计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
2025,34(3):201−209 [doi: 10.15888/j.cnki.csa.009789] [CSTR: 32024.14.csa.009789]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于可逆机制的端到端单阶段数字水印算法①

胡子寒, 林立霞, 白少杰, 曹　鹏

(北京印刷学院 信息工程学院, 北京 102627)
通信作者: 林立霞, E-mail: linlixia@bigc.edu.cn

摘　要: 数字水印算法因其在版权保护、内容认证、数据隐藏等领域的重要应用价值而受到广泛关注. 在实际应用中, 嵌入水印的图像往往会遭受图像扭曲、锐化模糊等可微噪声的影响, 同时也会面临 JPEG 压缩、传输错误等不可微噪声的干扰. 现有研究多集中于单一噪声环境下的方案设计, 或者尝试使用可导模型来近似模拟不可微噪声, 这些方法在一定程度上限制了水印算法的鲁棒性. 针对这一问题, 本文提出了一种基于可逆神经网络的端到端单阶段数字水印方案. 该方案利用可逆神经网络模拟不可微噪声, 提高了算法对于实际噪声环境的适应性和鲁棒性. 与现有算法相比, 本算法在多噪声叠加情况下峰值信噪比 (*PSNR*) 提高了 3.12 dB, 平均提取精度 (ACC) 提高了 35.36%.

关键词: 数字水印; 可逆神经网络; 端到端; 不可微噪声; 抗叠加噪声

引用格式: 胡子寒,林立霞,白少杰,曹鹏.基于可逆机制的端到端单阶段数字水印算法.计算机系统应用,2025,34(3):201–209. http://www.c-s-a.org.cn/1003-3254/9789.html

End-to-end One-stage Digital Watermarking Algorithm Based on Invertible Mechanisms

HU Zi-Han, LIN Li-Xia, BAI Shao-Jie, CAO Peng

(College of Information Engineering, Beijing Institute of Graphic Communication, Beijing 102627, China)

Abstract: Digital watermarking algorithms attract widespread attention due to their important application value in the fields of copyright protection, content authentication, and data hiding. In practical applications, images with embedded watermarks are often affected by differentiable noises such as image distortion and sharpening blurring. At the same time, they also face interference from non-differentiable noises such as JPEG compression and transmission errors. Existing studies mostly focus on scheme design in a single noise environment, or attempt to use differentiable models to approximately simulate non-differentiable noises. These methods limit the robustness of watermarking algorithms to a certain extent. To solve this problem, this study proposes an end-to-end one-stage digital watermarking scheme based on an invertible neural network. The scheme uses an invertible neural network to simulate non-differentiable noise, enhancing the algorithm's adaptability and robustness to actual noisy environments. Compared with existing algorithms, this algorithm improves the peak signal-to-noise ratio (*PSNR*) by 3.12 dB and the average extraction accuracy (ACC) by 35.36% in the case of multiple noise superposition.

Key words: digital watermarking; invertible neural network; end-to-end; non-differentiable noise; anti-stacked noise resistance

数字水印最早由 van Schyndel 于 1994 年提出[1], 随后被广泛应用于图像、视频和音频的知识产权保护[2]. 具体来说, 图像数字水印的目的是以一种不可见的方式嵌入秘密信息, 即使图像被破坏干扰, 也能从编码后

① 基金项目: 北京市教委科技一般项目 (KM202410015001); 北京印刷学院校级项目 (Ea202302, 27170123033, Ea202301)
收稿时间: 2024-08-09; 修改时间: 2024-09-19; 采用时间: 2024-09-30; csa 在线出版时间: 2025-01-16
CNKI 网络首发时间: 2025-01-17

的图像中提取秘密信息. 通过这种方式, 数字水印技术为数字内容的版权保护提供一种高效且可靠的解决方案, 确保创作者和版权所有者的权益能够得到有效维护.

衡量数字水印算法性能的两个关键因素是鲁棒性和不可感知性. 鲁棒性是指嵌入在图像中的水印信息能够在各种失真下生存. 不可感知性则指嵌入水印后的图像可视效果应与原始图像一致. 最早的研究将水印信息编码在图像像素的最低有效位 (least significant bit, LSB) 上, 但此方法极易被攻击且鲁棒性差. 基于变换域的水印技术相继被提出, 在 DCT 域[3]和 DWT 域[4]对信息进行编码实现了更强的鲁棒性. 然而, 目前随着信息技术的迅速发展, 传统的安全措施已难以满足日益增长的版权保护需求, 例如, 手机拍摄会产生色彩失真、光源失真等“屏摄噪声”, 以及基于 AI 修图等应用日益广泛, 使得对多媒体内容的篡改和伪造变得更加精细和难以察觉, 这均对图像水印的鲁棒性和不可感知性提出了更高的要求. 为解决上述问题, 基于深度学习的数字水印算法得到广泛应用[5–7], 取得了比传统技术更优越的隐蔽性和鲁棒性.

在实际场景中, 嵌入水印的图像不仅容易受到图像扭曲、锐化、模糊等可微噪声的干扰, 还会遭受 JPEG 压缩、传输错误等不可微噪声的冲击. 现有基于深度学习的方法在解决各种情况下噪声攻击的问题上取得了一定进展. 其中, 一类方法[8–19]通过可导模型来近似模拟各种类型的噪声攻击, 另一类方法[20,21]通过引入额外解码器或是分阶段对模型先进行预训练. 但是, 现有研究存在两个关键问题, 一是, 这些方法往往难以全面应对多变的实际应用场景; 二是大多数现有研究都仅仅关注算法对噪声的鲁棒性, 将水印信息嵌入固定区域, 无法智能地根据实际场景噪声攻击类型选择最佳的嵌入域. 上述方法在一定程度上限制了水印算法的鲁棒性.

为解决上述问题, 本文提出一种基于可逆机制的端到端单阶段数字水印算法. 所提出算法首先对原始图片进行离散小波变换从而将图片分为 4 个子频带, 使其能够根据预期的噪声环境和特性来选择最合适的嵌入策略. 此外, 利用可逆神经网络模拟 JPEG 压缩等不可微噪声, 进而提高算法在实际噪声环境下的适应性和鲁棒性. 可逆性保证了水印的准确提取, 即使在不利条件下也能保持水印的完整性. 端到端的设计意味着模型从输入的原始图像直接生成嵌入水印的图像, 省略了传统流程中的多个分离步骤, 使水印编码器和解码器实现单阶段训练, 降低了训练复杂度, 提升了处理效率. 此外, 训练模型直接从数据中学习, 增强了数字水印的泛化性和鲁棒性.

本文的主要工作总结如下.

1) 针对现有算法仅考虑单一种类噪声及利用可微模型模拟不可微噪声的局限性, 提出了一种基于可逆机制的端到端单阶段训练的水印框架, 利用可逆神经网络模拟不可微噪声, 提高了算法对于实际场景的复杂噪声环境的适应性和鲁棒性.

2) 结合传统数字水印算法的优势, 设计了基于离散小波变换的水印嵌入策略, 使得所设计算法能够根据噪声环境特性来选择最合适的嵌入域.

3) 大量的实验表明, 与现有基于深度神经网络的水印模型相比, 该方案在实际噪声情况下的不可感知性和鲁棒性均具有一定程度的提升.

1 相关工作

1.1 解决可微噪声攻击的数字水印算法

为应对各种不同类型噪声攻击的问题, 常见的方法是利用噪声层模拟各种攻击类型并通过“编码器-噪声层-解码器”的端到端单阶段训练使水印能够在噪声中存活. Ahmadi 等[8]提出了一个深度端到端扩散水印框 ReDMark, 利用全卷积神经网络和残差结构实现水印的嵌入与提取, 并通过模拟攻击的可微网络层进行端到端训练, 同时在图像宽域内扩散水印, 以提高安全性和鲁棒性. Fang 等[9]提出了一种创新的解码器驱动的水印网络 De-END, 解决了传统编码器驱动端到端架构中编码器可能嵌入不必要特征的问题. Fang 等[10]针对屏幕截图对水印带来的挑战, 提出了 PIMoG 噪声层设计, 通过模拟屏摄过程中的主要失真来生成噪声层, 简化了屏幕截图过程的模拟. Cao 等[11]提出一个计算残差水印消息并使用轻量级神经网络进行编码, 实现了水印的隐藏和提取. 同时, 设计了噪声层来模拟屏幕截图的光学和辐射效果, 平衡了水印的隐蔽性和鲁棒性. Tang 等[12]提出了一种深度学习医疗图像盲水印算法, 通过 BCH 编码加密、模拟屏幕截图生成训练数据、空间域嵌入加密水印, 并优化损失函数训练神经网络, 以实现高隐蔽性和鲁棒性. Cao 等[13]提出了一种适用于屏幕截图的通用图像水印算法. 通过在 DCT 域中集成通道注意力机制增强隐蔽性, 使用噪声层模拟

屏幕截图失真, 并通过训练模型实现水印的强鲁棒性, 从而能够为各种图像生成抵抗多种失真的水印.

1.2 解决混合噪声攻击的数字水印算法

上述方法[8–13]大多采用基于“编码器-噪声层-解码器”的架构, 其中嵌入和提取过程由编码器和解码器分别完成. 然而, 这种框架的一个潜在缺点是, 编码器和解码器不能很好地耦合, 这导致编码器可能在载体图像中嵌入一些冗余特征, 从而影响整个算法的不可见性和鲁棒性. 可逆神经网络 (invertible neural network, INN) 是首个用于复杂高维密度建模的基于学习的归一化流框架, 由 Dinh 等[14]于 2014 年提出. 基于 INN 的可逆性, 可以有效地实现相同网络、相同参数的前向和后向映射, 可以很好地满足归一化映射. Guan 等[15]提出了一个基于可逆神经网络的新型多图像隐藏框架, 通过前向传播和反向传播过程模拟隐藏和揭示过程, 实现两者的紧密耦合和可逆性. Luo 等[16]提出的 IRWArt 算法利用 INN 实现水印的嵌入和提取, 将水印嵌入到对图像质量影响最小的域, 使水印具有较强的抗抄袭能力. Fang 等[17]提出了一种基于流的水印框架, 通过参数共享策略实现了编码器和解码器的紧密协作, 并设计了一个可逆噪声层来模拟黑盒失真作为训练阶段的噪声层, 通过预训练好的噪声层来保证黑盒失真的鲁棒性. Zhu 等[18]通过联合训练编码器和解码器网络, 该方法利用可微近似克服 JPEG 非可微问题. Jia 等[19]结合 mini-batch real and simulated JPEG (MBRS) 策略以随机选择真实 JPEG、模拟 JPEG 或无噪声层作为训练中的噪声层, 有效提升模型在面对 JPEG 压缩时的表现. Fang 等[20]提出一个 3 阶段水印框架, 首先通过无噪声训练和 JND-mask 图像损失提高水印质量, 然后使用频域增强算法优化编码特征以抵抗不可微扭曲, 最后通过对抗性训练增强解码器的提取能力. Ma 等[21]提出了一个结合可逆和不可逆 (CIN) 机制的框架, 引入额外的解码器来提高解码器抵抗不可微噪声的性能.

1.3 小结

深度学习技术在数字水印领域的应用显著提升了隐蔽性和鲁棒性, 但现有研究[8–21]揭示了一些问题. 这些模型通常依赖于使用可导模型来近似模拟噪声, 这在处理不可微噪声时可能存在局限. 尽管端到端训练框架通过模拟可微噪声优化了水印的稳定性和安全性, 但在模拟和抵抗实际复杂噪声环境下的噪声叠加方面, 其效果还有待提高. 此外, 现有研究在解决 JPEG 压缩的方法上, 还尝试同时引入额外解码器或分阶段对模型进行预训练, 这些方法虽然在一定程度上增强了对特定噪声的抵抗能力, 但也使模型结构变得复杂且繁琐, 限制了水印算法的鲁棒性. 因此, 如何设计出既高性能又简洁的数字水印算法, 以满足实际应用场景的需求, 仍然是该研究领域中一个亟待解决的关键问题.

2 本文模型

基于可逆机制的端到端单阶段数字水印模型的整体架构如图 1 所示, 共分为以下 6 个部分: 图像预处理模块 (image preprocessing module, IPM)、水印扩散模块 (watermark diffusion module, WDM)、水印嵌入模块 (watermark embedding module, WEM)、融合模块 (Fusion Module, FM)、不可微噪声处理模块 (non-differentiable noise processing module, NNPM) 和可微噪声模块 (differentiable noise module, DNM).

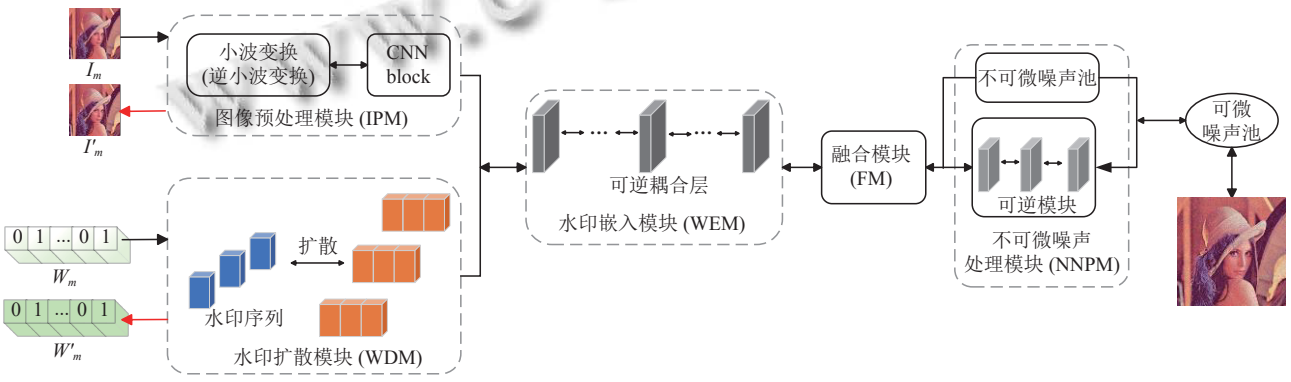

图 1 模型整体架构图

2.1 图像预处理模块

如图 1 所示, IPM 模块将载体图片 I_m 进行 Haar 小波变换以获取多尺度频率, 得到 4 个不同的子频带: 高频水平 (horizontal high, HH)、高频垂直 (horizontal high

and vertical low, HL)、高频对角线 (vertical high and horizontal low, LH) 和低频 (low low, LL). 然后利用卷积操作对图像进行下采样以减少图像的空间分辨率便于后续处理, 最后通过上采样恢复图像的细节和分辨率以便特征提取.

2.2 水印扩散模块

水印 W_m 是一个长度为 L 的二进制序列, 并将此水印嵌入到一张高度为 H、宽度为 W (H、W 默认为 256) 的载体图像 I_m 中. 将输入模型中的水印和图像分别表示为 $W_m \in R^{B\times L}$ 和 $I_m \in R^{B\times C\times H\times W}$, 其中, B 和 C 分别表示图片批次大小和通道数.

图 2 展示了水印扩散和提取过程. 在水印扩散过程中, 为了使水印与图像的通道数对齐, 首先复制 3 份水印 W_m. 不同的全连接 (fully connected, FC) 层分支会产生长度更长的冗余水印. 随后, 通过二维转置卷积进行重塑和上采样, 使其尺寸与覆盖图像相同. 经过 FC 层后, 水印的长度变为 $L=256$, 卷积的核大小和步长均为 2, 并设置 3 个这样的处理块. 最后, 拼接 3 个分支的输出, 并在经过 Haar 变换后送入水印嵌入模块. 水印前向嵌入操作定义为:

$$\Psi_{\mathrm{WDM}} = \Gamma_{\mathrm{haar}}(O_{\mathrm{cat}}(\Gamma_{\mathrm{convT}}(\Gamma_{\mathrm{fc}}(O_{\mathrm{copy}}(W_m))))) \tag{1}$$

其中, $O_{\mathrm{copy}} \in 3\times R^{B\times L}$ 代表复制、$\Gamma_{\mathrm{fc}} \in R^{B\times \hat{L}}$ 代表全连接、$\Gamma_{\mathrm{convT}} \in R^{B\times 1\times H\times W}$ 代表转置卷积、$O_{\mathrm{cat}} \in R^{B\times 3\times H\times W}$ 代表拼接和、$\Gamma_{\mathrm{haar}} \in R^{B\times 12\times H/2\times W/2}$ 代表 Haar 变换操作、Ψ_{WDM} 表示水印扩散模块的输出张量.

在水印提取过程中, 执行嵌入过程的逆操作 Ψ_{WDM}^{-1}. 与嵌入水印步骤中的复制操作不同, 最终结果通过平均池化输出. 其表达式为:

$$W'_m = \Psi_{\mathrm{WDM}}^{-1}(\cdot) \tag{2}$$

图 2 水印扩散模块 (WDM)

2.3 水印嵌入模块

可逆神经网络结构如图 3 所示, 水印的嵌入和提取分别对应于双射结构的正向和逆向过程. 在第 i 层耦合层中, u_{wm}^{l} 和 u_{im}^{l} 分别表示输入的水印和图像信息. 相应的 u_{wm}^{l+1} 和 u_{im}^{l+1} 表示通过当前耦合层处理后的输出水印和图像信息. WEM 模块可表述为:

$$u_{\mathrm{im}}^{l+1} = \varphi\left(u_{\mathrm{wm}}^{l}\right) + u_{\mathrm{im}}^{l} \tag{3}$$

$$u_{\mathrm{wm}}^{l+1} = u_{\mathrm{wm}}^{l} \odot \exp\left(\rho\left(u_{\mathrm{im}}^{l+1}\right)\right) + \eta\left(u_{\mathrm{im}}^{l+1}\right) \tag{4}$$

其中, $\exp(\cdot)$ 为指数运算符, $\rho(\cdot)$ 和 $\eta(\cdot)$ 为任意函数, $\odot$ 表示 Hadamard 乘积. 提取过程的相应反向传播公式为:

$$u_{\mathrm{im}}^{l} = u_{\mathrm{im}}^{l+1} - \varphi\left(u_{\mathrm{wm}}^{l}\right) \tag{5}$$

$$u_{\mathrm{wm}}^{l} = \left(u_{\mathrm{wm}}^{l+1} - \eta\left(u_{\mathrm{im}}^{l+1}\right)\right) \odot \exp\left(-\rho\left(u_{\mathrm{im}}^{l+1}\right)\right) \tag{6}$$

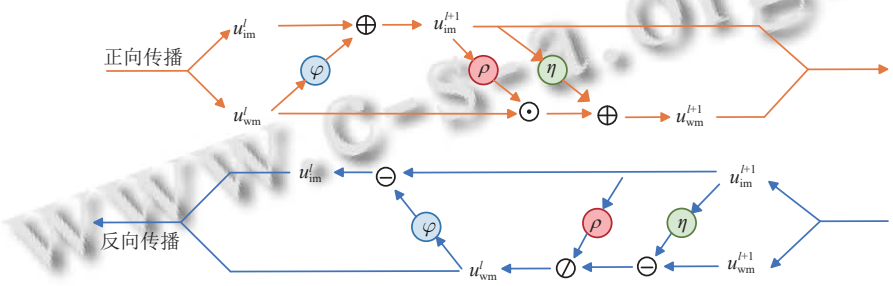

图 3 可逆神经网络结构图

图 4 展示了水印嵌入过程, 将 WEM 模块的输出标记为 $\Psi_{\mathrm{WEM}} \in R^{B\times 24\times \frac{H}{2}\times \frac{w}{y}}$, 其由 $\Psi_{\mathrm{WEM}}^{\mathrm{I}}(x;LR,HR)$ 和 $\Psi_{\mathrm{WEM}}^{\mathrm{II}}(x;LR\_,HR\_) \in R^{B\times 12\times \frac{H}{2}\times \frac{w}{y}}$ 两部分输出信息构成, 其中, LR 和 HR 分别代表 I_m 在 IPM 模块中经过 Haar 小波变换后的低频和高频信息.

2.4 融合模块

图 5 展示了 WEM 模块两部分输出融合的过程, 首先将 WEM 模块输出 $\Psi_{\mathrm{WEM}}^{\mathrm{I}}$ 和 $\Psi_{\mathrm{WEM}}^{\mathrm{II}}$ 的对应通道信息进行平均, 以便融合并挤压至图像尺寸. 然而, 要在水印鲁棒性和不可见性之间取得平衡较为困难. 因此, 在嵌入过程中, 本文丢弃 WEM 输出的图像部分, 仅保留映射后的水印部分, 然后按强度因子 S 缩放后加到图像上以得到最终的含水印图像.

FM 模块的计算过程如下:

$$WI_m = \Gamma_{\text{haar}}^{-1}\begin{pmatrix} \Psi_{\text{WEM}}^{\text{II}}(x; LR\_, HR\_)\times S \\ +\Gamma_{\text{haar}}(I_m(x; LR, HR)) \end{pmatrix} \tag{7}$$

其中, S 表示水印的强度, $\Gamma_{\text{haar}}^{-1}(\cdot)$表示逆 Haar 变换. 为了通过可逆分支恢复嵌入的水印, 输入为:

$$\hat{\Psi}_{\text{WEM}} = O_{\text{cat}}(O_{\text{copy}}(\Gamma_{\text{haar}}^{-1}(WI_m))) \tag{8}$$

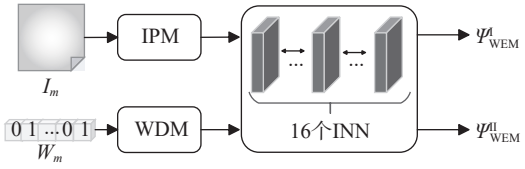

图 4 水印嵌入模块 (WEM)

2.5 不可微噪声处理模块

(1) JPEG 压缩噪声

联合图像专家组 (joint photographic experts group, JPEG) 压缩[22]是一种常用的图像压缩标准, 通过减少图像数据量来实现压缩. 图 6 展示了 JPEG 压缩的整体流程, JPEG 压缩通过 DCT、量化和熵编码来减小文件大小, 高压缩率可能损害图像细节并引入视觉失真, 量化步骤是一个非线性且不可逆的过程[23,24]. 现有研究尝试用可导模型模拟该噪声时会丢失关键图像特征, 尤其是高频区域信息, 进而影响数字水印的鲁棒性. 本文针对这类不可微噪声展开研究.

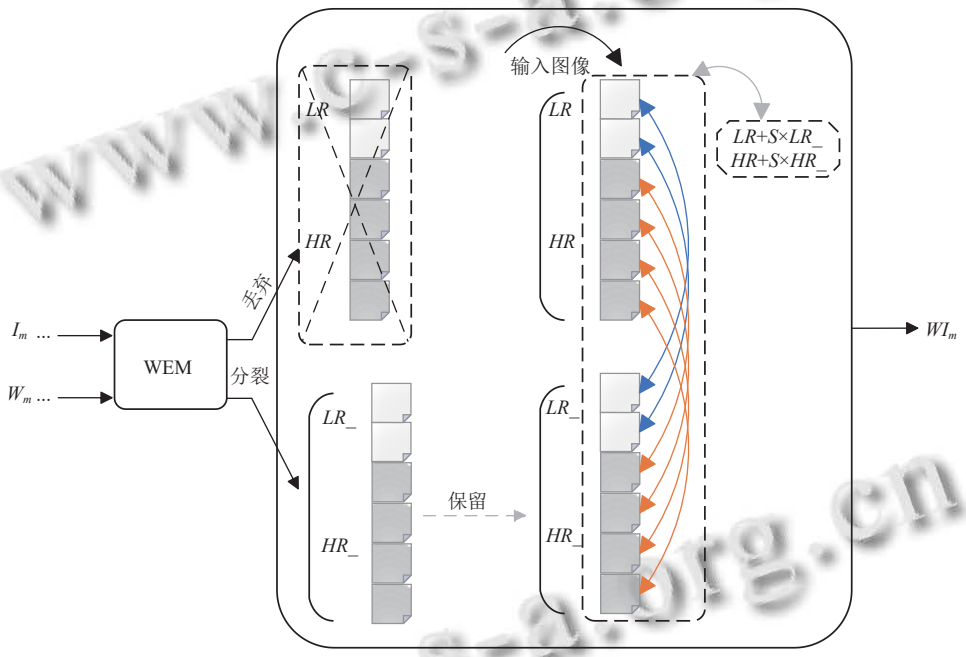

图 5 融合模块 (FM)

(2) 不可微噪声模拟模块

为了充分表示 JPEG 压缩等不可微噪声的特性, 引入可逆神经网络对其进行表征, 通过将可逆块与不可微噪声模块并联, 即使噪声图片包含不可微的部分, 也可以利用该可逆块进行逆向传播, 从而提取水印. 如图 7 所示, 可逆块包含 3 个子网络 F、G 和 H, 分别用于执行不同的变换. F 和 G 负责正向和反向传播中的变换, 而 H 用于计算可逆操作的雅可比行列式.

具体执行过程为, 将 FM 模块的输出 WI_m 标记为 x, 作为可逆块的输入; x 进而被分割为两部分 x_1 和 x_2, x_1 包含前 c_1 个通道, x_2 包含剩余的 c_2 个通道, $c_1 + c_2 = c$, c 是输入的总通道数. 正向传播中通过子网络 F 将 x_2 变换为:

$$y_1 = x_1 + F(x_2) \tag{9}$$

通过子网络 H 对 y_1 进行变换, 得到缩放因子 s, σ 为 *Sigmoid* 激活函数:

$$s = \sigma(H(x_1))\times 2 - 1 \tag{10}$$

其中, 缩放因子 s 被用于 x_2 的指数变换:

$$y_2 = x_2 \cdot \exp(s) + G(y_1) \tag{11}$$

FM 最终输出 y 为 y_1 和 y_2 的拼接:

$$y = concat(y_1, y_2) \tag{12}$$

在反向传播中, 需要逆操作来恢复 x_1 和 x_2, 首先通

过 G 的逆操作计算y_1的逆:

$$x_1 = y_1 - F(y_2) \tag{13}$$

然后使用 exp(s) 的逆操作来恢复x_2:

$$x_2 = (y_2 - G(x_1))/\exp(s) \tag{14}$$

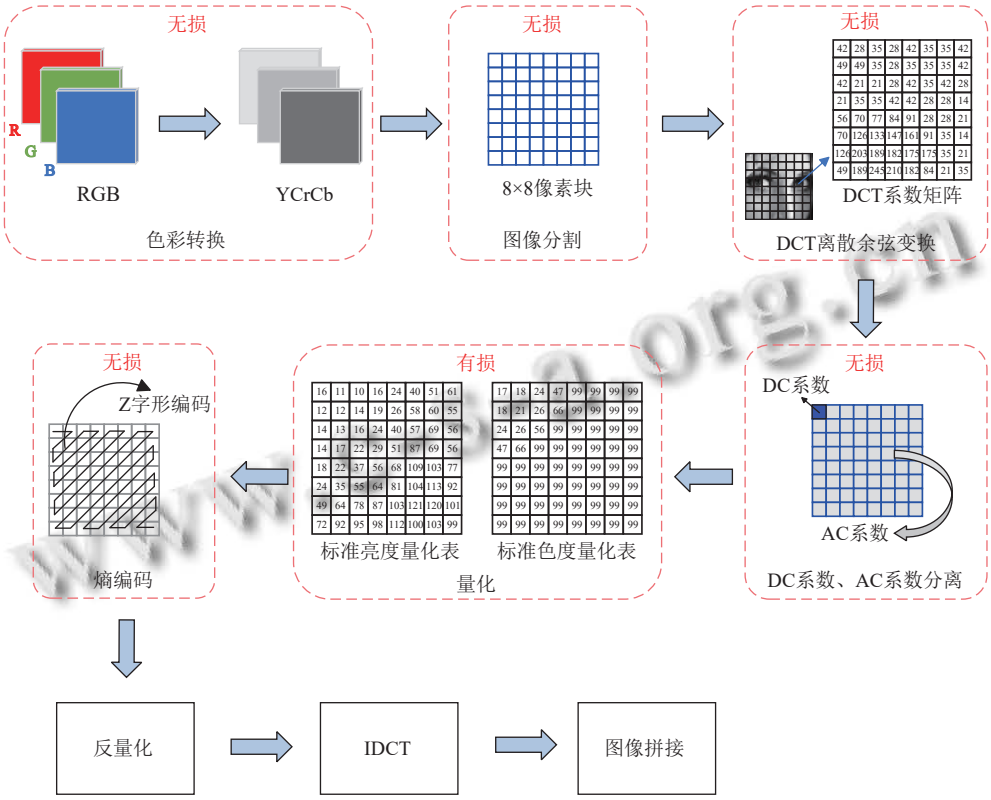

图 6 JPEG 压缩流程图

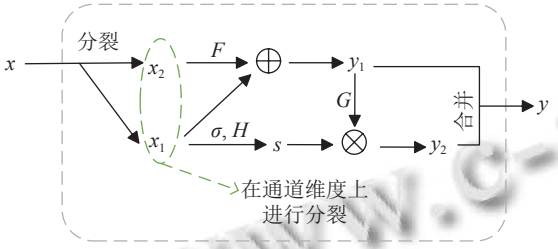

图 7 可逆块

2.6 可微噪声模块

此模块引入了 5 种常见的可微噪声, 包括高斯模糊、椒盐噪声、高斯噪声、裁剪和随机失活.

为测试模型对多个噪声同时叠加攻击的鲁棒性, 实验将选择在实际应用中较为常见的“高斯模糊”和“椒盐噪声”与不可微噪声进行叠加训练和测试[17,25,26].

2.7 损失函数

本文采用L_2来引导水印图像WI_m在不可见性上与载体图像I_m的相似度:

$$L_I = \|I_m - WI_m\|^2 \tag{15}$$

采用L_2来引导水印W_m与提取水印RW_m的相似度:

$$L_W = \|W_m - RW_m\|^2 \tag{16}$$

在使用恒等噪声层训练时, 采用L_2距离约束恢复图像RI_m与参考图像I_m之间的差值:

$$L_R = \|I_m - RI_m\|^2 \tag{17}$$

3 实验结果与分析

3.1 数据集

为了验证所提出方法的鲁棒性和不可感知性, 实验利用 COCO、DIV2K、ImageNet 数据集进行训练和评估. 随机选择 10 000 张图像进行训练, 并随机选择 5 000 张图像进行评估. 对于每个输入图像, 均嵌入相应水印信息, 该信息从二进制分布 W_m~$\{0,1\}^L$中随机采样.

3.2 评价指标

为客观地评估所提出水印框架的鲁棒性和不可感

知性, 将采用一系列客观的量化评估标准. 为了验证鲁棒性, 本文对照原始水印信息W_m, 评估了提取水印信息W'_m的精度. 对于每个输入图像$I_m(x_i)$, 其嵌入和恢复的水印分别为$W'_m(x_i)$和$W_m(x_i)$. ACC 根据误码率 (*BER*) 计算得到, 具体为 (1−*BER*).

$$BER=\left(\frac{1}{L}\times\sum_{k=1}^{L}\left|W'_m(x_i)-W_m(x_i)\right|\right)\times 100\% \tag{18}$$

对于水印图像的不可感知性, 本文采用峰值信噪比 (*PSNR*) 进行评价.

$$PSNR(I_m(x_i),I'_m(x_i)) = 20\times\log_{10}\frac{MAX(I_m(x_i),I'_m(x_i))-1}{MSE(I_m(x_i),I'_m(x_i))} \tag{19}$$

其中, $MAX(\cdot)$为图像的最大像素值, $MSE(\cdot)$为均方误差.

3.3 实验结果

为了保持公平的比较, 实验采用了与参考方法[17,19,21,25,26]完全相同的设置. 所有模型的图像大小都调整为 128×128, 水印长度为 30 或 64. 对于此模型, 使用 RTX3090 显卡进行训练, 批大小设置为 32, 采用默认超参数的 Adam 优化器.

在实验中, 分别在 JPEG 噪声和叠加噪声的情况下对模型进行训练和评估. 在叠加噪声中, 选取可微噪声池中的高斯噪声和椒盐噪声与不可微噪声 JPEG 压缩进行叠加.

表 1 为将本文方法分别与 CIN、FIN、TSDL、ARWGAN 在 JPEG 压缩情况下进行对比的测试结果, 表 2 为将本文方法分别与 CIN、FIN、MBRS 在噪声叠加情况下进行对比的测试结果, 表 3 为本文方法分别在 COCO、DIV2K、ImageNet 数据集上对叠加噪声的训练测试结果. 可以明显看出, 相较于其他方法, 本文方法无论是在只含有不可微噪声 (JPEG 压缩) 以及可微噪声 (高斯噪声、椒盐噪声) 与不可微噪声叠加情况下, 都有较好的鲁棒性和不可感知性. 此外, 本文方法在不同数据集上有较好的实验结果, 表明本文方法还有较好的泛化性.

表 1 不同方法下对 JPEG 压缩的鲁棒性和不可感知性结果

Method	*PSNR* (dB)	ACC (%)
CIN[21]	39.29	94.80
FIN[17]	38.21	93.71
TSDL[25]	33.5	88.8
ARWGAN[26]	35.87	93.98
Ours	**44.81**	**99.02**

表 2 不同方法下对叠加噪声的鲁棒性和不可感知性结果

Method	*PSNR* (dB)	ACC (%)
CIN[21]	36.20	61.43
FIN[17]	37.21	50.98
MBRS[19]	35.56	71.39
Ours	**39.32**	**96.79**

表 3 本文方法在不同数据集下对叠加噪声的鲁棒性和不可感知性结果

Dataset	*PSNR* (dB)	ACC (%)
COCO	**44.81**	99.02
DIV2K	44.76	**99.15**
ImageNet	44.52	98.93

在数字水印的鲁棒性与不可感知性研究中, 噪声攻击是不可避免的挑战. 噪声的类型和特性直接影响水印的稳定性和可靠性. 本文探讨了在 3 种不同噪声环境下, 将水印信息嵌入到图像的不同频带 (HH、HL、LH、LL) 的策略, 并对比分析其性能差异.

首先分析在仅有可微噪声存在的情况下, 表 4 展示了不同频带嵌入策略的鲁棒性. 本文选择在实际场景应用中较为常见的高斯噪声评估在这种环境下的水印如何保持其结构完整性和可提取性. 接着将对比在 JPEG 压缩情况下, 如表 5 所示, 展示了水印嵌入不同频带的性能. 最后考虑更复杂的实际应用场景, 即可微与不可微噪声的叠加情况, 实验结果如表 6 所示. 通过对不同噪声环境下的水印嵌入策略进行深入的分析和实验验证, 将揭示在特定噪声条件下, 哪些频带更有利于水印的隐藏和提取, 以及如何设计出能够综合抵抗各种噪声攻击的水印算法. 旨在为数字水印技术在复杂噪声环境下的应用提供理论依据和实践指导.

表 4 只含有可微噪声下不同频带下水印性能对比

Sub-bands	*PSNR* (dB)	ACC (%)
HH	**52.67**	**99.96**
HL	51.85	99.76
LH	52.48	99.34
LL	50.73	99.21

表 5 只含有不可微噪声下不同频带下水印性能对比

Sub-bands	*PSNR* (dB)	ACC (%)
HH	44.81	99.03
HL	44.35	99.56
LH	**44.96**	99.01
LL	43.73	**99.79**

上述实验表明, 当数据仅受到可微噪声的影响时, 将数字水印嵌入到高频区域 (HH 频带) 表现出了最佳的性能. 这一结果归功于高频区域对小扰动的敏感性,

以及可微噪声的特性, 使得通过优化技术能够有效地最小化其对水印的影响. 然而, 在面对包含不可微噪声的情况, 例如 JPEG 压缩, 实验结果如图 8、图 9 所示, 表明将水印嵌入到中低频区域能够带来更加稳健的效果. 其原因在于, 中低频区域相对压缩等非线性失真具有更好的容忍度, 这使得水印即便在经历常见的图像处理操作后, 仍能保持较高的识别度和完整性.

表 6 噪声叠加下不同频带下水印性能对比

Sub-bands	*PSNR* (dB)	ACC (%)
HH	39.32	96.79
HL	38.61	98.56
LH	**39.45**	97.01
LL	37.73	**98.79**

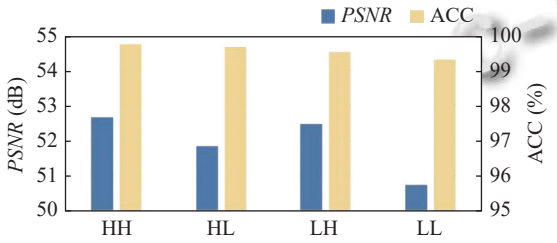

图 8 可微噪声下不同频带性能对比图

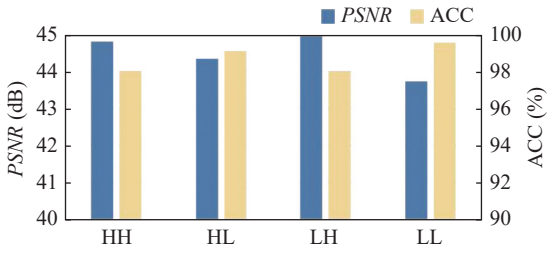

图 9 不可微噪声下不同频带性能对比图

这些发现提供了重要的指导意义: 在设计数字水印算法时, 应根据实际的噪声环境来选择合适的嵌入策略. 对于可微噪声, 倾向于利用图像的高频区域; 而对于不可微噪声, 策略则应转向利用中低频区域的鲁棒性. 这种灵活的适应性对于确保数字水印在多变的实际应用中的有效性至关重要.

4 结语

本文提出了一种基于可逆机制的端到端数字水印算法, 有效地提高了水印对叠加噪声的不可感知性和鲁棒性. 此外, 此模型能够智能地根据不同噪声环境选择最优的频带进行水印嵌入, 无论是在高频带还是低频带, 都能保持水印的稳定性和隐蔽性. 在 COCO、DIV2K、ImageNet 数据集上的大量实验表明, 此方法有效提升了数字水印图像的不可感知性和鲁棒性.

参考文献

1 van Schyndel RG, Tirkel AZ, Osborne CF. A digital watermark. Proceedings of the 1st International Conference on Image Processing. Austin: IEEE, 1994. 86–90.

2 Hsu CT, Wu JL. Hidden digital watermarks in images. IEEE Transactions on Image Processing, 1999, 8(1): 58–68. [doi: 10.1109/83.736686]

3 Guo HP, Georganas ND. Digital image watermarking for joint ownership verification without a trusted dealer. Proceedings of the 2003 International Conference on Multimedia and Expo. Baltimore: IEEE, 2003. II–497.

4 Hamidi M, Haziti ME, Cherifi H, *et al*. Hybrid blind robust image watermarking technique based on DFT-DCT and Arnold transform. Multimedia Tools and Applications, 2018, 77(20): 27181–27214. [doi: 10.1007/s11042-018-5913-9]

5 夏道勋, 王林娜, 宋允飞, 等. 深度神经网络模型数字水印技术研究进展综述. 科学技术与工程, 2023, 23(5): 1799–1811. [doi: 10.3969/j.issn.1671-1815.2023.05.002]

6 Mun SM, Nam SH, Jang HU, *et al*. A robust blind watermarking using convolutional neural network. arXiv: 1704.03248, 2017.

7 Mun SM, Nam SH, Jang H, *et al*. Finding robust domain from attacks: A learning framework for blind watermarking. Neurocomputing, 2019, 337: 191–202. [doi: 10.1016/j.neucom.2019.01.067]

8 Ahmadi M, Norouzi A, Karimi N, *et al*. ReDMark: Framework for residual diffusion watermarking based on deep networks. Expert Systems with Applications, 2020, 146: 113157. [doi: 10.1016/j.eswa.2019.113157]

9 Fang H, Jia ZY, Qiu YP, *et al*. De-END: Decoder-driven watermarking network. IEEE Transactions on Multimedia, 2023, 25: 7571–7581. [doi: 10.1109/TMM.2022.3223559]

10 Fang H, Jia ZY, Ma ZH, *et al*. PIMoG: An effective screen-shooting noise-layer simulation for deep-learning-based watermarking network. Proceedings of the 30th ACM International Conference on Multimedia. Barcelona: ACM, 2022. 2267–2275.

11 Cao F, Wang TJ, Guo DD, *et al*. Screen-shooting resistant image watermarking based on lightweight neural network in frequency domain. Journal of Visual Communication and Image Representation, 2023, 94: 103837. [doi: 10.1016/j.jvcir.2023.103837]

12 Tang ZW, Chai XL, Lu Y, *et al*. An end-to-end screen

shooting resilient blind watermarking scheme for medical images. Journal of Information Security and Applications, 2023, 76: 103547. [doi: 10.1016/j.jisa.2023.103547]

13 Cao F, Guo DD, Wang TJ, *et al*. Universal screen-shooting robust image watermarking with channel-attention in DCT domain. Expert Systems with Applications, 2024, 238: 122062. [doi: 10.1016/j.eswa.2023.122062]

14 Dinh L, Krueger D, Bengio Y. Nice: Non-linear independent components estimation. arXiv:1410.8516, 2014.

15 Guan ZY, Jing JP, Deng X, *et al*. DeepMIH: Deep invertible network for multiple image hiding. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(1): 372–390. [doi: 10.1109/TPAMI.2022.3141725]

16 Luo YJ, Zhou TQ, Liu F, *et al*. IRWArt: Levering watermarking performance for protecting high-quality artwork images. Proceedings of the 2023 ACM Web Conference. Austin: ACM, 2023. 2340–2348.

17 Fang H, Qiu YP, Chen KJ, *et al*. Flow-based robust watermarking with invertible noise layer for black-box distortions. Proceedings of the 37 AAAI Conference on Artificial Intelligence. Washington: AAAI Press, 2023. 5054–5061.

18 Zhu JR, Kaplan R, Johnson J, *et al*. HiDDeN: Hiding data with deep networks. Proceedings of the 15th European Conference on Computer Vision (ECCV). Milan: Springer, 2018. 657–672.

19 Jia ZY, Fang H, Zhang WM. MBRS: Enhancing robustness of DNN-based watermarking by mini-batch of real and simulated JPEG compression. Proceedings of the 29th ACM International Conference on Multimedia. ACM, 2021. 41–49.

20 Fang H, Jia ZY, Zhou H, *et al*. Encoded feature enhancement in watermarking network for distortion in real scenes. IEEE Transactions on Multimedia, 2023, 25: 2648–2660. [doi: 10.1109/TMM.2022.3149641]

21 Ma R, Guo MX, Hou Y, *et al*. Towards blind watermarking: Combining invertible and non-invertible mechanisms. Proceedings of the 30th ACM International Conference on Multimedia. Lisboa: ACM, 2022. 1532–1542.

22 段恒利, 梁晓燕. 基于 DCT 域 JPEG 图像压缩算法的研究与设计. 通信与信息技术, 2023(6): 10–12.

23 史艳琼, 王昌文, 卢荣胜, 等. 基于低秩稀疏矩阵分解和离散余弦变换实现多聚焦图像融合的算法. 激光与光电子学进展, 2024, 61(10): 1037010.

24 刘涤. 基于离散余弦变换的数字水印算法研究. 电脑编程技巧与维护, 2023(9): 138–140. [doi: 10.3969/j.issn.1006-4052.2023.09.043]

25 Liu Y, Guo MX, Zhang J, *et al*. A novel two-stage separable deep learning framework for practical blind watermarking. Proceedings of the 27th ACM International Conference on Multimedia. Nice: ACM, 2019. 1509–1517.

26 Huang JT, Luo T, Li L, *et al*. ARWGAN: Attention-guided robust image watermarking model based on GAN. IEEE Transactions on Instrumentation and Measurement, 2023, 72: 5018417. [doi: 10.1109/TIM.2023.3285981]

(校对责编: 王欣欣)