

基于可逆神经网络的小波域抗屏摄数字水印

程森茂¹, 郭玳豆¹, 栗风永², 韩彦芳¹, 秦川¹

(1. 上海理工大学 光电信息与计算机工程学院, 上海 200093; 2. 上海电力大学 计算机科学与技术学院, 上海 201306)

摘要: 屏摄信道传输中的多媒体安全问题是数字水印研究领域中的一大难点。鉴于屏摄过程中存在光照度差异和采样失真的问题, 基于数字信道设计的水印方法并不适用于屏摄信道, 为此, 提出了一种基于可逆神经网络的小波域抗屏摄水印方法, 以解决“跨媒介鲁棒性”问题。通过基于U-Net网络结构的预处理网络对水印信息和原始图像进行预编码生成残差图像, 将该残差图像与原始图像进行离散小波变换, 并利用可逆神经网络实现在小波域的水印嵌入和提取。在模型训练过程中加入噪声池, 以提升抵抗屏摄噪声攻击的鲁棒性。实验结果表明, 通过所提方法生成的含水印图像具有更好的视觉质量, 且在不同距离、角度和光照度的屏摄情况下, 均有着较高的水印信息提取准确率。

关键词: 鲁棒水印; 屏幕拍摄; 可逆神经网络; 小波域

中图分类号: TP 929 **文献标志码:** A

Screen-shooting resilient digital watermarking based on invertible neural network in wavelet domain

CHENG Senmao¹, GUO Daidou¹, LI Fengyong², HAN Yanfang¹, QIN Chuan¹

(1. School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China; 2. College of Computer Science and Technology, Shanghai University of Electric Power, Shanghai 201306, China)

Abstract: Multimedia security in screen-shooting channel transmission remains a major challenge in the digital watermarking research. In view of the problems of differences in light intensities and sampling distortions during the screen-shooting process, watermarking schemes designed based on digital channels are not suitable for screen-shooting channels. Therefore, a screen-shooting resilient watermarking scheme based on invertible neural networks in the wavelet domain was proposed to address the issue of "cross-media robustness". First, the watermark message and the original image were

收稿日期: 2024-01-10

基金项目: 国家自然科学基金资助项目 (62172280, U20B2051); 上海市自然科学基金资助项目 (21ZR1444600, 20ZR1421600)

第一作者: 程森茂 (1998—), 男, 硕士研究生。研究方向: 深度学习和鲁棒水印。E-mail: 212230406@st.usst.edu.cn

通信作者: 韩彦芳 (1974—), 女, 讲师。研究方向: 数字图像处理、多媒体信息安全。E-mail: hyf@usst.edu.cn

引文格式: 程森茂, 郭玳豆, 栗风永, 等. 基于可逆神经网络的小波域抗屏摄数字水印[J]. 上海理工大学学报, 2025, 47(2): 220-229.

Citation: CHENG Senmao, GUO Daidou, LI Fengyong, et al. Screen-shooting resilient digital watermarking based on invertible neural network in wavelet domain[J]. Journal of University of Shanghai for Science and Technology, 2025, 47(2): 220-229.

preprocessed using a preprocessing network based on the U-Net structure to generate the residual image. Next, the residual image was subjected to discrete wavelet transform with the original image, and the invertible neural network was used to embed and extract watermarks in the wavelet domain. Finally, a noise pool was integrated into the model training process to enhance the robustness against screen-shooting noise attacks. Simulation results demonstrate that the proposed scheme generates watermarked images with better visual quality and achieves high accuracy of watermark extraction at different distances, angles, and light intensities of screen-shooting.

Keywords: *robust watermarking; screen-shooting; invertible neural network; wavelet domain*

随着智能便携摄影设备拥有越来越高的拍照性能和普及率，拍照成为了日常生活中最简单高效的信息传递方式之一。人们只需要对显示在屏幕上的多媒体内容进行拍照就可以实现信息的获取，且不会留下任何记录痕迹。因此，屏幕拍照过程中的信息泄露问题亟需解决。抗屏摄数字水印的目标是通过在多媒体数据中嵌入水印信息，使得显示在屏幕上的多媒体数据即使被相机所捕获，水印信息仍然能够保证可提取，从而达到版权保护和追踪溯源的目的^[1-2]。

相较于人为设计嵌入规则的传统鲁棒水印方法，基于深度学习的鲁棒水印方法可以更好地发挥特征拟合能力，并通过端到端的联合训练实现高质高效嵌入，包括一般的深度鲁棒水印技术^[3-7]和针对跨媒介通信的深度鲁棒水印技术^[8-12]。然而，一般的深度鲁棒水印方法通常只考虑数字传输信道的鲁棒性，例如噪声、平移、缩放、JPEG 压缩和滤波等，忽略了跨媒介传输信道，特别是屏摄传输信道所带来的噪声攻击。与数字传输信道相比，屏摄传输信道中涉及的噪声更加复杂，例如屏幕显示和相机成像之间的数模转换、传感器以及后处理操作等，导致成像过程不仅受到硬件设备（如显示器类型和摄像头质量）的影响，还受到人为和环境因素（如环境光源、拍摄角度和拍摄距离）的影响。因此，真实屏摄环境下的噪声对含水印图像的干扰性更强，这就对水印信息的提取精度提出了更高的要求。

为了解决该问题，研究者们提出了一系列基于深度神经网络的抗屏摄数字水印方法。例如，Zhu 等^[13]提出了 HiDDeN 水印框架，通过在端到端的训练中引入噪声层来增强鲁棒性。Liu 等^[14]提出了一种基于深度学习的两阶段可分离水印框架。第一阶段采用多层特征编码策略来训练编码

器；第二阶段结合真实的噪声攻击生成噪声图像，利用噪声图像对解码器进行训练，以增强解码器的鲁棒性。Tancik 等^[15]提出了适用于“打印-拍照”过程的水印方法——StegaStamp，并使用生成对抗网络^[16]（generative adversarial network, GAN）对图像进行对抗训练。此外，还有 SSDeN^[17]、RIHOOP^[18]和 PIMoG^[19]等深度抗屏摄数字水印方法。上述抗屏摄水印方法基本上都遵循着编码器-噪声层-解码器的结构，虽然在不同程度上提升了抗屏摄攻击的能力，但是，该结构也存在着一些局限性。例如，解码器及其隐变量是由数据推断出的近似似然评估，这意味着模型训练目标不是一个精确的形式，会导致不可逆的特征丢失，不利于水印信息的提取。

针对上述问题，本文提出了一种基于可逆神经网络（invertible neural network, INN）的小波域抗屏摄水印方法，通过构建一个基于 U-Net 网络结构的预处理网络，对水印信息和原始图像进行预编码，提高含水印图像的视觉质量与鲁棒性。另外，基于可逆神经网络参数共享机制，在 Haar 小波域内实现水印信息的嵌入和提取，提升模型的训练效率和不可感知性，并增加了一个集成有数字攻击和真实屏摄攻击的噪声池，加强抵抗屏摄噪声攻击的能力。

1 基于可逆神经网络的小波域抗屏摄水印方法

如图 1 所示，本文所提出的基于可逆神经网络的小波域抗屏摄水印方法包括 5 个部分：预处理模块、可逆模块、噪声池、矫正模块和信息提取模块。图中：DWT 和 IWT 分别表示离散小波变换（discrete wavelet transform, DWT）和逆离散小波

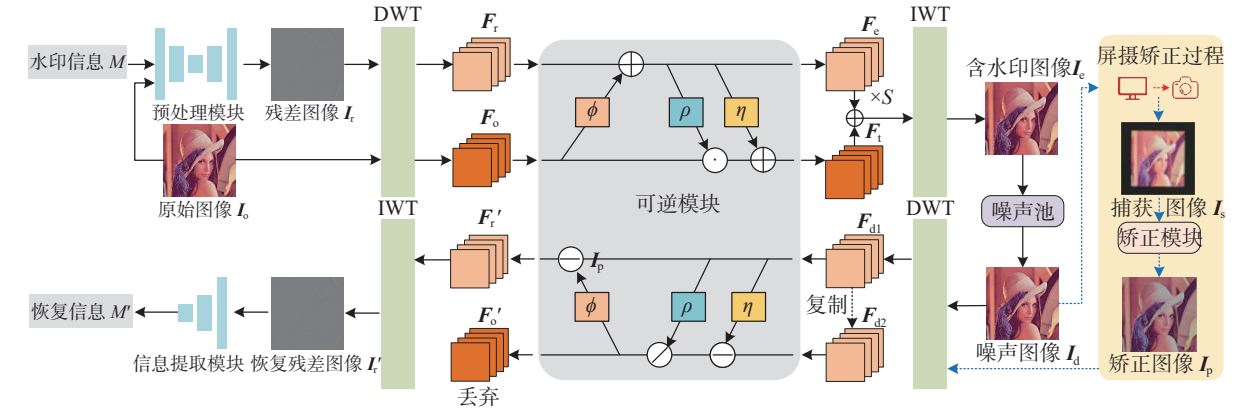


图 1 本文方法流程图
Fig.1 Flow chart of the method in this paper

变换 (inverse wavelet transform, IWT); F_o 和 F_r 分别表示原始图像特征和残差图像特征; F_e 和 F_i 分别表示含水印图像特征和原始图像特征 F_o 的输出; S 为缩放强度因子; F_{d1} 和 F_{d2} 分别为噪声图像特征和辅助变量; F'_r 和 F'_o 为恢复残差图像特征和恢复图像特征; $\phi(\cdot)$ 、 $\rho(\cdot)$ 和 $\eta(\cdot)$ 是任意函数。

1.1 预处理模块

预处理模块由一个全连接层 (fully connected layer, FC) 和一个简化版的 U-Net 网络^[20] 组成。简化版 U-Net 网络有 6 层网络结构, 其中包括: 2 个下采样层, 用于降低输入图像尺寸, 增强图像局部特征的提取; 中间的 1 个卷积层, 用于提取图像局部特征; 2 个上采样层, 用于将特征图恢复到原始输入图像的尺寸; 上采样和下采样之间的跳跃连接可以更多地传递上下文信息和保留细节; 最后 1 层为信息分类层。相较于常规的全连接层网络, U-Net 网络的设计使其在图像处理时能够更好地保留空间信息, 有效地融合了深层特征和浅层特征, 从而能够同时捕获图像的全局信息和细节信息。为了提升模型的训练速度, 本文设计了简化版的 U-Net 网络作为预处理模块, 减少模型的深度和卷积层中的参数数量, 在保证图像质量的同时降低了过拟合风险。

首先, 预处理网络将水印信息 $M \in (0, 1)^K$ 通过全连接层形成 $C \times (H/4) \times (W/4)$ 的张量; 然后, 再经过上采样操作生成与原始图像 $I_o \in (0, 255)^{C \times H \times W}$ 具有相同大小的张量; 最后, 将该张量和原始图像进行级联并输入到 U-Net 网络中, 得到含有水印信息的残差图像 I_r 。其中: C 表示通道数; H 表示图像的高度; W 表示图像的宽度; K 表示水印信息的长度。由于预处理模块实现了水印信息到原

始图像的预编码, 因此, 该模块的加入会增强含水印图像的视觉质量和鲁棒性, 其处理过程如下:

$$I_r = f_{PM}(M, I_o) = f_U(f_{US}(f_{FC}(M)), I_o) \quad (1)$$

式中, $f_{FC}(\cdot)$ 、 $f_{US}(\cdot)$ 、 $f_U(\cdot)$ 和 $f_{PM}(\cdot)$ 分别表示全连接操作、上采样操作、经过 U-Net 和经过预处理模块处理的操作。

1.2 基于频域的可逆神经网络

在 INN 中, 给定一个变量 x 和正向计算 $y=f(x)$, 可以通过反向计算 $x=f^{-1}(y)$ 直接恢复出变量 x , 正向函数 f 与反向函数 f^{-1} 共享相同的参数^[21-22]。受此启发, 为了提升模型的训练效率和降低训练难度, 本文使用 INN 作为骨干网络实现水印的嵌入和提取。此外, 根据人类视觉系统 (human visual system, HVS) 的掩蔽效应, 水印信息可以被嵌入到原始图像中不易被感知的区域, 特别是离散小波变换域中的高频区域。因此, 为了提高不可感知性, 本研究将 INN 与 Haar 小波域相结合。与传统的 U-Net 网络相比, INN 的设计允许在网络的正向和反向传播中精确地重建输入数据, 这意味着在频域上进行水印的嵌入和提取时, 它能够更好地保留图像的原始信息和细节。另外, INN 能够区分和处理图像信号的不同频率成分, 这使得它在面对噪声干扰时, 特别是在处理高频噪声时表现出更好的鲁棒性。INN 的逆向过程通过执行与正向过程相反的操作来进行, 其中, 正向传播和逆向传播共享参数。

在嵌入过程中, 大小为 $C \times H \times W$ 的原始图像 I_o 和残差图像 I_r 经过 DWT 后分别被转换成大小为 $4C \times (H/2) \times (W/2)$ 的张量, 得到 F_o 和 F_r , 其变换过程如下:

$$\mathbf{F}_o = f_{\text{DWT}}(\mathbf{I}_o) \quad (2)$$

$$\mathbf{F}_r = f_{\text{DWT}}(\mathbf{I}_r) \quad (3)$$

式中, $f_{\text{DWT}}(\cdot)$ 表示离散小波变换。

接着, 在将 $\mathbf{F}_o$ 和 $\mathbf{F}_r$ 输入到 INN 的正向传播过程中, 获得 $\mathbf{F}_e$ 和 $\mathbf{F}_t$ 。将 $\mathbf{F}_e$ 经强度因子 S 缩放并添加到 $\mathbf{F}_t$ 中, 通过 IWT 后得到含水印图像 $\mathbf{I}_e \in$

$(0, 255)^{C \times H \times W}$, 计算过程如下:

$$\mathbf{F}_e = \mathbf{F}_r \oplus \phi(\mathbf{F}_o) \quad (4)$$

$$\mathbf{I}_e = f_{\text{IWT}}(\mathbf{F}_e \times S \oplus (\exp(\rho(\mathbf{F}_e)) \odot \mathbf{F}_o \oplus \eta(\mathbf{F}_e))) \quad (5)$$

式中: $\oplus$ 表示加法运算; $\odot$ 表示点积运算; $f_{\text{IWT}}(\cdot)$ 表示逆离散小波变换。这里使用了 swin transformer^[23]的网络结构, 如图2所示。

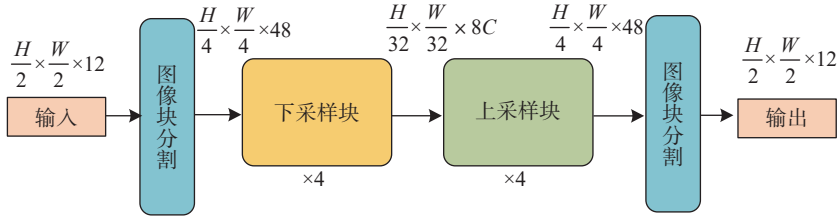


图2 swin transformer 结构

Fig.2 Structure of the swin transformer

提取过程中, 噪声图像 $\mathbf{I}_d$ 通过 DWT 后得到 $\mathbf{F}_{d1}$, 将其复制并作为辅助变量 $\mathbf{F}_{d2}$, 同时输入到 INN 的反向传播过程中, 得到 $\mathbf{F}'_r$ 和 $\mathbf{F}'_o$ 。由于所提方法的最终目标是恢复水印信息 M' , 本文将 $\mathbf{F}'_o$ 丢弃, 只保留 $\mathbf{F}'_r$ 。最后, 通过 IWT 得到恢复残差图像 $\mathbf{I}'_r$, 公式如下:

$$\mathbf{F}'_r = \mathbf{F}_{d1} \ominus \phi(\mathbf{F}'_o) \quad (6)$$

$$\mathbf{F}'_o = (\mathbf{F}_{d2} \ominus \eta(\mathbf{F}_{d1})) \odot \exp(-\rho(\mathbf{F}_{d1})) \quad (7)$$

$$\mathbf{I}'_r = f_{\text{IWT}}(\mathbf{F}'_r) \quad (8)$$

式中, $\ominus$ 表示减法运算。

1.3 噪声池

为了提升抵抗噪声攻击的能力, 本研究设计了一个噪声池来模拟屏摄过程中可能存在的攻击, 噪声池主要集成了5种攻击: 透视变换、裁剪攻击、JPEG压缩、组合噪声攻击、摩尔纹攻击。针对攻击透视变换-攻击组合噪声攻击, 本文使用可微函数进行近似模拟。例如: 使用一个随机单应性模拟与图像倾斜的相机效果模拟透视变换; 使用高斯噪声模型来解释成像噪声; 结合 Shin 等^[24]的方法运用分段函数逼近量化步长来模拟 JPEG 压缩。

由于无法使用可微函数对屏摄过程中产生的摩尔纹进行定量分析, 本文借鉴 Wengrowski 等^[25]的方法, 采用 CameraDisplay 1M 数据集对类似于 U-Net 网络的失真网络进行训练来模拟摩尔纹攻击, 具体结构如图3所示。

对含水印图像的攻击可表示为

$$\mathbf{I}_d = f_{\text{NP}}(\mathbf{I}_e) = A(\mathbf{I}_e) + Z \quad (9)$$

式中: $\mathbf{I}_d$ 表示受到攻击后的含水印图像, 即噪声图像; Z 表示叠加的失真摩尔纹; $A(\cdot)$ 表示5种随机噪声攻击; $f_{\text{NP}}(\cdot)$ 表示经过噪声池处理的操作。

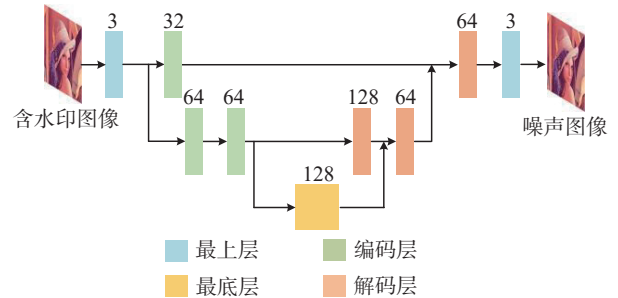


图3 摩尔纹噪声模拟网络结构

Fig.3 Structure of Moiré noise simulation network

1.4 矫正模块

在屏摄过程中, 通过相机捕获的屏摄图像 $\mathbf{I}_s$ 可能不仅包括含水印图像本身, 还包括部分多余的背景信息。这些背景信息的存在会严重干扰水印信息的提取精度。所以, 去除多余的背景信息非常必要。为了解决该问题, 本文采用了 DeepLabV3+^[26]语义分割模型实现目标区域的定位和分割, 如图4所示。同时, 使用基于中值滤波的 Canny 检测方法和 Hough 变换对分割区域的边缘进行平滑处理和角点定位。通过已定位角点对得到的分割图像进行透视变换处理, 最后获得矫正图像 $\mathbf{I}_p$, 计算过程如下:

$$\mathbf{I}_p = f_{\text{CN}}(\mathbf{I}_s) = (f_{\text{Hough}}(f_{\text{Canny}}(f_{\text{DM}}(\mathbf{I}_s)))) \quad (10)$$

式中, $f_{\text{DM}}(\cdot)$ 、 $f_{\text{Canny}}(\cdot)$ 、 $f_{\text{Hough}}(\cdot)$ 和 $f_{\text{CN}}(\cdot)$ 分别表示定位分割操作、Canny 检测、Hough 变换和经过

矫正模块处理的操作。

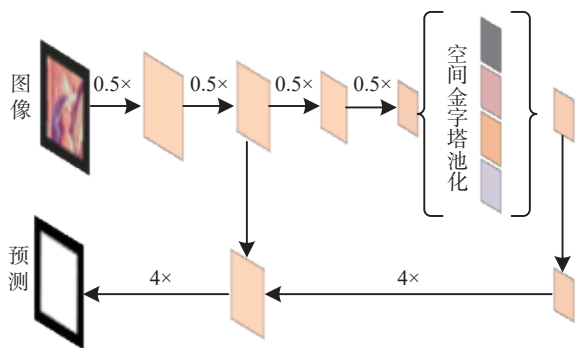


图4 基于DeepLabV3+的图像分割模型

Fig.4 Image segmentation model based on DeepLabV3+

1.5 信息提取模块

信息提取模块由4个卷积层和1个全连接层组成，其主要目标是从恢复的残差图像 I'_r 中提取出已嵌入的二进制水印信息 $M' \in (0,1)^K$ ，水印信息的提取过程如下：

$$M' = f_E(I'_r) \quad (11)$$

式中， $f_E(\cdot)$ 表示信息提取模块处理的操作。

1.6 损失函数

为了保证含水印图像的视觉质量以及水印信息的提取准确率，在训练过程中，分别使用了3种不同的损失函数对模型进行监督：缩小含水印图像 I_e 与原始图像 I_o 之间感知差异的损失函数 L_c ；提高信息恢复准确率的信息损失函数 L_R ；增强水印隐蔽性的低频小波损失函数 L_f 。

为了使编码后的含水印图像 I_e 与原始图像 I_o 在视觉感知上尽可能相似，设计了感知相似度损失函数 L_c ：

$$L_c = \frac{1}{C \times H \times W} \|I_o - I_e\|_2^2 \quad (12)$$

式中， $\|\cdot\|_2^2$ 表示对每个元素的平方求和，然后再计算平方根。

在信息提取过程中，给定含水印图像 I_e ，信息提取网络应该能够尽可能准确地恢复出长度为 K 的二进制水印信息 M' ，本文通过信息损失函数 L_R 监督该过程：

$$L_R = \frac{1}{K} \|M - M'\|_2^2 \quad (13)$$

由于人眼对图像高频子带噪声的敏感度较低，将水印信息嵌入到DWT域的高频成分中，可以提高隐蔽性。受此启发，文中提出了低频小波损失(low frequency wavelet loss, LL) L_f 来增强所提方法的隐蔽性。通过最小化原始图像和含水印

图像在 L_f 上的差值，使更多的水印信息嵌入到图像的高频成分中，低频小波损失函数的定义如下：

$$L_f = \frac{4}{C \times H \times W} \|I_{oLL} - I_{eLL}\|_2^2 \quad (14)$$

式中： I_{oLL} 表示原始图像在离散小波域的低频子带信息； I_{eLL} 表示含水印图像在离散小波域的低频子带信息。

最后，总损失函数 L_{total} 是 L_c 、 L_R 、 L_f 的加权和，即

$$L_{total} = \lambda_c L_c + \lambda_R L_R + \lambda_f L_f \quad (15)$$

式中， λ_c 、 λ_R 和 λ_f 是超参数。在模型训练初始阶段，只有 L_c 和 L_R 参与训练，随着模型逐渐收敛，最后将 L_f 添加到端到端的训练过程中。

2 实验结果及分析

2.1 实验数据集及参数设置

实验中所使用的原始图像均来自于COCO 2014数据集^[27]，其中，训练集包括40000张图像，测试集包括10000张图像，训练集中的图像数据与测试集中的图像数据均不相同。为了保证图像尺寸的统一性，使用中心裁剪的方式对原始图像进行统一裁剪，像素大小设置为256×256。二进制水印信息则采用随机生成的方式获取，长度设置为64 bit。整个模型使用双Nvidia 3090显卡进行训练，其中，批量大小设置为8，学习率设置为0.0001，优化算法选择Adam算法。图像分割模型DeepLabV3+所使用的数据集则为自建屏摄数据集，采集标准如下：相机拍摄角度控制在±25°的偏转范围内，拍摄距离小于等于50 cm，所拍摄的图像来自于ImageNet数据集^[28]，并对每张屏摄图像的角点进行标记，生成训练标签。

超参数 λ_c 、 λ_R 和 λ_f 取值分为以下3种情况：无噪声池参与的模型在训练过程中，参数分别设置为5.0、0.1和5.0；只添加数字攻击对模型进行训练时，参数分别设置为5.0、1.0和5.0；同时添加数字攻击和屏摄攻击对模型进行训练时，参数分别设置为1.0、1.0和1.0。

缩放强度因子 S 随着训练的进行而逐步减小，初始取值为1.5，每个训练周期之后衰减0.05。20个训练周期之后衰减到最小值0.5，之后不再变化。

2.2 评价指标

实验通过计算水印信息的提取精度

(accuracy, ACC)进行鲁棒性评估, 如式(16)所示, 误码率(bit error rate, BER)如式(17)所示。

$$A = 1 - B \quad (16)$$

$$B = \frac{1}{K} \sum_{i=1}^L |M_i - M'_i| \times 100\% \quad (17)$$

式中: A 为提取精度; B 为误码率; M_i 表示原始水印信息的第 i 个比特; M'_i 表示恢复水印信息的第 i 个比特。

通过使用峰值信噪比(peak signal-to-noise ratio, PSNR)和结构相似性(structural similarity index, SSIM)对含水印图像的图像质量进行评估:

$$P = 10 \lg \left(\frac{I_{\text{MAX}}^2}{E_{\text{MSE}}} \right) \quad (18)$$

$$Q_{\text{SSIM}}(I_o, I_e) = \frac{1}{N} \sum_{i=1}^N \varphi(o^{(i)}, e^{(i)}) \quad (19)$$

$$\varphi(o^{(i)}, e^{(i)}) = \frac{(2\mu_o^{(i)}\mu_e^{(i)} + C_1)(2\sigma_{oe}^{(i)} + C_2)}{\left\{ [\mu_o^{(i)}]^2 + [\mu_e^{(i)}]^2 + C_1 \right\} \left\{ [\sigma_o^{(i)}]^2 + [\sigma_e^{(i)}]^2 + C_2 \right\}} \quad (20)$$

式中: P 为峰值信噪比; E_{MSE} 为均方误差(mean squared error, MSE), $E_{\text{MSE}} = \frac{1}{HW} \sum_{x=1}^H \sum_{y=1}^W [I_o(x, y) - I_e(x, y)]^2$; Q_{SSIM} 为衡量结构相似性的指标; $I_o(x, y)$ 和 $I_e(x, y)$ 分别表示原始图像 I_o 和含水印图像

I_e 在坐标 (x, y) 处的像素; I_{MAX} 指图像的最大像素值; $o^{(i)}$ 和 $e^{(i)}$ 是 I_o 和 I_e 的第 i 个块; N 是指块的总数; $\mu_o^{(i)}$ 和 $\mu_e^{(i)}$ 指 $o^{(i)}$ 和 $e^{(i)}$ 的均值; $\sigma_{oe}^{(i)}$ 指 $o^{(i)}$ 和 $e^{(i)}$ 的协方差; $\sigma_o^{(i)}$ 和 $\sigma_e^{(i)}$ 指 $o^{(i)}$ 和 $e^{(i)}$ 的标准差; C_1 和 C_2 是两个趋近于 0 的小常数。

2.3 图像质量比较

对于含水印图像的质量分析, 实验分别与 StegaStamp、HiDDeN、CIN^[29]、UDH^[30] 和 PIMoG 方法进行比较。所有方法均使用来自于 COCO2014 数据集中相同的 1000 张原始图像, 水印信息的长度设置为 64 bit。图 5 是一组原始图像以及分别用 5 种方法所生成的含水印图像示例。通过图 5 可以看到, 与其他 5 种方法相比, 通过本文方法生成的含水印图像在大块纯色区域上, 阴影大幅度减少, 较好地保留了图像的细节特征, 在生成的含水印图像质量上也有着较为明显的优势, 详细实验结果见表 1。由表 1 可知, 本文方法生成的含水印图像, 其 PSNR 值和 SSIM 值分别达到 35.07 dB 和 0.9736, 相较于抗屏摄水印模型 StegaStamp, PSNR 值提高了 7.82 dB, SSIM 值提高了 0.0709。同时, 表 1 数据说明, 本文方法也分别优于 CIN、HiDDeN、UDH 和 PIMoG 这 4 种模型。

2.4 鲁棒性比较

为了更加全面客观地评价本文方法的鲁棒性, 实验分别对 6 种方法所生成的含水印图像进行了数字攻击, 对 StegaStamp、CIN、UDH、

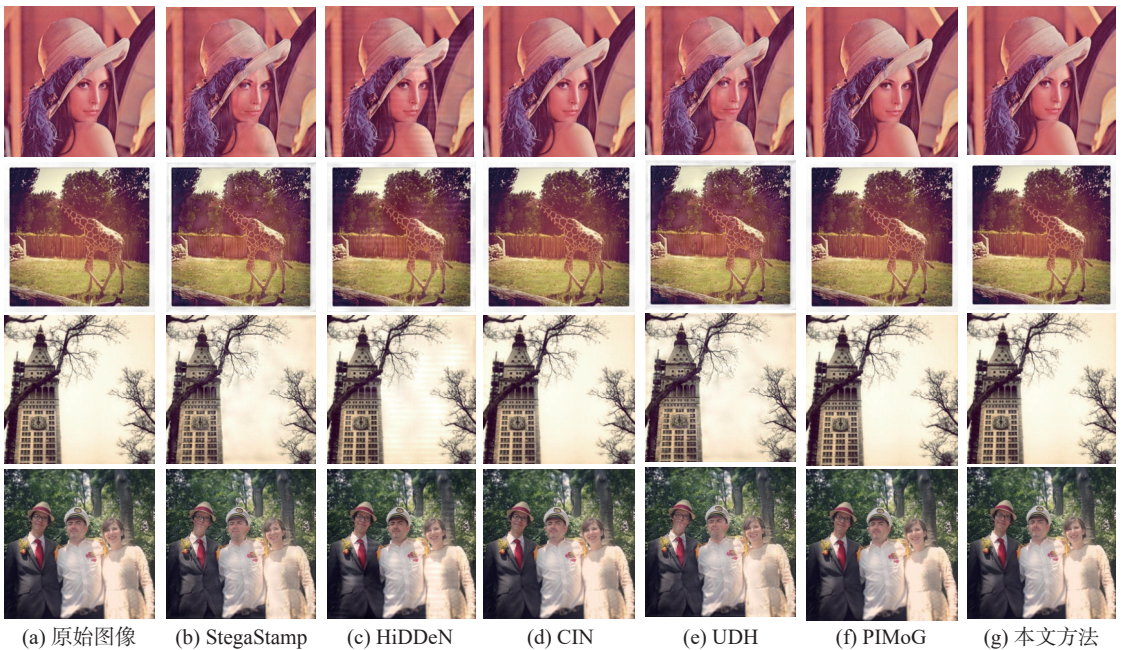


图 5 图像质量对比

Fig.5 Comparisons of image quality

表 1 含水印图像的质量比较

Tab.1 Comparisons of watermarked image quality

方法	PSNR/dB	SSIM
StegaStamp ^[13]	27.25	0.902 7
HiDDeN ^[11]	33.68	0.955 4
CIN ^[29]	34.16	0.969 4
UDH ^[30]	28.88	0.911 4
PIMoG ^[17]	33.43	0.938 4
本文方法	35.07	0.973 6

PIMoG 和本文方法进行了真实屏摄攻击，其中，水印信息的长度同样设置为 64 bit。

2.4.1 抗数字攻击的性能评估

实验随机选取了来自 COCO2014 数据集中的 1000 张图像作为原始图像，并使用 3 种常见的数字攻击，即裁剪、JPEG 压缩和组合噪声对含水印的图像进行测试。a. 裁剪攻击：现实屏摄过程中，异物遮挡镜头常导致拍摄到的图像不完整。为了模拟物体遮挡的情况，本研究选用不同比例大小的黑色像素块随机覆盖含水印图像，实验结果见图 6。b. JPEG 压缩：在图像存储和传输过程中，为了节省图像存储空间以及增加传输效率，一般会采用 JPEG 格式对图像进行有损压缩，本研究使用了不同的压缩质量因子对含水印图像进行测试，实验结果见图 7。c. 组合噪声：实验中主要考虑了最常见的椒盐噪声和高斯噪声。椒盐噪声在图像中表现为离散分布的纯白色或黑色像素点，高斯噪声是指噪声密度函数服从高斯分布的噪声，标准方差的大小代表了噪声强度。实验主要对含水印图像随机添加不同强度的椒盐噪声和

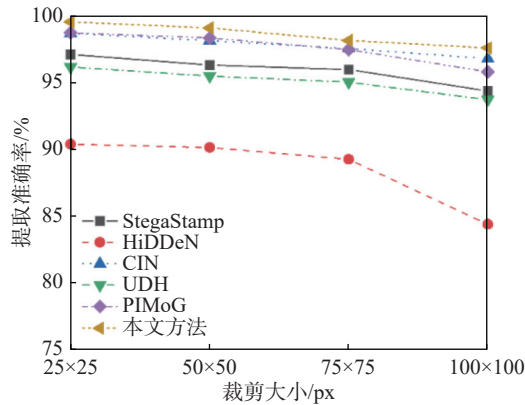


图 6 不同裁剪比例下的水印提取准确率

Fig.6 Accuracy of watermark extraction at different cropping ratios

高斯噪声进行测试，实验结果见图 8。通过图 6~8 可以看出，相较于 StegaStamp、HiDDeN、CIN、PIMoG 和 UDH，本文方法在不同类型和不同强度的数字攻击下均有着较高的提取准确率。

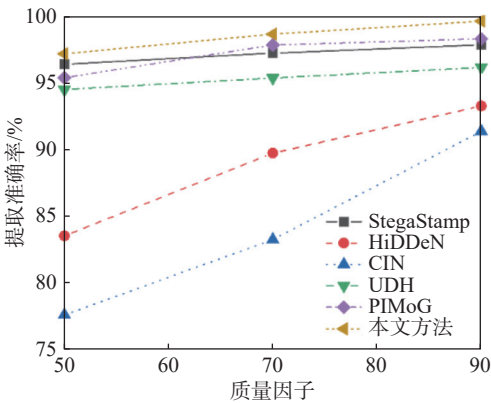


图 7 不同 JPEG 压缩质量因子下的水印提取准确率

Fig.7 Accuracy of watermark extraction at different JPEG quality factors

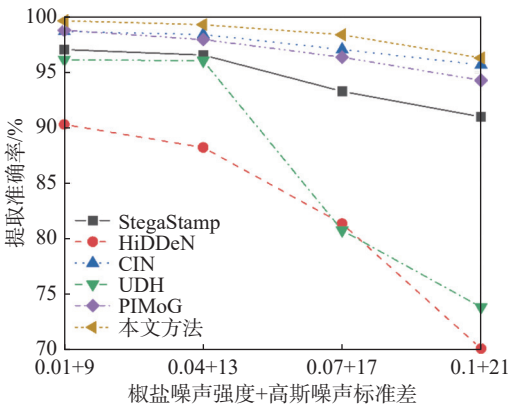


图 8 不同组合噪声强度下的水印提取准确率

Fig.8 Accuracy of watermark extraction at different intensities of combined noise

2.4.2 抗屏摄攻击的性能评估

针对真实的屏摄攻击，实验随机选取了来自 COCO2014 数据集中的 50 张图像作为原始图像，并对各方法生成的含水印图像在不同的拍摄距离、拍摄角度和光照度条件下进行测试。

a. 不同拍摄距离：实验分别在 20、35、50 cm 下对含水印图像进行拍摄。图 9 显示了在不同距离下的屏摄结果和对应的矫正图像示例，实验结果见表 2。通过表 2 可以看出，本文方法的水印信息提取准确率优于 StegaStamp、CIN、UDH 和 PIMoG 模型。

b. 不同的拍摄角度：实验选择了 60°、45°和 30°作为拍摄位置与屏幕所在位置的夹角，正号(+)和负号(-)分别代表左向拍摄和右向拍摄，垂

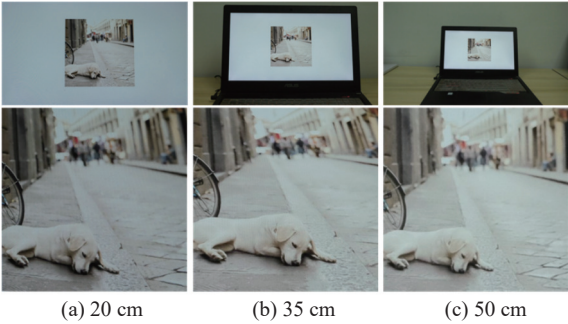


图 9 不同拍摄距离下屏摄图像示例

Fig.9 Examples of screen-shooting images at different shooting distances

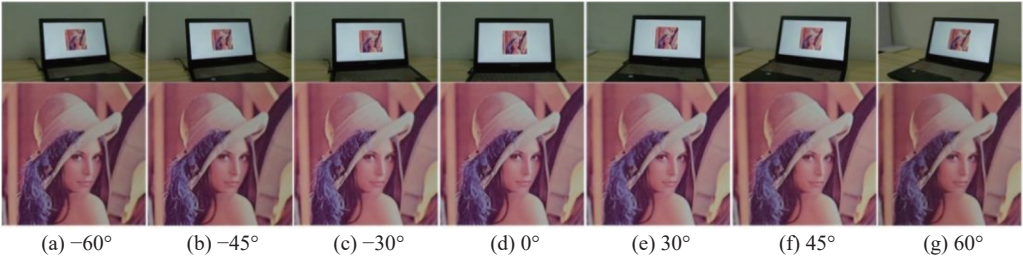


图 10 不同拍摄角度下屏摄图像示例

Fig.10 Examples of screen-shooting images at different shooting angles

表 3 不同拍摄角度下的水印提取准确率
Tab.3 Accuracy of watermark extraction at different shooting angles

角度	提取准确率/%				
	StegaStamp	CIN	PIMoG	UDH	本文方法
-60°	98.68	69.16	97.26	93.86	98.74
-45°	98.96	74.35	97.52	96.12	98.81
-30°	99.14	79.29	97.61	96.58	99.08
0°	99.22	83.56	97.87	97.34	99.34
30°	99.05	77.24	97.89	95.57	99.27
45°	98.31	72.17	97.54	94.29	98.67
60°	98.38	69.83	96.98	91.35	98.76

c. 不同的光照度：在低光照环境下拍摄图像时，为了实现理想曝光，通常需要提高感光度，但高感光度会显著增加画面噪点。本研究分别在光照强度为 50、150、300 lx 的环境下进行了实验，垂直拍摄距离为 35 cm。图 11 展示了不同光照强度下的屏幕拍摄图像及其对应的矫正图像示例。另外，通过表 4 可以看出，本文方法在不同光照度下的水印提取准确率同样优于 StegaStamp、CIN、UDH 和 PIMoG 方法。

2.5 消融实验

本研究主要对以下 3 种情况进行了消融实验：

表 2 不同拍摄距离下的水印提取准确率
Tab.2 Accuracy of watermark extraction at different shooting distances

距离/cm	提取准确率/%				
	StegaStamp	CIN	PIMoG	UDH	本文方法
20	97.64	76.38	97.64	95.43	98.88
35	98.58	75.72	98.23	94.77	99.32
50	98.43	70.14	97.81	94.91	98.53

直拍摄距离设置为 35 cm，实验结果见图 10 和表 3。根据表 3 可知，在不同的拍摄角度下，本文方法的水印提取准确率依然有着较为明显的优势。

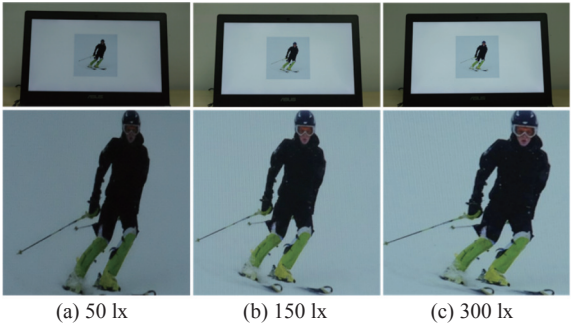


图 11 不同光照强度下屏摄图像示例

Fig.11 Examples of screen-shooting images at different light intensities

a. 对预处理模块 PM 的重要性讨论。本文方法的预处理模块 PM 主要用于提升含水印图像的质量。因此，为进一步验证预处理模块 PM 的重要性，对其进行了消融实验，实验的参数设置与上文保持一致，实验结果见表 5。通过表 5 可以看出，在未添加 PM 模块的情况下，PSNR 值和 SSIM 值分别为 33.96 dB 和 0.9685，而添加 PM 之后，PSNR 值和 SSIM 值均有明显提升。

b. 对 PM 中有/无原始图像 I_0 参与的讨论。对该问题进行两种情况的讨论：(a) 预处理模块 PM 的输入仅与水印信息 M 有关，原始图像 I_0 不参与此模块的输入；(b) 水印信息 M 和原始图像 I_0 均作为预处理模块 PM 的输入。通过对上述两种

表 4 不同光照度下的水印提取准确率
Tab.4 Accuracy of watermark extraction at different light intensities

光照度/lx	提取准确率/%				
	StegaStamp	CIN	PIMoG	UDH	本文方法
50	98.12	65.31	98.24	93.86	98.36
150	96.56	73.26	98.91	95.39	99.17
300	98.31	78.24	97.53	97.41	99.24

表 5 添加或者不添加 PM 情况下的含水印图像质量
Tab.5 Watermarked image quality with or without PM

评价指标	添加PM	未添加PM
PSNR/dB	35.07	33.96
SSIM	0.973 6	0.965 8

情况生成的含水印图像 I_e 进行分析, 实验结果见表 6。通过表 6 可以看出, 有原始图像 I_0 参与 PM 训练的情况下, PSNR 值和 SSIM 值均优于原始图像 I_0 不参与 PM 训练情况下的相应数值。

表 6 原始图像 I_0 是否参与 PM 训练的含水印图像质量
Tab.6 Watermarked image quality with or without I_0

评价指标	I_0 参与PM训练	I_0 不参与PM训练
PSNR/dB	35.07	34.43
SSIM	0.973 6	0.968 2

c. 对函数 $\phi(\cdot)$ 、 $\rho(\cdot)$ 和 $\eta(\cdot)$ 不同结构的讨论。实验针对不同结构的 $\phi(\cdot)$ 、 $\rho(\cdot)$ 和 $\eta(\cdot)$ 进行了比较, 候选结构分别为 RDB (residual dense block)^[28] 模块与文中使用的 ST (swin transformer) 模块, 实验结果见表 7。通过该表可以看出, 在使用 ST 结构的情况下, 含水印图像的 PSNR 值和 SSIM 值均优于使用 RDB 结构的相应数值。

表 7 使用 ST 或使用 RDB 的含水印图像质量
Tab.7 Watermarked image quality with ST or with RDB

评价指标	使用ST	使用RDB
PSNR /dB	35.07	34.28
SSIM	0.973 6	0.955 8

3 结 论

本文提出了一种基于可逆神经网络的小波域抗屏摄噪声攻击的水印方法。为了提高含水印图像的质量, 提出了一种基于 U-Net 网络结构的预处理网络。此外, 可逆神经网络用于预处理后的

水印信息的嵌入和提取。本文还在噪声池中加入真实的屏摄攻击, 使得模型对现实世界中不可微的噪声也具有较的鲁棒性。相较于抗屏摄数字水印方法 StegaStamp, 本文方法在含水印图像质量上有极大的提升。相较于基于可逆神经网络的 CIN 数字水印方法和基于深度学习的 UDH 和 HiDDen 鲁棒水印方法, 本文方法提升了抵抗多种噪声攻击的能力。

参考文献:

[1] 易开祥, 石教英, 孙鑫. 数字水印技术研究进展 [J]. 中国图象图形学报, 2001, 6(A)(2): 111–117.

[2] 徐军军, 毛倩, 董德存. 一种基于人类视觉系统的彩色图像水印算法 [J]. 上海理工大学学报, 2010, 32(5): 475–478,492.

[3] 付笛, 孔平, 周亮, 等. 一种密文域医学图像可逆信息隐藏算法 [J]. 上海理工大学学报, 2022, 44(3): 262–268.

[4] LIU G, XIANG R T, LIU J, et al. An invisible and robust watermarking scheme using convolutional neural networks[J]. [Expert Systems with Applications](#), 2022, 210: 118529.

[5] CU V L, NGUYEN T, BURIE J C, et al. A robust watermarking approach for security issue of binary documents using fully convolutional networks[J]. [International Journal on Document Analysis and Recognition](#), 2020, 23(3): 219–239.

[6] LI F, WAN C, HUANG F J. Adaptive robust watermarking method based on deep neural networks[C]//Proceedings of the 21st International Workshop on Digital Forensics and Watermarking. Cham: Springer, 2022: 162–173.

[7] 王馨雅, 华光, 江昊, 等. 深度学习模型的版权保护研究综述 [J]. 网络与信息安全学报, 2022, 8(2): 1–14.

[8] NAKAMURA T, KATAYAMA A, YAMAMURO M, et al. Fast watermark detection scheme for camera-equipped cellular phone[C]//Proceedings of the 3rd International Conference on Mobile and Ubiquitous Multimedia. College Park Maryland: ACM, 2004: 101–108.

[9] KIM W G, LEE S H, SEO Y S. Image fingerprinting scheme for print-and-capture model[C]//Proceedings of the 7th Pacific Rim Conference on Multimedia on Advances in Multimedia Information Processing. Hangzhou: Springer, 2006: 106–113.

[10] PRAMILA A, KESKINARKAUS A, SEPPÄNEN T. Toward an interactive poster using digital watermarking and a mobile phone camera[J]. [Signal, Image and Video Processing](#), 2012, 6(2): 211–222.

[11] YAMADA T, KAMITANI M. A method for detecting

- watermarks in print using smart phone: finding no mark[C]//Proceedings of the 5th Workshop on Mobile Video. Oslo: ACM, 2013: 49–54.
- [12] FANG H, ZHANG W M, ZHOU H, et al. Screen-shooting resilient watermarking[J]. *IEEE Transactions on Information Forensics and Security*, 2019, 14(6): 1403–1418.
- [13] ZHU J R, KAPIAN R, JOHNSON J, et al. HiDDeN: hiding data with deep networks[C]//Proceedings of the 15th European Conference on Computer Vision. Munich: Springer, 2018: 682–697.
- [14] LIU Y, GUO M X, ZHANG J, et al. A novel two-stage separable deep learning framework for practical blind watermarking[C]//Proceedings of the 27th ACM International Conference on Multimedia. Nice: ACM, 2019: 1509–1517.
- [15] TANCIK M, MILDENHALL B, NG R. StegaStamp: invisible hyperlinks in physical photographs[C]//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020: 2114–2123.
- [16] GOODFELLOW I J, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial nets[C]//Proceedings of the 28th International Conference on Neural Information Processing Systems. Montreal: MIT Press, 2014: 2672–2680.
- [17] BAI R, LI L, ZHANG S Q, et al. SSDeN: framework for screen-shooting resilient watermarking via deep networks in the frequency domain[J]. *Applied Sciences*, 2022, 12(19): 9780.
- [18] JIA J, GAO Z P, CHEN K, et al. RIHOOP: robust invisible hyperlinks in offline and online photographs[J]. *IEEE Transactions on Cybernetics*, 2022, 52(7): 7094–7106.
- [19] FANG H, JIA Z Y, MA Z H, et al. PIMoG: an effective screen-shooting noise-layer simulation for deep-learning-based watermarking network[C]//Proceedings of the 30th ACM International Conference on Multimedia. Lisboa: ACM, 2022: 2267–2275.
- [20] RONNEBERGER O, FISCHER P, BROX T. U-Net: convolutional networks for biomedical image segmentation[C]//Proceedings of the 18th International Conference on Medical Image Computing and Computer-Assisted Intervention. Munich: Springer, 2015: 234–241.
- [21] DINH L, KRUEGER D, BENGIO Y. NICE: non-linear independent components estimation[DB/OL]. [2014-10-30]. <https://arxiv.org/abs/1410.8516>.
- [22] DINH L, SOHL-DICKSTEIN J, BENGIO S. Density estimation using real NVP[C]//Proceedings of the 5th International Conference on Learning Representations. Toulon: ICLR, 2017.
- [23] LIU Z, LIN Y T, CAO Y, et al. Swin transformer: hierarchical vision transformer using shifted windows[C]//Proceedings of IEEE/CVF International Conference on Computer Vision. Montreal: IEEE, 2021: 9992–10002.
- [24] SHIN R, SONG D. Jpeg-resistant adversarial images[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017: 8.
- [25] WENGROWSKI E, DANA K. Light field messaging with deep photographic steganography[C]//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019: 1515–1524.
- [26] CHEN L C, ZHU Y K, PAPANDREOU G, et al. Encoder-decoder with atrous separable convolution for semantic image segmentation[C]//Proceedings of the 15th European Conference on Computer Vision. Munich: Springer, 2018: 833–851.
- [27] LIN T Y, MAIRE M, BELONGIE S, et al. Microsoft COCO: common objects in context[C]//Proceedings of the 13th European Conference on Computer Vision. Zurich, Switzerland: Springer, 2014: 740–755.
- [28] DENG J, DONG W, SOCHER R, et al. ImageNet: a large-scale hierarchical image database[C]//Proceedings of 2009 IEEE Conference on Computer Vision and Pattern Recognition. Miami: IEEE, 2009: 248–255.
- [29] MA R, GUO M X, HOU Y, et al. Towards blind watermarking: combining invertible and non-invertible mechanisms[C]//Proceedings of the 30th ACM International Conference on Multimedia. Lisboa: ACM, 2022: 1532–1542.
- [30] ZHANG C N, BENZ P, KARJAUV A, et al. UDH: universal deep hiding for steganography, watermarking, and light field messaging[C]//Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2020: 857.

(编辑: 丁红艺)