 1$，要么来自一组已知的非目标来源并带有标签 $y_i = 0$。这些已知来源可能包括真实图像，以及由训练时可用的数量有限的辅助生成模型生成的图像。

无标签野外数据。 在部署时，分类器会遇到来自更广泛且不受控制的来源混合分布的输入。我们通过一个野外分布来对这种情形进行建模，

$$\mathbb{P}_{\text{wild}} := (1 - \pi_k - \pi_u)\, \mathbb{P}_{\mathbb{G}_t}^{\mathcal{X}} \ + \ \pi_k\, \mathbb{P}_{\text{known}} \ + \ \pi_u\, \mathbb{P}_{\text{unknown}},$$

其中 $\pi_k,\ \pi_u \geq 0$ 和 $\pi_k + \pi_u \leq 1$ 是未知的混合比例。这里：

- **目标分布** $\mathbb{P}_{\mathbb{G}_t}^{\mathcal{X}}$ 表示由目标模型 $\mathbb{G}_t$生成的图像的边缘分布；
- **已知源分布** $\mathbb{P}_{\text{已知}}$对应于训练期间观测到的非目标来源；
- **未知源分布** $\mathbb{P}_{\text{未知}}$表示来自先前未见过的生成器或其他新颖来源的图像。

这一表述刻画了一个真实的开放世界部署场景：测试时输入可能来自已见过和未见过的生成过程。根据训练设置，真实图像可能出现在已知或未知的源分布中。核心挑战在于，即使无法确定野外数据的组成，也要学习一个能够可靠地区分目标模型输出与所有其他来源的决策函数。

学习框架。令 $f_\theta : \mathcal{X} \to \mathbb{R}$ 表示一个由 θ参数化的评分函数，分数越高，表示输入由目标模型生成的可能性越大。我们首先使用带标签的分布内数据集 $\mathcal{D}_{\text{labeled}}$训练$f_\theta$。随后，可以使用从 $\mathbb{P}_{\text{wild}}$ 抽取的无标签野外数据对模型进行微调，以提升其对未知生成器的鲁棒性。我们使用与阈值无关的指标评估性能，包括平均精度（AP）和 ROC 曲线下面积（AUROC）。

4 具备未知感知能力的人工智能生成内容溯源

我们提出了一种具备未知感知能力的特定生成器图像归因框架，即使面对尚未出现或新发布的生成器，也能保持稳健。该任务旨在判断给定图像是否由特定目标模型生成。在实验中，我们以 OpenAI 的 DALL·E 3 为代表性案例进行研究。为便于说明，我们在全文中使用 DALL·E 3，但所提出的框架与模型无关，可轻松应用于任何目标生成器。
我们的方法采用两阶段流程。首先，我们使用一个小型标注数据集训练基线归因分类器，该数据集包含来自目标生成器的图像以及少量已知非目标来源的图像。尽管该基线在分布内（ID）数据上表现出色，但往往无法泛化到由尚未出现或新发布的生成器生成的图像。为解决这一局限，我们提出了一种利用无标签野外数据的约束微调策略。其核心思想是让分类器接触广泛且多样的图像分布，同时明确约束其在可信带标签数据上的表现，从而在不牺牲分布内准确率的情况下提升泛化能力。

我们采用一种二元归因形式，适用于实际部署场景：组织希望确定内容是否由其自有的专有模型生成。虽然我们的实验聚焦于这一单一目标设置，但所提出的约束优化框架具有通用性，只需稍作修改即可扩展到多生成器归因或联合检测与归因任务。

本节余下部分将首先介绍使用带标签的分布内数据训练的基线分类器（第 4.1 节），然后介绍我们提出的、纳入无标签野外数据的约束微调方法（第 4.2 节）。

4.1 分布内数据训练

我们首先使用带标签的分布内（ID）数据训练基线归因分类器。该数据包含来自目标生成器（DALL·E 3）的图像，以及少量可用的非目标生成器生成的图像。我们将此数据集记为 D_{labeled}。每张图像均由 CLIP ViT-L/14图像编码器进行编码，并从倒数第二层提取一个 768 维特征向量。研究表明，CLIP 特征在合成图像检测与归因任务中效果显著 [8], ，近期研究也表明，使用其他CLIP变体也能取得相近的性能 [57]。
在冻结的 CLIP 特征之上，我们训练一个轻量级线性分类器来进行二元归因。目标模型生成的图像标记为 0，所有其他图像标记为 1。该分类器由单个线性层和后接的 Sigmoid 激活函数组成，输出图像不源自目标生成器的概率。训练采用二元交叉熵（BCE）损失和 Adam 优化器，并根据在 D_{labeled} 留出子集上计算的验证损失进行提前停止。

所得分类器作为我们的基线模型。我们记录其在带标签的分布内数据上的 BCE 损失，作为后续使用野外数据进行微调时约束性能的参考点。

4.2 通过约束优化使用野外数据进行微调

为了提升对未见生成器和具有挑战性的生成器的泛化能力，我们引入了从互联网等多种来源收集的无标签野外数据。这些野外数据可能包括真实图像、由已知非目标模型生成的图像、来自此前未见生成器的图像，以及目标生成器自身生成的一小部分图像。由于无法获得真实标签，这些样本无法以完全监督方式加以利用。

我们不对野外样本显式分配硬标签，而是将野外数据视为一种辅助信号，促使分类器将其决策边界扩展到远离目标生成器的一侧。仅对野外数据进行朴素优化会导致退化解（例如，将所有样本都预测为非目标）。为避免这种情况，我们引入了一种约束优化形式，在利用野外样本多样性的同时，明确保持在带标签的分布内数据上的性能。

具体而言，我们使用带标签的分布内数据和无标签野外数据对基线分类器进行微调，同时约束其在 D_{labeled} 上的损失。该约束确保微调不会削弱分类器原有的归因能力。在所有实验中，我们将约束阈值设为仅使用带标签的分布内数据训练的基线分类器所达到损失的两倍。

对野外数据应用相同的 CLIP 特征提取流程。提取特征后，使用下文所述的约束目标进行微调。由于分类器是在固定表征上运行的低容量线性模型，因此即使野外数据集规模庞大或高度多样，该流程仍能保持稳定。

学习目标 我们的目标是利用无标签野外数据，提升模型对未知或未见生成器的归因鲁棒性，同时明确保留模型在带标签的分布内（ID）数据上的性能。令 $D_{\text{labeled}} = \{(x_j,\ y_j)\}_{j=1}^{n}$ 表示带标签的 ID 数据集，其中 $y_j \in \{0, 1\}$ 表示 x_j 是否由目标模型生成（0）或并非由其生成（1）。令 $D_{\text{wild}} = \{\tilde{x}_i\}_{i=1}^{m}$ 表示无标签野外数据集。令 $f_\theta(x)$ 表示分类器的输出，令 $L(\cdot,\ \cdot)$ 表示二元交叉熵（BCE）损失。

我们将微调目标表述为以下约束优化问题：

$$\min_{\theta} \frac{1}{m} \sum_{i=1}^{m} L(f_\theta(\tilde{x}_i), 1) \tag{1}$$

$$\text{s.t. } \frac{1}{n} \sum_{j=1}^{n} L(f_\theta(x_j), y_j) \leq \alpha, \tag{2}$$

其中，目标函数促使野外样本被分类为非目标类，而约束条件确保带标签的分布内数据上的平均损失不超过预定义阈值 α。在所有实验中，我们将 α 设为仅使用 D_{labeled}训练的分类器所达到的二元交叉熵损失的两倍。

In p实践中，我们通过最小化两个损失的加权 d
sum 两个损失中的：

$$\mathcal{L}(\theta) = \frac{1}{n} \sum_{j=1}^{n} L(f_\theta(x_j), y_j) + \lambda \frac{1}{m} \sum_{i=1}^{m} L(f_\theta(\tilde{x}_i), 1)\,, \tag{3}$$

其中，$\lambda \geq 0$ 控制野外数据正则化的强度。λ 的取值经过选择，使得最终带标签的分布内数据损失接近约束阈值 α，从而避免退化解，即将所有样本都简单地分类为非目标类。

微调过程中，我们监控从与训练数据相同分布中抽取的留出验证集上的复合损失，并在损失趋于稳定后提前停止。该流程使分类器能够从野外数据的多样性中获益，同时保持在已知分布内来源上的可靠性能。

5 实验

我们的实验设计反映了现实的开放世界部署环境。我们使用来自一小部分当前可用生成器的图像训练归因分类器，并评估其对训练期间未见过的更新、更具挑战性的生成器的泛化能力。这一设置体现了现实世界归因系统面临的两个实际限制：（i）生成模型集合不断快速演变；（ii）收集大量带标签的归因数据集成本高昂，而来自真实世界的无标签图像却十分丰富。

在所有实验中，我们以 OpenAI 的 DALL·E 3 为目标生成器，并使用 CLIP 特征和线性分类器训练分类器，将 DALL·E 3 生成的图像与所有其他来源生成的图像区分开来。我们在三类来源上评估归因性能：（1）

有标签训练期间观察到的生成器（分布内）；（2）仅出现在用于微调的野外数据中的生成器；（3）既未出现在有标签训练集、也未出现在野外数据中的生成器。结果分别报告了纳入和未纳入野外数据的情况。

5.1 实验细节

数据集。 在主要实验中，带标签的分布内（ID）训练集由来自四种生成器的图像组成：DALL·E 3、悟空、Stable Diffusion v1.4 和 Stable Diffusion v1.5，其中 DALL·E 3 是目标生成器。我们使用 200 张 DALL·E 3 图像，以及来自每种辅助生成器的 67 张图像，从而构成一个平衡的二分类数据集。该设置表明，利用预训练 CLIP 表征，即使带标签数据有限，也能实现出色的归因性能，同时反映了现实的数据收集限制。

为模拟无标签野外数据，我们构建了一个数据集，其中包含以下各来源的 67 张图像：ImageNet（真实图像）、DALL·E3、ADM、BigGAN、GLIDE (4.5B)、Midjourney、Stable Diffusion v1.4、Stable Diffusion v1.5、VQDM、悟空、Firefly、Stable Diffusion XL、LDM_200_cfg 和 DALL·E 2。将 DALL·E 3 图像纳入野外数据集会引入标签噪声，从而反映现实场景：野外数据中可能包含来自目标生成器的无标签样本。

数据来源。 DALL·E 3 图像取自 Hugging Face 上的 DALL·E 3 数据集。[1] 真实图像来自 ImageNet。[2]

Midjourney、Stable Diffusion v1.4 和 v1.5、ADM、GLIDE (4.5B)、悟空、VQDM 和 BigGAN 的图像取自 GenImage 数据集 [58]。其他图像来源包括 GLIDE_50_27、GLIDE_100_10、GLIDE_100_27、Guided、LDM_100、LDM_200_cfg 和 LDM_200，这些图像均取自 [8]。我们还从 Synthbuster [9], 纳入了 DALL·E 2、Firefly、更多 Midjourney 样本，以及 Stable Diffusion 的 v1.3、v1.4、v2 和 XL 变体。

完整数据集汇集自多个来源，因此各生成器可用图像的数量差异显著，约从 1,000 张到 160,000 多张不等。尽管我们精心整理了涵盖开源和专有生成器的多样化且及时更新的集合，但模型快速发展的步伐意味着任何固定的数据集都不可避免地会过时。这进一步凸显了对未知类别感知归因方法的需求。

对于纳入野外集合的生成器，我们从每个来源随机抽取 67 张图像用于训练。评估时，无论图像是否属于训练集或野外集合，所有来源的每个生成器均使用 600 张测试图像，以确保不同评估之间的一致性。

我们报告平均精度（AP）和 ROC 曲线下面积（AUROC），这两项指标不依赖阈值，且比准确率更能反映归因性能。计算 AP 和 AUROC 时，我们将 DALL·E 3 测试图像与某一给定对比来源的图像对应的分类器输出及其真实标签拼接起来，并使用 scikit-learn 中的 average_precision_score 和 roc_auc_score 进行计算。

模型架构。我们使用 CLIP ViT-L/14 作为图像编码器，并从每张图像的倒数第二层提取一个 768 维特征向量。在这些冻结的表征之上，我们训练一个轻量级线性分类器，该分类器由单个全连接层和后接的 Sigmoid 激活函数组成。分类器输出图像不是由 DALL·E 3 生成的概率。

分类器使用 Adam 优化器进行训练，学习率为 10^{-3}，并采用二元交叉熵损失。我们使用较大的批次大小，以便每个训练步骤都能汇总所有可用带标签样本和野外样本的梯度。根据验证损失进行提前停止。在使用野外数据进行微调时，我们优化第 4 节所述的复合损失，并调整野外数据损失权重 λ，以保持在带标签的分布内数据上的性能。由于线性分类器的容量较低，且对 λ 进行显式控制，各次运行的训练过程都保持稳定。

主要的计算开销来自 CLIP 特征提取，该过程由图形处理器加速，即使面对大型数据集，也只需几小时即可完成。提取特征后，训练线性分类器的效率很高，通常几分钟内即可完成。

[1] https://huggingface.co/datasets/OpenDatasets/dalle-3-dataset [2] https://www.kaggle.com/c/imagenet-object-localization-challenge/overview/ 描述

5.2 主要结果

主要实验。表1总结了分布内（ID）生成器和一部分具有挑战性的生成器在纳入野外数据前后的归因性能。
即使不使用野外数据，基线分类器在大多数来源上也已表现出色，平均精度（AP）和 ROC 曲线下面积通常都超过 99%。不过，若干生成器——尤其是 Midjourney、Firefly 和 Stable Diffusion XL——仍然更具挑战性，得分明显较低。这一观察结果与 OpenAI 关于其内部 DALL·E 3 归因分类器的公开报告一致。报告指出，在区分 DALL·E 3 的输出与其他人工智能模型的输出时，分类器的性能会下降，约有 5–10% 的非 DALL·E 图像被误标为由 DALL·E 生成。[3]

这些生成器之所以难以区分，一种可能的解释是，它们可能与 DALL·E 3 共享架构组件、训练数据或风格特征，从而产生更相似的特征表示。然而，由于这些模型是闭源的，这种重叠的确切原因难以验证。因此，我们将分析重点放在提升这些较难案例的性能上。

表1：将图像归因为 DALL·E 3 或各对比来源时的平均精度（AP）和 ROC 曲线下面积，评估时分别纳入和不纳入野外数据。“SD”表示 Stable Diffusion。“cons. opt.”指约束优化，“pseudo”指伪标签。最后一列报告了具有挑战性的生成器 Midjourney、Firefly 和 Stable Diffusion XL 的平均性能。

指标	悟空	SD v1.4	SD v1.5	Midjourney	Firefly	SD XL	平均值（严格）
平均精度（不含野生样本）	0.9959	0.9848	0.9974	0.9198	0.9284	0.8604	0.9029
平均精度（伪样本）	**0.9967**	**0.9882**	**0.9979**	0.9328	0.9436	0.8927	0.9230
平均精度（一致性优化）	0.9936	0.9866	0.9957	**0.9346**	**0.9472**	**0.9015**	**0.9278**
*ROC*曲线下面积（不含野外数据）	0.9957	0.9856	0.9974	0.9252	0.9259	0.8619	0.9043
*ROC*曲线下面积（伪标签）	**0.9966**	**0.9882**	**0.9979**	0.9340	0.9378	0.8930	0.9216
*ROC*曲线下面积（约束优化）	0.9932	0.9866	0.9958	**0.9361**	**0.9423**	**0.9033**	**0.9272**

如表1所示，通过约束微调引入野外数据后，在 Midjourney、Firefly 和 Stable Diffusion XL 这些具有挑战性的生成器上，性能得到了一致提升。具体而言，这些生成器的平均精度从 0.9029 提高至 0.9278，平均 ROC曲线下面积从 0.9043 提高至 0.9272。同时，在分布内生成器上的性能仍然很高，仅有轻微波动，这表明引入野外数据并未明显降低对已知来源的归因准确率。
这些结果表明，约束微调能够有效利用无标签野外数据，提升对未见过的困难生成器的泛化能力，同时保持在带标签的分布内数据上的强劲性能。

我们认为，观察到的性能提升源于两种互补效应。首先，野外数据可能包含来自与困难测试样例重叠的生成器的样本，从而提供额外的接触机会，提升区分能力。其次，接触范围更广、更多样化的图像有助于分类器学习更稳健的决策边界，将 DALL·E 3 的输出与非目标内容区分开来，从而提升整体归因性能。

5.3 伪标签

我们用于纳入野外数据的框架支持多种策略。除了约束优化，我们还考虑了伪标签基线。在伪标签过程中，分类器会迭代地为其高置信度预测（使用 90% 的置信度阈值）的野外样本分配标签，并使用这些伪标签样本重新训练。训练会持续进行，直至无法获得更多置信度较高的伪标签。其他所有实验设置均与主要实验相同。

如表1所示，并且根据附录中报告的完整结果，伪标签和
约束优化均取得了出色的整体表现。伪标签的

[3]参见 https://openai.co m/index/understanding-the-source-of-what -we-see-and-hear-online/。

在分布内生成器上的表现方面，约束优化在具有挑战性的生成器 Midjourney、Firefly 和 Stable Diffusion XL 上始终取得更高的平均精度和曲线下面积。这些结果表明，伪标签主要强化标注数据中已有的模式，而约束优化更有效地将决策边界推向更难、此前未见过的样本。

5.4 与其他方法和基线方法的比较

据我们所知，本研究是首项明确探讨存在未知和未见生成器时的目标生成器归因问题的学术研究，尤其针对 DALL·E 3 等闭源商业模型。现有关于人工智能生成图像的学术文献主要聚焦于二元真假检测，这与本文研究的定向归因问题有本质区别，因此无法直接比较。

虽然可能存在用于目标生成器归因的行业系统，但其设计和评估协议仍属专有信息，尚未公开发表。此外，先前的学术研究并未探索如何利用无标签野外数据来提升开放世界条件下定向归因的鲁棒性。因此，目前没有可直接比较的基线方法能够同时处理目标生成器归因、未知生成器和野外数据利用。我们认为本研究是弥补这一空白的一步，并希望它能推动未来基准和未知生成器感知归因方法的发展。

5.5 消融研究

我们开展了一系列消融研究，以考察本方法在不同训练配置下的鲁棒性和表现。在所有设置中，我们始终观察到，通过约束优化纳入无标签野外数据，能够提升本方法在具有挑战性的生成器上的归因性能，同时保持较强的分布内（ID）性能。

野外数据规模的影响。 我们首先研究野外数据集规模的影响，同时保持带标签的分布内数据集规模不变（DALL·E 3：200 个样本；悟空、Stable Diffusion v1.4 和 v1.5：各 67 个样本）。将每个生成器的野外样本数量从较小规模增加到建议的较大规模，可带来稳定但幅度适中的性能提升，尤其是在 Midjourney、Firefly 和 Stable Diffusion XL 等具有挑战性的生成器上。

这些观察结果表明，扩大野外数据集规模能够进一步提升泛化能力，不过随着数据集增大，边际收益会逐渐减小。这一现象说明，基于 CLIP 表征构建的线性分类器在低数据场景中已经能够捕捉到强有力的判别信号，凸显了预训练特征在归因任务中的有效性。

值得注意的是，即使野外数据集的规模远超带标签的分布内数据集，训练仍然保持稳定。这种稳定性源于按样本归一化损失，并在优化过程中明确约束分布内损失；二者共同确保了模型能够抵御数据集不平衡。

野外数据中错误标记的目标生成器样本的影响。为评估对标签噪声的鲁棒性，我们改变野外数据混合集中的目标生成器图像（DALL·E 3）相对于其他来源的占比，其中还包括一种极端设置，即目标图像与非目标图像的比例大致相等。在所有设置下，所提出的方法始终保持了 ID 性能，同时提高了对具有挑战性的生成器的归因准确率。
即使野外数据中有相当一部分来自目标生成器，与不使用野外数据进行训练相比，该方法在困难来源上的归因性能仍有所提升。曲线下面积的变化趋势与平均精度的提升基本一致。这些结果表明，受约束的目标函数能够有效减轻标签噪声的影响，并使模型即使在来源构成不完美的情况下，也能从野外数据中获益。

更大的带标签分布内数据集的影响。 接下来，我们考察随着可用的带标签训练数据增多，纳入野外数据所带来的收益是否会减弱。我们逐步扩大分布内数据集的规模，从适度扩大的设置到大幅扩大的设置，其中每个生成器都有数千个带标签样本。

在这些数据规模下，使用野外数据进行微调仍能带来提升，尤其是对于 Midjourney 和 Stable Diffusion XL 等具有挑战性的生成器。这表明，所提出的框架在低数据量场景之外仍然有效，并为利用无标签数据提升泛化能力提供了一种可扩展的机制。

改变 ID 数据集中的生成器多样性。 为了进一步评估不同训练配置下的鲁棒性，我们通过在监督训练中直接纳入真实图像（例如来自 ImageNet 的图像）或 Midjourney 等具有挑战性的生成器，来改变带标签 ID 数据集的组成。人们或许会认为，接触这类多样化或难以应对的数据源会降低野外数据的边际效用，但我们发现，使用野外数据进行微调仍能提升泛化性能。

这些结果表明，野外数据能够为带标签数据集提供互补的覆盖范围，这可能归因于其分布范围更广、异质性更强。

ID 损失约束阈值的影响。 最后，我们研究了放宽带标签 ID 损失约束的影响。主要实验将 ID 损失限制在基线值的两倍以内；若允许更大的偏差，则在具有挑战性的生成器上会取得更显著的提升，但在其他一些数据源上的性能会略有下降。

这种权衡凸显了我们框架的灵活性：通过调整约束阈值，实践者可以根据部署需求，在提升对未知生成器的泛化能力与更严格地维持分布内性能之间进行明确权衡。

5.6 讨论

我们的实验表明，即使在标注数据较少的情况下，使用无标签野外数据进行约束微调，也能持续提升模型对未见过且具有挑战性的图像生成器的归因鲁棒性。该方法在广泛的训练配置下均保持有效，且能随着野外数据量的增加而取得更好的效果，与伪标签等替代策略相比也表现出优势。这些特性使其非常适用于现实部署场景：生成模型快速演进，而获取标注数据的成本高昂。

我们的实证分析得出以下几点发现：

- **大型野外数据集下的稳定性**。通过按样本归一化损失，并对分布内（ID）损失施加明确约束，即使野外数据集的规模远超已标注 ID 数据集，优化过程仍能保持稳定。
- **对野外数据中标签噪声的鲁棒性**。将野外样本视为非目标样本不可避免地会引入标签噪声，因为部分野外图像可能源自目标生成器。尽管如此，对 ID 损失施加明确约束能够减轻对这类噪声的过拟合，使模型在保持标注数据准确率的同时，受益于接触多样化的野外样本。
- **通过约束调节实现可控权衡**。 放宽对带标签分布内损失的约束（例如，从 $2\times$ 调整为 $3\times$ 基线损失），可以在保持分布内性能与提升对未知生成器的泛化能力之间实现有依据的权衡。这种灵活性使该方法能够适应特定部署场景的优先需求。
- **与伪标签方法的比较**。 在我们的半监督框架中，伪标签提供了一种自然的替代方案：将对野外数据的高置信度预测迭代地纳入标签。虽然伪标签略微提升了在分布内生成器上的性能，但约束优化在最具挑战性的未见生成器上始终带来更大的提升，这表明它能更有效地将决策边界推向更难处理的样本。

局限性。 所提出方法的有效性取决于野外数据的多样性。当野外数据集存在严重偏差，或未覆盖具有挑战性的来源时，所获得的泛化增益可能有限。此外，我们的实验使用 CLIP ViT-L/14 作为特征骨干网络；探索其他表征模型或可进一步提升归因性能。最后，我们未探讨对对抗性操纵或对生成器特定签名进行故意混淆的鲁棒性，这将作为未来研究的重要方向。

6 结论

我们研究了如何利用无标签野外数据提升目标生成器归因。尽管基于 CLIP 的表征结合线性分类器在许多来源上表现出色，但要准确区分目标生成器与某些具有挑战性的生成器（如 Midjourney、Firefly 和 Stable Diffusion XL）仍然困难。为应对这一挑战，我们提出了一种约束优化框架，在纳入未标注野生图像的同时，明确保持模型在带标签的分布内数据上的性能。

大量实验表明，所提出的方法在不牺牲已知来源准确率的前提下，始终能够提升对未见过的、难以识别的生成器的泛化能力。无论野外数据规模、标注数据配置如何变化，还是采用伪标签等其他半监督策略，这些提升都具有稳健性。通过按样本归一化损失，并明确约束分布内性能，即使野外数据量远远超过标注数据量，该方法仍能保持稳定。

更广泛地说，我们的框架能够有效利用无标签数据，无需人工重新标注，并支持在新的生成模型出现时进行增量微调。这些特性使其非常适用于生成器生态快速演变的真实开放世界图像归因场景。

参考文献

[1] Yang Song 和 Stefano Ermon。通过估计数据分布的梯度进行生成建模。《神经信息处理系统进展》，32，2019。[2] Aditya Ramesh、Mikhail Pavlov、Gabriel Goh、Scott Gray、Chelsea Voss、Alec Radford、Mark Chen 和 Ilya Sutskever。零样本文本到图像生成。载于国际机器学习会议。PMLR，2021。[3] Robin Rombach、Andreas Blattmann、Dominik Lorenz、Patrick Esser 和 Björn Ommer。利用潜在扩散模型进行高分辨率图像合成。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第10684–10695页，2022。[4] Arash Vahdat、Karsten Kreis 和 Jan Kautz。在潜在空间中进行基于分数的生成建模。《神经信息处理系统进展》，34：11287–11302，2021。[5] Aditya Ramesh、Prafulla Dhariwal、Alex Nichol、Casey Chu 和 Mark Chen。利用CLIP潜变量进行分层文本条件图像生成。*arXiv*预印本 *arXiv:2204.06125*，1(2)：3，2022。[6] Lucy Chai、David Bau、Ser-Nam Lim 和 Phillip Isola。是什么让伪造图像可被检测？理解具有泛化能力的特性。载于《计算机视觉—*ECCV 2020*：第*16*届欧洲会议，格拉斯哥，英国，*2020*年*8*月*23*日至*28*日，论文集，第二十六部分*16*》，第103–120页。施普林格，2020。[7] Sheng-Yu Wang、Oliver Wang、Richard Zhang、Andrew Owens 和 Alexei A Efros。CNN生成的图像出奇地容易识别……至少目前如此。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第8695–8704页，2020。[8] Utkarsh Ojha、Yuheng Li 和 Yong Jae Lee。迈向能够跨生成模型泛化的通用伪造图像检测器。第24480–24489页，2023。[9] Quentin Bammey。Synthbuster：迈向扩散模型生成图像的检测。*IEEE*开放信号处理期刊，5：1–9，2024。[10] James Betker、Gabriel Goh、Li Jing、Tim Brooks、Jianfeng Wang、Linjie Li、Long Ouyang、Juntang Zhuang、Joyce Lee、Yufei Guo 等。通过改进图像说明提升图像生成效果。计算机科学。*https://cdn. openai. com/papers/dall-e-3. pdf*，2(3)：8，2023。[11] Alec Radford、Jong Wook Kim、Chris Hallacy、Aditya Ramesh、Gabriel Goh、Sandhini Agarwal、Girish Sastry、Amanda Askell、Pamela Mishkin、Jack Clark 等。通过自然语言监督学习可迁移视觉模型。载于国际机器学习会议，第8748–8763页。PMLR，2021。[12] Ian Goodfellow、Jean Pouget-Abadie、Mehdi Mirza、Bing Xu、David Warde-Farley、Sherjil Ozair、Aaron Courville 和 Yoshua Bengio。生成对抗网络。《神经信息处理系统进展》，27，2014。[13] Yunjey Choi、Minje Choi、Munyoung Kim、Jung-Woo Ha、Sunghun Kim 和 Jaegul Choo。StarGAN：用于多域图像到图像转换的统一生成对抗网络。载于*IEEE*计算机视觉与模式识别会议论文集，第8789–8797页，2018。[14] Jun-Yan Zhu、Taesung Park、Phillip Isola 和 Alexei A Efros。使用循环一致对抗网络进行非配对图像到图像转换。载于*IEEE*国际计算机视觉会议论文集，第2223–2232页，2017。[15] Andrew Brock、Jeff Donahue 和 Karen Simonyan。用于高保真自然图像合成的大规模GAN训练。*arXiv*预印本 *arXiv:1809.11096*，2018。[16] Haoyue Bai、Ceyuan Yang、Yinghao Xu、S-H Gary Chan 和 Bolei Zhou。通过生成式插值提升分类器的分布外鲁棒性。*arXiv*预印本 *arXiv:2307.12219*，2023。

[17] Niki Parmar、Ashish Vaswani、Jakob Uszkoreit、Lukasz Kaiser、Noam Shazeer、Alexander Ku 和 Dustin Tran. 图像 Transformer. 载于国际机器学习会议，第 4055–4064 页. PMLR，2018年.[18] Bowen Zhang、Shuyang Gu、Bo Zhang、Jianmin Bao、Dong Chen、Fang Wen、Yong Wang 和 Baining Guo. Styleswin：用于高分辨率图像生成的基于 Transformer 的生成对抗网络. 载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第 11304–11314 页，2022年.[19] Huiwen Chang、Han Zhang、Jarred Barber、AJ Maschinot、Jose Lezama、Lu Jiang、Ming-Hsuan Yang、Kevin Murphy、William T Freeman、Michael Rubinstein 等. Muse：通过掩码生成式 Transformer 实现文本到图像生成. *arXiv* 预印本 *arXiv:2301.00704*，2023年.[20] Alex Nichol、Prafulla Dhariwal、Aditya Ramesh、Pranav Shyam、Pamela Mishkin、Bob McGrew、Ilya Sutskever 和 Mark Chen. GLIDE：借助文本引导的扩散模型实现照片级真实感图像生成与编辑. *arXiv* 预印本 *arXiv:2112.10741*，2021年.[21] Prafulla Dhariwal 和 Alexander Nichol. 扩散模型在图像合成方面优于生成对抗网络.《神经信息处理系统进展》，34：8780–8794，2021年.[22] Yuanzhi Zhu、Zhaohai Li、Tianwei Wang、Mengchao He 和 Cong Yao. 利用扩散模型实现条件文本图像生成. 载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第 14235–14245 页，2023年.[23] Falko Matern、Christian Riess 和 Marc Stamminger. 利用视觉伪影揭露深度伪造和人脸篡改. 载于*2019*年*IEEE*冬季计算机视觉应用研讨会（*WACVW*），第 83–92 页. IEEE，2019年.[24] Scott McCloskey 和 Michael Albright. 利用颜色线索检测生成对抗网络生成的图像. *arXiv* 预印本 *arXiv:1812.08247*，2018年.[25] Francesco Marra、Diego Gragnaniello、Luisa Verdoliva 和 Giovanni Poggi. 生成对抗网络会留下人工指纹吗？载于*2019*年*IEEE*多媒体信息处理与检索会议（*MIPR*），第 506–511 页. IEEE，2019年.[26] Xu Zhang、Svebor Karaman 和 Shih-Fu Chang. 检测和模拟生成对抗网络伪造图像中的伪影. 载于*2019*年*IEEE*信息取证与安全国际研讨会（*WIFS*），第 1–6 页. IEEE，2019年.[27] Ning Yu、Larry S Davis 和 Mario Fritz. 将伪造图像溯源至生成对抗网络：学习并分析 GAN 指纹. 载于*IEEE/CVF*国际计算机视觉会议，第 7556–7566 页，2019年.[28] Riccardo Corvi、Davide Cozzolino、Giada Zingarini、Giovanni Poggi、Koki Nagano 和 Luisa Verdoliva. 关于检测由扩散模型生成的合成图像. 载于*ICASSP 2023-2023*年*IEEE*声学、语音与信号处理国际会议（*ICASSP*），第 1–5 页. IEEE，2023年.[29] Davide Cozzolino、Justus Thies、Andreas Rössler、Christian Riess、Matthias Nießner 和 Luisa Verdoliva. Forensictransfer：用于伪造检测的弱监督域适应. *arXiv* 预印本 *arXiv:1812.02510*，2018年.[30] Haoyue Bai、Yiyou Sun、Wei Cheng 和 Haifeng Chen. 生成式时代，责任何在？对人工智能生成内容进行基于恢复的黑盒检测. 载于计算机视觉与模式识别会议论文集，第 28821–28830 页，2025年.[31] Zeyang Sha、Zheng Li、Ning Yu 和 Yang Zhang. DE-FAKE：检测并溯源由文本到图像生成模型生成的伪造图像. 载于*2023*年*ACM SIGSAC*计算机与通信安全会议论文集，第 3418–3432 页，2023年.[32] Zhendong Wang、Jianmin Bao、Wengang Zhou、Weilun Wang、Hezhen Hu、Hong Chen 和 Houqiang Li. 用于扩散模型生成图像检测的 DIRE. 载于*IEEE/CVF*国际计算机视觉会议论文集，第 22445–22455 页，2023年。

[33] Hany Farid. 文本生成绘画的光照一致性. *arXiv* 预印本 *arXiv:2207.13744*，2022年.
[34] Hany Farid. 文本生成绘画的透视一致性. *arXiv* 预印本 *arXiv:2206.14617*，2022年.
[35] Lydia Abady、Jun Wang、Benedetta Tondi 和 Mauro Barni. 用于合成图像开放集架构归因的孪生网络验证系统. 《模式识别快报》，180，2024年.[36] S. Girish、S. Suri、S. S. Rambhatla 和 A. Shrivastava. 探索开放世界生成对抗网络生成图像的发现与归因. 载于*IEEE/CVF*国际计算机视觉会议，第 14094–14103 页，2021年.[37] J. Wang、O. Alamayreh、B. Tondi 和 M. Barni. 通过基于视觉变换器的混合架构对基于生成对抗网络的图像篡改进行开放集分类. 载于*IEEE/CVF*计算机视觉与模式识别会议，第 953–962 页，2023年.[38] T. Yang、D. Wang、F. Tang、X. Zhao、J. Cao 和 S. Tang. 面向开放集模型归因的渐进式开放空间扩展. 载于*IEEE/CVF*计算机视觉与模式识别会议，2023年.[39] J. Wang、B. Tondi 和 M. Barni. BOSC：一种基于后门的开放集合成图像归因框架. 《*IEEE*信息取证与安全汇刊》，2025年.[40] Xiaojin Jerry Zhu. 半监督学习文献综述. 2005年.
[41] Jesper E. van Engelen 和 Holger H. Hoos. 半监督学习综述. 《机器学习》，109(2)：373–440，2020年2月.[42] Yassine Ouali、Céline Hudelot 和 Myriam Tami. 深度半监督学习概述，2020年.[43] Kihyuk Sohn、David Berthelot、Nicholas Carlini、Zizhao Zhang、Han Zhang、Colin A Raf- fel、Ekin Dogus Cubuk、Alexey Kurakin 和 Chun-Liang Li. FixMatch：通过一致性和置信度简化半监督学习. 载于 H. Larochelle、M. Ranzato、R. Hadsell、M.F. Balcan 和 H. Lin 编，《style id='28'>《神经信息处理系统进展》》，第 33 卷，第 596–608 页. Curran Associates, Inc.，2020年.[44] David Berthelot、Nicholas Carlini、Ekin D. Cubuk、Alex Kurakin、Kihyuk Sohn、Han Zhang 和 Colin Raffel. RemixMatch：结合分布匹配与增强锚定的半监督学习. 载于国际学习表征会议，2020年.[45] Bowen Zhang、Yidong Wang、Wenxin Hou、HAO WU、Jindong Wang、Manabu Okumura 和 Takahiro Shinozaki. FlexMatch：通过课程式伪标签提升半监督学习. 载于 M. Ranzato、A. Beygelzimer、Y. Dauphin、P.S. Liang 和 J. Wortman Vaughan 编，《style id='34'>《神经信息处理系统进展》》，第 34 卷，第 18408–18419 页. Curran Associates, Inc.，2021年.[46] Yidong Wang、Hao Chen、Qiang Heng、Wenxin Hou、Yue Fan、Zhen Wu、Jindong Wang、Marios Savvides、Takahiro Shinozaki、Bhiksha Raj 等. FreeMatch：半监督学习的自适应阈值方法. *arXiv* 预印本 *arXiv:2205.07246*，2022年.[47] Hao Chen、Ran Tao、Yue Fan、Yidong Wang、Jindong Wang、Bernt Schiele、Xing Xie、Bhiksha Raj 和 Marios Savvides. SoftMatch：解决半监督学习中数量与质量的权衡，2023年.[48] Zhaowei Cai、Avinash Ravichandran、Paolo Favaro、Manchen Wang、Davide Modolo、Rahul Bhotika、Zhuowen Tu 和 Stefano Soatto. 大规模半监督视觉变换器. 载于 Alice H. Oh、Alekh Agarwal、Danielle Belgrave 和 Kyunghyun Cho 编，《style id='41'>《神经信息处理系统进展》》，2022年.[49] Manuel Lagunas、Brayan Impata、Victor Martinez、Virginia Fernandez、Christos Georgakis、Sofia Braun 和 Felipe Bertrand. 利用半监督学习和视觉变换器进行细粒度分类的迁移学习. *arXiv* 预印本 *arXiv:2305.10018*，2023年。

[50] Zhen Xing、Qi Dai、Han Hu、Jingjing Chen、Zuxuan Wu 和 Yu-Gang Jiang。Svformer：用于动作识别的半监督视频变换器。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第 18816–18826 页，2023年。[51] Yidong Wang、Hao Chen、Yue Fan、Wang SUN、Ran Tao、Wenxin Hou、Renjie Wang、Linyi Yang、Zhi Zhou、Lan-Zhe Guo、Heli Qi、Zhen Wu、Yu-Feng Li、Satoshi Nakamura、Wei Ye、Marios Savvides、Bhiksha Raj、Takahiro Shinozaki、Bernt Schiele、Jindong Wang、Xing Xie 和 Yue Zhang。USB：用于分类的统一半监督学习基准。载于第三十六届神经信息处理系统会议数据集与基准专题，2022年。[52] Gilles Blanchard、Gyemin Lee 和 Clayton Scott。半监督新颖性检测。《机器学习研究杂志》，11：2973–3009，2010年。[53] Julian Katz-Samuels、Julia B Nakhleh、Robert Nowak 和 Yixuan Li。在自然环境中训练 OOD 检测器。载于国际机器学习会议，第 10848–10865 页。PMLR，2022年。[54] Haoyue Bai、Gregory Canal、Xuefeng Du、Jeongyeol Kwon、Robert D Nowak 和 Yixuan Li。一石二鸟：利用野外数据实现分布外泛化与检测。载于国际机器学习会议。PMLR，2023年。[55] Haoyue Bai、Jifan Zhang 和 Robert Nowak。Aha：人类辅助的分布外泛化与检测。《神经信息处理系统进展》，37：33863–33890，2024年。[56] Haoyue Bai、Xuefeng Du、Katie Rainey、Shibin Parameswaran 和 Yixuan Li。利用人类反馈进行分布外学习。*arXiv* 预印本 *arXiv:2408.07772*，2024年。[57] Davide Cozzolino、Giovanni Poggi、Riccardo Corvi、Matthias Nießner 和 Luisa Verdoliva。借助 CLIP 提升人工智能生成图像检测的水平。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第 4356–4366 页，2024年。[58] Mingjian Zhu、Hanting Chen、Qiangyu Yan、Xudong Huang、Guanyu Lin、Wei Li、Zhijun Tu、Hailin Hu、Jie Hu 和 Yunhe Wang。GenImage：用于检测人工智能生成图像的百万规模基准，2023年。