arXiv:2601.11978v1 [eess.IV] 17 Jan 2026

NiMark：一种抵御屏幕拍摄攻击的非侵入式水印框架

Yufeng Wu[1] , **Xin Liao**[1] , **Baowei Wang**[2] , **Han Fang**[3] , **Xiaoshuai Wu**[1] 和**Guiling Wang**[4]1湖南大学网络空间安全与技术学院2南京信息工程大学计算机科学学院3新加坡国立大学计算学院4新泽西理工学院计算机科学系 wuyufeng0523@163.com, xinliao@hnu.edu.cn, wbw.first@163.com, fanghan@nus.edu.sg, shinewu@hnu.edu.cn, gwang@njit.edu

摘要

未经授权的屏幕拍摄会带来严重的数据泄露风险。抵御屏幕拍摄攻击通常需要嵌入高强度水印，这不可避免地会降低载体图像的质量。为解决鲁棒性与保真度冲突，非侵入式水印通过构建逻辑验证密钥而不改变原始内容，成为一种解决方案。然而，现有的非侵入式方案无法抵御屏幕拍摄噪声。尽管深度学习有望解决这一问题，但我们发现，直接应用深度学习会导致一种此前鲜少探讨的失败模式——结构捷径：网络倾向于学习简单的恒等映射，而忽略图像与水印之间的绑定。此外，即使强制建立逻辑绑定，标准训练策略仍无法完全弥合噪声差距，因而在抵御物理失真方面的鲁棒性欠佳。本文提出端到端框架 NiMark，以应对这些挑战。首先，为消除结构捷径，我们引入 Sigmoid 门控异或（SG-XOR）估计器，使逻辑运算能够进行梯度传播，从而有效强制建立严格的图像—水印绑定。其次，为突破鲁棒性瓶颈，我们设计了一种两阶段训练策略，将恢复器纳入其中，以弥合屏幕拍摄噪声导致的域差距。实验表明，与具有代表性的最先进方法相比，NiMark 在抵御数字攻击和屏幕拍摄噪声方面始终表现更优，同时保持零视觉失真。

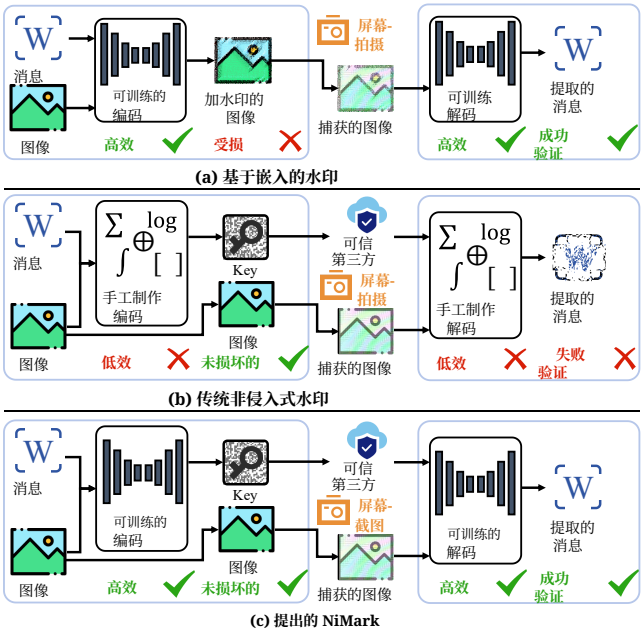


图 1：不同水印框架抵御屏幕拍摄攻击的对比。(a) 基于嵌入的水印：尽管能够抵御屏幕拍摄，但通过显式修改像素值，不可避免地会降低载体图像的质量。(b) 传统非侵入式水印：虽然能够保留原始图像，但由于手工设计特征的鲁棒性有限，无法抵御屏幕拍摄造成的严重失真。(c) 提出的 NiMark：借助深度学习的强大能力，我们的框架在非侵入式架构下实现了抵御屏幕拍摄攻击的强大鲁棒性，并保证零视觉失真。

1 引言

通过外部摄像机非法拍摄屏幕显示内容，对信息安全构成了严重威胁。数字水印长期以来一直是识别所有权和追踪泄漏的基础防御机制。

图像源。然而，与无损数字传输不同，这一过程会引入复杂的失真，例如莫尔纹和亮度失真 [Fang等，2018年；Wengrowski 和 Dana， 2019]，使大多数传统水印方案失效。当前的解决方案主要依赖基于嵌入的水印 [Tancik 等，2020年；Fang 等，2022年；Li 等， 2024]。尽管这些方法具有较强的鲁棒性，但它们通过修改像素值不可避免地降低视觉质量（图 1(a)）。在要求严格保障数据完整性的高安全性应用中，此类修改是不可接受的，因此亟需采用非侵入式替代方案。

为规避保真度-鲁棒性权衡，非

侵入式水印，也称为零水印[Wen等人， 2003], 已成为重要的研究方向。图 1(b) 展示了这种方法：算法不修改图像，而是提取关键特征，并将其与版权信息进行逻辑组合，以构建验证密钥。然而，当前非侵入式水印研究面临着重大瓶颈。绝大多数现有方案仍依赖传统的手工设计算法。这些手工特征提取器基于信号处理的先验知识设计，缺乏捕捉深层语义特征的自适应能力，因而难以抵御屏幕拍摄带来的严重非线性失真。因此，如图 1 中间一行所示，传统方法在真实的屏幕拍摄攻击下往往无法验证版权。

相反，尽管深度神经网络彻底改变了图像处理，但用于实现鲁棒非侵入式水印的端到端深度学习框架仍几乎是一片未被探索的领域。在探索这片未知领域时，我们发现了一个关键障碍：结构捷径。非侵入式水印要求载体图像与版权信息之间建立紧密绑定。然而，在深度学习框架中强制实现这种依赖关系颇具挑战。由于缺乏强制逻辑绑定的机制，网络可能会收敛到一种退化解，基本忽略图像特征，转而近似于一种简单的恒等映射。关键在于，如果密钥变成独立于图像的自包含载体，版权绑定就会被切断，从而削弱保护机制。

此外，即使避免了结构捷径，另一个挑战依然存在：噪声域差距。同时优化逻辑绑定与抵抗屏幕拍摄的能力，会使学习过程更加复杂。标准训练策略无法完全弥合域差距，导致模型抵御真实世界噪声的鲁棒性不足。

为应对这些挑战，我们提出了 NiMark——首个专为抵抗屏幕拍摄的非侵入式水印而设计的端到端深度学习框架，如图 1(c) 所示。不同于以往方法，NiMark 利用深度神经网络确保鲁棒性，同时保证零视觉失真。我们设计的方法旨在通过模型构建和定制化训练策略，规避结构捷径并实现鲁棒性。

首先，受传统非侵入式方案的启发，我们将逻辑异或运算确定为绑定图像特征与水印的关键约束。为将这一离散逻辑融入深度学习框架，我们提出了 Sigmoid门控异或（SG-XOR）估计器。通过为离散异或逻辑构建可微代理，该模块在反向传播过程中施加约束。这一设计显著增强了生成密钥对载体图像特征的依赖性，使得在优化过程中采用简单的恒等映射在实践中变得不利。

其次，为实现对屏幕拍摄的鲁棒性，我们设计了两阶段训练策略。鉴于屏幕拍摄噪声会造成严重的域差距，我们在流程中引入了专用恢复器。通过将逻辑绑定学习与物理恢复解耦，该策略确保模型首先学习正确的依赖关系，然后通过在解码前恢复图像质量来适应屏幕拍摄噪声。

本文的主要贡献总结如下：

- 我们提出了 NiMark，这是首个针对屏幕拍摄具有鲁棒性的端到端非侵入式水印深度学习框架，有效克服了传统手工设计特征的瓶颈。
- 我们设计了 SG-XOR 估计器，使梯度能够通过离散逻辑生成过程传播。该模块通过抑制网络收敛到无意义的结构性捷径特征，强化图像-水印绑定。
- 我们提出了一种整合独立恢复器的两阶段训练策略。该方法通过在解码前校正失真，弥合屏幕拍摄带来的域差距。
- 实验表明，NiMark 的鲁棒性始终优于现有最先进方法，同时在非侵入式设置下保持零视觉失真。

2 相关工作

2.1 基于嵌入的鲁棒水印

传统水印技术在空间域或变换域中运行 [Wolfgang 等，2002年；Chu，2003年；Zhang 和 Wei，2020年；Liu 和 Tan， 2002] 但难以应对非线性失真。深度学习推动了这一领域的发展[Wu 等， 2025], 并催生了 HiDDeN 等端到端框架[Zhu 等， 2018] 以及后续改进方法 [Liu 等，2019年；Jia 等，2021年；Sander等， 2025]。针对屏幕-相机信道，早期的模板匹配方法 [Fang 等，2018年；Fang 等， 2021] 已被采用可微噪声层的深度学习方法取代。StegaStamp[Tancik 等， 2020]和 PIMoG [Fang 等， 2022] 模拟透视和莫尔纹失真，而其他方法则探索逆向编码 [Wang 等， 2025]。近期研究致力于解决几何同步问题 [He 等， 2024], 局部遮挡问题 [Ma等，2025年；Chen 等， 2025], 以及数据驱动模拟问题 [Wengrowski 和 Dana，2019年；Gao 等，2025年；Wu 等，2026]。然而，这些方法都不可避免地会修改载体图像的像素以嵌入信号，从而损害视觉保真度，并成为要求严格保证数据完整性的应用的瓶颈。

2.2 非侵入式水印与捷径学习

非侵入式方案 [Wen 等人， 2003] 构建逻辑验证密钥以保持图像质量。传统算法依赖正交矩 [Gao 和 Jiang，2015年；Wang等人， 2017] 或其几何不变量 [Wang 等人，2016；Wang 等人，2019年；Yang 等人，2020年；Yang 等人，2021]。近期研究已将这些方法扩展到光场 [Wen等人，2025], Radon 空间 [Qi 等人， 2024], 以及基于区块链的协议 [Wang 等人，2022年；Hou 等人， 2025]。尽管

这些方法能够实现零失真，但这些手工设计特征无法抵御严重的屏幕拍摄攻击。

尽管深度学习方法已相继出现 [Li 等，2025年；Liu 等， 2025], 将其应用于这项任务可能会引发捷径学习 [Geirhos 等， 2020]。在一般视觉任务中，模型往往会利用背景等表面相关性 [Beery 等， 2018], 纹理 [Geirhos 等， 2018], 或高频伪影 [Jo 和 Bengio，2017；Smeu 等， 2025]。然而，这一现象在非侵入式水印中的具体表现仍未得到探索。在这项工作中，我们发现了一种特定领域的失败模式，称为结构捷径：网络通过完全绕过图像-水印绑定来最小化损失。NiMark 旨在解决这一关键漏洞，而它也是一种有别于以往研究的捷径特征的独特挑战。

3 方法

图2展示了 NiMark 框架。本节详细介绍所提出水印系统的四个核心组成部分：网络架构、用于强制实现逻辑绑定的 SG-XOR 估计器，以及用于处理屏幕拍摄失真的两阶段训练策略。

3.1 网络架构

如图 2 上方所示，整体框架由四个依次排列的模块组成：编码器、噪声层、恢复器和解码器。

编码器

编码器的目标是根据载体图像 $I_{co} \in \mathbb{R}^{C\times H\times W}$ 和二进制消息 $M \in \{0,1\}^{L}$生成密钥。为促进一维消息向量与二维图像特征之间的交互，我们设计了一种双分支架构。在图像特征提取方面，我们采用可微矩特征提取器。我们将 I_{co} 的灰度版本投影到极坐标系中定义的一组径向谐波核上，以直接提取紧凑的几何特征图。同时，在消息分支中，我们采用消息处理器。在该模块中，首先对消息向量进行重塑，然后使用卷积层进行处理，使其与提取的图像特征在空间分辨率上相匹配。令 $E(\cdot)$表示编码器。它将归一化后的矩特征与处理后的消息特征融合，生成编码后的密钥 $I_{en} \in \{0,1\}^{C'\times H'\times W'}$，其中$C'$、 H'、W'分别表示特征图的通道数、高度和宽度。

$$I_{en} = E(I_{co}, M) \tag{1}$$

噪声层

在基于深度学习的水印中，噪声层在训练过程中充当关键的对抗因素。通过准确模拟真实世界失真，它促使网络自适应地学习鲁棒表征，从而抵御此类攻击。尽管我们的框架采用非侵入式设计，我们仍沿用这一成熟策略来获得鲁棒性。我们集成了一个可微分的噪声层 $N(\cdot)$，用于模拟屏幕拍摄的模数转换过程，其中包含四种关键物理失真：

a) 透视失真：我们应用随机的四点透视变换矩阵来模拟几何变形。**b) 亮度失真：**按照 $I' = I + b$ 施加线性亮度偏移，其中 $b \sim U(0.5, 0.7)$。**莫尔′纹：**我们注入由随机方向的叠加余弦函数生成的莫尔′纹失真（$p = 0.2$）。**d) 高斯噪声：**加入加性噪声以模拟传感器颗粒噪声。

在非侵入式设置中，这些失真会直接施加到载体图像上，记作 $I_{no} = N(I_{co})$。值得注意的是，该模块是一个灵活的组件，可以替换为不同的噪声层，以针对不同的攻击场景。

恢复器

恢复器（记作 $R(\cdot)$）是用于缩小域差距的关键预处理模块。它旨在校正受攻击图像 I_{no} 中的噪声，以恢复原始载体图像的视觉内容。该恢复过程在数学上可表示为：

$$I_{re} = R(I_{no}) \tag{2}$$

其中， I_{re} 表示复原图像。理想情况下， $R(\cdot)$会学习噪声层的逆变换，使得 $I_{re} \approx I_{co}$，从而为后续的解码步骤提供更干净的参考图像。

解码器

解码器 $D(\cdot)$提取水印。在我们的非侵入式设置中，解码依赖于密钥。解码器接收生成的编码图像 I_{en} 和恢复的参考图像 I_{re}，以检索消息：

$$M_{pre} = D(I_{en}, I_{re}) \tag{3}$$

解码过程使用消息逆向处理模块，该模块采用步幅卷积对特征图进行下采样，使其恢复到消息维度 L。

3.2 Sigmoid门控异或估计器

我们的 NiMark 框架的核心组件涉及使用 round(·) 进行硬二值化，随后执行按位异或运算。我们明确设计该模块，以迫使模型学习依赖于内容的水印嵌入。通过异或运算将二值化的图像特征与二值化的消息特征不可分割地关联起来，模型会受到强烈促使，将载体图像特征纳入其中，因为忽略这些特征会导致明显的性能下降。这种逻辑依赖性明确降低了模型收敛到平凡结构捷径的倾向，确保形成强绑定，使验证密钥源自图像特征，而不是独立于特定的图像内容。图 3 展示了这种离散运算与我们的可微代理函数之间的视觉对比。

在前向传播中，该操作由复合离散函数形式化定义：

$$W = \text{round}(X_1) \oplus \text{round}(X_2) \tag{4}$$

其中， X_1， $X_2 \in [0,1]^{C'\times H'\times W'}$ 表示预处理后的连续特征张量（特征图）。round(·)表示

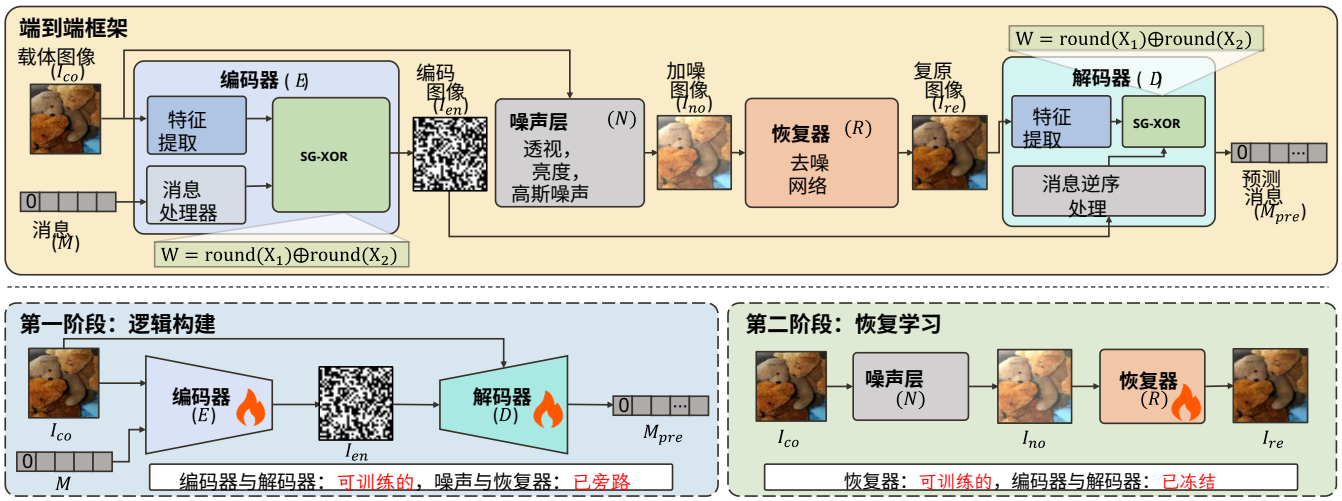


图 2：所提出的 NiMark 框架概览。顶部面板展示了详细的网络架构，其中包括与 SG-XOR 集成的编码器、噪声层、恢复器和解码器。底部面板展示了所提出的两阶段训练策略。第一阶段在无噪声环境中构建逻辑依赖关系。第二阶段独立优化恢复网络，以处理屏幕拍摄失真。

逐元素硬二值化函数，而 $\oplus$ 表示逐元素逻辑异或。如图 3a 所示，W的离散表面呈现出尖锐的不连续点和平坦的平台，这些特征会阻断梯度流，给基于梯度的学习带来根本性障碍。

该运算从根本上来说是不可微的。如图 3b 所示，一种常见的变通方法是使用直通估计器（STE） [Bengio等人, 2013], 其仅将梯度从输出直接复制到输入。实际上，这种方法将二值化视为恒等函数。然而，标准 STE 通常将二值化函数的梯度近似为恒等映射，即梯度为 1，因此并不适用于这种复杂的复合运算。如此粗糙的线性近似与真实取整（·） $\oplus$取整（·）逻辑高度非线性的梯度几乎没有相似之处。离散前向传播与过度简化的反向传播过程之间存在严重不匹配，导致训练不稳定以及性能显著下降。

为解决这一问题，我们设计了一种专用的可微代理梯度，能够准确模拟目标离散行为。我们将这种方法称为 SG-XOR 估计器。反向传播过程基于一个复合代理函数推导而来，该函数对二值化（即门控）和异或逻辑运算进行建模。如工作流程图（图 4）所示，我们旨在通过反向传播求出梯度 $\partial\mathcal{L}/\partial X_1$ 和 $\partial\mathcal{L}/\partial X_2$。

二值化近似

首先，我们使用参数化的 Sigmoid 函数 $\sigma_{m,\ n}(\cdot)$ 近似不可微的取整(·)函数：

$$p_1 = \sigma_{m,n}(X_1) = \frac{1}{1+\exp(-m(X_1+n))} \tag{5}$$

$$p_2 = \sigma_{m,n}(X_2) = \frac{1}{1+\exp(-m(X_2+n))} \tag{6}$$

其中， p_1 和 p_2 分别表示对 X_1 和X_2进行软二值化。该函数由两个超参数控制：陡峭度参数 m和二值化阈值 n。参数 m 控制 Sigmoid 曲线的陡峭程度；我们将其设置为 $m = 10$，以确保近似结果接近阶跃函数。参数 n 定义了过渡的中心位置。为了准确模拟取整(·)函数的行为——对于属于 $[0,1]$, 的输入，其决策边界位于 0.5——我们将阈值设置为 $n = -0.5$。

连续异或近似

接下来，我们必须为离散逻辑$p_1 \oplus p_2$定义一个连续代理函数。目标运算由异或真值表定义。我们选择满足这一条件的双线性多项式；它也等价于两个相互独立的伯努利变量进行异或运算的概率：

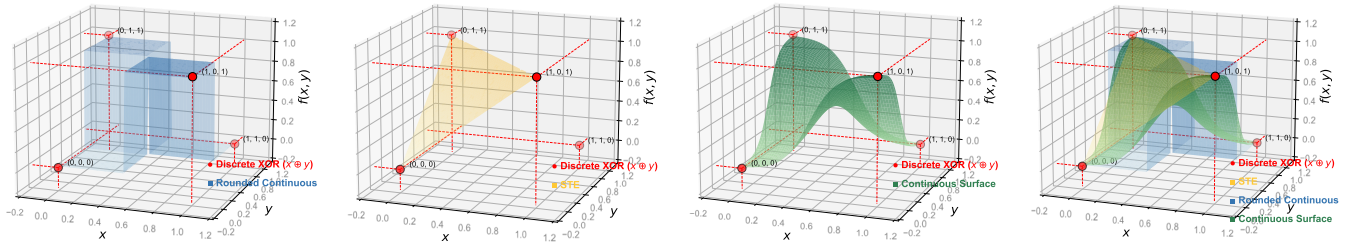
$$W_{proxy} = f(p_1, p_2) = p_1 + p_2 - 2p_1p_2 \tag{7}$$

该函数与离散异或逻辑相符。例如，在输入顶点 (1, 1) 处，它给出 $f(1,\ 1) = 1 + 1 - 2 = 0$，符合真值表。

梯度计算

在反向传播过程中，我们通过对复合代理函数W_{proxy}应用链式法则，计算损失 $\mathcal{L}$相对于原始输入 X_1 和 X_2 的梯度。有关此梯度流的详细数学推导和理论依据，请参阅补充材料。这里，我们直接给出用于反向传播的最终梯度，它们由逻辑梯度和 Sigmoid 门组成：

$$\frac{\partial\mathcal{L}}{\partial X_1} = \frac{\partial\mathcal{L}}{\partial W_{proxy}} \cdot \underbrace{(1-2p_2)}_{\text{Logic Gradient}} \cdot \underbrace{(m \cdot p_1(1-p_1))}_{\text{Sigmoid Gate}} \tag{8}$$



(a) 不可微函数　(b) STE代理函数　(c) 提出的 SG-XOR 代理函数　(d) 所有函数的比较

图 3：离散异或运算及其用于梯度估计的各种可微代理函数的可视化。(a) 复合运算 $W = \text{round}(X_1) \oplus \text{round}(X_2)$ 的真实不可微曲面，它定义了离散目标行为，并导致梯度峭壁，阻碍标准反向传播。(b) 标准线性 STE 代理函数，展示了粗略且不匹配的梯度近似。(c) 所提出的 SG-XOR 代理函数，提供了一个平滑的可微流形，与离散异或逻辑高度吻合，同时保持可微性。(d) 直接比较 SG-XOR 代理函数（绿色）、STE 代理函数（黄色）和目标离散函数（蓝色），展示了我们的方法具有更优的一致性。

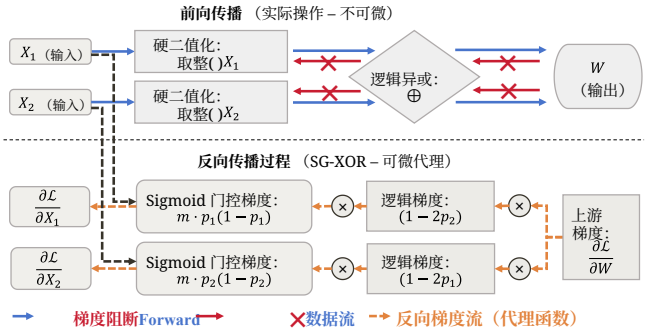


图 4：所提出的 SG-XOR 估计器的机制。前向传播（上方）执行严格的离散运算 $W = \text{round}(X_1) \oplus \text{round}(X_2)$，其中梯度会被硬二值化自然阻断。反向传播过程（下方）利用我们提出的可微代理函数来估计梯度。上游梯度由两个不同的组成部分调制：逻辑梯度用于强制遵循异或真值表，而 Sigmoid 门控梯度则用于稳定决策边界附近的训练。

$$\frac{\partial \mathcal{L}}{\partial X_2} = \frac{\partial \mathcal{L}}{\partial W_{proxy}} \cdot \underbrace{(1-2p_1)}_{\text{Logic Gradient}} \cdot \underbrace{(m \cdot p_2(1-p_2))}_{\text{Sigmoid Gate}} \tag{9}$$

其中$p_1 = \sigma_{m,n}(X_1)$ 和 $p_2 = \sigma_{m,n}(X_2)$ 会在反向传播过程中利用保存的输入 X_1 和 X_2重新计算。这种表述既能确保忠实近似离散逻辑，又能实现利于梯度传播的学习动态。

3.3 两阶段训练策略

每个训练阶段的具体数据流和组件状态如图 2 底部面板所示。由于域差距较大，直接在严重的屏幕拍摄失真下训练深度网络颇具挑战。我们提出了一种两阶段训练策略，将逻辑绑定学习与物理恢复解耦，确保模型同时捕获正确的依赖关系以及对噪声的鲁棒性。

第一阶段：逻辑构建

该阶段的主要目标是在无噪声环境中通过 SG-XOR 构建正确的特征-水印依赖关系。

在第一阶段，我们绕过噪声层和恢复器。目标是训练编码器（E）和解码器（D），使其执行基础的非侵入式水印任务。训练仅由消息恢复损失 $\mathcal{L}_{de}$ 引导，因为在我们的非侵入式设置中，载体图像的视觉质量自然得以保留。解码器损失 $\mathcal{L}_{de}$ 衡量预测消息 M_{pre} 与真实值 M 之间的误差：

$$\mathcal{L}_{stage1} = \mathcal{L}_{de} = \|M_{pre} - M\|_2^2 \tag{10}$$

其中，$\|\cdot\|_2^2$ 表示均方误差（MSE）。在这一阶段，SG-XOR 估计器负责实现逻辑绑定。通过学习利用异或运算解码消息，网络被迫使用图像特征，从而避免结构捷径（恒等映射）。

第二阶段：恢复学习

这一阶段旨在训练恢复器，使其能够独立校正物理屏幕拍摄失真。

在第二阶段，我们冻结编码器和解码器，仅专注于优化恢复器（R）。该网络接收失真的图像，并学习恢复原始载体图像：

$$I_{no} = N(I_{co}), \quad I_{re} = R(I_{no}) \tag{11}$$

损失函数 $\mathcal{L}_{stage2}$ 通过均方误差和多尺度结构相似性来确保恢复质量：

$$\mathcal{L}_{stage2} = \lambda_{pix} \cdot \|I_{re} - I_{co}\|_2^2 + \lambda_{str} \cdot (1 - \text{SSIM}(I_{re}, I_{co})) \tag{12}$$

其中，我们将 $\lambda_{pix} = 10$和 $\lambda_{str} = 1$ 设为相应值，以确保像素级保真度和结构保真度的优化目标一致。通过恢复图像质量，恢复器有效弥合了域差距，使第一阶段预训练的编码器和解码器即使在攻击条件下也能正常工作。

方法	标准		水平视角：左侧				水平视角：右侧				拍摄距离（变化）					
	(0°, 30cm)		60°		30°		30°		60°		40cm		50cm		60cm	
	比特错误率	U-BER	比特错误率	U-BER	比特错误率	U-BER	比特错误率	U-BER	比特错误率	U-BER	比特错误率	U-BER	比特错误率	U-BER	比特错误率	U-BER
RHFM	21.10	24.24	22.62	22.75	20.38	22.98	20.56	23.29	19.74	23.16	21.23	22.86	21.04	24.08	21.13	23.85
EFM	19.24	20.40	19.91	20.67	18.23	20.95	18.64	20.76	18.55	20.36	19.01	19.97	19.44	20.60	19.63	21.22
PCET	20.04	20.91	20.44	22.03	19.16	21.33	19.82	21.38	18.91	20.32	20.25	21.07	20.55	21.76	20.52	21.86
BFM	20.51	22.66	21.90	22.85	19.82	21.85	20.18	22.40	19.85	23.45	20.70	23.53	20.42	23.54	21.16	22.60
FJFM	21.83	23.62	22.55	23.45	21.34	23.75	21.81	24.32	21.37	24.02	22.51	23.76	21.92	23.63	22.40	22.92
GPCET	19.26	20.69	19.39	20.93	18.25	20.70	19.05	20.28	17.96	20.50	19.57	21.09	19.77	20.42	19.66	21.23
FJFMR	8.58	9.68	8.70	9.36	8.36	9.77	8.40	9.58	8.38	8.73	8.42	9.27	8.47	9.22	8.65	9.42
GPCETR	14.94	16.31	16.08	17.15	14.91	16.61	14.98	17.70	14.69	16.31	15.37	16.08	15.15	16.81	15.32	15.62
GRHFMR	8.72	14.77	13.30	14.51	12.58	14.25	12.59	13.93	12.26	13.96	12.85	13.64	12.52	14.20	12.87	13.75
NiMark	**0.46**	**49.41**	**1.13**	**53.82**	**0.97**	**40.25**	**0.68**	**48.16**	**1.28**	**47.32**	**0.57**	**51.98**	**0.58**	**51.64**	**0.59**	**51.03**

表1：不同水平角度和拍摄距离下的鲁棒性（BER %）与绑定完整性（U-BER %）。「标准」列表示基线设置（0°, 30cm），其余列则相对于该设置仅改变一个参数。

4 实验

本节评估 NiMark 的性能。我们首先详细介绍实验设置，随后将其与最先进的方法进行比较，最后通过消融研究验证我们的设计选择。

4.1 实验设置

数据集与实现

遵循 [Wu 等人的方案， 2024], 我们使用 1,000 张 COCO 图像进行训练（另有两个不重叠的测试集，分别包含 500 张和 100 张图像），并采用 AdamW 优化器，权重衰减设为 0.02，批量大小设为 16。关于输入配置，与 [Wen 等人的设置一致， 2025],我们将所有图像调整为 512×512 ，并将消息长度设为 $L = 1024$ 位，生成大小为$1 \times 32 \times 32$的编码密钥 I_{en}。该框架使用 PyTorch 在英伟达 RTX 3060 GPU 上实现。训练采用我们提出的两阶段策略，总计进行 300 个 epoch，并由带预热期的余弦退火调度器控制。

基线

为了评估性能，我们选取了一组全面的基线方法。

1）非侵入式方法：我们实现了 9 种基于正交矩的代表性方案，包括 RHFM[Wang 等人， 2019], EFM [Wang 等人， 2016], PCET [Wang 等人， 2017], BFM [Gao 和 Jiang， 2015], FJFM [Yang 等人，2021], GPCET [Yang 等人， 2020], FJFMR [Qi 等人，2024],GPCETR [Qi 等人， 2024], 以及 GRHFMR [Wen 等人，2025]。2）抗屏摄方法：我们将 NiMark 与专为屏幕-相机信道设计的领先基于嵌入的框架进行比较，包括 StegaStamp [Tancik 等人， 2020], PIMoG [Fang 等人， 2022], SRWSS [Gao 等人，2025], SSDS [Li 等人， 2024] 以及 S2R [Wu 等人， 2026]。

评估指标

1）鲁棒性：通过比特错误率（BER，%）进行衡量。数值越低越好。

2) 视觉质量：以 PSNR（dB）和 SSIM 衡量。对于非侵入式方法，这些数值在理论上已达到最优（PSNR $\to \infty$，SSIM $= 1.0$）。

3) 非配对比特错误率（U-BER）：为检测结构捷径，我们验证验证密钥是否在逻辑上与图像内容绑定，还是已退化为自包含载体。U-BER 用于评估特定有效密钥 I_{en} 与错误且不相关的参考图像配对时的解码性能。至关重要的是，为确保该指标反映模型在攻击条件下的行为，非配对参考图像会经过与正确载体完全相同的噪声**和复原**处理过程。单个样本的 U-BER 通过对 K 个随机非配对参考图像上的错误率取平均值来计算：

$$\text{U-BER} = \frac{1}{K}\sum_{k=1}^{K}\left(\frac{1}{L}\sum_{i=1}^{L}\left|m_i - D(I_{en}, I'^{(k)}_{rand})_i\right|\right) \times 100\% \tag{13}$$

其中， $K = 10$ 表示用于取平均值以确保稳定性的随机参考图像数量， L为消息长度， m_i 是真实消息的第 i位，$I'^{(k)}_{rand}$ 表示恢复后的第 k张随机参考图像。理想情况下，U-BER $\approx 50\%$，这表明仅凭密钥不足以在缺少正确图像特征的情况下进行解码。

4.2 与非侵入式方案的比较：屏幕拍摄鲁棒性

表1展示了关于水平透视角度和不同拍摄距离的综合评估结果。除非另有说明，实验均使用联想拯救者Y9000P显示器和三星S20FE拍摄设备进行，默认设置为距离30cm且为 0° 角（垂直于屏幕）。垂直角度（向上和向下）的结果见补充材料。

如表1所示，传统的基于矩的方法在不同距离下表现不稳定，比特错误率在19%至21%之间波动。相比之下，NiMark 展现出更强的鲁棒性，比特错误率始终保持在3.5%以下。尤为重要的是，NiMark 在实现这一点的同时，在所有测试条件下都将无条件比特错误率维持在接近50%，这证实了即使在严重的几何失真下，逻辑绑定仍能得到稳健保留。

方法	峰值信噪比	结构相似性指数	角度（右） 30° 距离 30cm	
StegaStamp	33.07	0.904	1.27	1.13
PIMoG	36.06	0.967	8.27	6.94
SRWSS	33.98	0.963	22.97	12.4
S2R	42.27	0.962	6.68	3.51
NiMark	∞	**1**	**0.68**	**0.46**

表 2：与最先进的抗屏幕拍摄方案的比较。鲁棒性（比特错误率 %）与视觉质量。

显示设备	Samsung S20 FE		iPhone 16		Vivo S17t	
	比特错误率	U-BER	比特错误率	U-BER	比特错误率	U-BER
Lenovo Y9000P	0.46	49.41	0.73	49.43	0.62	51.89
Envision G249G	0.37	48.52	1.31	51.39	0.72	49.41
ASUS ROG Strix	1.06	47.92	0.88	50.80	0.47	50.31

表 3：跨设备鲁棒性评估（比特错误率 % / U-BER %）。结果表明，该方法能够推广至不同的显示设备与采集硬件组合。

4.3 与非侵入式方案的比较：数字攻击

为验证其通用性，我们还测试了 NiMark 在标准数字攻击（JPEG、噪声、滤波）下的表现。补充材料中的结果表明，NiMark 具有较强的抗干扰能力，无条件比特错误率始终接近50%，证实了逻辑绑定的稳固性。

4.4 与抗屏幕拍摄方案的比较

我们进一步将 NiMark 与具有代表性的抗屏幕拍摄方法进行比较：StegaStamp、PIMoG、SRWSS 和 S2R。表 2 展示了这些方法在鲁棒性和视觉质量方面的结果。大多数基线方法通常处理 128×128 张图像，承载 30 比特有效载荷（≈ 0.0018 bpp）。S2R 将其提高到 64 比特（≈ 0.0039 bpp），而 NiMark 处理 512×512 个输入并承载 1024 比特消息，同时保持相近的比特密度。实验结果表明，NiMark 对屏幕拍摄具有很强的鲁棒性。此外，作为一种非侵入式框架，它保证零视觉失真，为要求严格数据完整性的场景提供了可行的替代方案。

4.5 跨设备鲁棒性评估

为验证 NiMark 在异构硬件上的泛化能力，我们使用一套全面的测试矩阵进行了跨设备评估，其中包含三种不同的显示终端（联想拯救者 Y9000P、Envision G249G、华硕 ROG Strix SCAR Edition 8）和三种不同的拍摄设备（三星 Galaxy S20 FE、iPhone 16、Vivo S17t）。该设置模拟了现实世界中的验证场景，在这些场景中，屏幕和摄像头传感器的物理特性差异显著。表3报告了各种硬件组合下的比特错误率表现。所有组合的比特错误率始终较低，证明 NiMark 在多样化硬件环境下具有良好的泛化能力和稳定性。

模块	变体	比特错误率 (%)	U-BER（%）
架构	Cat +Conv 融合	0.00	0.00
梯度	STE	4.14	46.62
噪声层	StegaStamp	2.55	47.64
	PIMoG	1.26	50.53
	SSDS	1.96	47.36
	S2R	0.62	52.84
	NiMark	0.46	49.41

表4：关于架构、梯度估计器和噪声层的消融实验。

策略	描述	比特错误率 (%)
仅S1	逻辑预训练	4.12
联合训练	单阶段	0.68
NiMark	**两阶段**	**0.46**

表5：训练策略的消融研究。我们将两阶段方法与仅逻辑预训练和端到端联合训练进行比较。

4.6 消融研究

我们通过受控实验考察了 NiMark 各关键组件的有效性。表4总结了架构变体的影响。首先，将我们的 SG-XOR 替换为“Concat+Conv”融合后，U-BER 接近于零，这表明密钥变成了自包含的；这解释了为何比特错误率为微不足道的 0%，因为模型绕过了图像特征，使其失真变得无关紧要。此外，与 SG-XOR 估计器相比，采用标准 STE 的变体比特错误率高得多，因为其过度简化的梯度无法准确模拟离散异或逻辑。最后，与现有鲁棒水印方案（StegaStamp、PIMoG、SSDS、S2R）中的噪声层相比，我们的噪声层在非侵入式任务中表现最佳，在现实世界攻击下实现了最低的比特错误率 0.46%。

训练策略的影响

表5比较了我们的两阶段策略与单阶段基线。结果表明：1）仅第一阶段：不使用噪声层进行训练，只能获得有限的鲁棒性。2）联合训练：虽然端到端训练可行，但同时优化逻辑绑定与恢复，其鲁棒性不如我们的渐进式策略。3）NiMark（S1 $\rightarrow$ S2）：通过将两项任务解耦，我们的策略使网络能够在每个阶段专注于特定目标，从而实现最佳的整体鲁棒性。

5 结论

本文提出了 NiMark，这是一种能够抵御屏幕拍摄的非侵入式框架。我们发现了结构捷径这一失败模式，即网络可能会学习平凡的恒等映射。为减轻其影响，我们引入了 SG-XOR 估计器，以弥合不可微间隙。此外，为弥合屏幕拍摄造成的严重域差距，我们设计了一种两阶段训练策略，

将逻辑学习与物理恢复解耦。实验表明，与最先进的方法相比，NiMark 的鲁棒性具有竞争力，且往往更胜一筹，同时保持零视觉失真。未来的工作将聚焦于实时移动端优化和视频保护。

参考文献

[Beery 等人， 2018] Sara Beery、Grant Van Horn 和 Pietro Perona。未知领域识别。载于《欧洲计算机视觉会议（ECCV）论文集》，第 456–473 页，2018年。[Bengio 等人， 2013] Yoshua Bengio、Nicholas L´eonard 和 Aaron Courville。针对条件计算，通过随机神经元估计或传播梯度。arXiv 预印本 arXiv:1308.3432，2013年。[Chen 等人， 2025] Mingyue Chen、Xin Liao、Han Fang、Jin- lin Guo、Yanxiang Chen 和 Xiaoshuai Wu。具有可证明鲁棒性的灵活局部屏幕拍摄水印。IEEE《电路与系统汇刊：视频技术》，2025年。[Chu， 2003] Wai C Chu。使用子采样的基于DCT的图像水印。IEEE《多媒体汇刊》，5(1):34–38，2003年。[Fang 等人， 2018] Han Fang、Weiming Zhang、Hang Zhou、Hao Cui 和 Nenghai Yu。抗屏幕拍摄的水印。IEEE《信息取证与安全汇刊》，14(6):1403–1418，2018年。[Fang 等人， 2021] Han Fang、Dongdong Chen、Feng Wang、Zehua Ma、Honggu Liu、Wenbo Zhou、Weiming Zhang 和 Nenghai Yu。Tera：具有透明性、效率、鲁棒性和适应性的屏幕到相机图像码。IEEE《多媒体汇刊》，24:955–967，2021年。[Fang 等人， 2022] Han Fang、Zhaoyang Jia、Zehua Ma、Ee- Chien Chang 和 Weiming Zhang。PIMoG：一种用于基于深度学习的水印网络的有效截屏噪声层模拟。载于 ACM 多媒体国际会议（MM）论文集，第 2267–2275 页，2022年。[Gao 和 Jiang， 2015] Guangyong Gao 和 Guoping Jiang。基于贝塞尔-傅里叶矩的鲁棒图像零水印。Multimed. Tools Appl., 74(3):841–858，2015年。[Gao 等人， 2025] Guangyong Gao、Xiaoan Chen、Li Li、Zhi- hua Xia、Jianwei Fei 和 Yun-Qing Shi。基于风格迁移和结构重参数化的抗屏幕拍摄水印。IEEE《信息取证与安全汇刊》，2025年。[Geirhos 等人， 2018] Robert Geirhos、Patricia Rubisch、Claudio Michaelis、Matthias Bethge、Felix A Wichmann 和 Wieland Brendel。ImageNet 训练的卷积神经网络偏向纹理；增强形状偏置可提高准确率和鲁棒性。2018年。arXiv:1811.12231。[Geirhos 等人， 2020] Robert Geirhos、J¨orn-Henrik Jacob- sen、Claudio Michaelis、Richard Zemel、Wieland Brendel、Matthias Bethge 和 Felix A Wichmann。深度神经网络中的捷径学习。《自然·机器智能》，2(11):665–673，2020年。[He 等人， 2024] Mingjin He、Bingwen Feng、Yizhi Guo、Jian Weng 和 Wei Lu。图像实例级抗摄像拍摄水印。IEEE《电路与系统汇刊：视频技术》，34(11):10874–10887，2024年。[Hou 等人， 2025] Zhaoyang Hou、Haowen Yan、Liming Zhang、Na Ren、Rongjuan Ma 和 Ruitao Qu。基于多通道PCNN的零水印方法及用于遥感图像交易证书和版权保护的区块链。IEEE《地球科学与遥感汇刊》，2025年。[Jia 等人， 2021] Zhaoyang Jia、Han Fang 和 Weiming Zhang。Mbrs：通过真实和模拟 JPEG 压缩的小批量增强基于深度神经网络的水印鲁棒性。载于 ACM 国际多媒体会议（MM）论文集，第 41–49 页，2021年。[Jo 和 Yoshua Bengio， 2017] Jason J o 和 Yoshua Bengio。衡量卷积神经网络学习表面统计规律的倾向。载于 arXiv 预印本 arXiv:1711.11561，2017年。[Li 等人， 2024] Yiyi Li、Xin Liao 和 Xiaoshuai Wu。采用灰度偏差仿真的抗屏幕拍摄水印。IEEE《多媒体汇刊》，2024年。[Li 等人， 2025] Jingyou Li、Rongle Wei、Xiaotian Xi、Guangda Zhang、Zixin Yang 和 Fengshan Zhang。用于医学图像鲁棒零水印的双分支卷积神经网络-Transformer网络。《信息科学》，第 122511 页，2025年。[Liu 和 Tieniu Tan， 2002] Ruizhen Liu 和 Tieniu Tan。用于保护合法所有权的基于奇异值分解的水印方案。IEEE《多媒体汇刊》，4(1)：121–128，2002年。[Liu 等人， 2019] Yang Liu、Mengxi Guo、Jian Zhang、Yuesheng Zhu 和 Xiaodong Xie。面向实际盲水印的新型两阶段可分离深度学习框架。载于 ACM 国际多媒体会议（MM）论文集，第 1509–1517 页，2019年。[Liu 等人， 2025] Xiyao Liu、Cundian Yang、Jianbiao He、Hui Fang、Gerald Schaefer、Jian Zhang、Yuesheng Zhu 和 Shichao Zhang。用于体积医学图像零水印的攻击防御对比学习。ACM《多媒体计算、通信与应用汇刊》，21(2)：1–23，2025年。[Ma 等人， 2025] Zehua Ma、Han Fang、Xi Yang、Kejiang Chen 和 Weiming Zhang。Ropass：面向局部屏摄场景的鲁棒水印。载于 AAAI 人工智能会议（AAAI）论文集，第 39 卷，第 19332–19339 页，2025年。[Qi 等人， 2024] Shuren Qi、Yushu Zhang、Chao Wang、Tao Xiang、Xiaochun Cao 和 Yong Xiang。无需去噪地表示噪声图像。IEEE《模式分析与机器智能汇刊》，46(10)：6713–6730，2024年。[Sander 等人， 2025] Tom Sander、Pierre Fernandez、Alain Durmus、Teddy Furon 和 Matthijs Douze。使用局部化消息为任何内容添加水印。载于国际学习表征会议（ICLR），2025年。[Smeu 等人， 2025] Stefan Smeu、Dragos-Alexandru Boldisor、Dan Oneata 和 Elisabeta Oneata。利用无监督学习规避视听深度伪造检测数据集中的捷径特征。载于 IEEE/CVF 计算机视觉与模式识别会议（CVPR）论文集，第 18815–18825 页。IEEE，2025年。[Tancik 等人， 2020] Matthew Tancik、Ben Mildenhall 和 Ren Ng。StegaStamp：物理照片中的隐形超链接。载于 IEEE 计算机视觉与模式识别会议（CVPR）论文集，第 2117–2126 页，2020年。

[Wang 等人， 2016] Chun-peng Wang、Xing-yuan Wang、Zhi-qiu Xia、Chuan Zhang 和 Xing-jun Chen。基于四元数指数矩的几何鲁棒彩色图像零水印算法。《视觉通信与图像表示》，41：247–259，2016年。[Wang 等人， 2017] Chun-peng Wang、Xing-yuan Wang、Xing-jun Chen 和 Chuan Zhang。基于极坐标复指数变换和逻辑映射的鲁棒零水印算法。《多媒体工具与应用》，76(24)：26355–26376，2017年。[Wang 等人， 2019] Chunpeng Wang、Xingyuan Wang、Zhiqiu Xia 和 Chuan Zhang。基于三元径向谐波傅里叶矩的鲁棒立体图像零水印算法。《信息科学》，470：109–120，2019年。[Wang 等人， 2022] Baowei Wang、Jiawei Shi、Weishen Wang 和 Peng Zhao。基于区块链和零水印的图像版权保护。IEEE《网络科学与工程汇刊》，9(4)：2188–2199，2022年。[Wang 等人， 2025] Heng Wang、Hongxia Wang、Fei Zhang、Zhenhao Shi 和 Xinyi Huang。Moir´e 水印：利用 Moir´e 图案实现抗屏幕拍摄的鲁棒水印。IEEE《多媒体汇刊》，2025年。[Wen 等人， 2003] Quan Wen、T-F Sun 和 S-X Wang。零水印的概念与应用。《电子学报》，31(2)：214–216，2003年。[Wen 等人， 2025] Wenying Wen、Yu Ye、Ziye Yuan、Baolin Qiu 和 Dingli Hua。LFIZW-GRHFMR：采用 GRHFMR 的光场图像鲁棒零水印。ACM《多媒体计算、通信与应用汇刊》，21(4)：1–17，2025年。[Wengrowski 和 Dana， 2019] Eric Wengrowski 和 Kristin Dana。利用深度摄影隐写进行光场信息传递。载于 IEEE 计算机视觉与模式识别会议（CVPR）论文集，第 1515–1524 页，2019年。[Wolfgang 等人， 2002] Raymond B Wolfgang、Christine I Podilchuk 和 Edward J Delp。面向数字图像与视频的感知水印。《IEEE 会刊》，87(7)：1108–1126，2002年。[Wu 等人， 2024] Yufeng Wu、Baowei Wang、Guiling Wang 和 Xin Liao。MCFN：用于高性能水印的多尺度交叉前馈网络。《神经计算》，第 129282 页，2024年。[Wu 等人， 2025] Shaowu Wu、Wei Lu 和 Xiangyang Luo。基于多层水印特征融合的鲁棒水印。IEEE《多媒体汇刊》，2025年。[Wu 等人， 2026] Yufeng Wu、Xin Liao、Baowei Wang、Han Fang、Xiaoshuai Wu、Mingyue Chen 和 Guiling Wang。仿真到现实：用于提升屏幕-摄像机水印鲁棒性的无监督噪声层。载于 AAAI 人工智能会议（AAAI）论文集，第 40 卷，2026年。

[Yang 等， 2020] Hong-ying Yang、Shu-ren Qi、Pan-pan Niu 和 Xiang-yang Wang。基于快速四元数通用极坐标复指数变换的彩色图像零水印。《信号处理：图像通信》，82:115747，2020年。[Yang 等， 2021] Hongying Yang、Shuren Qi、Jialin Tian、Panpan Niu 和 Xiangyang Wang。鲁棒且具判别性的图像表示：分数阶 Jacobi-Fourier 矩。《模式识别》，115:107898，2021年。[Zhang 和 Wei， 2020] Lina Zhang 和 Deyun Wei。基于矩阵分解和旋转变换的不变整数小波域图像水印。《信号处理》，169:107421，2020年。[Zhu 等， 2018] Jiren Zhu、Russell Kaplan、Justin Johnson 和李飞飞。HiDDeN：利用深度网络隐藏数据。载于欧洲计算机视觉会议（*ECCV*）论文集，第657–672页，2018年。