软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2023,34(9):4351−4361 [doi: 10.13328/j.cnki.jos.006661]
©中国科学院软件研究所版权所有.
E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

基于残差学习的新型不可感知水印攻击方法*

李 琦[1,2], 王春鹏[1,2], 王晓雨[1,2], 李 健[1,2], 夏之秋[1,2], 高 锁[3], 马 宾[1,2]

[1](齐鲁工业大学 (山东省科学院) 网络空间安全学院, 山东 济南 250353)

[2](齐鲁工业大学 (山东省科学院) 山东省计算中心 (国家超级计算济南中心), 山东 济南 250353)

[3](哈尔滨工业大学 计算学部, 黑龙江 哈尔滨 150001)

通信作者: 马宾, E-mail: sddxmb@126.com

摘 要: 传统的水印攻击方法虽然能够干扰水印信息的正确提取, 但同时会对含水印图像的视觉质量造成较大损失, 为此提出了一种基于残差学习的新型不可感知水印攻击方法. 首先, 通过构建基于卷积神经网络的水印攻击模型, 在含水印图像和无水印图像之间进行端到端非线性学习, 完成含水印图像映射到无水印图像的任务, 达到水印攻击的目的; 其次, 根据水印信息的嵌入区域选择合适数目的特征提取块以提取含水印信息的特征图. 鉴于含水印图像和无水印图像之间的差异过小, 水印攻击模型在训练过程中的可学习性受到限制, 导致模型很难收敛. 引入残差学习机制来提升水印攻击模型的收敛速度和学习能力, 通过减少残差图像 (含水印图像和提取的特征图像做差) 与无水印图像之间的差异来提升被攻击图像的不可感知性. 此外, 还根据 DIV2K2017 超分辨率数据集以及所攻击的基于四元数指数矩的鲁棒彩色图像水印算法构建了训练水印攻击模型的数据集. 实验结果表明该水印攻击模型能够在不破坏含水印图像视觉质量的前提下以高误码率实现对鲁棒水印算法的攻击.

关键词: 残差学习; 不可感知; 卷积神经网络; 水印攻击模型; 鲁棒水印算法

中图法分类号: TP393

中文引用格式: 李琦, 王春鹏, 王晓雨, 李健, 夏之秋, 高锁, 马宾. 基于残差学习的新型不可感知水印攻击方法. 软件学报, 2023, 34(9): 4351–4361. http://www.jos.org.cn/1000-9825/6661.htm

英文引用格式: Li Q, Wang CP, Wang XY, Li J, Xia ZQ, Gao S, Ma B. Novel Imperceptible Watermarking Attack Method Based on Residual Learning. Ruan Jian Xue Bao/Journal of Software, 2023, 34(9): 4351–4361 (in Chinese). http://www.jos.org.cn/1000-9825/6661.htm

Novel Imperceptible Watermarking Attack Method Based on Residual Learning

LI Qi[1,2], WANG Chun-Peng[1,2], WANG Xiao-Yu[1,2], LI Jian[1,2], XIA Zhi-Qiu[1,2], GAO Suo[3], MA Bin[1,2]

[1](School of Cyber Security, Qilu University of Technology (Shandong Academy of Sciences), Jinan 250353, China)

[2](Shandong Computer Science Center (National Supercomputing Center in Jinan), Qilu University of Technology (Shandong Academy of Sciences), Jinan 250353, China)

[3](Faculty of Computing, Harbin Institute of Technology, Harbin 150001, China)

Abstract: Although traditional watermarking attack methods can obstruct the correct extraction of watermark information, they reduce the visual quality of watermarked images greatly. Therefore, a novel imperceptible watermarking attack method based on residual learning is proposed. Specifically, a watermarking attack model based on a convolutional neural network is constructed for the end-to-end nonlinear learning between a watermarked image and an unwatermarked one. A mapping from the watermarked image to the unwatermarked one is thereby accomplished to achieve the purpose of watermarking attack. Then, a proper number of feature extraction blocks are selected

* 基金项目: 国家自然科学基金 (61802212, 61872203); 山东省自然科学基金 (ZR2019BF017, ZR2020MF054); 山东省高校科研计划 (J18KA331); 山东省重大科技创新工程 (2019JZZY010127, 2019JZZY010132, 2019JZZY010201); 济南市“高校 20 条”引进创新团队 (2019GXRC031)
收稿时间: 2021-04-23; 修改时间: 2021-07-30; 采用时间: 2022-02-25; jos 在线出版时间: 2022-12-22
CNKI 网络首发时间: 2022-12-26

according to the embedding region of watermark information to extract a feature map containing watermark information. As the difference between the two images is insignificant, the learning ability of the watermarking attack model is limited in the training process, making it difficult for the model to reach a convergence state. A residual learning mechanism is thus introduced to improve the convergence speed and learning ability of the watermarking attack model. The imperceptibility of the attacked image can be improved by reducing the difference between the residual image (the subtraction between the watermarked image and the extracted feature map) and the unwatermarked one. In addition, a dataset for training the watermarking attack model is constructed with the super-resolution dataset DIV2K2017 and the attacked robust color image watermarking algorithm based on quaternion exponent moments. The experimental results show the proposed watermarking attack model can attack a robust watermarking algorithm with a high bit error rate (BER) without compromising the visual quality of watermarked images.

Key words: residual learning; imperceptibility; convolutional neural network (CNN); watermarking attack model; robust watermarking algorithm

随着互联网技术的不断发展, 信息获取变得越来越便捷, 随之而来的是海量信息不受限制地在网络上进行存储和传输. 如何有效地保护信息安全是一个亟待解决的关键问题, 也一直是科学研究中最重要的课题之一. 数字水印技术[1,2]是数字图像版权保护的关键技术, 通过把一些标识信息 (即数字水印) 嵌入到需保护的图像中, 以达到确认图像版权归属的目的. 目前关于数字水印技术的研究主要集中在两个方面, 水印方法和水印攻击方法. 扮演“守方”的水印方法通过提高算法的鲁棒性增强对各种水印攻击方法的抵抗能力; 而作为“攻方”的水印攻击方法则通过对数字水印系统进行各种攻击, 试图让水印方法无法正确提取出嵌入的水印信息.

为了抵抗水印攻击方法, “守方”近些年提出了多种应对策略, 并持续产生了大量的成果. 为了抵抗信号处理攻击, 设计了基于图像空域[3,4]、变换域[5,6]和特征空间[7,8]的水印方法; 为了抵抗几何攻击, 设计了基于几何不变量[9,10]、同步校正[11,12]、局部特征区域方法等策略的水印方法; 此外, 设计了多种水印方法以抵抗降梯度攻击、敏感性攻击和扰乱攻击[13,14]. 近几年, 随着深度学习的发展, 研究人员将深度神经网络扩展到了图像水印领域. Haribabu 等人[15]于 2015 年提出了一种基于自编码的神经网络数字图像水印算法, 其根本思想是使用标准梯度下降反向传播算法学习给定图像的自动编码网络的权重, 基于该思想将水印不可见地嵌入到给定的图像中. 2018 年, Zhu 等人[16]提出了针对水印算法的 HiDDeN 架构, 利用神经网络学习使用细微扰动编码大量有用信息来完成水印嵌入的任务. 同年, Ahmadi 等人[17]提出了一种深度端到端差分水印框架, 该框架能够在嵌入容量和鲁棒性两者之间进行适当的调整和权衡, 并具有自适应性和灵活性. 2020 年, Hao 等人[18]提出了一种基于生成对抗网络的图像水印算法, 通过该方法得到的含水印图像具有更好的视觉效果, 其抗噪声能力也更有优势. 同年, Lee 等人[19]提出一种基于神经网络的图像盲水印算法, 其能够在不使用任何分辨率相关层或组件的情况下执行嵌入网络的操作.

近年来, 对数字水印技术的研究主要集中于“守方”, 旨在提升现有水印方法的鲁棒性, 但是现有的水印攻击体系已无法满足水印方法的需求, 含水印图像在被攻击后能以极低的误码率甚至于无损提取水印信息. 并且目前的水印攻击方式并没有考虑被攻击后水印图像的视觉质量, 这对于很多需要被保护的信息来说是不切实际的.

随着计算机硬件和网络带宽的快速发展, 人工智能和深度学习领域引起了研究学者的广泛关注. 到目前为止, 深度学习和卷积神经网络已为图像识别、语音识别和自然语言处理等多个领域提供了很多完美的解决方案. 在深度学习时代, 卷积神经网络为改变传统的水印攻击方式提供了契机. 2020 年, Nam 等人[20]提出一种水印攻击网络 (watermarking attack network), 他们意识到目前存在的攻击方案仍然不能够作为一个测试水印方案鲁棒性的基准, 并且存在很多问题. 他们指出了目前的水印攻击方式仅是单一地对水印图像进行干扰, 而忽略了目标水印的具体特性. 他们提出使用一种基于残差密集块 (residual dense blocks) 的网络架构来用于学习水印图像的局部和全局特征, 在尽可能保护水印图像质量不被干扰的前提下使得各类水印方案失效, 即提取出正常但是相反的水印信息. 他们对 9 种较为主流的水印方案都进行了攻击, 与传统的攻击方式相比, 他们的攻击方案能够很好地控制水印图像的信息损失, 并达到水印信息被破坏的效果. Sharma 等人[21]提出一种鲁棒混合水印技术, 能够抵抗基于 CNN 的对抗性攻击. 该水印方案首次假想到了基于 CNN 的攻击方式, 并结合了对抗的思想. 他们首先对基于变换域 (DWT、DCT 和 SVD 等) 的图像水印方案的鲁棒性进行了传统攻击方式的混合测试. 然后他们提出了一种基于深度卷积神经网络的自编码器新型水印攻击, 其能通过网络中间层的低维度投影来表示水印图像的内容 (空间和结构). 在

训练环节, CAFAR 10 数据集被用来作为图像库, 他们的目的是为了和传统的水印方案相比, 其基于 DWT+SVD 的水印嵌入方案能抵抗更加现代化的攻击方式, 同时具有更好的提取效果, 但是其新型攻击方式的提出也是一个很不错的创新. Geng 等人[22]提出一种基于 CNN 的对鲁棒水印方案的实时攻击方案, 以提升水印方案鲁棒性为前提, 并指出现有的攻击方式并不能够很好地平衡图像质量和水印破坏能力, 并基于此提出一种以 CNN 为主的去除攻击方式, 他们主要攻击具有高鲁棒性和不需要宿主图像的盲水印方案, 该攻击方案能够在不具备任何先验知识的前提下对水印图像进行预处理操作, 进而阻碍水印的提取; 甚至在水印方案未知的情况下, 仍然能够利用水印图像的一些共同特征来破坏水印. Quiring 等人[23]研究了一种基于对抗学习的黑盒攻击方法, 专门用于针对数字水印. 首先, 该方案阐明, 尽管机器学习和数字水印都是独立的领域, 但是存在着某种共性 (易受攻击性). 他们使用神经网络来代替水印检测工具, 并利用神经网络来去除水印. 并且实现了能够在水印方案未知的前提下完成对嵌入水印的攻击操作. 尽管目前少数的研究成果表明深度学习技术能够作为一种新型的水印攻击方式达到干扰水印提取并同时保证水印的图像质量, 但是攻击体系仍然不够成熟, 并且目前大多数的水印攻击方案是以提升水印图像的质量 (PSNR、SSIM 等) 为目标, 忽略了水印提取 (bit error rate, BER) 的问题.

为了解决目前水印攻击方法所存在的问题, 本文将卷积神经网络引入水印攻击领域, 提出了一种基于残差学习的新型不可感知水印攻击方法. 首先, 将含水印图像和不含水印图像作为不可感知水印攻击模型的输入和输出, 通过卷积神经网络在含水印图像和无水印图像之间进行端到端非线性学习, 完成含水印图像映射到无水印图像的任务, 从而达到去除水印信息的目的; 其次, 根据水印信息的嵌入区域选择合适数目的特征提取块以提取出包含大量水印信息的特征图. 由于含水印图像和无水印图像之间的差异很小, 水印攻击模型在训练过程中的可学习性受到限制, 导致模型很难收敛, 引入残差学习机制, 通过学习含水印图像和无水印图像之间的差异来提升水印攻击模型的收敛速度以及学习能力; 此外, 为了体现提出的水印攻击模型的性能, 本文采用了对常规信号处理以及几何攻击都同样具有较好鲁棒性能的基于四元数指数矩的彩色图像水印算法, 并根据 DIV2K2017 以及所攻击的彩色图像水印算法[24]构建了训练水印攻击模型的数据集. 水印攻击模型输出的是不含水印图像, 因此和传统的水印攻击方法相比, 被攻击的含水印图像的不可感知性得到了大幅度提升. 该深度水印攻击模型验证了在不破坏含水印图像视觉质量的前提下实现对鲁棒水印算法的攻击是可行的.

本文第 1 节对攻击的基于四元数指数矩的鲁棒彩色图像水印算法进行了介绍. 第 2 节对提出的新型水印攻击模型进行了详细的介绍. 第 3 节对实验结果进行了分析和总结. 最后总结了全文, 并对本文的研究意义以及未来值得关注的研究方向进行了初步探讨.

1 攻击的水印方案

本文攻击的水印方案是基于四元数指数矩的鲁棒彩色图像水印算法, 该算法借助四元数和指数矩理论将水印嵌入方案从灰度图像延伸到了彩色图像, 定义并推导了可用于彩色图像的高精度四元数指数矩, 同时构造了相对应的四元数指数矩几何不变量, 与目前存在的绝大多数彩色图像水印方案相比, 其不仅可以有效抵抗常规信号处理攻击, 并且有效增强了对几何攻击的鲁棒性. 此外, 推导出的四元数指数矩可广泛用于图像识别、检索等多个领域[24]. 四元数可以视为复数的推广形式, 假设大小为 $M\times N$ 的彩色图像的一个像素点使用 $f(x,y)$ 表示, x 和 y 分别表示该像素点在矩阵中行列的位置信息. 四元数的 3 个虚部可以分别看作彩色图像的 RGB 三个通道, 那么 RGB 彩色图像可以表示为无实部的纯虚四元数:

$$f(x,y)=f_{\mathrm{R}}(x,y)i+f_{\mathrm{G}}(x,y)j+f_{\mathrm{B}}(x,y)k \tag{1}$$

其中, $f_{\mathrm{R}}(x,y)$, $f_{\mathrm{G}}(x,y)$ 和 $f_{\mathrm{B}}(x,y)$ 分别为彩色图像的 RGB 分量, i,j 和 k 表示与其对应的虚数单位.

2011 年指数矩首次被提出, 图像的重建任务可以使用较少的矩就能够完成. 与其他矩相比, 指数矩还具有对噪声不敏感、数值计算稳定等特点, 因此被广泛用于鲁棒水印领域. 在极坐标系 (r,θ) 中, 阶数为 n, 重复度为 m 的指数矩 E_{nm}:

$$E_{nm}=\frac{1}{4\pi}\int_0^{2\pi}\int_0^1 f(r,\theta)A_n^*(r)\exp(-jm\theta)r\mathrm{d}r\mathrm{d}\theta \tag{2}$$

其中, 图像函数 $A_n(r)=\sqrt{2/r}\exp(j2n\pi r)$ 使用 $f(r,\theta)$ 表示, $A_n^*(r)$ 为图像函数的共轭. 指数矩的优势是可使用少数矩

便能重构图像, 假设图像指数矩的最高阶为 $n_{\max}$, 最大重复度为 $m_{\max}$, 则其重构公式如下:

$$f'(r,\theta)=\sum_{n=-\infty}^{+\infty}\sum_{m=-\infty}^{+\infty}E_{nm}A_n(r)\exp(jm\theta)\approx\sum_{n=-n_{\max}}^{n_{\max}}\sum_{m=-m_{\max}}^{m_{\max}}E_{nm}A_n(r)\exp(jm\theta) \tag{3}$$

基于四元数指数矩的鲁棒彩色图像水印算法首先将指数矩延申到四元数领域, 定义了彩色图像的四元数指数矩. 假设 $f(r,\theta)$ 为极坐标系下的彩色图像, 根据四元数和指数矩理论 (见公式 (1) 和公式 (2)) 可定义四元数指数矩:

$$E_{nm}^{R}=\frac{1}{4\pi}\int_0^{2\pi}\int_0^1 f(r,\theta)A_n^*(r)\exp(-\mu m\theta)rdrd\theta \tag{4}$$

其中, μ 表示的是一个单位纯四元数, 即 $\mu=(i+j+k)/\sqrt{3}$, 同理可得彩色图像的四元数指数矩的重建公式:

$$f'(r,\theta)=\sum_{n=-\infty}^{+\infty}\sum_{m=-\infty}^{+\infty}E_{nm}^{R}A_n(r)\exp(\mu m\theta)\approx\sum_{n=-n_{\max}}^{+n_{\max}}\sum_{m=-m_{\max}}^{+m_{\max}}E_{nm}^{R}A_n(r)\exp(\mu m\theta) \tag{5}$$

具体的水印嵌入方案可见算法 1.

算法 1. 基于四元数指数矩的彩色图像水印算法.

输入: 大小为 $C\times H\times W$ 的原始不含水印图像 I_c, 大小为 $M\times N$ 的水印图像 I_s.

1. 使用 Arnold 变换对二值水印图像 I_s 进行置乱并将其转换为一维序列 $S=\{s(k),1\leqslant k\leqslant M\times N\}$.

2. 根据公式 (4) 计算出原始不含水印图像 I_c 的四元数指数矩 E_1.

3. 利用密钥 K_1 随机从 E_1 选取 $M\times N$ 个四元数指数矩 E^R, 其中 $E^R=(E_{n_1m_1}^R,E_{n_2m_2}^R,\ldots,E_{n_{M\times N}m_{M\times N}}^R)$. 其对应的幅值为 $A=(A_{n_1m_1},A_{n_2m_2},\ldots,A_{n_{M\times N}m_{M\times N}})$.

4. 使用以下量化规则将水印信息 S 嵌入到幅值 A 中:

$$\hat{A}_{n_km_k}=\begin{cases}(\lambda_k-1/2)\Delta, & \bmod(\lambda_k+s(k),2)=1\\(\lambda_k-1/2)\Delta, & \bmod(\lambda_k+s(k),2)=0\end{cases},$$

其中, $\lambda_k=\text{round}(A_{n_km_k}/\Delta)$, round$(\cdot)$ 表示为四舍五入函数, Δ 表示为量化步长, $\bmod(x,y)$ 表示为 x 除 y 所得余数. $\hat{A}=(\hat{A}_{n_1m_1},\hat{A}_{n_2m_2},\ldots,\hat{A}_{n_{M\times N}m_{M\times N}})$ 为嵌入水印信息后的幅值, $\hat{E}^R=(\hat{E}_{n_1m_1}^R,\hat{E}_{n_2m_2}^R,\ldots,\hat{E}_{n_{M\times N}m_{M\times N}}^R)$ 则是对应的四元数指数矩.

5. 使用未修改的四元数指数矩计算重建图像 $f^*(h,w)$:

$$f^*(h,w)=f_o(h,w)-f_s(h,w),$$

其中, $f_o(h,w)$ 代表原始不含水印图像, $f_s(h,w)$ 表示使用 E^R 重建的图像.

6. 由以下公式可得到嵌入水印信息的含水印图像 $f'(h,w)$:

$$f'(h,w)=f^*(h,w)+f_s'(h,w),$$

其中, $f_s'(h,w)$ 表示的是使用 $\hat{E}^R$ 重建的图像 (I_c 为 256×256, I_s 为 32×32).

2 深度水印攻击模型

基于残差学习的深度水印攻击方法将原始不含水印图像作为水印攻击模型的优化目标, 并充分利用卷积神经网络的非线性拟合能力, 将受攻击的含水印图像映射回原始不含水印图像, 以此方式移除含水印图像中嵌入的水印信息, 达到攻击的目的, 图 1 展示了水印攻击模型的整体框架结构. 对于绝大多数的数字水印技术而言, 水印的不可感知性是重要的评价指标, 通过水印技术得到的含水印图像与原始不含水印图像之间并无显著的视觉差异. 正因如此, 将原始不含水印图像作为攻击含水印图像的优化目标, 让攻击后的含水印图像更加接近原始不含水印图像是一种简单且有效的训练手段, 具体有 3 方面原因: (1) 训练过程无需先验知识, 即不需要提前知道水印方法中水印嵌入和提取的具体策略, 因此具有普适性和优化过程简单的优点; (2) 训练好的水印攻击模型将含水印图像转化为不含水印图像, 移除了含水印图像中由于水印嵌入带来的内容改变, 导致水印提取器无法有效提取出水印信息; (3) 训练好的水印攻击模型输出的攻击后含水印图像、原始含水印图像及不含水印图像三者间具有高度视觉相似性, 满足对水印攻击的不可感知需求.

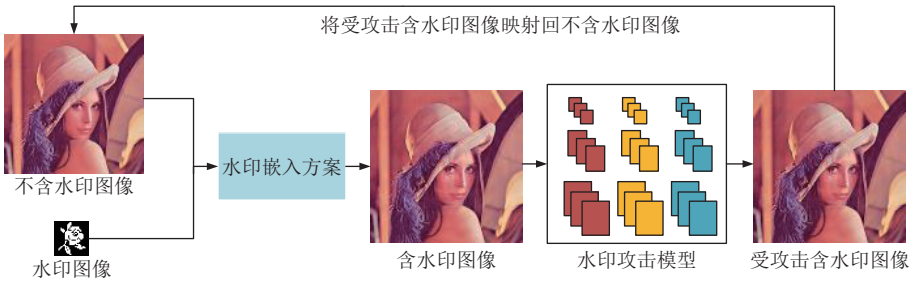

图 1　深度水印攻击模型的流程图

2.1　攻击模型设计

对于基于卷积神经网络的深度水印攻击模型, 攻击能力与卷积神经网络的学习能力密切相关. 如果卷积神经网络能够挖掘出更有效的特征, 有利于原始不含水印图像和含水印图像之间映射关系的学习, 从而提高模型的攻击能力. 目前大多数的数字水印技术都具有较高的不可感知性, 即原始不含水印图像和含水印图像之间存在高度相似性, 那么模型如果直接在原始不含水印图像和含水印图像上进行特征提取进行训练将很难收敛. 受到深度超分辨率技术 (very deep super resolution, VDSR)[25]的启发, 我们将残差学习的思想引入到水印攻击模型的设计之中, 水印攻击网络直接对残差图像建模, 减少了对图像的大部分冗余信息的处理, 以便更快地达到收敛状态. 此外, 模型采用深层网络结构, 合理地加深了特征提取模块, 进而提升了模型的攻击能力. 特征提取块主要用来提取含有水印信息且与原始不含水印图像相同大小的低频特征图像, 因此提取块的卷积层都采用了 padding 操作, 进而模型能够适应原始不含水印图像的任意尺寸. 模型的设计结构图如图 2 所示.

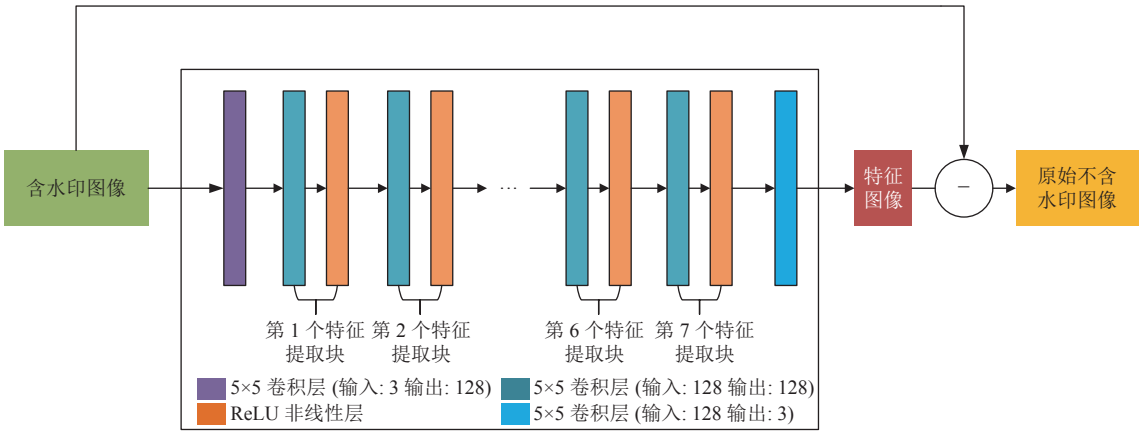

图 2　深度水印攻击模型的网络架构设计图

在本文中, 我们通过实验证明了基于四元数指数矩的水印方案将水印信息大部分嵌入到了原始不含水印图像的低频区域以保证其鲁棒性, 而理论上卷积神经网络的层数越深, 所提取的特征带有高频信息就越多, 反而不会对水印信息产生破坏, 因此本文提出的水印攻击模型由 2 层卷积层和 7 个特征提取块构成. 每个特征提取块都包含 128 个 5×5 的卷积核和 ReLU 非线性层, 以提升模型的非线性拟合能力. 为了保证输出的特征图像和含水印图像大小相同, 水印模型中的卷积层都采用了 padding 操作.

2.2　损失函数

由图 2 可知, 水印攻击模型的目的是经过特征提取层将能够表示水印信息的低频特征图像 I_L 提取出来, 为了破坏含水印图像 I_W 中的水印信息, 将含水印图像 I_W 与低频特征图像做差得到残差图像 I_R:

$$I_R = I_W - I_L \tag{6}$$

同时, 为了保证被攻击含水印图像的高度不可感知性, 本文采用 MSE (均方误差) 来减少残差图像与原始不含水印图像 I_O 之间的差异, 即:

$$\mathcal{L}_{loss}=\frac{1}{2}\|I_R-I_O\|^2 \tag{7}$$

2.3 训练设置

本文的数据集为 800 幅 256×256 的彩色图像, 750 张为训练数据集, 50 张为测试集. 本文的非线性模块都采用的 ReLU 层, 因此采用 kaiming 均匀分布 (He initialization) 来初始化模型参数, 即 $\mu=(-bound,bound)$ 见公式 (8). 此外, 本文的实验环境为装有 NVIDA Tesla V100 32 GB 显卡的 Windows 版服务器, 深度学习框架为 Python 3.6, PyTorch 1.60 版本. 实验中, 训练集的批次大小设置为 50, 学习率设置为 0.005, 优化器采用了 Adam.

$$bound=\sqrt{\frac{6}{(1+a^2)\times fan\_in}} \tag{8}$$

3 实验结果与分析

3.1 水印信息的嵌入区域

基于四元数指数矩的彩色图像鲁棒水印方法将水印信息嵌入到原始不含水印图像的具体区域决定了水印攻击模型的结构设计. 如果水印信息嵌入到了原始不含水印图像的高频区域, 那么水印攻击模型的特征提取模块主要用来提取含水印图像的高频信息即可, 那么残差图像 (含水印图像和提取的特征图像作差) 的水印信息就能够最大限度地受到破坏; 同理, 水印信息嵌入到原始不含水印图像的低频区域也是以上述方式破坏水印信息. 因此, 我们首先对水印的嵌入区域进行了分析, 利用 Haar 小波将含水印图像进行频域分解, 分别可以得到含水印图像的低频信息, 水平高频信息, 垂直高频信息以及对角高频信息如图 3 所示.

图 3 含水印图像小波分解后对应的频域信息

为了验证水印信息存在于含水印图像的具体区域, 我们将含水印图像小波分解后对应的频域信息分别设置为 0, 然后进行重构, 然后根据随机选择重构的含水印图像提取水印信息来计算误码率 (bit error rate, BER), 其实验结果如表 1 所示.

表 1 重构的含水印图像对应的误码率

水印图像	缺少低频	缺少水平高频	缺少垂直高频	缺少对角高频
01_watermark	**0.3630**	0.2783	0.2685	0.2724
04_watermark	**0.3955**	0.2713	0.2666	0.2656
08_watermark	**0.4336**	0.3154	0.3291	0.3154
14_watermark	**0.4601**	0.2490	0.2314	0.2314

从表 1 的实验结果可以看出, 缺少低频信息重构出的含水印图像提取水印信息得到的误码率远高于缺少其他频域信息, 因此我们得出结论, 即基于四元数指数矩的水印嵌入方案将水印信息大部分都嵌入到了无水印图像的低频区域. 该结论为模型的网络架构提供了思路, 即要尽可能提升网络架构提取含水印图像的低频信息的能力.

3.2 网络架构深度

为了验证网络架构的深度对模型攻击能力的影响, 本文的特征提取块分别选取 7、11、15 以及 20, 并随机在

测试集中挑选了 25 幅含水印图像来计算提取水印信息的误码率, 实验结果如图 4 所示.

从图 4 的实验结果可以看出, 当特征提取块的层数 N 取 2 时, 水印信息提取的误码率的平均值为 0.117 16, 表示水印攻击模型具有一定的破坏能力; 而特征提取块的层数 N 取 7 时, 水印信息提取的误码率的平均值高达 0.138 85, 这说明, 随着模型深度的增加, 水印攻击模型的攻击能力在不断增强, 即模型提取的含水印图像的特征包含大部分低频信息, 即残差图像中的水印信息得到了大幅度的破坏, 误码率得到了提升; 当特征提取块的层数 N 取 11 到 15 层时, 水印信息提取的误码率的平均值分别为 0.008 60 和 0.060 07, 误码率反而降低了, 这说明, 具有更多特征提取块的模型提取的含水印图像特征包含的是大部分高频信息, 导致残差图像中包含了大部分的水印信息, 因此其误码率不降反增. 因此, 我们得出结论, 水印攻击模型的网络架构深度需要依据水印信息的嵌入区域来决定, 当水印信息嵌入到了低频区域, 层数不宜过多; 反之, 则可适当增加水印攻击模型的网络层数.

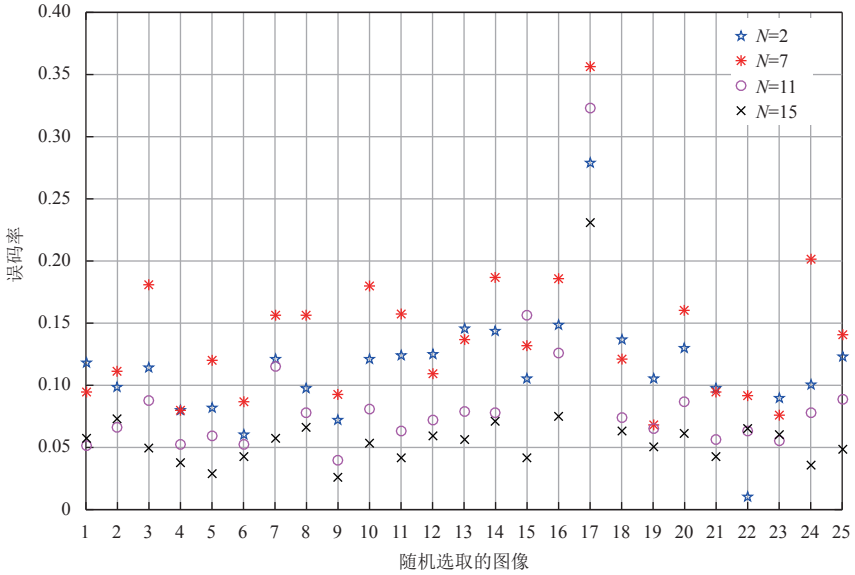

图 4 使用不同层数的特征提取块计算出的误码率

3.3 高度不可感知性

保证被攻击含水印图像的高度不可感知性是水印攻击模型的一个重要目标, 即尽可能降低含水印图像视觉质量的损失. 目前的传统水印攻击方法 (滤波、锐化以及组合攻击等) 没有针对水印信息的特性进行有效的攻击, 而是通过对含水印图像的完整性进行破坏, 进而对水印提取起到一定的干扰作用, 但同时会对含水印图像的视觉质量造成严重的破坏. 为此, 本文使用不同的攻击方法对含水印图像进行了干扰, 并计算了被攻击含水印图像和原始含水印图像的残差图像 (为了显示效果, 残差图像放大 15 倍), 实验结果如图 5 所示.

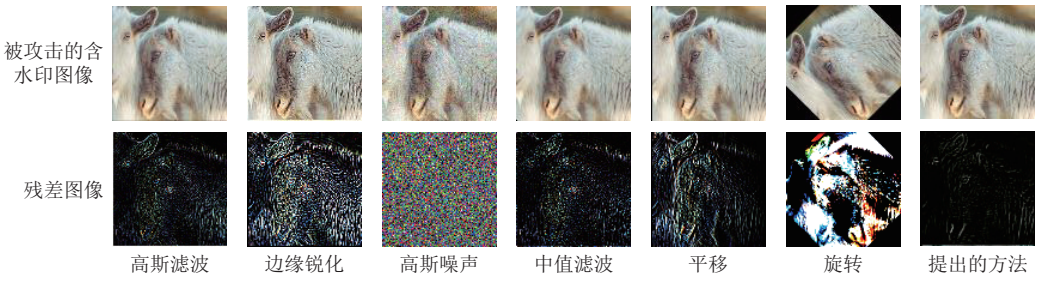

图 5 使用不同攻击方法得到的被攻击含水印图像及残差图像 (残差图像放大 15 倍)

从图 5 的实验结果可以看出, 与目前传统的水印攻击方法相比, 本文提出的方法能够充分利用嵌入水印信息的特征属性, 减少对含水印图像的内容改动, 从而更好地保证含水印图像的高度不可感知性. 为了更加直观体现被

攻击的含水印图像的不可感知性, 分别用被攻击图像和原始水印图像的峰值信噪比 (peak signal to noise ratio, PSNR) 以及结构相似度 (structural similarity, SSIM) 来衡量, 实验结果如表 2 所示.

表 2　使用不同攻击方法得到的被攻击含水印图像对应的 PSNR 以及 SSIM

水印图像	指标	高斯噪声	椒盐噪声	随机噪声	边缘锐化	高斯滤波	旋转45°	平移	提出方法
001_watermarker	PSNR (dB)	20.1512	25.1514	26.8890	22.9069	30.7975	8.9284	25.1610	**32.4647**
	SSIM	0.4541	0.7891	0.9194	0.8068	0.9392	0.2650	0.9339	**0.9698**
024_watermarker	PSNR (dB)	20.1579	25.1479	26.8668	21.3129	29.5370	9.4586	25.8150	**31.3935**
	SSIM	0.7104	0.8901	0.9576	0.8507	0.9440	0.2845	0.9515	**0.9695**
032_watermarker	PSNR (dB)	20.1811	25.3401	26.8602	20.3784	27.5865	12.4761	27.5865	**33.2992**
	SSIM	0.6240	0.8668	0.9436	0.8376	0.9440	0.3191	0.9440	**0.9830**
048_watermarker	PSNR (dB)	21.3325	24.1367	26.9311	24.2208	31.0952	14.3154	30.1180	**32.5973**
	SSIM	0.7110	0.8612	0.8930	0.9078	0.9795	0.3188	0.9641	**0.9824**
050_watermarker	PSNR (dB)	21.1220	24.1366	26.8629	21.8707	30.4770	14.2464	28.5057	**34.3261**
	SSIM	0.4353	0.7521	0.7370	0.8334	0.9407	0.3023	0.9201	**0.9746**

表 2 是从 50 幅测试图像中随机选取的 5 幅图像, 从实验结果可以看出, 我们提出的方法在 PSNR 和 SSIM 都优于传统的水印攻击方法. 根据 50 幅测试图像的 PSNR 和 SSIM 的平均值的实验数据统计, 本文提出方法的 PSNR 平均值为 33.579 6 dB, SSIM 平均值为 0.978 4. 而被高斯噪声攻击的含水印图像的 PSNR 平均值为 20.384 8 dB, SSIM 平均值为 0.567 9, 远远低于我们提出的攻击方法所得到的被攻击含水印图像的质量. 其中, 传统水印攻击方法表现最好的是高斯滤波, 其 PSNR 平均值为 28.984 6 dB, SSIM 平均值为 0.924 1, 但是本文提出的水印攻击模型对含水印图像的水印信息的破坏能力远远优于高斯滤波攻击, 具体的实验结果可参考第 4 节. 从图 5 以及表 2 的实验结果可以看出, 与传统的水印攻击方法相比, 水印攻击模型能够充分利用嵌入水印信息的特征属性, 减少对含水印图像的内容改动, 更好地保证含水印图像的高度不可感知性.

3.4　攻击能力

传统的水印攻击方法在牺牲图像视觉质量的前提下能够对水印信息提取产生一定的干扰作用, 但是由于其没有充分利用水印信息的具体特性, 往往攻击效果不佳. 本文提出的水印攻击模型能够根据水印信息的嵌入区域针对性地对其进行破坏, 合理选择模型的特征提取块对含水印图像进行攻击, 从而达到最优的攻击效果. 图 6 表示的是不同的单一攻击方法得到的被攻击含水印图像以及提取的水印信息, 图 7 则表示的是不同的组合攻击方法得到的被攻击含水印图像以及提取的水印信息.

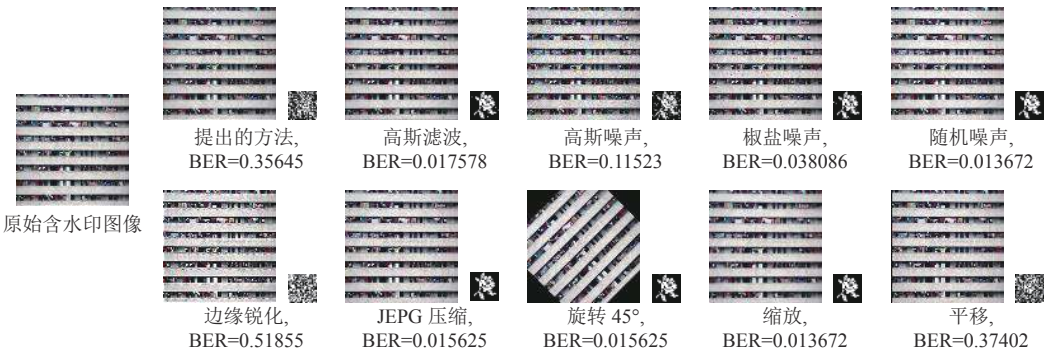

图 6　不同的单一攻击方法得到的被攻击含水印图像以及提取的水印信息

从图 6, 图 7 的实验结果可以看出, 与传统的水印攻击方法相比, 本文提出的水印攻击模型能够很好地保证含水印图像被攻击后的高度不可感知性. 从图 6 的实验结果分析来看, 与单一的传统水印攻击方法相比, 水印攻击模型的攻击能力远高于大多数水印攻击方法, 其水印信息提取的误码率高达 0.356 45, 仅低于边缘锐化和平移. 从图 7 的实验结果分析来看, 水印攻击模型的攻击能力仍远高于大多数的组合水印攻击方法, 仅低于边缘锐化和

JPEG70 的组合攻击方法. 实验结果表明水印攻击模型能够能很好地控制含水印图像的信息损失, 并有效破坏含水印图像的水印信息. 为了更加直观地体现水印攻击模型地攻击能力, 我们从 50 幅测试图像中随机选取了 25 幅图像, 并在不同的攻击方法下对误码率进行了统计, 实验结果如图 8 所示.

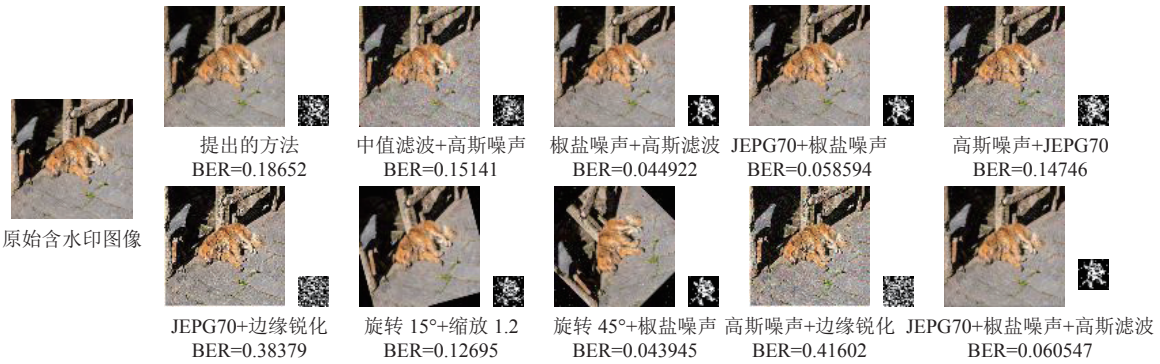

图 7　不同的组合攻击方法得到的被攻击含水印图像以及提取的水印信息

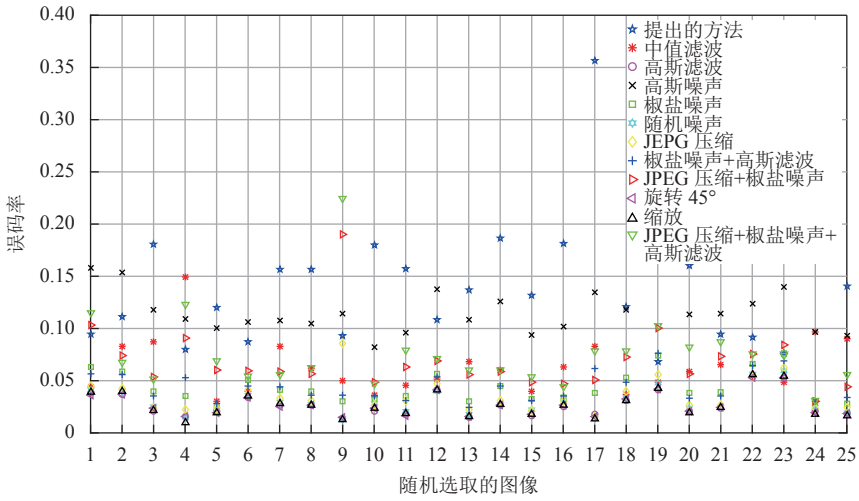

图 8　使用不同的水印攻击方法计算出的误码率

从图 8 的实验结果可以看出, 50 幅测试图像的水印信息提取误码率的平均值为 0.138 9, 而组合攻击 (JEPG 压缩+椒盐噪声+高斯滤波) 的水印信息提取误码率的平均值为 0.076 2, 而中值滤波、高斯噪声等单一的水印攻击方法的提取误码率分别为 0.063 1 和 0.043 7, 实验结果表明水印攻击模型的攻击能力优于目前大多数的水印攻击方法, 并且是一种有效的攻击手段.

4　结　论

数字水印技术的研究主要集中在两个方面, 水印方法和水印攻击方法. 尽管传统的水印攻击方法能够干扰水印信息的正确提取, 但同时会对含水印图像的视觉质量造成较大损失, 因此本文提出一种基于残差学习的新型不可感知水印攻击方法. 本文将卷积神经网络引入水印攻击领域, 首先, 将含水印图像和不含水印图像作为不可感知水印攻击模型的输入和输出, 通过卷积神经网络在含水印图像和无水印图像之间进行端到端非线性学习, 完成含水印图像映射到无水印图像的任务, 从而达到去除水印信息的目的; 其次, 根据水印信息的嵌入区域选择合适数目的特征提取块以提取出包含大量水印信息的特征图. 由于含水印图像和无水印图像之间的差异很小, 水印攻击模型在训练过程中的可学习性受到限制, 导致模型很难收敛, 引入残差学习机制, 通过学习含水印图像和无水印图像之间的差异来提升水印攻击模型的收敛速度以及学习能力; 此外, 我们根据 DIV2K2017 以及所攻击的水印算法构

建了训练水印攻击模型的数据集. 经过本文实验验证, 基于深度神经网络构造新型的水印攻击方式是完全可行的, 能够实现保持含水印图像视觉质量的同时干扰水印信息的正常提取.

扮演“守方”的水印方法通过提高算法的鲁棒性增强对各种水印攻击方法的抵抗能力; 而作为“攻方”的水印攻击方法则通过对数字水印系统进行各种攻击, 试图让水印方法无法正确提取出嵌入的水印信息. 近年来, 对数字水印技术的研究主要集中于“守方”, 旨在提升现有水印方法的鲁棒性, 现有水印方法能够很好地抵抗各类水印攻击, 含水印图像在被攻击后能以极低的误码率甚至于无损提取水印信息, 使得水印方法在现有的水印攻击评价体系下无法得出客观且有效的评价. “攻方”的发展陷入停滞, 无疑会对水印方法的发展产生重要影响, 进而阻碍整个水印领域的进步, 因此从逆向思维的角度出发, 以“攻”促“守”, 势必会促进数字水印领域“攻守”双方的良性发展. 在未来的研究工作中, 我们会尝试将图像复原、超分辨率等图像增强技术应用到水印攻击领域, 以进一步提升水印攻击模型的高度不可感知性和攻击能力. 同时会致力于构建出完整的一体化不可感知水印攻击系统, 并为完善水印攻击基准评价体系提供理论支撑.

References:

[1] Ma ZH, Zhang WM, Fang H, Dong XY, Geng LF, Yu NH. Local geometric distortions resilient watermarking scheme based on symmetry. IEEE Trans. on Circuits and Systems for Video Technology, 2021, 31(12): 4826–4839. [doi: 10.1109/TCSVT.2021.3055255]

[2] Yue Z, Li ZC, Yang YX, You FC, Liu FP. A histogram-based 2Bin M-ary image digital watermarking algorithm. Acta Electronica Sinica, 2020, 48(3): 531–537 (in Chinese with English abstract). [doi: 10.3969/j.issn.0372-2112.2020.03.016]

[3] Hua G, Xiang Y, Zhang LY. Informed histogram-based watermarking. IEEE Signal Processing Letters, 2020, 27: 236–240. [doi: 10.1109/LSP.2020.2965331]

[4] Zong TR, Xiang Y, Natgunanathan I, Guo S, Zhou WL, Beliakov G. Robust histogram shape-based method for image watermarking. IEEE Trans. on Circuits and Systems for Video Technology, 2015, 25(5): 717–729. [doi: 10.1109/TCSVT.2014.2363743]

[5] Shen YX, Tang C, Xu M, Chen MM, Lei ZK. A DWT-SVD based adaptive color multi-watermarking scheme for copyright protection using AMEF and PSO-GWO. Expert Systems with Applications, 2021, 168: 114414. [doi: 10.1016/j.eswa.2020.114414]

[6] Liu XL, Han GN, Wu JS, Shao ZH, Coatrieux G, Shu HZ. Fractional krawtchouk transform with an application to image watermarking. IEEE Trans. on Signal Processing, 2017, 65(7): 1894–1908. [doi: 10.1109/TSP.2017.2652383]

[7] Wang CP, Wang XY, Zhang C, Zhu XQ, Xia ZQ. Stereo image zero-watermarking algorithm based on ternary polar harmonic Fourier moments and chaotic mapping. SCIENTIA SINICA Informationis, 2018, 48(1): 79–99 (in Chinese with English abstract). [doi: 10.1360/N112017-00109]

[8] Yamni M, Karmouni H, Sayyouri M, Qjidaa H. Image watermarking using separable fractional moments of charlier–meixner. Journal of the Franklin Institute, 2021, 358(4): 2535–2560. [doi: 10.1016/j.jfranklin.2021.01.011]

[9] Hu RW, Xiang SJ. Cover-lossless robust image watermarking against geometric deformations. IEEE Trans. on Image Processing, 2021, 30: 318–331. [doi: 10.1109/TIP.2020.3036727]

[10] Ma B, Chang LL, Wang CP, Li J, Wang XY, Shi YQ. Robust image watermarking using invariant accurate polar harmonic Fourier moments and chaotic mapping. Signal Processing, 2020, 172: 107544. [doi: 10.1016/j.sigpro.2020.107544]

[11] Wang CP, Wang XY, Zhang C, Xia ZQ. Geometric correction based color image watermarking using fuzzy least squares support vector machine and Bessel K form distribution. Signal Processing, 2017, 134: 197–208. [doi: 10.1016/j.sigpro.2016.12.010]

[12] Wang XY, Zhang SY, Wen TT, Xu H, Yang HY. Synchronization correction-based robust digital image watermarking approach using bessel *K*-form PDF. Pattern Analysis and Applications, 2020, 23(2): 933–951. [doi: 10.1007/s10044-019-00828-w]

[13] Shahdoosti HR, Salehi M. Transform-based watermarking algorithm maintaining perceptual transparency. IET Image Processing, 2018, 12(5): 751–759. [doi: 10.1049/iet-ipr.2017.0898]

[14] Zhang XP, Wang SZ. Watermarking scheme capable of resisting sensitivity attack. IEEE Signal Processing Letters, 2007, 14(2): 125–128. [doi: 10.1109/LSP.2006.882092]

[15] Haribabu K, Subrahmanyam GRKS, Mishra D. A robust digital image watermarking technique using auto encoder based convolutional neural networks. In: Proc. of the 2015 IEEE Workshop on Computational Intelligence: Theories, Applications and Future Directions (WCI). Kanpur: IEEE, 2015. 1–6. [doi: 10.1109/WCI.2015.7495522]

[16] Zhu JR, Kaplan R, Johnson J, Fei-Fei L. HiDDeN: Hiding data with deep networks. In: Proc. of the 15th European Conf. on Computer Vision (ECCV). Munich: Springer, 2018. 682–697. [doi: 10.1007/978-3-030-01267-0_40]

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

[18] Hao KL, Feng GR, Zhang XP. Robust image watermarking based on generative adversarial network. China Communications, 2020, 17(11): 131–140. [doi: 10.23919/JCC.2020.11.012]

[19] Lee JE, Seo YH, Kim DW. Convolutional neural network-based digital image watermarking adaptive to the resolution of image and watermark. Applied Sciences, 2020, 10(19): 6854. [doi: 10.3390/app10196854]

[20] Nam SH, Yu IJ, Mun SM, Kim D, Ahn W. WAN: Watermarking attack network. arXiv:2008.06255, 2020.

[21] Sharma SS, Chandrasekaran V. A robust hybrid digital watermarking technique against a powerful CNN-based adversarial attack. Multimedia Tools and Applications, 2020, 79(43): 32769–32790. [doi: 10.1007/s11042-020-09555-5]

[22] Geng LF, Zhang WM, Chen HZ, Fang H, Yu NH. Real-time attacks on robust watermarking tools in the wild by CNN. Journal of Real-time Image Processing, 2020, 17(3): 631–641. [doi: 10.1007/s11554-020-00941-8]

[23] Quiring E, Rieck K. Adversarial machine learning against digital watermarking. In: Proc. of the 26th European Signal Processing Conf. (EUSIPCO). Rome: IEEE, 2018. 519–523. [doi: 10.23919/EUSIPCO.2018.8553343]

[24] Wang XY, Yang HY, Niu PP, Wang CP. Quaternion exponent moments based robust color image watermarking. Journal of Computer Research and Development, 2016, 53(3): 651–665 (in Chinese with English abstract). [doi: 10.7544/issn1000-1239.2016.20148177]

[25] Kim J, Lee JK, Lee KM. Accurate image super-resolution using very deep convolutional networks. In: Proc. of the 2016 IEEE Conf. on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 1646–1654. [doi: 10.1109/CVPR.2016.182]

附中文参考文献:

[2] 岳桢, 李子臣, 杨义先, 游福成, 刘福平. 直方图2Bin多进制图像数字水印算法的研究. 电子学报, 2020, 48(3): 531–537. [doi: 10.3969/j.issn.0372-2112.2020.03.016]

[7] 王春鹏, 王兴元, 张川, 朱晓强, 夏之秋. 基于三元数极谐–Fourier矩和混沌映射的立体图像零水印算法. 中国科学: 信息科学, 2018, 48(1): 79–99. [doi: 10.1360/N112017-00109]

[24] 王向阳, 杨红颖, 牛盼盼, 王春鹏. 基于四元数指数矩的鲁棒彩色图像水印算法. 计算机研究与发展, 2016, 53(3): 651–665. [doi: 10.7544/issn1000-1239.2016.20148177]

李琦(1992－), 男, 博士, 主要研究领域为信息安全, 图像隐写, 图像水印.

夏之秋(1992－), 女, 博士, 主要研究领域为信息安全, 图像水印.

王春鹏(1989－), 男, 博士, 副教授, CCF 专业会员, 主要研究领域为图像水印, 深度学习.

高锁(1995－), 男, 博士, 主要研究领域为图像加密, 信息安全.

王晓雨(1994－), 女, 博士, 主要研究领域为信息安全, 可逆信息隐藏.

马宾(1973－), 男, 博士, 教授, 博士生导师, 主要研究领域为信息安全, 深度学习.

李健(1982－), 男, 博士, 副教授, 主要研究领域为信息安全, 可逆信息隐藏.