arXiv:2303.09272v1 [cs.LG] 15 Mar 2023

生成式人工智能的版权保护与问责：攻击、水印与归因

Haonan Zhong[1§]、Jiamin Chang[1§]、Ziyue Yang[2§]、Tingmin Wu[3],Pathum Chamikara Mahawaga Arachchige[3], Chehara Pathmabandu[3], 和Minhui Xue[3]1新南威尔士大学，澳大利亚2澳大利亚国立大学，澳大利亚3澳大利亚联邦科学与工业研究组织 Data61

摘要

近年来，生成式人工智能（例如生成对抗网络——GAN）日益普及。然而，生成式人工智能引发了重大关切：如何保护其生成的图像（以及有害图像所涉及的模型问责）和模型（以及中毒模型所涉及的模型问责）相关的知识产权。在本文中，我们提出了一种评估框架，旨在全面概述当前生成对抗网络版权保护措施的现状，评估这些措施在各种生成对抗网络架构上的性能，并找出影响其性能的因素及未来的研究方向。我们的研究结果表明，针对输入图像的当前知识产权保护方法、模型水印和归属网络，对于广泛的生成对抗网络而言总体上令人满意。我们强调，应进一步关注训练集保护，因为当前方法无法为训练集提供稳健的知识产权保护和来源追踪。

CCS 概念

• **安全　与隐私 → 软件和应用安全　　性。**

ACM 参考格式：
Haonan Zhong、Jiamin Chang、Ziyue Yang、Tingmin Wu、Pathum Chamikara Mahawaga Arachchige、Chehara Pathmabandu 和 Minhui Xue。2023年。《生成式人工智能的版权保护与问责：攻击、水印和归因》。载于《*ACM Web Conference 2023* 伴随论文集（*WWW '23 Companion*）》，*2023*年*4*月*30*日至*5*月*4*日，美国德克萨斯州奥斯汀。ACM，纽约，纽约州，美国，5页。https://doi.org/10.1145/3543873.3587321

1 引言

近年来，生成式人工智能（例如生成对抗网络——生成对抗网络）领域取得了显著进展。这些进展提升了生成式人工智能生成高度逼真内容（如艺术作品和图像）的能力 [20]。因此，生成式人工智能在科研和产业领域得到了广泛应用，也引发了人们对知识产权保护的重大担忧。基于生成对抗网络的图像生成模型使用激增，可用于创作侵犯现有版权、欺骗安全系统的内容，并通过散布虚假信息损害内容所有者的权利，影响个人声誉。因此，人工智能生成图像正因涉嫌版权侵权而面临诉讼。Getty Images 与一群艺术家起诉 AIart 生成器，是涉及人工智能生成图像的著名法律纠纷案例之一。[1] 此外，生成最先进的生成模型需要大量计算资源和大型数据集。这些问题促使人们日益关注如何保护和验证生成对抗网络模型及训练集的所有权。

此前针对图像和生成对抗网络的版权保护方法包括攻击（例如对抗噪声）、水印和归因技术。然而，相关文献主要关注深度神经网络（DNN），对开发生成对抗网络的版权保护框架关注有限 [1]。

我们的研究旨在通过全面分析现有生成对抗网络版权保护方法的有效性来填补这一空白。我们提出的框架通过分析各种生成对抗网络模型，帮助确定影响版权保护技术有效性的关键因素。为实现这一目标，我们探讨并回答以下三个研究问题。

• **研究问题1：**当前的对抗攻击能否有效防止输入图像遭到版权侵害?

• **研究问题2：**能否通过水印保护并验证生成对抗网络及其训练集的知识产权?

• **研究问题3：**当前的归因方法能否有效支持知识产权保护，并通过对输出图像进行来源归因，追溯有害生成式人工智能模型的问责责任?

2 实验设置

本节详细介绍了本文的实验设置。图1以示意图的形式展示了用于评估生成对抗网络知识产权保护方法的框架。为了使用所提出的评估框架评估不同知识产权保护方法的有效性，我们采用了多个数据集，包括用于无条件翻译模型的 Horse2Zebra [21]，以及用于条件转换和生成模型的 CelebA [12] 和 CelebAHQ [11]。我们的主要目标之一是

[§]前三位作者贡献相同，且本研究在澳大利亚 CSIRO 的 Data61 完成。

允许免费制作本作品的全部或部分内容的数字或纸质副本，供个人或课堂使用，但前提是这些副本不得用于营利或商业利益，也不得为此目的进行分发，并且副本首页须附有本声明和完整引文。对于本作品中由作者以外的其他人拥有版权的部分，必须尊重其版权。注明出处后可以摘录。若要以其他方式复制、再版、发布到服务器或分发至列表，则须事先获得特别许可和/或支付费用。请向 permissions@acm.org 申请许可。*WWW '23 Companion，2023*年*4*月*30*日至*5*月*4*日，美国德克萨斯州奥斯汀 © 2023年版权归所有者/作者所有。出版权已授权给 ACM。
ACM ISBN 978-1-4503-9419-2/23/04...$15.00
https://doi.org/10.1145/3543873.3587321

[1]https://www.businessinsider.com/ai-art-artists-getty-images-lawsuits-stable-diffusion-2023-1?op=1

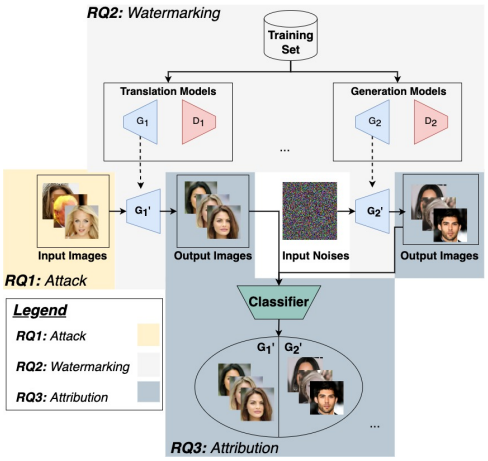


图1：所提出的生成对抗网络、图像版权保护与问责方法评估框架。

表1：针对不同生成对抗网络的不可感知对抗攻击性能。我们遵循了[15]中采用的实验配置，并在所有实验中将正则化系数（ϵ）设为0。05，以确保公平比较。FID：Fréchet Inception 距离 [7]。

模型	数据集	L1	L2	FID
CUT	Horse2Zebra	0.5229	0.4278	37.2135
AttentionGAN		0.5590	0.6153	227.4701
CycleGAN		0.4512	0.3084	74.4684
StarGANv2	CelabAHQ 256×256	0.3553	0.2901	20.1240
StarGAN		1.0459	1.3470	251.7915
StarGAN	CelabA 128×128	0.9288	1.1167	166.7310

本研究旨在对生成对抗网络的版权保护技术及其相关输入集（图像）和输出进行基准测试。我们针对以下生成对抗网络架构开展了此项分析 —*(i)* **生成模型：** StyleGANv2 [10], StyleGAN3 [9], StyleSwin [20],DDGAN [17]; *(ii)* **翻译模型：** *(a)* 无条件： Cycle- GAN [21], AttentionGAN-v2 (AttGAN) [16], CUT [14]; *(b)* 条件式： StarGAN [2], StarGAN-v2 [3]。

3 研究问题1：对抗攻击

保护输入图像的版权对于防止利用生成对抗网络恶意修改图像至关重要。一种颇具前景的应对伪造图像生成的方法，是对输入图像施加对抗扰动。

相关工作。 关于将不可见的对抗扰动应用于输入图像，以阻止生成模型被非法用于生成虚假图像，已有大量讨论。先前研究已成功对无条件和有条件图像转换模型实施攻击。因此，对抗攻击已成为一种很有前景的工具，可确保受版权保护的图像不会遭到生成对抗网络的未经授权修改。因此，我们考察当前的对抗水印能否有效保护输入图像的版权，防止其被非法使用。为此，我们采用 Ruiz 等人的研究 [15] 所提出的方法，使用通过投影梯度下降（PGD）算法生成的对抗水印，并以模型输出作为训练依据。该技术已广泛应用于图像水印领域，

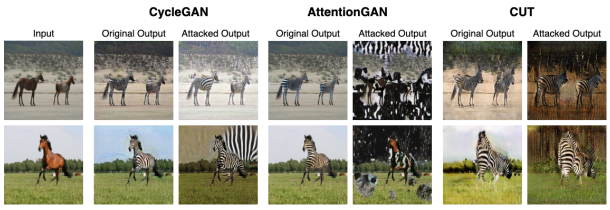


图2：使用 Horse2Zebra 数据集的无条件翻译生成对抗网络的原始输出与攻击后输出。

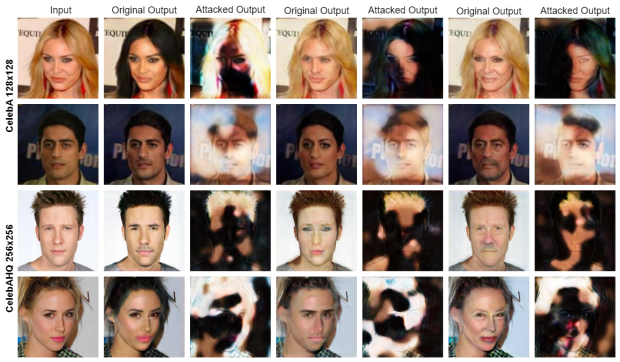


图3：不同图像质量下 StarGAN 的原始输出与攻击后输出。



图4：使用 CelebAHQ 数据集的 StarGANv2 原始输出与攻击后输出。

并且常被用作许多近期攻击中的基准方法之一 [8]。

实验结果与讨论。 我们使用对抗水印生成对抗样本，以攻击尚未经过定性或定量评估的模型。我们通过计算生成网络的原始输出与使用对抗输入样本生成的输出之间的 L1 范数、L2 范数和 FID [7]，来衡量扰动程度。

如表1所示，对抗攻击会显著扰乱输出。然而，其中涉及多种复杂因素的相互作用，此前尚未得到研究。因此，为了更深入地了解原始图像与生成图像之间的相似度，我们结合 FID [7]、L1 和 L2 范数进行计算。实验结果表明，与针对低分辨率图像的攻击相比，针对高分辨率图像的对抗攻击具有更好的攻击性能（图3）。此外，性能更好的模型（AttentionGAN 和 CycleGAN）会造成更容易察觉的扰动（图2）。

图 2 还显示，AttentionGAN 受到的失真最为严重，图像背景受到的影响尤为显著。我们推测，这与 AttentionGAN 的注意力机制有关，并且是由于存在

噪声干扰了模型区分目标物体与嘈杂背景的能力。

由于 FID 值较低，StarGANv2 生成的输出引起了广泛关注。图4表明，当人脸清晰可见时，对抗攻击造成的失真较小。不过，在这两种情况下，源输入的风格都得到了保留。因此，我们认为，这种干扰更针对源特征保留 [3]，而对阻止风格重建的效果较弱。不过，要证实这一假设，还需要进一步研究相关损失函数。此外，这也表明，当图像风格与源图像相似时，FID 在评估失真方面存在局限性。

研究问题1. 要点：
• 近期，基于不可见对抗攻击的输入图像水印取得了进展，有助于有效防止图像遭到未经授权的使用。• 我们需要新的指标来全面评估攻击的有效性，因为现有指标在准确反映失真方面存在局限性。• 与图像到图像转换中侧重于源特征保留的应用相比，对抗攻击水印在侧重于风格迁移的应用中效果可能较差。

4 研究问题2：水印

鉴于开发先进生成对抗网络架构、收集全面的数据集并严格训练模型所需投入的成本和资源，水印对于生成对抗网络模型和训练集的版权保护至关重要。生成对抗网络水印的目标是将水印嵌入模型中，之后再从生成内容中提取出来。如果发生未经授权使用受版权保护的模型和数据集的情况，这些水印可作为证据。

相关工作。 实现水印有两种潜在方法。一种经过广泛研究的方法是利用深度神经网络创建水印网络。该方法包括在训练模型时训练水印编码器以嵌入水印，并训练解码器以从生成内容中提取水印。Yu 等人 [19] 采用这一思路，先训练编码器和解码器，再将指纹叠加到用于生成对抗网络训练的训练集上。水印将附加到生成对抗网络中，因此可以在生成图像中检测到指纹。Fei 等人 [5] 通过将解码器和编码器集成到生成对抗网络训练中，扩展了这一方法。在我们的分析中，我们选择了Yu 等人的方法 [19] 进行评估，因为该方法为训练集和模型都提供了知识产权保护策略。

Ong等人 [13] 提出了另一种方案：将所有权信息作为正则化项嵌入模型，使得向可疑模型输入触发输入后，模型会生成带有水印的输出。这实质上为版权所有者创建了一个后门，使其能够利用触发图像验证黑盒API是否使用了他们的模型。

实验结果与讨论。 我们采用 Yu 的方法 [19]作为基于卷积神经网络的水印技术的代表，并且

(a) Bit Accuracy on Different Models

Model	Bit Accuracy
CycleGAN	0.4851
CUT	0.5448
AttentionGANv2	0.4820
StarGAN	0.4960
StyleGANv2	0.7621

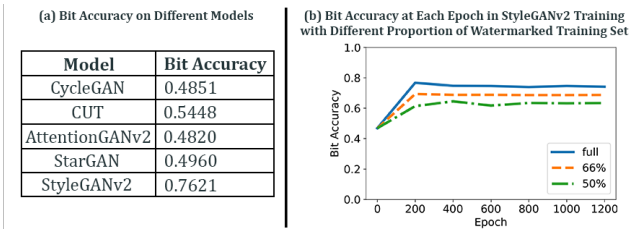


图5：基于 Yu 等人[19]的方法进行水印检测时的比特准确率。(a) 在 Horse2Zebra（Cycle-GAN、CUT 和 AttentionGAN-v2）及 CelebA（StarGAN 和 Style-GANv2）上提取水印的比特准确率。(b) 生成对抗网络训练不同轮次下的比特准确率。

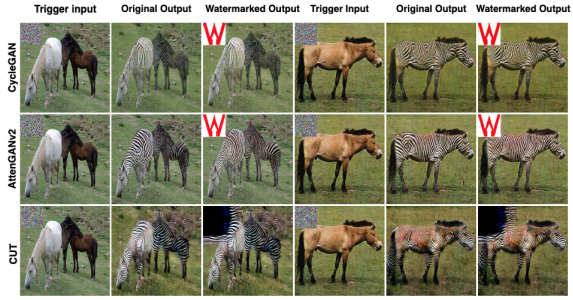


图6：在无条件翻译模型中采用水印注入方法得到的原始图像、触发图像和带水印图像。

采用 Ong [13] 创建模型水印后门。在水印的实施过程中，我们评估了其在多个训练轮次中的性能，以确定最佳配置。

图5展示了在不同模型上进行数据集水印的结果。我们使用带水印数据集训练模型，并进行了水印检测。我们发现，为翻译模型添加水印颇具挑战，而生成模型的比特准确率有所提高。我们推测，这是因为与翻译模型相比，生成模型对原始训练集的反映程度更高。此外，无论训练集中带水印图像的比例如何，生成对抗网络生成模型训练过程中，水印往往会在早期训练阶段被学习到。这一特性很重要，因为它表明，为模型添加水印的工作量可能低于预期：只需在带水印数据集上重新训练，当前模型就很容易被添加水印。

我们还在翻译模型上采用了正则化方法，并使用触发输入测试了全部三种无条件翻译模型的性能。总体而言，正则化方法得到的结果表明，触发图像能够成功生成带水印输出，且不会影响模型的性能。然而，CUT 模型无法重现水印图像。我们推测这是由其逐块训练机制导致的。这个问题凸显了深入研究当前水印方法的局限性对于实现知识产权保护的重要性。

表 2：使用 Horse2Zebra 并通过频谱分析对“翻译生成对抗网络”进行归因的混淆矩阵（行：真实值，列：预测值）。

	AttGAN	*CUT*	*CycleGAN*	真实
AttGAN	**209**	3	0	25
CUT	5	**221**	7	4
CycleGAN	0	2	**233**	2
Real	19	2	0	**216**
精确率	**0.8970**	**0.9693**	**0.9708**	**0.8745**

表 3：使用频谱分析对“生成对抗网络”在 CelebAHQ 上生成的 5000 张图像进行归因的混淆矩阵（行：真实值，列：预测值）。

	DD-GAN	*StarGAN*	*StarGANv2*	*StyleGAv2*	*StyleSwin*	真实
DD-GAN	**934**	13	0	33	0	20
StarGAN	4	**996**	0	0	0	0
StarGANv2	0	1	**992**	0	3	4
StyleGANv2	3	1	0	**992**	2	2
StyleSwin	0	0	12	10	**962**	16
Real	36	6	43	61	77	**777**
精确率	**0.956**	**0.9794**	**0.9475**	**0.9051**	**0.9215**	**0.9487**

研究问题2. 要点：

- 水印技术的最新进展提升了对生成对抗网络模型的保护。然而，训练集的知识产权保护仍是一个普遍关注的问题。
- 在带水印训练集上训练的生成模型，其水印检测能力优于翻译模型。

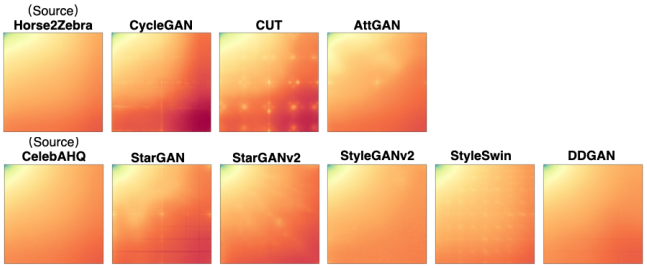
5 研究问题3：归因

生成对抗网络架构的归因在版权保护中也发挥着至关重要的作用。它使知识产权所有者能够区分真实图像和伪造图像，并根据底层模型的架构特征对伪造图像进行归因，从而确定其源模型。归因方法主要分为两大类：频谱归因和对比归因。

相关工作。 研究已证明，频谱分析在深度伪造检测和模型归因中非常有效。因此，研究者提出并验证了基于离散余弦变换（DCT）[6] 的检测器。对比归因是另一种用于确定生成对抗网络模型对生成图像所做修改的方法。该方法通过分析原始图像与生成图像之间的局部特征差异来实现。Yang等人 [18] 探讨了逐块对比学习，以基于模型指纹生成预测。

实验结果与讨论。 按照 Frank 等人的研究方法并拓展其研究范围 [6],，我们生成了用于归因频谱分析的混淆矩阵。在这项分析中，我们针对每个图像转换模型，从 Horse2Zebra 数据集中生成了 237 张伪造图像；针对每个生成模型，从 CelebAHQ 中生成了 5,000 张伪造图像。

结果表明，在不同数据集上进行训练时，频谱分析生成的准确率很高。这一结果符合预期，因为离散余弦变换会为每个模型生成不同的指纹（图7）。实证结果进一步证实，能力更强的模型在生成更逼真的伪造图像方面性能更佳（参见表3）。



F图 7：使用 D 生成对抗网络生成的频谱印记 CT.

我们正在对 Dong 等人 [4]提出的对比方法和归因规避进行持续实验。然而，规避流程需要修改网络，这与知识产权保护期间水印必须保持完整的要求相矛盾。因此，作者没有动力将频谱损失纳入其模型。

研究问题3：要点：

- 利用频域指纹的分类器可以通过对输出图像进行归因，实现对模型的知识产权保护。
- 当模型生成的输出越逼真时，频谱特征就越难追踪。

6 结论

本研究初步评估了针对生成对抗网络及其训练集（图像）的最先进知识产权保护方法。实验结果表明，近期开发的知识产权保护技术，如对抗攻击、水印和归因，在保护版权以及追踪生成对抗网络模型的问责方面大体有效。然而，我们的研究结果表明，当前用于保护训练集（原始受版权保护图像）知识产权的技术尚不充分。我们还强调，进一步分析近期能够利用知识产权所有者未经授权的风格和作品生成逼真图像的生成式人工智能模型十分重要。此外，本文强调有必要开发统一流程，以评估对抗攻击和水印的性能，并衡量生成式人工智能的知识产权保护水平和问责追踪能力。我们正在开展的研究旨在整合多个领域中针对大量生成式人工智能模型的知识产权保护研究成果，从而制定一套全面的指标。本项初步研究得出的发现将为立法者和从业者采用合乎伦理的生成式人工智能提供参考。

参考文献

[1] Jialuo Chen、Jingyi Wang、Tinglan Peng、Youcheng Sun、Peng Cheng、Shouling Ji、Xingjun Ma、Bo Li 和 Dawn Song。2022年。《复制，权利？深度学习模型版权保护测试框架》。*SP*。[2] Yunjey Choi、Minje Choi、Munyoung Kim、Jung-Woo Ha、Sunghun Kim 和 Jaegul Choo。2018年。StarGAN：用于多域图像到图像转换的统一生成对抗网络。*CVPR*。[3] Yunjey Choi、Youngjung Uh、Jaejun Yoo 和 Jung-Woo Ha。2020年。StarGAN v2：面向多个域的多样化图像合成。*CVPR*。[4] Chengdong Dong、Ajay Kumar 和 Eryun Liu。2022年。在频谱域印记中检测生成对抗网络生成的伪造图像前请三思。*CVPR*。[5] Jianwei Fei、Zhihua Xia、Benedetta Tondi 和 Mauro Barni。2022年。用于知识产权保护的有监督生成对抗网络水印。*IEEE* 信息取证与安全研讨会。

[6] Joel Frank、Thorsten Eisenhofer、Lea Schönherr、Asja Fischer、Dorothea Kolossa 和 Thorsten Holz。2020年。利用频率分析进行深度伪造图像识别。*ICML*。[7] Martin Heusel、Hubert Ramsauer、Thomas Unterthiner、Bernhard Nessler 和 Sepp Hochreiter。2017年。采用双时间尺度更新规则训练的生成对抗网络收敛至局部纳什均衡。*NeurIPS*。[8] Hao Huang、Yongtao Wang、Zhaoyu Chen、Yuze Zhang、Yuheng Li、Zhi Tang、Wei Chu、Jingdong Chen、Weisi Lin 和 Kai-Kuang Ma。2022年。CMUA-Watermark：用于对抗深度伪造的跨模型通用对抗水印。*AAAI*。[9] Tero Karras、Miika Aittala、Samuli Laine、Erik Härkönen、Janne Hellsten、Jaakko Lehtinen 和 Timo Aila。2021 年。无混叠生成对抗网络。*NeurIPS*。[10] Tero Karras 、Samuli Laine、Miika Aittala、Janne Hellsten、Jaakko Lehtinen 和 Timo Aila。2020年。分析并改进 StyleGAN 的图像质量。*CVPR*。[11] Cheng-Han Lee、Ziwei Liu、Lingyun Wu 和 Ping Luo。2020年。MaskGAN：迈向多样化且交互式的人脸图像操控。*CVPR*。[12] Ziwei Liu、Ping Luo、Xiaogang Wang 和 Xiaoou Tang。2015年。野外人脸属性的深度学习。*ICCV*。[13] Ding Sheng Ong、Chee Seng Chan、Kam Woh Ng、Lixin Fan 和 Qiang Yang。2021年。保护生成对抗网络的知识产权，使其免受

歧义攻击。*CVPR*。[14] Taesung Park、Alexei A. Efros、Richard Zhang 和 Jun-Yan Zhu。2020年。《用于非配对图像到图像转换的对比学习》。*ECCV*。[15] Nataniel Ruiz、Sarah Adel Bargal 和 Stan Sclaroff。2020年。《扰乱深度伪造：针对条件图像转换网络和人脸操控系统的对抗攻击》。*CVPR*。[16] Hao Tang、Hong Liu、Dan Xu、Philip HS Torr 和 Nicu Sebe。2021年。《Attention-GAN：利用注意力引导生成对抗网络进行非配对图像到图像转换》。*IEEE TNNLS*（2021年）。[17] Zhisheng Xiao、Karsten Kreis 和 Arash Vahdat。2022年。《利用去噪扩散生成对抗网络应对生成式学习三难困境》。*ICLR*。[18] Tianyun Yang、Ziyao Huang、Juan Cao、Lei Li 和 Xirong Li。2022年。《深度伪造网络架构归因》。*AAAI*。[19] Ning Yu、Vladislav Skripniuk、Sahar Abdelnabi 和 Mario Fritz。2021年。《生成模型的人工指纹：将深度伪造归因根植于训练数据》。*ICCV*。[20] Bowen Zhang、Shuyang Gu、Bo Zhang、Jianmin Bao、Dong Chen、Fang Wen、Yong Wang 和 Baining Guo。2022年。《StyleSwin：用于高分辨率图像生成的基于Transformer的生成对抗网络》。*CVPR*。[21] Jun-Yan Zhu、Taesung Park、Phillip Isola 和 Alexei A. Efros。2017年。《利用循环一致对抗网络进行非配对图像到图像转换》。*ICCV*。