结合可逆神经网络和逆梯度注意力的抗屏摄攻击水印方法

李谢华* 娄 芹 杨俊雪 廖 鑫
(湖南大学信息科学与工程学院 长沙 410082)

摘 要：随着智能设备的普及，数字媒体内容的传播和分享变得更加便捷，人们可以通过手机拍摄屏幕等简单方式轻松获取未经授权的信息，导致屏幕拍摄传播成为版权侵权的热点问题。为此，该文针对屏幕盗摄版权保护任务提出一种端到端的基于可逆神经网络和逆梯度注意力的抗屏摄攻击图像水印框架，实现屏幕盗摄场景下版权维护的目标。该文将水印的嵌入和提取视为相互关联的逆问题，利用可逆神经网络实现编解码网络的一体化，有助于减少信息传递损失。进一步地，通过引入逆梯度注意模块，捕捉载体图像中鲁棒性强且视觉质量高的像素值，并将水印信息嵌入到载体图像中不易被察觉和破坏的区域，保证水印的不可见性和模型的鲁棒性。最后，通过可学习感知图像块相似度(LPIPS)损失函数优化模型参数，指导模型最小化水印图像感知差异。实验结果表明，所提方法在鲁棒性和水印图像视觉质量上优于目前同类的基于深度学习的抗屏摄攻击水印方法。

关键词：数字水印；可逆神经网络；逆梯度注意力；屏幕拍摄
中图分类号：TN911.73; TP309 **文献标识码**：A **文章编号**：1009-5896(2024)00-0001-08
DOI: 10.11999/JEIT230953

Screen-Shooting Resilient Watermarking Scheme Combining Invertible Neural Network and Inverse Gradient Attention

LI Xiehua LOU Qin YANG Junxue LIAO Xin
(*College of Computer Science and Electronic Engineering*, *Hunan University*, *Changsha* 410082, *China*)

Abstract: With the growing use of smart devices, the ease of sharing digital media content has been enhanced. Concerns have been raised about unauthorized access, particularly via screen shooting. In this paper, a novel end-to-end watermarking framework is proposed, employing invertible neural networks and inverse gradient attention, to tackle the copyright infringement challenges related to screen content leakage. A single invertible neural network is employed by the proposed method for watermark embedding and extraction, ensuring information integrity during network propagation. Additionally, robustness and visual quality are enhanced by an inverse gradient attention module, which emphasizes pixel values and embeds the watermark in imperceptible areas for better invisibility and model resilience. Model parameters are optimized using the Learnable Perceptual Image Patch Similarity (LPIPS) loss function, minimizing perception differences in watermarked images. The superiority of this approach over existing learning-based screen-shooting resilient watermarking methods in terms of robustness and visual quality is demonstrated by experimental results.

Key words: Digital watermarking; Invertible neural network; Inverse gradient attention; Screen shooting

收稿日期：2023-08-31；改回日期：2024-03-21；网络出版：2024-04-12
*通信作者： 李谢华 beverly@hnu.edu.cn
基金项目：国家自然科学基金(U22A2030, 61972142)，湖南省自然科学基金(2021JJ30140)，湖南省杰出青年科学基金项目(2024JJ2025)，长沙市科技重大专项经费资助(kh2205033)
Foundation Items: The National Natural Science Foundation of China (U22A2030, 61972142), Hunan Provincial Natural Science Foundation (2021JJ30140), Hunan Provincial Funds for Distinguished Young Scholars (2024JJ2025), Changsha Science and Technology Major Project (kh2205033)

1 引 言

智能手机的普及为数字媒体内容传播和分享提供了显著便利，却促使屏幕拍摄成为一种获取未经授权数字内容的常见方法。 数字水印[1]是解决屏幕拍摄盗版问题的有效方案，但从传统电子信道传输转换到屏摄信道传输为数字水印技术带来了新的需求和挑战[2]。以往的数字水印方案[3–5]大多仅涉及数字域，将图像和水印作为数字信号传输和处理，无法准确提取经过屏摄后的水印标识信息。这样是由于进行屏幕图像拍摄时，其过程包含两类电子设备之间的一系列模拟-数字信号转换过程。而电子设备之间的工作频率差异，在拍摄屏幕时产生的摩尔纹失真、模糊失真，以及相机在拍摄屏幕时产生的水印在空间上的不同步失真等造成水印提取困难。而抗屏摄水印要求显示于屏幕上的含水印图像经摄屏后，仍能正确提取其中的水印信息。

抗屏幕拍摄水印的鲁棒性和含水印图像的不可感知性是抗屏摄水印方案亟需解决的问题。传统的抗屏摄水印方案[6]往往通过统计特征的定性分析融合水印通用嵌入方法来嵌入水印。文献[7]提出了帧混合模块用于建模失真场景，在水印提取过程中使水印与载体在一定程度上同步，提高了水印的鲁棒性。文献[8]利用强化尺度不变特征变换定位水印嵌入区域，并将水印嵌入到每个特征区域的离散余弦变换(Discrete Cosine Transform, DCT)域中，以提高水印在屏显拍摄下的鲁棒性，然而这些方法无法充分利用载体图像的特征，同时也会导致水印图像局部出现较为明显的视觉失真[9]，因此这类方法往往缺乏灵活性和通用性。近年来，随着深度学习的快速发展，研究人员充分利用图像特征设计了基于深度神经网络的抗屏摄攻击水印架构。通过提取图像的高维特征提高数字水印嵌入和提取的鲁棒性。文献[10]首次将卷积神经网络引入到非盲水印中。文献[11]提出了一种端到端对抗训练神经网络，可用于隐写和鲁棒盲水印。文献[12]结合注意力机制，为不同的像素值分配不同的权重比例以此达到鲁棒性的目的。文献[13]设计了一个网络用于模拟屏摄失真，并自制了一个大型图像数据集用于训练失真模拟网络，但该算法需要拍摄大量的屏幕-图像对，难以推广。文献[14]通过模拟添加水印图像在跨设备传输中面临的各种失真，克服了数据集难以捕获的问题。文献[15]提出了一种新的见解，即只模拟对屏幕拍摄影响最大的失真而无需模拟屏幕拍摄的整个过程。并将噪声层分为透视失真、光照失真和摩尔失真3类。但目前基于深度学习的抗屏摄鲁棒水印算法都采用了编码器-噪声层-解码器的框架。在这种体系结构中，编码器将水印信息嵌入到载体图像中，解码器从经过屏摄失真的水印图像中提取水印信息。这样的框架要求对编码器和解码器都进行复杂的设计和训练。同时编码器与解码器相互独立的设计分离了水印信息的嵌入与提取过程，易造成水印信息丢失，同时无法生成高质量的水印图像，导致目前提出的抗摄屏水印提取方法在鲁棒性和水印不可感知性方面仍存在一定局限性。

可逆神经网络(Invertible Neural Network, INN)[16]在前向传播和反向传播中具有双射的性质，即每个输入都有唯一对应的输出。作为一种有效的图像可逆变换方法，将前向和反向传播包含在同一网络模型中，减少信息传递损失。研究者提出了多种可逆网络的架构，如NICE[16], GLOW[17]等。这些网络在图像压缩[18]、图像缩放[19,20]等任务中取得了显著成果。近年来，随着文献[21]将INN引入图像隐写领域，涌现了一系列深度隐写水印方案[22,23]。这些研究表明现有的INN在处理与图像视觉相关的任务和在建模信息嵌入/提取过程方面表现良好。

为解决现有深度学习抗屏摄攻击水印框架编解码器分离的结构导致的信息传递损失和引入可见伪影的问题，本文提出了一种新的基于可逆神经网络和逆梯度注意力的抗屏摄攻击水印方法，来实现在真实拍摄场景下的版权保护与追踪溯源。本文的方法基于INN实现，通过将水印的嵌入和提取视为相互关联的逆问题，紧密耦合载体图像和水印信息，在共享网络参数和模块的INN的正向映射和逆向恢复过程中实现水印信息的嵌入和提取，以此缓解传统编码器与解码器相互独立带来的信息损失。为了保证水印的鲁棒性和不可见性，本文设计了逆梯度注意模块，该模块定位鲁棒性强且视觉质量更高的像素值，并将水印信息嵌入到载体图像中不易被察觉和破坏的区域。通过使用感知上的图像相似性指标(Learned Perceptual Image Patch Similarity, LPIPS)[24]损失函数来优化水印图像的主观视觉质量。另外，本文还引入了一个关于屏摄失真的噪声层增强模型，以提升水印的鲁棒性。实验结果表明，本文提出的方法在鲁棒性和不可感知性方面都表现出良好的性能。

2 基于INN和逆梯度注意力的抗屏摄攻击水印框架

2.1 整体框架

图1为整体的水印框架，整体模型由逆梯度注意力模块、数据处理模块、可逆嵌取模块、融合和划分模块、噪声层共同构建。前向映射将水印信息

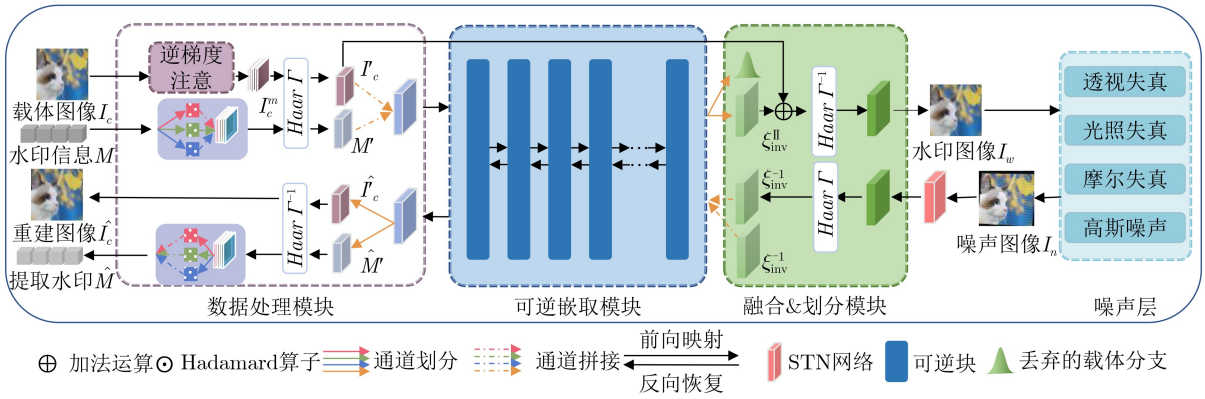


图 1 基于INN和逆梯度注意力的抗屏摄攻击水印框架

嵌入到载体图像中，反向恢复实现从屏摄失真图像中恢复出带有版权标识的水印信息。对于水印嵌入过程，逆梯度注意模块将载体图像$\boldsymbol{I}_c \in \mathbb{R}^{C\times H\times W}$处理得到注意力图像特征$\boldsymbol{I}_c^m \in \mathbb{R}^{C\times H\times W}$。其中$C$表示通道数，$H$表示高度，$W$表示宽度。数据处理模块将载体图像的注意力图像特征$\boldsymbol{I}_c^m$转换为频域特征$\boldsymbol{I}'_c \in \mathbb{R}^{4C\times H/2\times W/2}$。并将长度为$L$的待嵌入的水印信息$M \in \{0,1\}^L$处理成和载体图像频域特征相同大小的维度的水印信息频域特征$\boldsymbol{M}' \in \mathbb{R}^{4C\times H/2\times W/2}$。可逆嵌取模块将输入接收载体图像和水印信息频域特征生成水印残差$\boldsymbol{\xi}_{\text{inv}}^{\text{II}} \in \mathbb{R}^{4C\times H/2\times W/2}$。融合和划分模块将水印残差$\boldsymbol{\xi}_{\text{inv}}^{\text{II}}$和$\boldsymbol{I}'_c$相加后经过频域转换得到水印图像$\boldsymbol{I}_w \in \mathbb{R}^{C\times H\times W}$。屏摄噪声层模拟屏摄过程带来的各种失真，得到噪声图像$\boldsymbol{I}_n \in \mathbb{R}^{C\times H\times W}$。水印的提取过程为嵌入过程的逆过程，在水印提取之前，首先通过轻量级空间变换网络(Spatial Transformer Network, STN)实现空间变换不变性，以校正模拟或真实屏摄场景引起的透视变换失真。然后将校正后的噪声图像输入到模型反向恢复过程提取出嵌入的水印信息$\hat{M} \in \{0,1\}^L$。

2.2 逆梯度注意模块

本文设计了一个保证模型屏摄鲁棒性和水印图像质量的逆梯度注意模块，如图2所示。

通过在载体图像中定位对水印信息重建具有鲁棒性和水印图像复杂纹理区域的像素，然后为这些像素值分配更多的权重并嵌入更多的信息，从而提高模型的鲁棒性和水印图像视觉质量。首先基于待嵌入的水印消息M和重构后的消息$\hat{M}$计算信息重构损失$\text{MSE}(M,\hat{M})$，MSE表示均方误差(Mean Squared Error)。然后通过反向传播计算载体图像$\boldsymbol{I}_c$对水印重建损失$\text{MSE}(M,\hat{M})$的逆归一化梯度生成注意力掩膜mask_A。梯度值显示每个像素对消息重建的鲁棒性。此过程可以表示为

$$\text{MSE}(M,\hat{M}) = \frac{1}{L}\sum_p (M-\hat{M})^2 \tag{1}$$

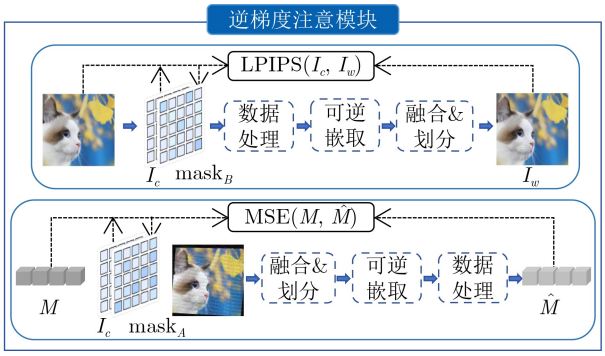


图 2 逆梯度注意模块

$$\text{mask}_A = \Psi - \text{Normalize}(\Delta_{\boldsymbol{I}_c}\text{MSE}(M,\hat{M})) \tag{2}$$

其中，p表示单个水印比特信息位，Ψ表示一个全1张量矩阵，$\text{Normalize}(\cdot)$表示将梯度值归一化到0到1之间。此外基于载体图像$\boldsymbol{I}_c$和水印图像$\boldsymbol{I}_w$计算图像视觉损失$\text{LPIPS}(\boldsymbol{I}_c,\boldsymbol{I}_w)$，再通过反向传播计算载体图像$\boldsymbol{I}_c$对图像视觉损失$\text{LPIPS}(\boldsymbol{I}_c,\boldsymbol{I}_w)$的逆归一化梯度生成注意力掩膜$\text{mask}_B$，此过程可以表示为

$$\text{LPIPS}(\boldsymbol{I}_c,\boldsymbol{I}_w) = \sum_l \frac{1}{H_l W_l}\sum_{h,w} w_l \odot (\boldsymbol{I}_c^l - \boldsymbol{I}_w^l)_2^2 \tag{3}$$

$$\text{mask}_B = \Psi - \text{Normalize}(\Delta_{\boldsymbol{I}_c}\text{LPIPS}(\boldsymbol{I}_c,\boldsymbol{I}_w)) \tag{4}$$

其中，$\odot$表示Hadamard乘积算子$\boldsymbol{I}_w^l, \boldsymbol{I}_c^l \in \mathbb{R}^{H_l\times W_l\times C_l}$表示从VGG网络的$l$层提取的特征并在通道维度中进行单元标准化，通过权重向量$w_l \in \mathbb{R}^{C_l}$缩放激活通道维度并计算特征层的L_2距离。然后在空间维度计算载体图像和水印图像的平均值，最后Σ表示对所有层进行求和。将载体图像$\boldsymbol{I}_c$和梯度注意掩膜mask_A和mask_B通过Hadamard乘积算子逐像素相乘得到载体图像逆梯度注意力特征$\boldsymbol{I}_c^m$。

$$\boldsymbol{I}_c^m = ((\boldsymbol{I}_c \odot \text{mask}_A) \odot \text{mask}_B) \tag{5}$$

失真代价会被反馈到可逆嵌取模块的前向映射

过程，定位载体图像中对消息重建具有鲁棒性的像素，同时约束可逆嵌取模块将信息嵌入到复杂纹理区域，以此提升模型鲁棒性和水印图像视觉质量。

2.3 数据处理模块

数据处理模块在水印嵌入过程中，处理水印信息和载体图像注意力特征以适应可逆嵌取模块的输入；在水印的提取过程中，处理可逆嵌取模块的输出，提取出嵌入的水印信息和恢复载体图像。

在水印嵌入过程中，首先将载体图像的逆梯度注意力特征$\boldsymbol{I}_c^m$通过Haar小波变换得到频域特征$\boldsymbol{I}_c^{'}$，并将其作为后续可逆嵌取模块的载体分支的输入。此过程可表示为 $\boldsymbol{I}_c^{'}=\Gamma_{\text{haar}}(\boldsymbol{I}_c^m)$ ，$\Gamma_{\text{haar}}(\cdot)$表示Haar小波变换。在水印编码过程中，需要将一维水印信息M处理为与载体图像的频域特征相同的2维水印特征$M^{'}$。这个过程参考了文献[23]提出的方法，流程如图1所示。为了适应载体图像的维度，首先将水印信息M复制成3份，并通过不同的全连接层FC(Fully Connections)增长水印信息的长度以产生更多的冗余位。随后，通过反卷积上采样这3个分支。最后，将3个分支拼接后经过 Haar小波变换得到最终的水印信息表示，并将其作为后续可逆嵌取模块的水印分支的输入。此过程可表示为

$$\boldsymbol{M}^{'}=\Gamma_{\text{haar}}(\Gamma_{\text{cat}}(\Gamma_{\text{conv}}(\Gamma_{\text{FC}}(\Gamma_{\text{copy}}(M))))) \tag{6}$$

其中，$\Gamma_{\text{cat}}(\cdot)$表示通道拼接，$\Gamma_{\text{conv}}(\cdot)$表示反卷积上采样，$\Gamma_{\text{FC}}(\cdot)$表示全连接，$\Gamma_{\text{copy}}(\cdot)$表示复制操作。

在水印提取过程中，数据处理模块将可逆嵌取模块的输出划分为恢复的水印信息分支$\hat{M}^{'}$和载体图像分支$\hat{\boldsymbol{I}}_c^{'}$，再通过数据处理模块的逆操作得到嵌入的水印信息$\hat{M}$和恢复载体图像$\hat{\boldsymbol{I}}_c$。由于在数据处理的正向中采取复制操作将水印信息进行扩散，因此在逆向提取过程中将3个水印信息通道通过平均池化后得到提取的水印信息$\hat{M}\in\{0,1\}^L$。

2.4 可逆嵌取模块

INN的可逆性使得可以将水印的嵌入和提取视为一对相互关联的逆问题，本文使用INN的前向过程进行水印嵌入，并利用反向过程进行水印的提取。如图3所示，本文基于IRN[19]构建了包含相同结构的可逆块$f_\theta=f_\theta^1\circ f_\theta^2\circ\cdots\circ f_\theta^N$，$N\in\{1,2,\cdots,16\}$。可逆块的级联操作使得当前块的输出为下一个可逆块的输入。具体来说，对于每个可逆块f_θ^k，$k\in\{1,2,\cdots,N\}$由加性仿射耦合层组成，其输入在通道维度分为两部分，分别表示载体图像和水印信息。对于第l个可逆块的前向过程，输入为$S(B)$，输出为$C(b_{\text{im}}^{l+1},b_{\text{wm}}^{l+1})$，$B$由$b_{\text{im}}^l$，$b_{\text{wm}}^l$拼接组成。其中$S(\cdot)$表示通道划分函数，$C(\cdot)$表示通道拼接函数；$b_{\text{im}}^l$，$b_{\text{wm}}^l$表示输入的载体图像和水印信息，$b_{\text{im}}^{l+1}$，$b_{\text{wm}}^{l+1}$表示经过可逆变换输出的载体图像和水印信息分支。对载体图像分支使用加性变换，对水印信息分支采用仿射变换。前向过程的计算可表示为

$$\begin{aligned}b_{\text{im}}^{l+1}&=b_{\text{im}}^l+\varphi(b_{\text{wm}}^l)\\ b_{\text{wm}}^{l+1}&=b_{\text{wm}}^l\odot\exp(\rho(b_{\text{im}}^{l+1}))+\eta(b_{\text{im}}^{l+1})\end{aligned} \tag{7}$$

其中，$\exp(\cdot)$表示指数函数；$\varphi(\cdot)$，$\rho(\cdot)$，$\eta(\cdot)$可由任意卷积运算函数表示，其参数在前向和反向过程中被共享，本文采用可学习参数的5个级联的Denseblock子模块来表示，利用其良好的非线性变换能力，以提高其在水印嵌取任务中的效率。相应地，第l个可逆块的反向过程可表示为

$$\begin{aligned}b_{\text{im}}^{l}&=b_{\text{im}}^{l+1}-\varphi(b_{\text{wm}}^l)\\ b_{\text{wm}}^{l}&=(b_{\text{wm}}^{l+1}-\eta(b_{\text{im}}^{l+1}))\odot\exp(-\rho(b_{\text{im}}^{l+1}))\end{aligned} \tag{8}$$

2.5 融合和划分模块

融合和划分模块在水印嵌入过程中，处理可逆嵌取模块的输出得到含水印图像；在水印提取过程中，处理屏摄噪声图像得到可逆嵌取模块的输入。载体图像和水印信息经过可逆嵌取模块后将输出张量矩阵$\boldsymbol{\xi}_{\text{inv}}\in\mathbb{R}^{B\times24\times H/2\times W/2}$，将其在通道维度分割成两部分$\boldsymbol{\xi}_{\text{inv}}^{\text{I}}$，$\boldsymbol{\xi}_{\text{inv}}^{\text{II}}\in\mathbb{R}^{B\times12\times H/2\times W/2}$分别表示耦合后的载体图像和水印信息分支，其中$B$表示训练批次大小。在嵌入过程中，丢弃载体图像分支，只保留映射的水印分支，并将其与载体图像频域特征$\boldsymbol{I}_c^{\prime}$相加，然后通过Haar小波逆变换后得到最终的水印图像。这个过程可以表示为 $\boldsymbol{I}_w=\Gamma_{\text{haar}}^{-1}(\boldsymbol{\xi}_{\text{inv}}^{\text{II}}+\boldsymbol{I}_c^{\prime})$。其中，$\Gamma_{\text{haar}}^{-1}(\cdot)$表示Haar逆变换，为了保持可逆嵌取模块输入输出的一致性，在逆向过程先将含噪的水印图像通过haar小波变化后得到频域特征，然后再将频域特征复制后一同馈送到可逆网络中，此过程可表示为 $\xi_{\text{inv}}^{-1}=\Gamma_{\text{haar}}(I_n)$ ，$\Gamma_{\text{copy}}(\xi_{\text{inv}}^{-1},\ \xi_{\text{inv}}^{-1})$。

2.6 噪声层

为了提高在屏摄失真场景下的鲁棒性，在模型训练阶段，本文引入一个可微分的噪声层对输出的

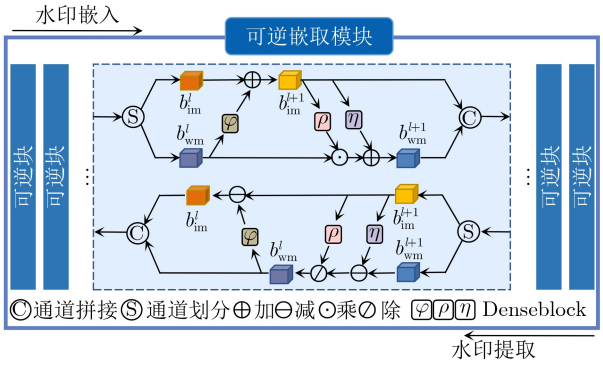


图 3 可逆嵌取模块

含水印图像$\boldsymbol{I}_w$进行处理，从而得到屏摄噪声图像$\boldsymbol{I}_n$。本文借鉴了文献[15]所提出的噪声层，通过选择屏摄影响最大的失真，并将他们混合形成一个组合噪声层用于训练。这种组合噪声层使得模型在屏摄失真场景中表现出更强的鲁棒性。屏摄噪声层包含透视失真、光照失真、摩尔失真、高斯噪声。

2.7 损失函数

本文模型损失函数可分为以下3个部分。水印不可见性损失：以最小化MSE损失指导含水印图像$\boldsymbol{I}_w$向载体图像$\boldsymbol{I}_c$趋近：$\ell_{\mathrm{mc}}=\mathrm{MSE}(\boldsymbol{I}_w,\boldsymbol{I}_c)$。因为MSE损失集中于像素级差异，可能导致生成的水印图像和原始载体图像的结构存在较大的差异。因此，本文使用结构相似性指数 (Structural Similarity Index, SSIM)损失来约束载体图像和水印图像之间的结构相似性。设$x=\{x_p|p=1,2,\cdots,N\}$和$y=y_p|p=1,2,\cdots,N\}$分别从载体图像$\boldsymbol{I}_c$和水印图像$\boldsymbol{I}_w$中提取的两个对应的图像块。SSIM损失计算为$\ell_{\mathrm{ms}}=1-\mathrm{SSIM}(x,y)$。此外，为了进一步增强模型的不可感知性，引入注重图像高维特征和感知效果LPIPS损失来优化水印图像的视觉质量，将其表示为$\ell_{\mathrm{ml}}=\mathrm{LPIPS}(\boldsymbol{I}_c,\boldsymbol{I}_w)$。

水印鲁棒性损失：将恢复的水印信息$\hat{M}$与原水印信息M构MSE损失，$\ell_{\mathrm{mm}}=\mathrm{MSE}(M,\hat{M})$。

重构图像损失：在原始载体图像和重构图像之间MSE损失来优化整个网络：$\ell_{\mathrm{mr}}=\mathrm{MSE}(\boldsymbol{I}_c,\hat{\boldsymbol{I}}_c)$。其中，$\hat{\boldsymbol{I}}_c$表示模型逆向过程重建得到的载体图像。本文方法的联合损失为

$$\mathcal{L}_{total}=\lambda_1\ell_{\mathrm{ml}}+\lambda_2\ell_{\mathrm{mc}}+\lambda_3\ell_{\mathrm{ms}}+\lambda_4\ell_{\mathrm{mm}}+\lambda_5\ell_{\mathrm{mr}} \quad (9)$$

超参数$\lambda_i,i=\{1,2,\cdots,5\}$用于权衡视觉质量和鲁棒性。

3 实验结果与分析

3.1 数据集和实验设置

本文选择了mirflickr数据集作为训练集，并采用经典的USC-SIPI图像数据集作为测试集。此外，为了验证生成的含水印图像的视觉质量以及评估模型在不同失真强度下的鲁棒性，本文在COCO图像数据集中选择300幅图像进行验证，涵盖了自然风景、人物肖像以及动物图像等多个类别。整体框架由PyTorch实现，并在NVIDIA RTX 3080上执行。为了统一训练图像的尺寸，所有训练图像都被重采样为(128×128×3)大小，并且随机比特信息M的长度被设置为30。在参数优化方面，采用Adam优化器，lr=1×10^{-4}，Batchsize=2，将本文提出的模型与最先进的传统方法SSRW[8]和3种基于深度学习的方法HiDDeN[11], Stegastamp[14], PIMoG[15]进行比较。

3.2 视觉质量测试

为了验证水印图像的视觉质量，本文采用3个评价指标来评估本文模型的性能：峰值信噪比(Peak Signal-to-Noise Ratio, PSNR), SSIM以及LPIPS。PSNR独立测量载体图像和水印图像每对像素之间的距离；SSIM测量两幅图像的结构相似性；相应地，依赖于深层特征的LPIPS能够更好地反映人类视觉判断。比较不同方法的水印图像的视觉质量，结果如图4和表1所示。

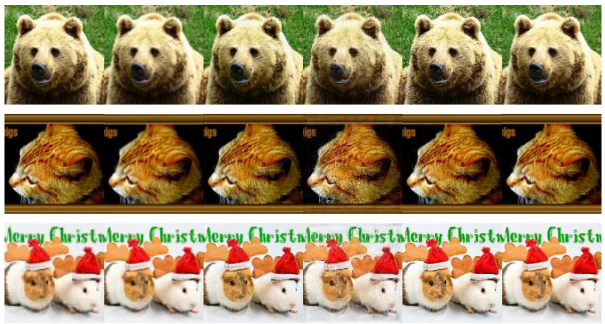
从表1可知，本文提出的方法的客观评价指标PSNR、SSIM值都大于其他方法，同时主观评价指标LPIPS值远小于文献[8,14]，具有良好的视觉不可感知性。此外，从图4中可以看出，文献[11,14]会在嵌入区域产生较为明显的视觉失真，但本文所提方法生成的水印图像能生成更加自然和平滑的水印图像，减少视觉伪影。这主要归功于本文模型的可逆结构能减少信息传递损失，提高了水印图像的不可感知性。

3.3 模型鲁棒性测试

本文使用提取水印提取准确率作为模型鲁棒性评估的标准，水印提取准确率越高模型鲁棒性越好。

(1)屏幕拍摄鲁棒性测试

本文首先将随机比特消息位嵌入到USI-SIPI图像数据集中，随后在屏幕上显示分辨率统一放大到(1024 × 1024)的水印图像，并在不同的条件下使用拍摄设备进行拍摄。对于捕获的图像，先检测并透视校正每张照片中的水印图像，再使用解码器提取



载体图像　文献[8]　文献[11]　文献[14]　文献[15]　本文

图 4 不同方法的可视化水印图像

表 1 视觉质量指标测试

方法	PSNR↑	SSIM↑	LPIPS↓
SSRW[8]	37.776	0.969	0.033
HiDDeN[11]	31.152	0.945	0.011
Stegastamp[14]	29.419	0.906	0.063
PIMoG[15]	36.411	0.975	0.009
本文	**39.306**	**0.984**	**0.001**

校正后图像中的水印，水印图像的检测和校正过程与文献[14]设置相同。屏幕显示器选择LG-22M37A和ENVISION-G249G，拍摄设备包括HuaWei Nova7、HuaWei mate30和Nikon Z30。

拍摄距离测试：在ENVISION屏幕上显示带水印的图像并使用HuaWei Nova7手机在不同的距离下进行拍摄，拍摄距离在20～60 cm之间，相应的水印提取准确率如表2所示。

由表2可知，本文的方法在拍摄距离增加的情况的鲁棒性略有降低，但仍然能保证 98%的水印提取准确率，具有较强的鲁棒性。

拍摄角度测试：为验证模型在不同水平/垂直拍摄角度下的鲁棒性，本文固定拍摄距离固定为40 cm，在ENVISION屏幕上显示带水印的图像并使用HuaWei Nova7手机在不同的水平/垂直角度进行拍摄，相应的水印提取精度如表3、表4所示。

分析表3、表4可知，本文提出的方法在水平角度下拍摄水印图像的提取准确率都在98%以上，垂直角度下拍摄水印图像的提取准确率相对于其他方法也具有一定的优势。

拍摄设备测试：为了测试本文方法实际应用场景的适用性，在不同类型的屏幕和拍摄设备上进行了屏幕拍摄实验，拍摄距离为20 cm。测试结果如表5所示。

分析表5可知，本文提出的方法在不同的屏幕显示设备和拍摄设备上的水印提取准确率都在90%以上，与最新的基于深度学习的抗屏摄方法PIMoG水印提取准确率相当，具有良好的稳定性。

(2)不同失真强度鲁棒性测试

为了全面深入地探究模型的鲁棒性，针对对屏摄影响最大的失真类型，通过构建准确度曲线来评估模型在不同水平噪声强度下的性能。结果如图5所示，横坐标表示不同失真类型，纵坐标表示对应的水印提取准确率。其中，透视失真的随机扰动强度表示随机扰动水印图像4个角的像素值大小，摩尔失真率表示摩尔失真添加到水印图像中的比例值。可以看出，文献[15]能应对不同强度的光照失真、摩尔失真，但是在透视变换扰动强度变大时，水印提取准确率急剧下降。同样文献[14]在透视失真扰动强度变大时，水印提取准确率也明显降低。文献[11]在透视失真、光照失真中具有较为稳定的表现，但很容易受到摩尔失真的干扰。而文献[8]不能抵抗透视失真以及光照失真。相比之下，本文的方法不仅能够抵抗较大的透视扰动强度，同时还能在不同强度下的光照失真、摩尔失真条件下维持较高的水印提取准确率。这证明本文使用的噪声层有助于提升模型的鲁棒性。

3.4 消融实验

本文提出的逆梯度注意模块和引入LPIPS损失使得模型获得更好的性能，在本节中进行消融分析来验证其合理性。

本文分别训练有和没有逆梯度注意模块的两个模型，并在模型收敛后执行视觉质量测试和拍摄距离为20～60 cm的屏幕拍摄实验。相应的结果如表6和表7所示，为了简化表示，使用w/o IGA表示未使用逆梯度注意模块，w IGA表示使用该模块。

分析表6、表7可知，当使用逆梯度注意模块时，在视觉质量方面PSNR和SSIM值显著增加，同时LPIPS值也有所降低。拍摄距离较远时，水印提取准确率远高于未使用逆梯度注意模块的模型，这

表 2 不同拍摄距离下的水印提取准确率(%)

距离(cm)	20	30	40	50	60
SSRW[8]	92.38	90.00	71.19	95.00	82.97
HiDDeN[11]	80.22	80.44	82.00	73.11	85.00
Stegastamp[14]	99.93	89.99	90.93	85.33	97.10
PIMoG[15]	97.78	**99.78**	**99.33**	94.52	**99.17**
本文	**100**	99.52	99.29	**98.91**	99.08

表 3 不同水平拍摄角度下的水印提取准确率(%)

角度(°)	左45	左30	左15	右15	右30	右45
SSRW[8]	83.21	88.45	79.04	92.38	81.91	77.86
HiDDeN[11]	72.26	76.67	87.02	88.21	71.19	82.74
Stegastamp[14]	97.13	99.03	97.20	97.48	86.47	96.37
PIMoG[15]	96.91	94.64	**98.21**	**99.88**	96.79	**98.93**
本文	**98.57**	**96.55**	**98.21**	98.81	**98.33**	98.81

表 4 不同垂直拍摄角度下的水印提取准确率(%)

角度(°)	上45	上30	上15	下15	下30	下45
SSRW[8]	76.67	95.12	98.33	82.62	93.69	84.05
HiDDeN[11]	74.05	77.50	73.93	88.81	70.95	75.24
Stegastamp[14]	95.40	98.03	99.17	98.41	98.13	88.67
PIMoG[15]	**99.52**	**99.76**	95.78	99.05	96.43	**96.19**
本文	99.29	99.05	**99.22**	**99.17**	**98.57**	94.40

表 5 不同显示/拍摄设备下的水印提取准确率(ENVISION/LG)

方法	HuaWei Nova7	Nikon Z30	HuaWei mate30
SSRW[8]	92.38/79.52	83.81/72.14	75.24/84.76
HiDDeN[11]	80.22/66.22	81.78/77.11	73.56/69.33
Stegastamp[14]	99.93/86.86	86.53/**99.00**	81.00/97.00
PIMoG[15]	97.77/96.00	**100**/96.89	**94.44**/95.33
本文	**100/97.11**	99.76/98.10	90.95/**99.29**

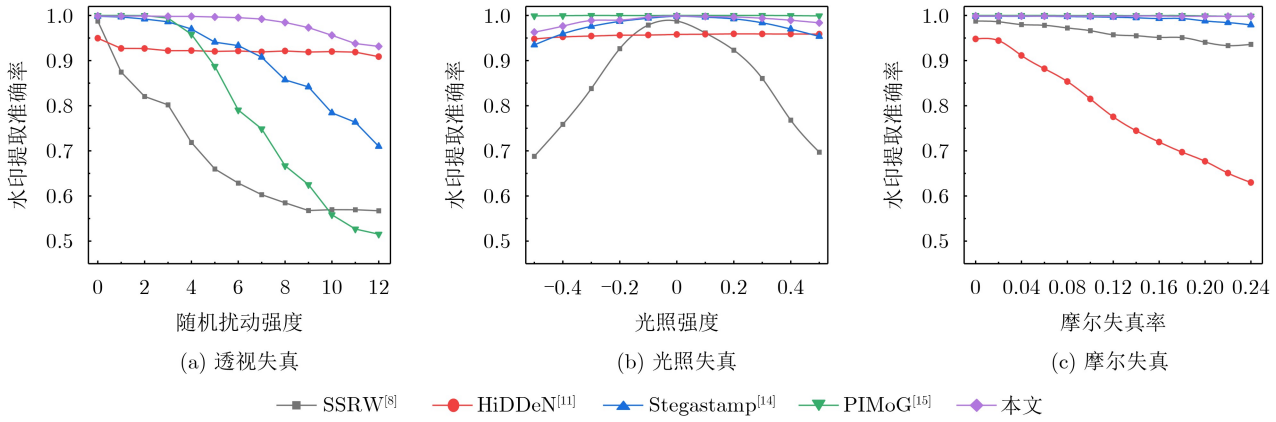


图 5 不同强度失真条件下的水印提取准确率

表 6　视觉质量指标测试($w\&w/o$ IGA)

视觉评价指标	PSNR↑	SSIM↑	LPIPS↓
w/o IGA	37.464	0.972	0.006
w IGA	**39.306**	**0.984**	**0.001**

表 7　不同拍摄距离的水印提取精度($w\&w/o$ IGA)

距离(cm)	20	30	40	50	60
w/o IGA	99.52	97.63	95.95	90.95	97.62
w IGA	**100**	**99.52**	**99.29**	**98.91**	**99.08**

表明引入逆梯度注意模块能帮助网络有效学习更好的特征，提升模型鲁棒性和水印图像的视觉质量。

本文通过引入LPIPS损失优化水印嵌入过程，以获得更好的视觉质量。对比了有和没有使用LPIPS损失的两个模型，并通过输出两个模型生成的水印图像来展示LPIPS损失的效果。$w/o\ell_{\text{lpips}}$表示未使用LPIPS损失，$w\ell_{\text{lpips}}$表示使用LPIPS损失。从图6可以看出，LPIPS损失的使用能够减小水印残差块的大小，生成更符合人类视觉感知的水印图像。这

图 6 有/没有LPIPS损失引导的水印图像及水印残差

证明LPIPS损失在水印嵌入中起到了关键作用，提高了水印图像的质量。

4　结束语

本文提出了一种基于INN和逆梯度注意力的抗屏摄攻击水印方法，该方法将水印的嵌入和提

取视为相互关联的逆问题，实现了编解码网络的一体化，减少了水印信息传递损失。通过引入逆梯度注意模块和LPIPS损失，在保持水印图像高视觉质量的同时实现了水印的鲁棒性。通过与同类抗屏摄水印算法进行实验对比，表明本文提出的方法能够生成高视觉感知的水印图像，同时能够达到较高的水印提取精度，确保屏摄水印的鲁棒性。这为屏摄失真场景下的数字媒体内容的版权保护、防伪溯源提供了一种有效的解决方案，未来可进一步拓展研究，探索更多的优化策略和应用场景，推动深度水印技术的发展和应用。

参 考 文 献

[1] VAN SCHYNDEL R G, TIRKEL A Z, and OSBORNE C F. A digital watermark[C]. 1st International Conference on Image Processing, Austin, USA, 1994, 2: 86–90. doi: 10.1109/ICIP.1994.413536.

[2] 方涵. 屏摄鲁棒水印方法研究[D]. [博士论文], 中国科学技术大学, 2021. doi: 10.27517/d.cnki.gzkju.2021.000591.
FANG Han. Research on screen shooting resilient watermarking[D]. [Ph. D. dissertation], University of Science and Technology of China, 2021. doi: 10.27517/d.cnki.gzkju.2021.000591.

[3] WAN Wenbo, WANG Jun, ZHANG Yunming, *et al*. A comprehensive survey on robust image watermarking[J]. *Neurocomputing*, 2022, 488: 226–247. doi: 10.1016/j.neucom.2022.02.083.

[4] 项世军, 杨乐. 基于同态加密系统的图像鲁棒可逆水印算法[J]. 软件学报, 2018, 29(4): 957–972. doi: 10.13328/j.cnki.jos.005406.
XIANG Shijun and YANG Le. Robust and reversible image watermarking algorithm in homomorphic encrypted

domain[J]. *Journal of Software*, 2018, 29(4): 957–972. doi: 10.13328/j.cnki.jos.005406.

[5] 张天骐, 周琳, 梁先明, 等. 基于Blob-Harris特征区域和NSCT-Zernike的鲁棒水印算法[J]. 电子与信息学报, 2021, 43(7): 2038–2045. doi: 10.11999/JEIT200164.
ZHANG Tianqi, ZHOU Lin, LIANG Xianming, *et al*. A robust watermarking algorithm based on Blob-Harris and NSCT-Zernike[J]. *Journal of Electronics & Information Technology*, 2021, 43(7): 2038–2045. doi: 10.11999/JEIT200164.

[6] KANG Shuangyong, JIN Biao, LIU Yuxin, *et al*. Research on screen shooting resilient watermarking based on dot-matric[C]. 2023 2nd International Conference on Big Data, Information and Computer Network (BDICN), Xishuangbanna, China, 2023: 194–199. doi: 10.1109/BDICN58493.2023.00048.

[7] SCHABER P, KOPF S, WETZEL S, *et al*. CamMark: Analyzing, modeling, and simulating artifacts in camcorder copies[J]. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 2015, 11(2s): 42. doi: 10.1145/2700295.

[8] FANG Han, ZHANG Weiming, ZHOU Hang, *et al*. Screen-shooting resilient watermarking[J]. *IEEE Transactions on Information Forensics and Security*, 2019, 14(6): 1403–1418. doi: 10.1109/TIFS.2018.2878541.

[9] DONG Li, CHEN Jiale, PENG Chengbin, *et al*. Watermark-preserving keypoint enhancement for screen-shooting resilient watermarking[C]. 2022 IEEE International Conference on Multimedia and Expo, Taipei, China, 2022: 1–6. doi: 10.1109/ICME52920.2022.9859950.

[10] KANDI H, MISHRA D, and GORTHI S R K S. Exploring the learning capabilities of convolutional neural networks for robust image watermarking[J]. *Computers & Security*, 2017, 65: 247–268. doi: 10.1016/j.cose.2016.11.016.

[11] ZHU Jiren, KAPLAN R, JOHNSON J, *et al*. HiDDeN: Hiding data with deep networks[C]. 15th European Conference on Computer Vision, Munich, Germany, 2018: 682–697. doi: 10.1007/978-3-030-01267-0_40.

[12] ZHANG Honglei, WANG Hu, CAO Yuanzhouhan, *et al*. Robust data hiding using inverse gradient attention[EB/OL]. https://doi.org/10.48550/arXiv.2011.10850, 2020.

[13] WENGROWSKI E and DANA K. Light field messaging with deep photographic steganography[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, USA, 2019: 1515–1524. doi: 10.1109/CVPR.2019.00161.

[14] TANCIK M, MILDENHALL B, and NG R. StegaStamp: Invisible hyperlinks in physical photographs[C]. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, USA, 2020: 2114–2123. doi: 10.1109/CVPR42600.2020.00219.

[15] FANG Han, JIA Zhaoyang, MA Zehua, *et al*. PIMoG: An effective screen-shooting noise-layer simulation for deep-learning-based watermarking network[C]. 30th ACM International Conference on Multimedia, Lisbon, Portugal, 2022: 2267–2275. doi: 10.1145/3503161.3548049.

[16] DINH L, KRUEGER D, and BENGIO Y. NICE: Non-linear independent components estimation[C]. 3rd International Conference on Learning Representations, San Diego, USA, 2015.

[17] KINGMA D P and DHARIWAL P. Glow: Generative flow with invertible 1×1 convolutions[C]. Proceedings of the 32nd International Conference on Neural Information Processing Systems, Montréal, Canada, 2018: 10236–10245.

[18] XIE Yueqi, CHENG K L, and CHEN Qifeng. Enhanced invertible encoding for learned image compression[C]. 29th ACM International Conference on Multimedia, Chengdu, China, 2021: 162–170. doi: 10.1145/3474085.3475213.

[19] XIAO Mingqing, ZHENG Shuxin, LIU Chang, *et al*. Invertible image rescaling[C]. 16th European Conference on Computer Vision, Glasgow, UK, 2020: 126–144. doi: 10.1007/978-3-030-58452-8_8.

[20] GUO Mengxi, ZHAO Shijie, LI Yue, *et al*. Invertible single image rescaling via steganography[C]. 2022 IEEE International Conference on Multimedia and Expo (ICME), Taipei, China, 2022: 1–6. doi: 10.1109/ICME52920.2022.9859915.

[21] LU Shaoping, WANG Rong, ZHONG Tao, *et al*. Large-capacity image steganography based on invertible neural networks[C]. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, USA, 2021: 10811–10820. doi: 10.1109/CVPR46437.2021.01067.

[22] GUAN Zhenyu, JING Junpeng, DENG Xin, *et al*. DeepMIH: Deep invertible network for multiple image hiding[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023, 45(1): 372–390. doi: 10.1109/TPAMI.2022.3141725.

[23] MA Rui, GUO Mengxi, HOU Yi, *et al*. Towards blind watermarking: Combining invertible and non-invertible mechanisms[C]. 30th ACM International Conference on Multimedia, Lisbon, Portugal, 2022: 1532–1542. doi: 10.1145/3503161.3547950.

[24] ZHANG R, ISOLA P, EFROS A A, *et al*. The unreasonable effectiveness of deep features as a perceptual metric[C]. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, USA, 2018: 586–595. doi: 10.1109/CVPR.2018.00068.

李谢华：女，助理教授，研究方向为大数据安全，云计算与存储系统安全；图像与文本可搜索加密.

娄 芹：女，硕士生，研究方向为数字水印、人工智能安全.

杨俊雪：女，博士生，研究方向为信息隐藏、人工智能安全.

廖 鑫：男，教授，研究方向为多媒体安全、数字取证、人工智能安全.

责任编辑：马秀强