arXiv:2305.06321v2 [cs.CV] 1 Aug 2023

SepMark：用于统一来源追踪与深度伪造检测的深度可分离水印

Xiaoshuai Wu shinewu@hnu.edu.cn 湖南大学计算机科学与电子工程学院，中国长沙

廖昕*xinliao @hnu.edu.cn 湖南大学计算机科学与电子工程学院，中国长沙

欧博 oubo@hnu.edu.cn 中国长沙湖南大学计算机科学与电子工程学院

摘要

恶意深度伪造引发了关于区分真实人脸与伪造人脸的激烈争论。尽管人们已开发出许多事后检测深度伪造的对策，但毋庸置疑，被动取证并未考虑在可预见的篡改发生前对原始人脸采取任何预防措施。为完善这一取证生态系统，我们提出了名为 SepMark 的主动式解决方案，它为来源追踪和深度伪造检测提供了统一框架。SepMark 源自基于编码器—解码器的深度水印技术，但采用了两个可分离的解码器。作为首个深度可分离水印方案，SepMark 为既有的深度水印研究带来了新范式：单个编码器优雅地嵌入一个水印，而两个解码器可分别以不同的鲁棒性级别提取该水印。名为 Tracer 的鲁棒解码器能够抵御各种失真，其鲁棒性可能过高，使水印在深度伪造前后都能留存。名为 Detector 的半鲁棒解码器则会选择性地对恶意失真敏感，从而使水印在深度伪造后消失。只有由 Tracer 和 Detector 共同组成的 SepMark，才能可靠地追踪标记人脸的可信来源，并检测其自标记以来是否遭到篡改；单独使用其中任一解码器都无法做到这一点。大量实验表明，所提出的 SepMark 对典型深度伪造技术均有效，包括人脸交换、表情重演和属性编辑。代码将在 https://github.com/sh1newu/SepMark. 提供。</final>

CCS 概念

• **安全与隐私 → 数字版权管理**.

关键词

深度水印、深度伪造取证、水印鲁棒性

ACM 引用格式：

Xiaoshuai Wu、Xin Liao 和 Bo Ou。2023。SepMark：用于统一来源追踪和深度伪造检测的深度可分离水印技术。载于 论文集

*通讯作者。

允许免费出于个人或课堂用途制作本作品全部或部分内容的数字版或纸质版，前提是这些副本不得用于营利或商业利益，且每份副本首页均须附有本声明及完整引文。必须尊重本作品中归他人而非 ACM 所有的组成部分的版权。注明出处后允许摘录。若要以其他方式复制、再版、发布到服务器或重新分发至列表，则须事先获得明确许可和/或支付费用。请向 permissions@acm.org 申请许可。*MM '23*，*2023*年*10*月*29*日至*11*月*3*日，渥太华，安大略省，加拿大。© 2023 计算机协会。ACM ISBN 979-8-4007-0108-5/23/10...$15.00
https://doi.org/10.1145/3581783.3612471

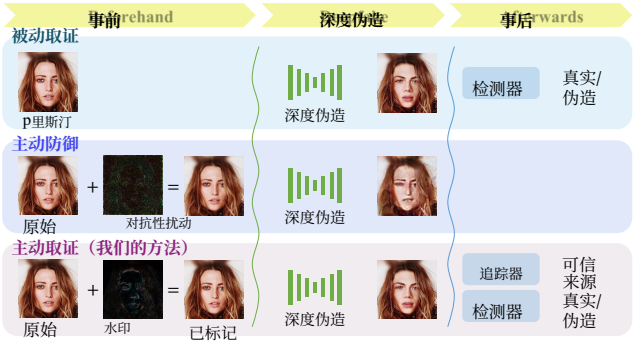


图 1：不同类型深度伪造对策的示意图：被动取证关注深度伪造检测任务，而原始人脸不受保护；主动防御专注于深度伪造破坏任务；我们的主动取证则同时关注来源追踪和深度伪造检测任务。

第31届 ACM 多媒体国际会议（MM '23）论文集，2023年10月29日至11月3日，加拿大安大略省渥太华。ACM，纽约州纽约市，美国，12页。
https://doi.org/10.1145/3581783.3612471

1 引言

迄今为止，人工智能生成内容一直受到深度学习社区的全面推动，例如生成式预训练变换器[61, 62] 和稳定扩散 [63]。在这个时代，每个人都可以成为一名顶尖操作者，只需点击一下即可创作出看似真实的内容 [1]。然而，多媒体生成领域的突破性进展再次将深度伪造推到了聚光灯下；深度伪造主要指利用深度学习伪造人脸的技术 [68, 76, 77], 。一般而言，伪造生成技术的发展通常在时间上领先于其取证技术 [13, 36]。尽管这项技术对于人脸去身份化 [4, 40]和隐私保护 [53, 82], 具有重要意义，但一些人仍会故意制作并传播具有高度视觉欺骗性的深度伪造，以歪曲事实、操纵公众舆论。在“眼见不一定为实”的背景下，无法控制的深度伪造最终将引发前所未有的信任危机和道德恐慌。因此，深度伪造取证正变得愈发重要且紧迫。

自深度伪造出现以来，被动取证技术便应运而生 [88]。眼部眨动 [43],头部姿态 [78], 颜色不一致性 [41], 等人工设计的特征在早期阶段能够有效检测此类伪造。然而，越来越高明的创作者利用对抗性训练，可以轻易修复这些浅层篡改痕迹。

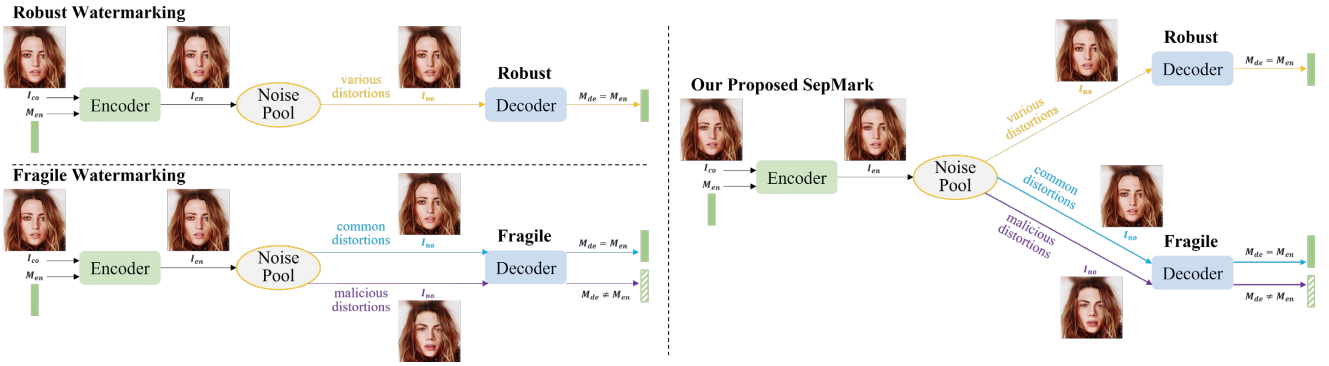


图2：鲁棒水印（左上）、脆弱水印（左下）以及我们提出的 SepMark（右）的处理流程。为便于比较和清晰展示，此处省略了额外的对抗判别器。在先前的编码器-解码器结构中，编码器和解码器通过从噪声池中采样各种失真进行联合训练，其中鲁棒解码器能够抵抗所有这些失真，并以较高的鲁棒性提取水印；相比之下，脆弱解码器对恶意失真具有选择性敏感性，因此只能以较低的鲁棒性提取水印。SepMark 创新性地将单个编码器与两个可分离解码器相结合，使嵌入的水印能够以不同的鲁棒性级别分别提取。

[12, 29]. 研究表明，传统的深度学习特征具有良好的有效性 [64],，但在遇到未见过的伪造方法时，其泛化能力会受到影响。因此，一些数据驱动模型旨在学习通用伪造痕迹，例如融合边界 [42],、频谱异常 [15],、时间不一致性 [23], 等，同时避免过度拟合特定的伪造方法。然而，当待查询人脸经历多种后处理操作或遭受对抗性扰动时，这些模型的性能仍不明确 [5, 24, 30, 45, 55],，这表明由于鲁棒性有限，真实场景中的取证仍然面临挑战 [22, 28, 91]。尽管已有少数采用对抗攻击的主动防御方法，致力于干扰或消除深度伪造内容的生成 [1, 20, 26, 27, 65, 71, 79, 80],，但这些方法可能有助于检测任务 [9, 73],，却无法解决来源追踪任务 [47, 59, 72]，而后者也是我们主动取证研究的另一个重点。我们在图1中重点展示了这些对策之间的差异，主动取证的应用将在第3.1节中进一步详细介绍。

据我们所知，无形水印技术旨在不影响媒体原有视觉特征的情况下，将额外消息以不可察觉的方式隐藏在有价值的媒体中 [11]。根据水印提取过程中是否需要原始图像，无形水印可进一步分为盲水印和非盲水印。由于提取过程仅需要标记图像，盲水印在现代多媒体生态系统中的应用日益广泛 [6, 18, 81],，包括版权保护、来源追踪、内容认证等。因此，本文采用盲水印技术来对抗深度伪造。得益于深度学习在盲水印领域所取得的进展 [89],，基于编码器-解码器的深度水印技术因其有效性和灵活性而受到越来越多的关注 [34, 56],，相比之下，传统水印技术则相形见绌 [57, 75]。然而，如图 2 所示，让我们深入分析当前单一功能的深度水印技术为何无法很好地满足主动取证的需求。

鲁棒水印技术高度重视鲁棒性，强调嵌入的水印在经过……后仍然能够被提取出来。

严重失真 [89]。这对于保护版权和追溯来源至关重要，即使标记图像的大部分区域被遮挡，或通过截图捕获也是如此 [33, 70]。然而，过高的鲁棒性也可能使水印在深度伪造前后都得以保留，从而难以区分真实图像和伪造图像。更糟的是，即使图像被伪造，原始来源标识仍会保留。因此，仅依靠鲁棒水印无法在追溯深度伪造图像的可信来源的同时检测出深度伪造。

脆弱水印侧重于完全脆弱性或选择性脆弱性，这与鲁棒水印相反；它强调嵌入的水印会随着标记图像在遭受任意或特定失真时相应改变 [2]。事实上，它主要用于指示内容是否遭到恶意失真篡改，但由于水印已消失，无法提供原始可信图像来源的信息 [56]。更重要的是，在缺乏先验知识的情况下，很难区分伪造图像与原始图像，因为它们都没有正确的水印 [69]。因此，仅依赖脆弱水印并不总能追溯可信来源或检测深度伪造。

为了同时实现来源追踪和深度伪造检测这两项任务，最直观的方法是先嵌入一个鲁棒水印，再嵌入另一个脆弱水印 [66],，然而，这两种水印可能会显著相互影响，并同时降低图像的视觉质量。另一种直观的思路是将这些相互独立的水印嵌入人脸图像中不同的感兴趣区域，例如前景和背景。然而，我们认为，这种区域划分可能需要借助额外的复杂人脸解析模型。更重要的是，利用感兴趣区域的相关知识，水印更容易被感知和攻击。为克服多重嵌入方法的这些局限性，本文提出了一种整体式的、基于学习的端到端 SepMark，对应于一种全新范式，即通过一种优雅的嵌入方式实现两种截然不同的提取方式。总而言之，我们的主要贡献如下：

1) 我们首次提出了深度可分离水印（SepMark），为当前的深度水印技术带来了全新范式：单个编码器能够以优雅的方式嵌入一个水印，而两个解码器则能够分别以不同的鲁棒性等级提取该水印。此外，我们提出的 SepMark 还自然地提供了一种新颖的主动取证解决方案，可统一实现来源追踪与深度伪造检测。

2）这是首个用于深度可分离水印的端到端学习架构，由一个编码器、一个判别器和两个可分离解码器组成。训练时，从随机前向噪声池（RFNP）中采样不同类型的失真。鲁棒解码器能够抵御所有常见失真和恶意失真，而半鲁棒解码器则会选择性地对恶意失真产生敏感反应。

3）针对人脸图像开展的大量实验表明，鲁棒解码器在面对多种失真时具有很强的鲁棒性，而半鲁棒解码器面对恶意深度伪造失真时则表现出选择性脆弱性。因此，可分离解码器能够可靠地追踪带水印人脸的可信来源，并检测其自添加水印后是否遭到篡改。

2 相关工作

2.1 深度伪造被动取证

被动取证，尤其是自监督深度伪造检测，近年来受到广泛关注 [85]。其中的基本原则是利用正样本驱动合成自构造的深度伪造。通常，这些方法的检测性能取决于合成数据生成过程中引入的先验知识。FWA [44] 针对内外人脸分辨率不一致导致的仿射人脸扭曲伪影。Face X-ray [42] 关注融合边界处的伪影，以判断待检测图像是否可分解为来自不同来源的两张人脸的融合结果。类似地，I2G [86] 旨在通过融合流程生成来源特征不一致的融合人脸。SBIs [67] 对单张图像应用两种随机变换，分别作为融合过程的来源图像和目标图像，以生成更具挑战性的负样本。此外，ICT [14] 考虑待检测图像中的身份一致性，将不同身份的内侧人脸与外侧人脸互换，以构造合成数据。借助这些额外线索[7, 21, 87], ，对深度伪造失真进行更通用的建模，有望进一步提升对未见伪造内容的泛化能力。

2.2 深度伪造主动防御

与主动取证类似，主动防御需要提前向原始图像中添加不可感知的信号。通过注入对抗性扰动，可以破坏或抵消深度伪造模型的作用，从而实现深度伪造破坏。前者可视为一种非定向攻击，即经过深度伪造处理后的受攻击图像，与原始伪造图像相比，在视觉上应尽可能表现出异常和不自然的特征 [26, 27, 65, 71, 80]。相反，后者属于定向攻击，即经过深度伪造处理后的受攻击图像，应尽可能与原始图像保持相似 [20, 79]。受通用对抗扰动的启发 [54]，CMUA [26]生成适用于图像的通用跨模型扰动，以攻击多个深度伪造模型，而不是.....

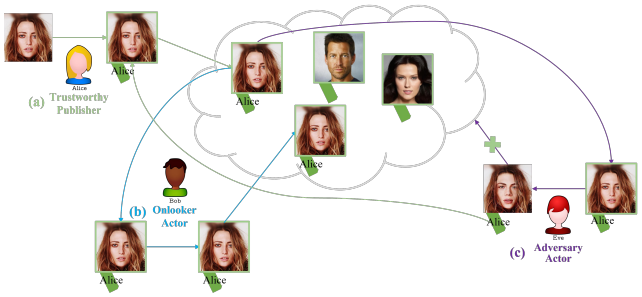


图3：主动取证示意图。(a) 可信发布者 Alice 将其身份标识嵌入已发布的人脸图像中。(b) 旁观者 Bob 转发该人脸图像，但不影响其内容语义。(c) 恶意攻击者 Eve 篡改该人脸图像，以歪曲事实。

而非特定的那一种。Anti-Forgery [71] 进一步增强了感知感知扰动对各种失真的鲁棒性。然而，由于被动检测器与人眼的观察视角并不相同 [73]，视觉上可感知的伪影仍可能逃过检测。此外，诸如 MagDR[9] 之类的方法能够检测并移除对抗性扰动；即使是深度伪造内容，也仍然可以被进一步正常重建。

2.3 深度水印

受深度学习在计算机视觉任务中取得的成功启发，研究人员开始探索基于 DNN 的盲水印模型，以提升鲁棒性。HiDDeN [89] 是首个用于鲁棒水印的端到端学习架构，由编码器、判别器、噪声层和解码器组成。保障鲁棒性的关键在于利用差分噪声层对标记图像进行数据增强。为抵御非差分 JPEG 失真，MBRS [34]在每个小批次训练中随机切换使用真实 JPEG 和模拟的可微分 JPEG。CIN [52]引入可逆神经网络来学习水印嵌入与提取之间的联合表示，并额外设计了一种基于注意力的不可逆模块，以抵御非加性量化噪声。在另一条与失真无关的研究路线中， [51], TSDL [48] 提出了一种两阶段学习架构：首先以无噪声的方式训练编码器和解码器，随后仅微调解码器，使其适应黑盒失真。ASL [83]将前向噪声建模为与标记图像交互的伪噪声，因此在反向传播时，无需经过噪声层即可将梯度传递给编码器。FIN [17]通过类似于可逆去噪 [49]的可逆噪声层处理黑盒失真；该噪声层旨在模拟失真，在训练期间充当噪声层，也会在提取前作为预处理步骤，用于消除失真。此外，LFM [74], StegaStamp [70], RIHOOP [32], 和 PIMoG [16] 致力于抵御打印机—扫描仪、打印机—相机和屏幕—相机等跨媒介物理失真。据我们所知，尽管深度鲁棒水印已取得长足发展，但深度脆弱水印仍鲜有探索。

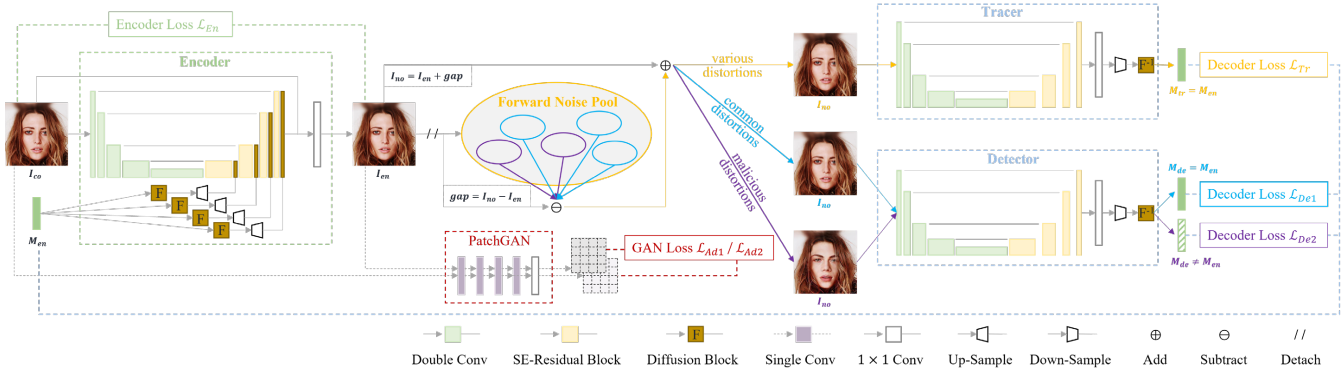


图 4：SepMark 的整体架构。编码器 En 接收封面图像 I_{co} 和编码消息 M_{en} 作为输入，并生成编码图像 I_{en}。随机前向噪声池 $RFNP$在分离式前向传播中采样不同类型的伪噪声 gap，并在标准前向传播和反向传播过程中与 I_{en} 交互，从而生成常规失真图像或恶意失真图像 I_{no}。追踪器 Tr从任意失真图像中提取与 M_{en}相同的消息 M_{tr}。检测器 De 从常规失真图像中提取正确消息，而从恶意失真图像中提取错误消息。此外，采用 PatchGAN Ad 来区分原始图像块和编码图像块。

3 主动取证：SEPMAR K

3.1 问题陈述

SepMark 并非在事后检测哪些人脸是伪造的，而是致力于提前确保哪些人脸是真实的，并追溯其可信来源。这种主动取证应当预先作用于原始人脸，其中，“原始”是指在本文中尚未经过 SepMark 处理的人脸。在现实世界中，值得信赖的内容发布者会将其原始人脸上传至社交网络，并向所有人公开。然而，一些恶意攻击者可能会下载这些人脸并对其进行篡改，以传播虚假信息，从而对社会信任和公共利益造成严重影响。此外，依赖事后取证的方法，无法在深度伪造内容得到确认之前立即阻止其传播。

以下通过一个更具说明性的示例，介绍图 3 所示的主动取证方案。为了对未篡改人脸采取预防措施并提供更可靠的证据，SepMark 会预先在人脸中嵌入一个水印（图 3-a），例如发布者标识和时间戳，以识别标记人脸的可信来源，同时表明该人脸已由 SepMark 处理。随后，可以分别以不同的鲁棒性级别提取嵌入的水印。具体而言，在对图像语义影响较小的常见失真下（图 3-b），例如 JPEG 压缩、缩放、对比度调整等，鲁棒水印和半鲁棒水印都可以正确提取。相反，在本文主要关注的恶意失真（如深度伪造）下（图 3-c），仍可正确提取鲁棒水印，但半鲁棒水印会被错误提取，这表明人脸图像自标记后已被篡改。这样一来，任何个人或社交网络都可以验证该人脸的原始来源是否来自可信发布者。此外，即使不知道原始水印，也可以比较提取出的两个水印，以判断标记人脸在发布后是否遭到操纵，并阻止继续传播水印不匹配的可疑图像。

总之，无需获知原始图像，SepMark 即可实现深度伪造溯源追踪和盲检测。

按照图4，我们详细阐述 SepMark。给定任意未篡改人脸（也称为封面图像）I_{co}，编码器 En 将编码消息 M_{en} 嵌入图像 I_{co}中，从而得到编码图像 I_{en}。之后，可分离的追踪器 Tr 和检测器 De 分别从普通图像或恶意篡改图像 I_{no}中提取消息 M_{tr}和 M_{de}。因此，根据可分离解码器 Tr 和 De，我们可以将主动取证分为以下几种情况。

情况 1：如果鲁棒解码器 Tr 可用，则无论标记图像受到常见失真还是恶意失真的影响，提取的 M_{tr}都应始终与嵌入的 M_{en}完全相同，从而可靠地追踪可信来源。

情况 2：如果存在半鲁棒解码器 De，则它可以从常规失真图像中提取正确消息，但从受到恶意失真处理的图像中提取错误消息，以区分失真类型。

情况 3：只有当解码器 Tr 和 De 均可用时，才能同时实现可信来源追踪和恶意失真检测。具体而言，如果$M_{de} \approx M_{en}$或 $M_{de} \approx M_{tr}$，SepMark 将判定该人脸是真实的，且来自可信来源 M_{tr}。否则，如果$M_{de} \not\approx M_{en}$ 或 $M_{de} \not\approx M_{tr}$，则表明该人脸自标记以来已被伪造，并且在必要时，可以进一步从来源 M_{tr} 获取原始可信人脸。

3.2 模型架构

为实现深度可分离水印的端到端训练，我们提出的 SepMark 由一个编码器 En、一个随机前向噪声池 $RFNP$、两个可分离解码器 Tr和 De，以及一个额外的对抗判别器 Ad组成，如图4所示。简而言之，1) En 接收一批 $I_{co} \in \mathbb{R}^{B\times3\times H\times W}$ 和一批 $M_{en} \in \{-\alpha, \alpha\}^{B\times L}$，以生成 I_{en}；2) $RFNP$随机采样不同的失真作为伪噪声，并与 I_{en}交互，

从而得到 I_{no}；3) Tr 和 De 分别从 I_{no} 中提取 M_{tr} 和 M_{de}，且具有不同程度的鲁棒性；4) Ad 根据 I_{co}，将 I_{en} 中每个图像块分类为原始或已编码。

3.2.1 单编码器。 在主干路径中， I_{co} 被输入到一个类似 UNet 的结构中。首先，在下采样阶段，步长为 2 的“conv-in-relu”将空间维度减半，而“Double Conv（2× conv-in-relu）”则使特征通道数加倍。在并行路径中， M_{en} 被输入到多个“Diffusion Block（FC layer-Conv）”中以注入冗余信息，其长度 L 被扩展至 $L\times L$。其次，在上采样阶段，最近邻插值使空间维度加倍，“conv-in-relu”则使特征通道数减半，随后在每个分辨率下，将其与下采样阶段对应的特征图以及经过插值操作后的冗余消息进行拼接。此外，嵌入 Squeeze-and-Excitation [25] 的“SE-Residual Block”采用实例归一化替换了原始瓶颈残差块中的批归一化，并作用于拼接后的特征，以增强不同通道之间的信息交互。最后，我们进一步通过跳跃连接将上述特征图与封面图像进行拼接，并将输出输入到一个 1×1 卷积层中，以生成编码图像 I_{en}。

3.2.2 随机前向噪声池。考虑到大量失真和单一噪声层无法适应所有情况，我们提出将各种失真引入噪声池进行联合训练。此外，鉴于模拟失真难以逼近真实的复杂失真，我们在分离的前向传播中直接获取任意真实失真 $gap = I_{no} - I_{en}$。这样，标准的前向和反向传播都可以通过可微的 $I_{no} = I_{en} + gap$ 实现。值得注意的是， $RFNP$ 与 ASL [83] 的区别在于：前者在每个小批次中随机采样不同的失真，而非仅采样一种失真。从经验上看，这一变化有利于稳定训练，也解释了我们在编码器设计中采用实例归一化的原因。具体而言，从常见失真和恶意失真、仅常见失真以及仅恶意失真中随机采样，会形成三批 I_{no}。

3.2.3 可分离解码器。我们直接借用了编码器中的类 UNet 结构来构建两个解码器。需要注意的是，在每个分辨率下拼接相应的特征图后，我们将最终输出输入一个 1×1 卷积层，以生成单通道残差图像。随后，将残差图像下采样至 $L\times L$，并使用“逆扩散块（全连接层）”提取长度为 L 的解码消息。更重要的是，由于训练期间输入和优化目标不同，两个解码器采用相同的结构设计，但不共享模型参数；后文将介绍我们精心设计的损失函数。具体而言，鲁棒 Tr 接收经过各种失真处理的图像 I_{no}，而半鲁棒 De 则在每次小批次训练中分别接收经过常见失真和恶意失真处理的图像。此外， Tr 在任意失真下通过目标函数 $M_{tr} = M_{en}$ 进行优化。相比之下， De 在常见失真下通过目标函数 $M_{de} = M_{en}$ 进行优化，并在恶意失真下通过目标函数 $M_{de} \neq M_{en}$ 进行优化。

3.2.4 对抗判别器。 为了监督编码图像 I_{en} 的视觉质量，我们还采用对抗训练，其中一个附加判别器 Ad 与主编码器-解码器交替进行训练。具体而言，我们用 Patch-GAN [31, 90] 替代传统 GAN，将每个图像块分类为原始或编码状态。具体来说，判别器采用全卷积结构，由若干个“conv-in-leakyrelu”模块组成，最后通过一个 1×1 卷积层输出最终预测图。然而，研究发现，使用标准 GAN 损失 [19] 时，GAN 训练过程通常不稳定，甚至无法收敛。这种现象也出现在我们的训练过程中。我们推测，除了对抗生成和判别之外，这可能还与消息提取任务有关。为了提高训练的稳定性，我们改用 RaLSGAN [35] 的 GAN 损失，并发现这一调整使我们的训练过程更加稳定。这可能主要归因于将绝对真假判断转变为相对真假判断。

3.3 损失函数

我们发现，简单的 L_2 约束足以支持 SepMark 的端到端训练。请注意， θ 表示可训练参数，而 θ_{co} 和 θ_{ma} 分别仅表示由 $RFNP$ 采样得到的常见失真和恶意失真。我们采用以下损失函数。

对于编码器，其编码图像 I_{en} 应在视觉上与输入图像 I_{co} 相似：

$$\mathcal{L}_{En} = L_2(I_{co}, En(\theta, I_{co}, M_{en})) = L_2(I_{co}, I_{en}) \tag{1}$$

对于解码器，鲁棒的追踪器 Tr 应能够从经过任意常见失真和恶意失真处理的图像 I_{no} 中提取 M_{tr}，且该内容与 M_{en} 完全相同：

$$\mathcal{L}_{Tr} = L_2(M_{en}, Tr(\theta, RFNP(\theta_{co} + \theta_{ma}, I_{en}))) = L_2(M_{en}, M_{tr}) \tag{2}$$

其中

$$I_{en} = En(\theta, I_{co}, M_{en}) \tag{3}$$

半鲁棒检测器 De 应能容忍常规失真图像 I_{no}：

$$\mathcal{L}_{De1} = L_2(M_{en}, De(\theta, RFNP(\theta_{co}, I_{en}))) = L_2(M_{en}, M_{de}) \tag{4}$$

相比之下， De 应对恶意失真 θ_{ma} 具有选择性敏感性：

$$\mathcal{L}_{De2} = L_2(0, De(\theta, RFNP(\theta_{ma}, I_{en}))) = L_2(0, M_{de}) \tag{5}$$

其中，优化 0 与 M_{de} 之间的 L_2 距离。需要强调的是，二进制消息 $\{0,1\}$ 已被转换为零均值的 $\{-\alpha, \alpha\}$。因此，如果提取消息大于 0，则将其恢复为二进制 1；否则，恢复为二进制 0。该损失项使提取的 M_{de} 接近随机猜测。

对于判别器，其参数通过以下方式更新：

$$\mathcal{L}_{Ad1} = L_2(Ad(\theta, I_{co}) - \overline{Ad}(\theta, I_{en}), 1) + L_2(Ad(\theta, I_{en}) - \overline{Ad}(\theta, I_{co}), -1) \tag{6}$$

其中

$$\overline{Ad}(I) = \frac{1}{B}\sum_{i=1}^{B} Ad(I^{i\times 3\times H\times W}) \tag{7}$$

$$Ad(\theta, I_{en}) = Ad(\theta, En(I_{co}, M_{en})) \tag{8}$$

这里的相对平均判别器估计封面图像 I_{co} 比编码图像 I_{en} 更接近原始状态的概率

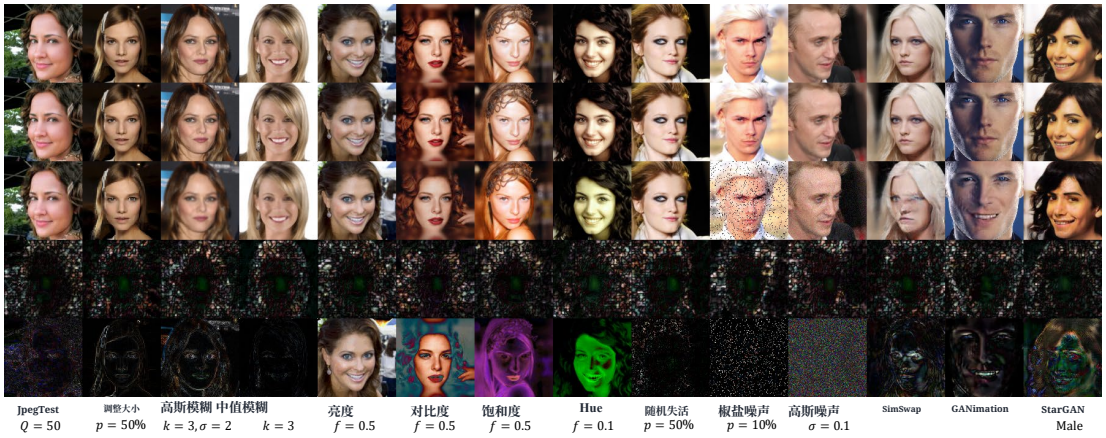


图5：典型失真下的主观视觉质量。从上到下依次为封面图像 I_{co}、编码图像 I_{en}、加噪图像 I_{no}、$\mathcal{N}(|I_{co} - I_{en}|)$ 的残差信号，以及 $\mathcal{N}(|I_{no} - I_{en}|)$，其中 $\mathcal{N}(I) = (I - min(I))/(max(I) - min(I))$。每列对应一种失真。图像大小： 128×128。**更多恶意失真请参见附录A.1。**

平均而言；详情请参阅文献 [35] 。请注意，判别器也参与编码器 En的更新：

$$\mathrm{L}_{Ad2} = L_2(Ad(I_{co}) - \overline{Ad(I_{en})} - 1) + L_2(Ad(I_{en}) - \overline{Ad(I_{co})}, \quad (9)$$

其中

$$Ad(I_{en}) = Ad(En(\theta, I_{co}, M_{en})) \quad (10)$$

综上，主编码器-解码器的总损失可表示为：

$$\mathrm{L}_{Total} = \lambda_1 \mathrm{L}_{Ad2} + \lambda_2 \mathrm{L}_{En} + \lambda_3 \mathrm{L}_{Tr} + \lambda_4 \mathrm{L}_{De1} + \lambda_5 \mathrm{L}_{De2} \quad (11)$$

其中 λ_1、λ_2、λ_3、λ_4、λ_5 分别为各损失项的权重。

4 实验

4.1 实现细节

实验主要在 CelebA-HQ 上进行 [37, 39], ，其中 24183/2993/2824 张人脸图像分别用于训练、验证和测试，划分方式参照 CelebA 的官方划分 [50]。此外，还使用包含 19962 张人脸图像的 CelebA 测试集，以及包含 5000 张常见物体图像的 COCO 验证集 [46], ，以评估泛化能力。除非另有说明，由于计算资源有限，所有图像均调整为 128×128 和 256×256 两种分辨率。对于恶意失真，选择 SimSwap [8],、GANimation [60], 和 StarGAN [10] ，因为它们是典型的深度伪造，分别代表人脸交换、表情重演和属性编辑。具体而言，用于交换的目标人脸从 CelebA 验证集中随机选取；用于重演的目标表情从驱动图像"eric_andre"[1] 中随机选取；编辑的属性集为 $\{Male,\ Young,\ BlackHair,\ BlondHair,\ BrownHair\}$。为生成逼真的深度伪造，所有方法均基于其公开发布的预训练模型实现。请注意，仅在测试 GANimation 时，才会使用 OpenFace [3] 进一步去除人脸图像的背景，以符合输入要求。在不失一般性的前提下，常见失真集合为 $\{Identity,\ JpegTest,\ Resize,$ $GaussianBlur,\ MedianBlur,\ Brightness,\ Contrast,\ Saturation,\ Hue,\ Dropout\ SaltPepper\ GaussianNoise\}$，。上述失真的详细参数见图 5 底部。请注意，集合中不包括裁剪，因为很难将其视为适用于完整人脸的常见失真。

[1]https://github.com/vipermu/ganimation/tree/master/animations/eric_andre/attribute_images

表 1：编码图像的客观视觉质量 I_{en}

模型	图像 Size	消息长度	峰值信噪比 ↑	结构相似性指数 ↑	学习感知图像块相似度 ↓
MBRS [34]	128×128	30	33.0456	0.8106	0.0141
CIN [52]	128×128	30	42.4135	0.9628	0.0006
PIMoG [16]	128×128	30	37.7271	0.9470	0.0086
SepMark	128×128	30	38.5112	0.9588	0.0028
FaceSigns [56]	256×256	128	32.3319	0.9211	0.0260
SepMark	256×256	128	38.5646	0.9328	0.0080

我们的 SepMark 使用 PyTorch [58] 实现，并在 NVIDIA RTX 4090 上运行。由于我们处理的是多种失真，而非某一种特定失真，因此 SepMark 采用联合训练方式，其中噪声池包含所有常见失真和恶意失真，即 SimSwap、GANimation 和 StarGAN（男性）。整个训练过程持续了 100 个训练轮次，批量大小为 16。作为一项经验法则，我们对 Adam优化器 [38] 进行了调整，其中 $lr = 0.0002$、$\beta_1 = 0.5$，以实现稳定的对抗训练。此外，在我们的训练中，公式（11）中的 λ_1、λ_2、λ_3、λ_4、λ_5 分别设置为 0.1、1、10、10、10。直观而言，我们将 SepMark 中的消息范围 α 设置为 0.1，并通过消融实验表明，α 对于视觉质量与鲁棒性之间的权衡至关重要。

基线方法。 据我们所知，这是首项关于深度可分离水印的研究，因此，我们必须采用 MBRS [34], 、CIN [52], 、PIMoG [16], 等鲁棒水印方法，以及 FaceSigns [56] 这一脆弱水印方法作为我们的基线方法。我们训练了两个独立模型：一个使用尺寸为 128×128 的图像和长度为 30 的消息进行训练，另一个使用尺寸为 256×256 的图像和长度为 128 的消息进行训练，并将它们与基线方法的预训练模型进行比较，以证明 SepMark 的有效性。

表 2：常规失真图像 I_{no} 的鲁棒性测试，其中"追踪器"表示 M_{tr} 与 M_{en} 之间的比特错误率，"检测器"表示 M_{de} 与 M_{en} 之间的比特错误率，"两者"表示 M_{tr} 与 M_{de} 之间的比特错误率。

	128 × 128						256 × 256			
失真	MBRS [34]	CIN [52]	PIMoG [16]	SepMark			FaceSigns [56]	SepMark		
				追踪器	检测器	Both		追踪器	检测器	Both
身份	0.0000%	0.0000%	0.0366%	0.0000%	0.0000%	0.0000%	0.0136%	0.0000%	0.0000%	0.0000%
JpegTest	0.2597%	2.7514%	19.5562%	0.2136%	0.2172%	0.2656%	0.8258%	0.0075%	0.0069%	0.0133%
调整大小	0.0000%	0.0000%	0.0767%	0.0059%	0.0212%	0.0201%	1.0726%	0.0000%	0.0000%	0.0000%
高斯模糊	0.0000%	22.7786%	0.1169%	0.0024%	0.0035%	0.0012%	0.1671%	0.0000%	0.0274%	0.0274%
中值模糊	0.0000%	0.0307%	0.0992%	0.0012%	0.0012%	0.0000%	0.0977%	0.0000%	0.0000%	0.0000%
亮度	0.0000%	0.0000%	1.3443%	0.0059%	0.0106%	0.0142%	10.8196%	0.0017%	0.0030%	0.0047%
对比度	0.0000%	0.0000%	0.8121%	0.0012%	0.0024%	0.0012%	0.0334%	0.0000%	0.0000%	0.0000%
饱和度	0.0000%	0.0000%	0.0803%	0.0000%	0.0000%	0.0000%	0.7113%	0.0000%	0.0000%	0.0000%
Hue	0.0000%	0.0000%	0.1523%	0.0000%	0.0012%	0.0012%	8.3780%	0.0000%	0.0000%	0.0000%
随机失活	0.0000%	0.0000%	0.4828%	0.0000%	0.0000%	0.0000%	17.5615%	0.0058%	0.0000%	0.0058%
椒盐噪声	0.0000%	0.0378%	2.3667%	0.0413%	0.0106%	0.0472%	12.3238%	0.0008%	0.0003%	0.0011%
高斯噪声	0.0000%	0.0000%	12.7396%	0.7460%	0.8735%	0.8994%	7.0697%	0.0578%	0.0622%	0.0930%
平均	0.0216%	2.1332%	3.1553%	0.0848%	0.0951%	0.1042%	4.9228%	0.0061%	0.0083%	0.0121%

表 3：针对恶意失真图像 I_{no} 的鲁棒性测试。

	128 × 128						256 × 256			
失真	MBRS [34]	CIN [52]	PIMoG [16]	SepMark			FaceSigns [56]	SepMark		
				追踪器	检测器	Both		追踪器	检测器	Both
SimSwap	19.3744%	48.5068%	8.6745%	13.8255%	50.8829%	50.8558%	49.9463%	7.9068%	45.9117%	45.6713%
GANimation	0.0000%	0.0000%	0.4802%	0.0000%	36.7938%	36.7938%	45.4172%	0.0020%	43.8524%	43.8527%
StarGAN（男性）	18.3133%	67.2568%	9.2044%	0.1157%	52.6003%	52.6027%	50.2617%	0.0033%	45.4624%	45.4641%
StarGAN（年轻）	17.0562%	69.0805%	8.7465%	0.1074%	52.3678%	52.3737%	50.3649%	0.0030%	45.5319%	45.5333%
StarGAN（黑发）	19.2233%	58.7913%	10.5312%	0.1416%	49.2434%	49.2599%	50.2576%	0.0019%	45.1299%	45.1296%
StarGAN（金发）	18.4478%	72.9733%	10.3506%	0.1712%	48.6084%	48.5930%	50.9829%	0.0022%	44.6953%	44.6953%
StarGAN（棕发）	17.6381%	79.5857%	9.0675%	0.0980%	52.3820%	52.3855%	50.7884%	0.0017%	45.9950%	45.9939%
平均值	15.7219%	56.5992%	8.1507%	2.0656%	48.9827%	48.9806%	49.7170%	1.1316%	45.2255%	45.1915%

评估指标。对于客观视觉质量的评估，我们报告整个测试集上编码图像的平均 PSNR、SSIM 和 LPIPS [84] 。对于鲁棒性测试，采用平均误码率（BER）作为默认评估指标。在我们的 SepMark 中，半鲁棒检测器在恶意失真下的 BER 应接近 50%；在其他情况下，追踪器和检测器的 BER 均应接近 0%。假设嵌入消息 M_E 和提取消息 M_D 分别形式化表示为，

$$BER(M_E, M_D) = \frac{1}{B} \times \frac{1}{L} \times \sum_{i=1}^{B} \sum_{j=1}^{L} (\mathcal{B}(M_E^{i \times j}) - \mathcal{B}(M_D^{i \times j})) \times 100\% \quad (12)$$

其中

$$\mathcal{B}(M) = \begin{cases} 1 & M > 0, \\ 0 & M \leq 0. \end{cases} \quad (13)$$

请注意，我们已严格按照原始设置调整了基线方法中的 B(·)，以避免影响其性能。

4.2 视觉质量

我们从客观和主观两个角度评估编码图像的视觉质量。如表 1所示，就峰值信噪比、结构相似性指数和学习感知图像块相似度而言，SepMark 的表现略优于 PIMoG，且大幅领先 MBRS 和 FaceSigns。此外，如图5所示，肉眼无法从编码图像中观察到明显伪影。我们推测，这种令人满意的性能主要源于我们稳定的对抗训练。尽管 CIN 的客观视觉质量令人印象深刻，但其可逆神经网络缺乏灵活性，无法与我们的深度可分离水印架构兼容，而且在遇到量化噪声时鲁棒性不理想。

4.3 鲁棒性测试

首先，稳健的追踪器应在任意失真下都具有较低的比特错误率，以追踪可信来源；而半稳健的检测器应在常见失真下具有较低的比特错误率，但在恶意失真下具有较高的比特错误率，以区分失真类型。

在表 2 所示的常见失真条件下，与基线方法相比，Tracer 的平均 BER 接近 0%，实现了近乎最优的鲁棒性，同时也达到了与 MBRS 相当的竞争力。此外，与 CIN、PIMoG 和 FaceSigns 相比，Detector 的 BER 也有所降低。通过观察"Both"列可以发现，在遇到常见失真时，可分离解码器提取的消息几乎相同。我们的目的并非证明 SepMark 在鲁棒性方面一定优于基线方法，而是表明其对常见失真具有足够的鲁棒性。

在表 3 所示的恶意失真条件下，与基线方法相比，Tracer 同样表现出理想的鲁棒性，其比特错误率（BER）平均值达到最低。正如预期的那样，Detector 的比特错误率（BER）平均值接近 50%，与随机猜测的结果相近。我们的检测器

表 4：消融实验。消息 M_{en} 的范围为 $\alpha \in \{0.15,0.05\}$。图像尺寸： 128×128。**增益："↓"表示更好。**

	$\alpha = 0.15$		$\alpha = 0.05$	
	峰值信噪比 = 38.1806		峰值信噪比 = 40.8058	
失真	追踪器	检测器	追踪器	检测器
身份	0.0000%	0.0000%	0.0000%	0.0000%
JpegTest	0.1062%	0.1204%	0.6941%	0.7495%
调整尺寸	0.0012%	0.0000%	0.0614%	0.0696%
高斯模糊	0.0000%	0.0000%	0.1464%	0.5076%
中值模糊	0.0000%	0.0000%	0.0141%	0.0106%
亮度	0.0012%	0.0012%	0.0035%	0.0083%
对比度	0.0012%	0.0000%	0.0224%	0.0224%
饱和度	0.0000%	0.0000%	0.0000%	0.0000%
Hue	0.0000%	0.0000%	0.0000%	0.0000%
随机失活	0.0012%	0.0012%	0.0024%	0.0047%
椒盐噪声	0.0059%	0.0083%	0.4308%	0.3778%
高斯噪声	0.4273%	0.4403%	2.0515%	2.2238%
SimSwap	15.8015%	54.1714%	20.0802%	48.3239%
GANimation	0.0037%	53.9691%	0.0012%	29.5494%
StarGAN（男性）	0.0331%	58.0914%	0.3175%	45.4544%
StarGAN（年轻）	0.0378%	57.7679%	0.3671%	45.2325%
StarGAN（黑发）	0.0425%	60.0248%	0.4355%	45.3978%
StarGAN（金发）	0.0354%	56.9795%	0.4202%	45.3872%
StarGAN（棕发）	0.0189%	59.2694%	0.2927%	45.0342%
平均公共	0.0454%	0.0476%	0.2856%	0.3312%
	↓0.0394%	**↓0.0475%**	**↑0.2008%**	**↑0.2961%**
平均恶意程度	2.2818%	57.1819%	3.1306%	43.4828%
	↑0.2162%	——	**↑1.0650%**	——

表 5：消融实验。先训练一个解码器，再训练另一个。图像尺寸： 128×128。

	追踪器→检测器		检测器→追踪器	
	峰值信噪比 = 39.3359		峰值信噪比 = 42.6461	
失真	追踪器	检测器	追踪器	检测器
身份	0.0000%	0.0000%	0.0012%	0.0000%
JpegTest	0.6645%	0.5430%	2.0267%	2.4906%
调整大小	0.0295%	0.0212%	0.1039%	0.0720%
高斯模糊	0.0012%	0.0000%	0.0260%	0.0189%
中值模糊	0.0012%	0.0012%	0.0059%	0.0201%
亮度	0.0000%	0.0012%	0.0035%	0.0047%
对比度	0.0035%	0.0035%	0.0106%	0.0071%
饱和度	0.0000%	0.0000%	0.0000%	0.0000%
Hue	0.0000%	0.0000%	0.0000%	0.0000%
随机失活	0.0035%	0.0059%	0.0047%	0.0024%
椒盐噪声	0.1333%	0.0024%	0.0342%	0.2998%
高斯噪声	1.5557%	1.5321%	3.5057%	4.1218%
SimSwap	17.9438%	52.1966%	29.1678%	48.5045%
GANimation	0.0062%	49.2074%	0.0198%	47.3840%
StarGAN（男性）	0.2491%	54.3366%	0.4001%	43.1256%
StarGAN（年轻）	0.2656%	53.9731%	0.4603%	43.0961%
StarGAN（黑发）	0.3376%	53.2153%	0.5453%	43.0146%
StarGAN（金发）	0.4084%	54.2292%	0.5382%	42.6039%
StarGAN（棕发）	0.2077%	53.5694%	0.4344%	42.9792%
平均公共	0.1994%	0.1759%	0.4769%	0.5865%
	↑0.1146%	**↑0.0808%**	**↑0.3921%**	**↑0.4914%**
平均恶意程度	2.7741%	52.9611%	4.5094%	44.3868%
	↑0.7085%	——	**↑2.4438%**	——

成功对恶意失真保持选择性脆弱性，而过于鲁棒的 MBRS、PIMoG 和追踪器则不然。还可以观察到，可分离解码器在遇到恶意失真时会提取不匹配消息。

有趣的是，即使对于未嵌入消息的原始图像，上述结论也成立。附录中的表 6 显示，在恶意失真下，提取出的两条消息并不匹配；而在常见失真下，它们在一定程度上匹配。此外，附录中的表 7 展示了我们的观察：得益于将水印融合到图像特征时所采用的插值，在尺寸为 128×128 的图像上训练的模型能够处理任意尺寸的输入图像。

关于 CelebA 上的跨数据集设置，请参阅附录中的表 8。与 CelebA-HQ 上的结果相比，各种失真对应的比特错误率以及编码图像的峰值信噪比仅有轻微波动，从而验证了模型出色的泛化能力。附录中的表 9 还显示，在人脸图像上训练的模型竟然也能处理 COCO 中的自然图像。

4.4 消融实验

我们声明，架构中缺少任何一个组件都可能导致训练效果不理想。为了在有效设置下进行消融实验，我们考察了嵌入消息的范围，即超参数 α。因此，如表 4 所示，较小的 α 可以获得更高的视觉质量，但鲁棒性会有所降低；而较大的 α 可以实现更强的鲁棒性，但视觉质量会有所下降。请注意，α 在视觉质量与鲁棒性之间起到了权衡作用。顺便说明一下，表中的粗体数值表示与原始对应数值相比所取得的性能提升（基本上指表 2 和表 3 中的数值）。省略的术语表示恶意失真情况下的比特错误率（BER）比较，因为它们接近 50%（随机猜测），但数值越高并不一定越好。

值得一提的是，我们还尝试了两阶段学习：先将一个解码器与左侧组件联合训练，然后仅使用另一个解码器再训练 100 个训练轮次。从表 5 可以看出，无论是先训练鲁棒解码器、再训练半鲁棒解码器，还是反过来，其鲁棒性都不如 SepMark 的端到端训练（见表 2 和表 3）。一种可能的解释是，这可能是由于视觉质量方面的权衡不均衡所致。此外，我们还进行了一个玩具示例：先由 MBRS 在图像中嵌入鲁棒水印，再由 FaceSigns 嵌入第二个脆弱水印，反之亦然；这导致鲁棒解码失败，视觉质量也有所下降，如附录表 10 所示。这些结果再次证明了所提出的 SepMark 的有效性。

5 结论

本文提出了首个端到端学习的深度可分离水印架构，称为 SepMark，并将其自然地应用于主动式深度伪造取证，实现来源追踪与深度伪造检测的统一。得益于能够以不同鲁棒性级别提取嵌入水印的可分离解码器，鲁棒解码器能够抵御大量失真，而半鲁棒解码器则对恶意失真具有选择性敏感。实验结果表明，SepMark 在应对典型深度伪造方面取得了令人鼓舞的初步成果，可为可信人脸提供预防措施和更可靠的证据。未来的工作将涉及对深度伪造失真进行更通用的建模，从而使水印技术能够泛化到未见过的深度伪造样本。

致谢

本研究得到中国国家自然科学基金（项目编号：U22A2030、61972142）和中国国家重点研发计划（项目编号：2022YFB3103500）的资助。

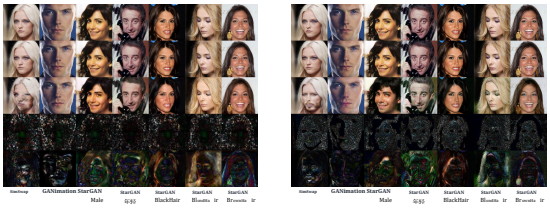
参考文献

[1] Shivangi Aneja、Lev Markhasin 和 Matthias Nießner。2022 年。TAFIM：针对人脸图像篡改的定向对抗性攻击。载于计算机视觉——*ECCV 2022*：第 *17* 届欧洲计算机视觉会议，以色列特拉维夫，*2022* 年 *10* 月 *23—27* 日，会议论文集，第 *XIV* 部分。Springer，58–75。[2] Vishal Asnani、Xi Yin、Tal Hassner、Sijia Liu 和 Xiaoming Liu。2022 年。主动式图像篡改检测。载于 *IEEE/CVF* 计算机视觉与模式识别会议论文集。15386–15395。[3] Tadas Baltrusaitis、Amir Zadeh、Yao Chong Lim 和 Louis-Philippe Morency。2018 年。OpenFace 2.0：人脸行为分析工具包。载于*2018* 年第 *13* 届 *IEEE* 国际自动人脸与手势识别会议（*FG 2018*）。IEEE，59–66。[4] Jingyi Cao、Bo Liu、Yunqian Wen、Rong Xie 和 Li Song。2021 年。通过解耦身份信息操控实现个性化且可逆的人脸去身份化。载于*IEEE/CVF* 国际计算机视觉会议论文集。3334–3342。[5] Nicholas Carlini 和 Hany Farid。2020 年。利用白盒和黑盒攻击规避深度伪造图像检测器。载于*IEEE/CVF* 计算机视觉与模式识别会议研讨会论文集。658–659。[6] Qinwei Chang、Leichao Huang、Shaoteng Liu、Hualuo Liu、Tianshu Yang 和 Yexin Wang。2022 年。基于自适应区域选择和信道参考的盲鲁棒视频水印技术。载于第 *30* 届 *ACM* 国际多媒体会议论文集。2344–2350。[7] Liang Chen、Yong Zhang、Yibing Song、Lingqiao Liu 和 Jue Wang。2022 年。对抗样本的自监督学习：迈向深度伪造检测的良好泛化能力。载于*IEEE/CVF* 计算机视觉与模式识别会议论文集。18710–18719。[8] Renwang Chen、Xuanhong Chen、Bingbing Ni 和 Yanhao Ge。2020 年。SimSwap：一种实现高保真人脸交换的高效框架。载于第 *28* 届 *ACM* 国际多媒体会议论文集。2003–2011。[9] Zhikai Chen、Lingxi Xie、Shanmin Pang、Yong He 和 Bo Zhang。2021 年。MAGDR：用于防御深度伪造的掩码引导检测与重建。载于*IEEE/CVF* 计算机视觉与模式识别会议论文集。9014–9023。[10] Yunjey Choi、Minje Choi、Munyoung Kim、Jung-Woo Ha、Sunghun Kim 和 Jaegul Choo。2018 年。StarGAN：用于多领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。8789–8797。[11] Ingemar Cox、Matthew Miller、Jeffrey Bloom、Jessica Fridrich 和 Ton Kalker。2007 年。数字水印与隐写术。Morgan Kaufmann。[12] Feng Ding、Guopu Zhu、Yingcan Li、Xinpeng Zhang、Pradeep K Atrey 和 Siwei Lyu。2021 年。通过对抗训练实现针对人脸交换视频的反取证。*IEEE* 多媒体汇刊 24（2021），3429–3441。[13] Brian Dolhansky、Joanna Bitton、Ben Pflaum、Jikuo Lu、Russ Howes、Menglin Wang 和 Cristian Canton Ferrer。2020 年。深度伪造检测挑战赛（DFDC）数据集。*arXiv* 预印本 *arXiv:2006.07397*（2020）。[14] Xiaoyi Dong、Jianmin Bao、Dongdong Chen、Ting Zhang、Weiming Zhang、Nenghai Yu、Dong Chen、Fang Wen 和 Baining Guo。2022 年。利用身份一致性 Transformer 保护名人免受深度伪造侵害。载于*IEEE/CVF* 计算机视觉与模式识别会议论文集。9468–9478。[15] Ricard Durall、Margret Keuper 和 Janis Keuper。2020 年。注意你的上采样卷积：基于 CNN 的生成式深度神经网络无法重现频谱分布。载于*IEEE/CVF* 计算机视觉与模式识别会议论文集。7890–7899。[16] Han Fang、Zhaoyang Jia、Zehua Ma、Ee-Chien Chang 和 Weiming Zhang。2022 年。PIMoG：一种用于基于深度学习的水印网络的有效屏幕拍摄噪声层模拟方法。载于第 *30* 届 *ACM* 国际多媒体会议论文集。2267–2275。[17] Han Fang、Yupeng Qiu、Kejiang Chen、Jiyi Zhang、Weiming Zhang 和 Ee-Chien Chang。2023 年。利用可逆噪声层实现针对黑盒失真的基于流的鲁棒水印技术。载于*AAAI* 人工智能会议论文集，第 37 卷。[18] Jiayun Fu、Bin B Zhu、Haidong Zhang、Yayi Zou、Song Ge、Weiwei Cui、Yun Wang、Dongmei Zhang、Xiaojing Ma 和 Hai Jin。2022 年。ChartStamp：面向真实世界应用的鲁棒图表嵌入技术。载于第 *30* 届 *ACM* 国际多媒体会议论文集。2786–2795。[19] Ian Goodfellow、Jean Pouget-Abadie、Mehdi Mirza、Bing Xu、David Warde-Farley、Sherjil Ozair、Aaron Courville 和 Yoshua Bengio。2020 年。生成对抗网络。*ACM* 通讯 63，11（2020），139–144。[20] Ziwen He、Wei Wang、Weinan Guan、Jing Dong 和 Tieniu Tan。2022 年。通过对抗性视觉重建击败深度伪造。载于第 *30* 届 *ACM* 国际多媒体会议论文集。2464–2472。<escape>[21] </escape>Kai Hong 和 Xiaoyu Du。2021。通过发现伪造痕迹差异进行自监督深度伪造检测。（2021）。<escape>[22] </escape>Yang Hou、Qing Guo、Yihao Huang、Xiaofei Xie、Lei Ma 和 Jianjun Zhao。2023。通过对抗性统计一致性规避深度伪造检测器。arXiv 预印本 arXiv:2304.11670（2023）。<escape>[23] </escape>Juan Hu、Xin Liao、Jinwen Liang、Wenbo Zhou 和 Zheng Qin。2022。Finfer：基于帧推理的高视觉质量视频深度伪造检测方法。载于《AAAI 人工智能会议论文集》，第 36 卷。951–959。<escape>[24] </escape>Juan Hu 、 Xin Liao 、 Wei Wang 和 Zheng Qin。2021。利用帧时序双流卷积网络检测社交网络中的压缩深度伪造视频。《IEEE 视频技术电路与系统汇刊》32，3（2021），1089–1102。<escape>[25] </escape>Jie Hu、Li Shen 和 Gang Sun。2018。压缩与激励网络。载于《IEEE 计算机视觉与模式识别会议论文集》。7132–7141。<escape>[26] </escape>Hao Huang、Yongtao Wang、Zhaoyu Chen、Yuze Zhang、Yuheng Li、Zhi Tang、Wei Chu、Jingdong Chen、Weisi Lin 和 Kai-Kuang Ma。2022。Cmua-watermark：一种用于对抗深度伪造的跨模型通用对抗性水印。载于《AAAI 人工智能会议论文集》，第 36 卷。989–997。<escape>[27] </escape>Qidong Huang、Jie Zhang、Wenbo Zhou、Weiming Zhang 和 Nenghai Yu。2021。针对人脸操纵的主动防御。载于《AAAI 人工智能会议论文集》，第 35 卷。1619–1627。<escape>[28] </escape>Yihao Huang、Felix Juefei-Xu、Qing Guo、Lei Ma、Xiaofei Xie、Weikai Miao、Yang Liu 和 Geguang Pu。2021。通过隐式空间域陷波滤波规避深度伪造检测。arXiv 预印本 arXiv:2009.09213（2021）。<escape>[29] </escape>Yihao Huang、Felix Juefei-Xu、Run Wang、Qing Guo、Lei Ma、Xiaofei Xie、Jianwen Li、Weikai Miao、Yang Liu 和 Geguang Pu。2020。Fakepolisher：通过浅层重建使深度伪造更难被检测。载于《第 28 届 ACM 国际多媒体会议论文集》。1217–1226。<escape>[30] </escape>Shehzeen Hussain、Paarth Neekhara、Malhar Jere、Farinaz Koushanfar 和 Julian McAuley。2021。对抗性深度伪造：评估深度伪造检测器对抗性样本的脆弱性。载于《IEEE/CVF 冬季计算机视觉应用会议论文集》。3348–3357。<escape>[31] </escape>Phillip Isola、Jun-Yan Zhu、Tinghui Zhou 和 Alexei A Efros。2017。利用条件对抗网络实现图像到图像的转换。载于《IEEE 计算机视觉与模式识别会议论文集》。1125–1134。<escape>[32] </escape>Jun Jia、Zhongpai Gao、Kang Chen、Menghan Hu、Xiongkuo Min、Guangtao Zhai 和 Xiaokang Yang。2020。RIHOOP：适用于离线和在线照片的鲁棒隐形超链接。《IEEE 控制论汇刊》52，7（2020），7094–7106。<escape>[33] </escape>Jun Jia、Zhongpai Gao、Dandan Zhu、Xiongkuo Min、Guangtao Zhai 和 Xiaokang Yang。2022。为离线到在线摄影中的隐藏代码学习隐形标记。载于《IEEE/CVF 计算机视觉与模式识别会议论文集》。2273–2282。<escape>[34] </escape>Zhaoyang Jia、Han Fang 和 Weiming Zhang。2021。MBRS：通过真实与模拟 JPEG 压缩的小批量训练增强基于深度神经网络的水印鲁棒性。载于《第 29 届 ACM 国际多媒体会议论文集》。41–49。<escape>[35] </escape>Alexia Jolicoeur-Martineau。2018。相对判别器：标准 GAN 中缺失的关键要素。arXiv 预印本 arXiv:1807.00734（2018）。<escape>[36] </escape>Felix Juefei-Xu、Run Wang、Yihao Huang、Qing Guo、Lei Ma 和 Yang Liu。2022。应对恶意深度伪造：综述、攻防战场与未来展望。《国际计算机视觉杂志》130，7（2022），1678–1734。<escape>[37] </escape>Tero Karras、Timo Aila、Samuli Laine 和 Jaakko Lehtinen。2017。通过 GAN 渐进式增长提升质量、稳定性和多样性。arXiv 预印本 arXiv:1710.10196（2017）。<escape>[38] </escape>Diederik P Kingma 和 Jimmy Ba。2014。Adam：一种随机优化方法。arXiv 预印本 arXiv:1412.6980（2014）。<escape>[39] </escape>Cheng-Han Lee、Ziwei Liu、Lingyun Wu 和 Ping Luo。2020。Maskgan：实现多样化和交互式人脸图像操纵。载于《IEEE/CVF 计算机视觉与模式识别会议论文集》。5549–5558。<escape>[40] </escape>Dongze Li、Wei Wang、Kang Zhao、Jing Dong 和 Tieniu Tan。2023。RiDDLE：利用潜在加密器实现可逆且多样化的去身份化。arXiv 预印本 arXiv:2303.05171（2023）。<escape>[41] </escape>Haodong Li、Bin Li、Shunquan Tan 和 Jiwu Huang。2020。利用颜色分量差异识别深度网络生成的图像。《信号处理》174（2020），107616。<escape>[42] </escape>Lingzhi Li、Jianmin Bao、Ting Zhang、Hao Yang、Dong Chen、Fang Wen 和 Baining Guo。2020。利用 Face X-ray 实现更通用的人脸伪造检测。载于《IEEE/CVF 计算机视觉与模式识别会议论文集》。5001–5010。<escape>[43] </escape>Yuezun Li、Ming-Ching Chang 和 Siwei Lyu。2018。转瞬之间：通过检测眨眼揭露人工智能生成的虚假视频。载于《2018 年 IEEE 信息取证与安全国际研讨会（WIFS）》。IEEE，1–7。<escape>[44] </escape>Yuezun Li 和 Siwei Lyu。2018。通过检测人脸扭曲痕迹揭露深度伪造视频。arXiv 预印本 arXiv:1811.00656（2018）。<escape>[45] </escape>Xin Liao、Yumei Wang、Tianyi Wang、Juan Hu 和 Xiaoshuai Wu。2023。FAMM：利用面部肌肉运动检测社交网络中的压缩深度伪造视频。《IEEE 视频技术电路与系统汇刊》（2023）。<escape>[46] </escape>Tsung-Yi Lin、Michael Maire、Serge Belongie、James Hays、Pietro Perona、Deva Ramanan、Piotr Dollár 和 C La wrence Zitnick。2014。Microsoft COCO：上下文中的常见物体。载于《计算机视觉——ECCV 2014：第 13 届欧洲计算机视觉会议》，

苏黎世，瑞士，2014年9月6日至12日，论文集，第V部分 13。Springer，740–755。[47] Yuzhen Lin、Han Chen、Emanuele Maiorana、Patrizio Campisi 和 Bin Li。2022。Source-ID-Tracker：人脸交换中的来源人脸身份保护。载于《2022年IEEE国际多媒体与博览会会议（ICME）》。IEEE，01–06。[48] Yang Liu、Mengxi Guo、Jian Zhang、Yuesheng Zhu 和 Xiaodong Xie。2019。一种用于实际盲水印的新型两阶段可分离深度学习框架。载于《第27届ACM国际多媒体会议论文集》。1509–1517。[49] Yang Liu、Zhenyue Qin、Saeed Anwar、Pan Ji、Dongwoo Kim、Sabrina Caldwell 和 Tom Gedeon。2021。可逆去噪网络：一种用于真实噪声去除的轻量化方案。载于《IEEE/CVF计算机视觉与模式识别会议论文集》。13365–13374。[50] Ziwei Liu、Ping Luo、Xiaogang Wang 和 Xiaoou Tang。2015。野外人脸属性的深度学习。载于《IEEE国际计算机视觉会议论文集》。3730–3738。[51] Xiyang Luo、Ruohan Zhan、Huiwen Chang、Feng Yang 和 Peyman Milanfar。2020。与失真无关的深度水印。载于《IEEE/CVF计算机视觉与模式识别会议论文集》。13548–13557。[52] Rui Ma、Mengxi Guo、Yi Hou、Fan Yang、Yuan Li、Huizhu Jia 和 Xiaodong Xie。2022。迈向盲水印：结合可逆与不可逆机制。载于《第30届ACM国际多媒体会议论文集》。1532–1542。[53] Vahid Mirjalili、Sebastian Raschka 和 Arun Ross。2020。PrivacyNet：用于多属性人脸隐私保护的半对抗网络。《IEEE图像处理汇刊》29（2020），9400–9412。[54] Seyed-Mohsen Moosavi-Dezfooli、Alhussein Fawzi、Omar Fawzi 和 Pascal Frossard。2017。通用对抗扰动。载于《IEEE计算机视觉与模式识别会议论文集》。1765–1773。[55] Paarth Neekhara、Brian Dolhansky、Joanna Bitton 和 Cristian Canton Ferrer。2021。深度伪造检测面临的对抗性威胁：实践视角。载于《IEEE/CVF计算机视觉与模式识别会议论文集》。923–932。[56] Paarth Neekhara、Shehzeen Hussain、Xinqiao Zhang、Ke Huang、Julian McAuley 和 Farinaz Koushanfar。2022。FaceSigns：用于媒体认证和抵御深度伪造的半脆弱神经水印。arXiv预印本 arXiv:2204.01960（2022）。[57] Zhicheng Ni、Yun-Qing Shi、Nirwan Ansari 和 Wei Su。2006。可逆数据隐藏。《IEEE视频技术电路与系统汇刊》16，3（2006），354–362。[58] Adam Paszke、Sam Gross、Francisco Massa、Adam Lerer、James Bradbury、Gregory Chanan、Trevor Killeen、Zeming Lin、Natalia Gimelshein、Luca Antiga 等。2019。PyTorch：一种命令式风格的高性能深度学习库。《神经信息处理系统进展》32（2019）。[59] Pengfei Pei、Xianfeng Zhao、Yun Cao、Jinchuan Li 和 Xuyuan Lai。2021。基于视觉Transformer的视频哈希检索，用于追踪伪造视频的来源。arXiv预印本 arXiv:2112.08117（2021）。[60] Albert Pumarola、Antonio Agudo、Aleix M Martinez、Alberto Sanfeliu 和 Francesc Moreno-Noguer。2018。GANimation：根据单张图像生成具备解剖感知能力的人脸动画。载于《欧洲计算机视觉会议（ECCV）论文集》。818–833。[61] Alec Radford、Jong Wook Kim、Chris Hallacy、Aditya Ramesh、Gabriel Goh、Sandhini Agarwal、Girish Sastry、Amanda Askell、Pamela Mishkin、Jack Clark 等。2021。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》。PMLR，8748–8763。[62] Alec Radford、Karthik Narasimhan、Tim Salimans、Ilya Sutskever 等。2018。通过生成式预训练提升语言理解能力。（2018）。[63] Robin Rombach、Andreas Blattmann、Dominik Lorenz、Patrick Esser 和 Björn Ommer。2022。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》。10684–10695。[64] Andreas Rossler、Davide Cozzolino、Luisa Verdoliva、Christian Riess、Justus Thies 和 Matthias Nießner。2019。Faceforensics++：学习检测经过篡改的人脸图像。载于《IEEE/CVF国际计算机视觉会议论文集》。1–11。[65] Nataniel Ruiz、Sarah Adel Bargal 和 Stan Sclaroff。2020。扰乱深度伪造：针对条件图像转换网络和人脸篡改系统的对抗攻击。载于《计算机视觉——ECCV 2020研讨会：英国格拉斯哥，2020年8月23日至28日，论文集，第IV部分 16》。Springer，236–251。[66] Hong Shen 和 Bo Chen。2012。从单水印到双水印：一种新的图像水印方法。《计算机与电气工程》38，5（2012），1310–1324。[67] Kaede Shiohara 和 Toshihiko Yamasaki。2022。利用自混合图像检测深度伪造。载于《IEEE/CVF计算机视觉与模式识别会议论文集》。18720–18729。[68] Jingxiang Sun、Xuan Wang、Yong Zhang、Xiaoyu Li、Qi Zhang、Yebin Liu 和 Jue Wang。2022。Fenerf：神经辐射场中的人脸编辑。载于《IEEE/CVF计算机视觉与模式识别会议论文集》。7672–7682。

[69] Pu Sun、Yuezun Li、Honggang Qi 和 Siwei Lyu。2022。Faketracer：通过训练数据污染揭露深度伪造。发表于《2022 年 IEEE 国际图像处理会议（ICIP）》。IEEE，1161–1165。[70] Matthew Tancik、Ben Mildenhall 和 Ren Ng。2020。Stegastamp：物理照片中的隐形超链接。发表于《IEEE/CVF 计算机视觉与模式识别会议论文集》。2117–2126。[71] Run Wang、Ziheng Huang、Zhikai Chen、Li Liu、Jing Chen 和 Lina Wang。2022。Anti-Forgery：通过对抗性感知扰动实现隐蔽且鲁棒的深度伪造干扰攻击。arXiv 预印本 arXiv:2206.00477（2022）。[72] Run Wang、Felix Juefei-Xu、Meng Luo、Yang Liu 和 Lina Wang。2021。Faketagger：通过溯源追踪防范深度伪造传播的鲁棒保护机制。发表于《第 29 届 ACM 国际多媒体会议论文集》。3546–3555。[73] Xueyu Wang、Jiajun Huang、Siqi Ma、Surya Nepal 和 Chang Xu。2022。Deepfake disrupter：深度伪造检测器是我的朋友。发表于《IEEE/CVF 计算机视觉与模式识别会议论文集》。14920–14929。[74] Eric Wengrowski 和 Kristin Dana。2019。利用深度摄影隐写术进行光场信息传递。发表于《IEEE/CVF 计算机视觉与模式识别会议论文集》。1515–1524。[75] Xiaoshuai Wu、Tong Qiao、Yanli Chen、Ming Xu、Ning Zheng 和 Xiangyang Luo。2022。重新审视符号隐写术：采用鲁棒的域选择方法。《信号处理》196（2022），108522。[76] Zhiliang Xu、Zhibin Hong、Changxing Ding、Zhen Zhu、Junyu Han、Jingtuo Liu 和 Errui Ding。2022。Mobilefaceswap：一种轻量级视频人脸交换框架。发表于《AAAI 人工智能会议论文集》，第 36 卷。2973–2981。[77] Zhiliang Xu、Hang Zhou、Zhibin Hong、Ziwei Liu、Jiaming Liu、Zhizhi Guo、Junyu Han、Jingtuo Liu、Errui Ding 和 Jingdong Wang。2022。Styleswap：基于风格的生成器赋能鲁棒人脸交换。发表于《欧洲计算机视觉会议》。Springer，661–677。[78] Xin Yang、Yuezun Li 和 Siwei Lyu。2019。利用不一致的头部姿态揭露深度伪造。发表于《ICASSP 2019——2019 年 IEEE 声学、语音与信号处理国际会议（ICASSP）》。IEEE，8261–8265。[79] Chin-Yuan Yeh、Hsi-Wen Chen、Hong-Han Shuai、De-Nian Yang 和 Ming-Syan Chen。2021。以攻为守：通过具有约束感知能力的对抗攻击使图像到图像转换 GAN 失效。发表于《IEEE/CVF 国际计算机视觉会议论文集》。16188–16197。[80] Chin-Yuan Yeh、Hsi-Wen Chen、Shang-Lun Tsai 和 Sheng-De Wang。2020。利用对抗攻击破坏基于图像转换的深度伪造算法。发表于《IEEE/CVF 冬季计算机视觉应用会议研讨会论文集》。53–62。[81] Ning Yu、Vladislav Skripniuk、Sahar Abdelnabi 和 Mario Fritz。2021。生成模型的人工指纹标记：将深度伪造归因追溯至训练数据。发表于《IEEE/CVF 国际计算机视觉会议论文集》。14448–14457。[82] Zhuowen Yuan、Zhengxin You、Sheng Li、Zhenxing Qian、Xinpeng Zhang 和 Alex Kot。2022。关于可识别虚拟人脸的生成。发表于《第 30 届 ACM 国际多媒体会议论文集》。1465–1473。[83] Chaoning Zhang、Adil Karjauv、Philipp Benz 和 In So Kweon。2021。面向实际盲水印的非可微失真条件下鲁棒深度信息隐藏。发表于《第 29 届 ACM 国际多媒体会议论文集》。5158–5166。[84] Richard Zhang、Phillip Isola、Alexei A Efros、Eli Shechtman 和 Oliver Wang。2018。深度特征作为感知度量的惊人有效性。发表于《IEEE 计算机视觉与模式识别会议论文集》。586–595。[85] Hanqing Zhao、Wenbo Zhou、Dongdong Chen、Weiming Zhang 和 Nenghai Yu。2022。用于深度伪造检测的自监督 Transformer。arXiv 预印本 arXiv:2203.01265（2022）。[86] Tianchen Zhao、Xiang Xu、Mingze Xu、Hui Ding、Yuanjun Xiong 和 Wei Xia。2021。学习自一致性以检测深度伪造。发表于《IEEE/CVF 国际计算机视觉会议论文集》。15023–15033。[87] Jiaran Zhou 和 Yuezun Li。2022。基于模拟的检测：通过模拟人脸重建伪造揭露深度伪造。发表于《2022 年 IEEE 第五届多媒体信息处理与检索国际会议（MIPR）》。IEEE，210–215。[88] Peng Zhou、Xintong Han、Vlad I Morariu 和 Larry S Davis。2017。用于篡改人脸检测的双流神经网络。发表于《2017 年 IEEE 计算机视觉与模式识别会议研讨会（CVPRW）》。IEEE，1831–1839。[89] Jiren Zhu、Russell Kaplan、Justin Johnson 和 Li Fei-Fei。2018。Hidden：利用深度网络隐藏数据。发表于《欧洲计算机视觉会议（ECCV）论文集》。657–672。[90] Jun-Yan Zhu、Taesung Park、Phillip Isola 和 Alexei A Efros。2017。利用循环一致性对抗网络实现无配对图像到图像转换。发表于《IEEE 国际计算机视觉会议论文集》。2223–2232。[91] Bojia Zi、Minghao Chang、Jingjing Chen、Xingjun Ma 和 Yu-Gang Jiang。2020。Wilddeepfake：用于深度伪造检测的真实世界挑战性数据集。发表于《第 28 届 ACM 国际多媒体会议论文集》。2382–2390。

A 附录 附录A.1 补充视觉示例



(a) 图像尺寸： 128×128 **(b) 图像尺寸：** 256×256

图 6：更多恶意失真。从上到下依次为封面图像 I_{co}、编码图像 I_{en}、加噪图像I_{no}、$\mathcal{N}(|I_{co} - I_{en}|)$ 的残差信号，以及 $\mathcal{N}(|I_{no} - I_{en}|)$。

A.2 补充实验结果

表 6：关于原始图像 I_{co} 的更多研究。"两者"表示 M_{tr} 和 M_{de} 之间的比特错误率。

	128×128	256×256
	峰值信噪比 = +∞	峰值信噪比 = +∞
失真	Both	Both
身份	18.3534%	15.8034%
JpegTest	14.3272%	16.7296%
调整大小	17.1341%	15.4153%
高斯模糊	18.0217%	17.9925%
中值模糊	17.6121%	17.0536%
亮度	18.7901%	16.7889%
对比度	18.6202%	16.8386%
饱和度	18.3392%	15.9481%
Hue	18.2649%	15.8679%
随机失活	18.3534%	15.8034%
椒盐噪声	27.8199%	23.0745%
高斯噪声	16.0966%	18.7992%
SimSwap	50.5406%	47.4623%
GANimation	48.6370%	49.8244%
StarGAN（男性）	50.4049%	49.0832%
StarGAN（年轻）	50.6209%	49.2378%
StarGAN（黑发）	49.5609%	48.9125%
StarGAN（金发）	48.7028%	48.9916%
StarGAN（棕发）	50.1771%	49.0907%
平均公共	18.4777%	17.1763%
平均恶意	49.8063%	48.9432%

表 7：关于 128×128 训练模型在任意图像输入尺寸 I_{co} 下的进一步研究。

	256×256		512×512	
	峰值信噪比 = 35.4984		峰值信噪比 = 35.2925	
失真	追踪器	检测器	追踪器	检测器
身份	0.0000%	5.2467%	0.0000%	25.1287%
JpegTest	0.1735%	0.9679%	0.1062%	0.7177%
调整大小	0.0508%	2.9592%	0.0165%	18.6532%
高斯模糊	0.0106%	11.0127%	0.0035%	33.1704%
中值模糊	0.0071%	9.0061%	0.0035%	30.9301%
亮度	0.0153%	5.8735%	0.0165%	25.0118%
对比度	0.0177%	6.2028%	0.0106%	25.3234%
饱和度	0.0000%	5.4958%	0.0000%	25.3966%
Hue	0.0000%	5.5902%	0.0000%	26.1036%
随机失活	0.0165%	7.6546%	0.0224%	26.1674%
椒盐噪声	0.1558%	2.0491%	0.2325%	5.9832%
高斯噪声	0.5323%	3.9341%	0.3895%	8.1161%
SimSwap	2.5484%	49.9115%	1.9677%	49.1407%
GANimation	0.0679%	46.8494%	0.1136%	44.9160%
StarGAN（男性）	0.2668%	50.2101%	0.0873%	49.8430%
StarGAN（年轻）	0.2821%	50.0390%	0.0814%	49.4653%
StarGAN（黑发）	0.2986%	50.6232%	0.0755%	50.8263%
StarGAN（金发）	0.3553%	49.5574%	0.0921%	48.2177%
StarGAN（棕发）	0.2644%	50.3057%	0.0767%	50.3836%
平均公共	0.0816%	5.4994%	0.0668%	20.8919%
		↑5.4043%	**↓0.0180%**	**↑20.7968%**
平均恶意行为	0.5834%	49.6423%	0.3563%	48.9704%
	↓1.4822%	——	**↓1.7093%**	——

表 8：跨数据集 CelebA 上图像 I_{no} 在常见失真和恶意失真下的鲁棒性测试。

	128×128		256×256	
	峰值信噪比 = 37.8865		峰值信噪比 = 38.7051	
失真	追踪器	检测器	追踪器	检测器
身份	0.0000%	0.0417%	0.0063%	0.0002%
JpegTest	0.2233%	0.2934%	0.0247%	0.0194%
调整大小	0.0134%	0.0272%	0.0001%	0.0002%
高斯模糊	0.0058%	0.0416%	0.0004%	0.0022%
中值模糊	0.0037%	0.0730%	0.0003%	0.0002%
亮度	0.0094%	0.0579%	0.0160%	0.0016%
对比度	0.0068%	0.0698%	0.0034%	0.0002%
饱和度	0.0000%	0.0519%	0.0056%	0.0002%
Hue	0.0000%	0.0401%	0.0070%	0.0002%
随机失活	0.0020%	0.0625%	0.4647%	0.0004%
椒盐噪声	0.0406%	0.0242%	0.0017%	0.0005%
高斯噪声	0.7461%	0.8952%	0.1594%	0.1708%
SimSwap	12.2428%	50.7506%	10.2216%	34.2429%
GANimation	0.0007%	39.6265%	0.0001%	45.5291%
StarGAN（男性）	0.1209%	50.8483%	0.0019%	45.5910%
StarGAN（年轻）	0.1358%	50.6294%	0.0025%	45.6882%
StarGAN（黑发）	0.1479%	48.0443%	0.0021%	46.1099%
StarGAN（金发）	0.1840%	46.7939%	0.0020%	44.6107%
StarGAN（棕发）	0.1196%	50.8680%	0.0014%	45.9944%
平均公共	0.0876%	0.1399%	0.0575%	0.0163%
	↑0.0028%	**↑0.0448%**	**↑0.0514%**	**↑0.0080%**
平均恶意程度	1.8502%	48.2230%	1.4617%	43.9666%
	↓0.2154%	——	**↑0.3301%**	——

表9：跨数据集 COCO 上常规失真图像 I_{no}的鲁棒性测试。

	128×128		256×256	
	峰值信噪比 = 35.1500		峰值信噪比 = 35.5506	
失真	追踪器	检测器	追踪器	检测器
身份	0.2100%	1.4440%	0.0420%	0.0475%
JpegTest	1.6027%	3.6627%	0.1691%	0.1439%
调整大小	3.9873%	2.8900%	0.2661%	0.2502%
高斯模糊	1.1447%	4.0467%	0.0698%	0.1478%
中值模糊	0.7460%	3.4340%	0.1005%	0.0878%
亮度	0.4560%	1.9213%	0.1745%	0.1500%
对比度	0.3860%	1.7367%	0.0756%	0.0639%
饱和度	0.2333%	1.5240%	0.0438%	0.0567%
Hue	0.2293%	1.6453%	0.0428%	0.0495%
随机失活	0.8800%	2.8593%	0.2595%	0.1906%
椒盐噪声	1.8013%	1.9353%	0.1944%	0.1355%
高斯噪声	1.0473%	2.2393%	0.2186%	0.2402%
平均公共	1.0603%	2.4449%	0.1381%	0.1303%
	↑0.9755%	**↑2.3498%**	**↑0.1320%**	**↑0.1220%**

表 10：玩具示例。先嵌入一种水印，随后再嵌入另一种水印。图像尺寸： 256×256。

	MBRS→人脸Signs		人脸标记→MBRS	
	峰值信噪比 = 32.1240		峰值信噪比 = 32.0076	
失真	MBRS	FaceSigns	MBRS	FaceSigns
身份	0.3378%	0.0133%	0.0000%	0.0340%
JpegTest	5.0135%	0.8648%	0.1546%	1.5589%
调整大小	5.7166%	1.1027%	1.6975%	1.5783%
高斯模糊	6.9225%	0.1328%	3.6559%	0.1676%
中值模糊	5.1384%	0.1151%	2.3408%	1.5160%
亮度	1.7552%	0.0764%	1.0235%	0.0943%
对比度	1.7989%	0.0318%	0.9658%	0.0515%
饱和度	0.3329%	0.7660%	0.0000%	0.8219%
Hue	0.9683%	8.0228%	0.0296%	9.3728%
随机失活	4.5914%	17.6177%	38.8410%	5.9916%
椒盐噪声	40.2750%	12.4776%	32.2704%	13.3133%
高斯噪声	25.7147%	0.8963%	12.5065%	1.2598%
SimSwap	46.5361%	49.8661%	45.8514%	49.8559%
GANimation	2.7499%	45.2734%	0.2568%	45.5127%
StarGAN（男性）	11.8858%	50.1740%	7.7480%	50.2683%
StarGAN（年轻）	11.7510%	50.3737%	7.8313%	50.4302%
StarGAN（黑发）	13.0742%	50.1859%	8.9936%	50.2606%
StarGAN（金发）	10.3177%	50.9066%	6.8103%	51.1755%
StarGAN（棕发）	11.0078%	50.7384%	7.1055%	50.9381%
平均公共	8.2138%	3.5098%	7.7905%	2.9800%
	↑8.2077%	**↑3.5015%**	**↑7.7844%**	**↑2.9717%**
平均恶意值	15.3318%	49.6454%	12.0853%	49.7773%
	↑14.2002%	——	**↑10.9537%**	——

附录中的每幅图均按从上到下的顺序展示封面图像 I_{co}、编码图像 I_{en}、加噪图像 I_{no}、N 的残差信号（$|I_{co} - I_{en}|$）和 N（$|I_{no} - I_{en}|$）。



图 7：SimSwap。数据集：CelebA-HQ。图像尺寸： 256×256。

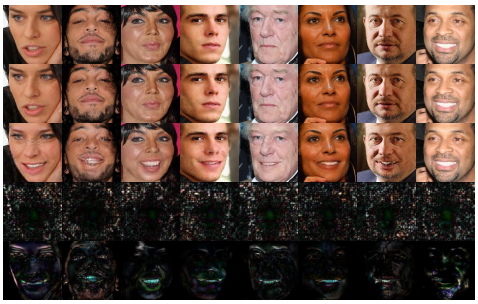


图 8：GANimation：仅在测试时，才使用 Open- Face 处理人脸图像，以满足输入要求。数据集：CelebA-HQ。图像尺寸： 128×128。



图 9：StarGAN（男性）。数据集：CelebA-HQ。图像尺寸： 256×256。



图 10：StarGAN（年轻）。数据集：CelebA-HQ。图像尺寸： 256×256。

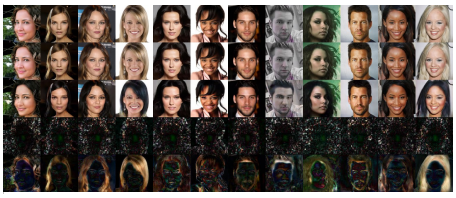


图 11：StarGAN（黑发）。数据集：CelebA-HQ。图像尺寸： 128×128。



图 12：StarGAN（金发）。数据集：CelebA-HQ。图像尺寸： 128×128。



图 13：StarGAN（棕发）。数据集：CelebA-HQ。图像尺寸： 128×128。

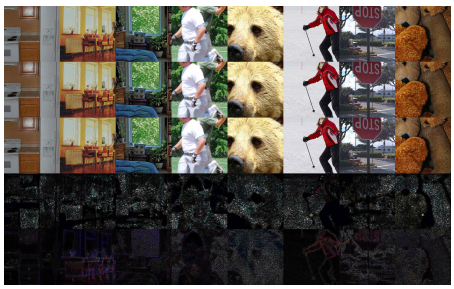


图 14：JpegTest。数据集：COCO。图像尺寸： 256×256。

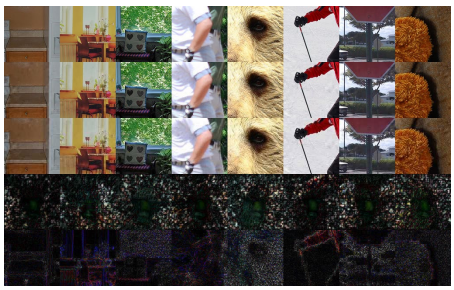


图 15：JpegTest。数据集：COCO。图像尺寸： 128×128。