

基于显著性检测的抗剪切攻击的 图像数字水印方法

黄 睿¹, 郁若雪¹, 樊 玮¹, 邢 艳², 陈梓尹³

(1. 中国民航大学 计算机与人工智能学院, 天津 300300; 2. 中国民航大学 安全科学与工程学院, 天津 300300;
3. 中国民航大学 电子信息与自动化学院, 天津 300300)

摘 要: 为了保证水印嵌入的区域在剪切时以较大的概率被保留, 提出了基于显著性检测的抗剪切攻击的图像数字水印方法。首先, 使用显著性检测网络确定嵌入区域; 其次, 使用 Transformer 网络将变换后的水印嵌入该区域中; 再次, 为了保证水印的鲁棒性, 在提取水印之前, 使用噪声层攻击嵌入水印的图像; 最后, 使用全卷积网络提取显著性区域中的水印。实验结果表明: 在常见的 9 种攻击方法下, 本文方法具有更好的鲁棒性; 在基于显著性目标检测的剪切攻击时, 本文方法依然可以准确地提取出水印。

关键词: 计算机应用; 图像水印; 显著性检测; 剪切

中图分类号: TP309 **文献标志码:** A **文章编号:** 1671-5497(2025)12-4000-10

DOI: 10.13229/j.cnki.jdxbgxb.20240301

Anti-cropping digital image watermarking based on saliency detection

HUANG Rui¹, YU Ruo-xue¹, FAN Wei¹, XING Yan², CHEN Zi-yin³

(1. School of Computer and Artificial Intelligence, Civil Aviation University of China, Tianjin 300300, China; 2. School of Safety Science and Engineering, Civil Aviation University of China, Tianjin 300300, China; 3. College of Electronic Information and Automation, Civil Aviation University of China, Tianjin 300300, China)

Abstract: To ensure that the embedded area of the watermark is preserved with a high probability during cropping, we propose an effective and anti-cropping attacked watermarking method. Firstly, we use a saliency detection network to select the embedding region; then we use a Transformer network to embed the transformed watermark into this region; to ensure the robustness of the watermark, a noise layer is used to attack the embedded image before extracting the watermark; finally, we use a fully convolutional network to extract watermark from the salient region. The experimental results show that the proposed method has higher robustness than compared methods under common nine attack methods. Our method can accurately extract the embedded watermark under saliency detection-based cropping attack.

Key words: computer application; image watermarking; saliency detection; cropping

收稿日期: 2024-03-24.

基金项目: 天津市科技计划项目(20JCQNJC00720).

作者简介: 黄睿(1987-), 男, 讲师, 博士. 研究方向: 计算机视觉, 图像处理, 机器学习. E-mail: rhuang@cuac.edu.cn

通信作者: 樊玮(1968-), 男, 教授, 博士. 研究方向: 人工智能, 大数据分析. E-mail: wfan@cauc.edu.cn

0 引 言

1994年, van Schyndel等^[1]提出了数字水印的概念,后被广泛用于多媒体产权的保护中。传统的水印方法通常基于经典的数字信号处理技术,如离散余弦变换^[2]、离散傅里叶变换^[3]、小波变换^[4,5]等。李喜艳等^[6]对原始图像进行离散小波变换和最优离散余弦变换后将水印信息嵌入其中,提高了水印的有效嵌入量;Hsu等^[7]提出了一种基于QR码的数字水印方案,水印的每个位都有4个副本,隐藏在不同的图像块中,对剪切攻击有很好的鲁棒性;Liu等^[8]通过多级离散小波变换(Discrete wavelet transform, DWT)将原始图像分解为多个子带,同时对水印图像进行奇异值分解(Singular value decomposition, SVD)操作,通过由果蝇优化算法计算得到的缩放因子,将水印嵌入原始图像的低频-低频子带(Low-low subband, LL)子带中。传统水印算法虽然可以嵌入一定的水印信息,但是此类方法嵌入的水印信息容量有限,并且鲁棒性依赖于算法的设计,通常情况下不能很好地抵抗剪切攻击。

随着深度学习的不断发展,基于深度学习的水印方法在对抗各种攻击时表现出较好的鲁棒性。Zhu等^[9]首次将用于水印嵌入的编码器和用于水印提取的解码器联合训练,并在训练过程中直接添加噪声,使得嵌入的水印具有较好的鲁棒性;Lu等^[10]将DWT与卷积网络结合,使模型在训练的过程中学习到更稳定的特征表示;Liu等^[11]提出了两阶段可分离深度学习框架,包括无噪声的端到端对抗训练和有噪声的解码器训练两部分;Zhang等^[12]通过逐步融合水印图像和原始图像的频率特征及深度特征实现水印的嵌入。虽然神经网络的应用使得水印方法的鲁棒性有了进一步的发展,但面对大尺度的剪切攻击(如图像缩略图^[13-15])时,现有方法无法有效提取出水印信息。

为了解决现有方法对基于显著性的剪切攻击鲁棒性较差的问题,本文设计了一种具有较强抗剪切攻击的数字图像水印方法。由于图像显著性从语义层面上将图像划分为包含主要内容的前景区域和信息较少的背景区域,将水印信息嵌入显著性区域可以极大地减少由于图像剪切导致信息丢失的可能性。为了保证嵌入的水印信息经过大尺度裁剪后仍然能被有效提取出来,本文设计了一种基于排列重组的水印冗余方法,当部分水印信息受损后,可通过

嵌入的冗余水印恢复。实验结果表明:在常见的9种攻击方法下,本文方法具有更高的鲁棒性。特别是在基于显著性检测的剪切攻击下,现有方法基本上无法提取出水印信息,而本文方法仍然可以准确地提取出水印。本文的主要贡献如下:①将显著性检测用于确定水印的嵌入区域,有效解决了基于图像内容的裁剪攻击导致无法提取出水印的问题;②利用排列重组的简单策略实现水印图像冗余,进一步提升了水印嵌入方法的抗剪切攻击能力。

1 相关工作

目前针对抵抗剪切攻击的水印方法较少,现有水印方法仅是在训练过程中添加剪切的扰动以抵抗剪切攻击。但是这种做法忽略了大尺度剪切攻击对水印图像带来的影响。传统的水印方法通过将水印嵌入图像的特定位置以抵抗剪切攻击。例如Thahab等^[16]采用一种被称为“同心矩形”的嵌入方法,从图像中心开始,沿着边界向外嵌入数据,以克服剪裁和旋转攻击。基于深度学习的方法通过将水印进行冗余操作和在噪声层中增加剪切的扰动提高水印方法对剪切攻击的鲁棒性。例如Jia等^[17]将水印进行冗余后通过消息处理器将其扩散到整张图像上以提高水印对剪切攻击的鲁棒性。Ding等^[18]改进了StegaStamp方法^[19],在模型的训练过程中引入剪切的扰动,增强了模型抵抗剪切攻击的能力。虽然该方法具有较强的鲁棒性,但嵌入的水印信息仅20 bit。Ahmadi等^[20]提出了一种端到端的深度扩散水印框架,扩大了水印嵌入的容量,水印信息以图像的形式被嵌入,通过扩散水印的方式,在多个图像块之间共享水印数据,提高了水印鲁棒性。虽然这些深度方法可以在一定程度上抵御剪切攻击,但是在面对基于显著性检测的剪切时,现有方法提取的水印图像质量较差。本文将显著性区域作为水印嵌入区域,从语义层面上降低了由于剪切造成水印信息丢失的可能性,从而极大地提高了本文水印方法的抗剪切攻击能力。

2 本文方法

本文提出的基于显著性检测的抗剪切攻击的图像数字水印方法的框架如图1所示。首先使用显著性网络确定水印的嵌入区域,然后将冗余的水印图像嵌入图像中。为了提高鲁棒性,使用噪声层添加噪声,最终使用水印提取网络提取嵌入的水印。

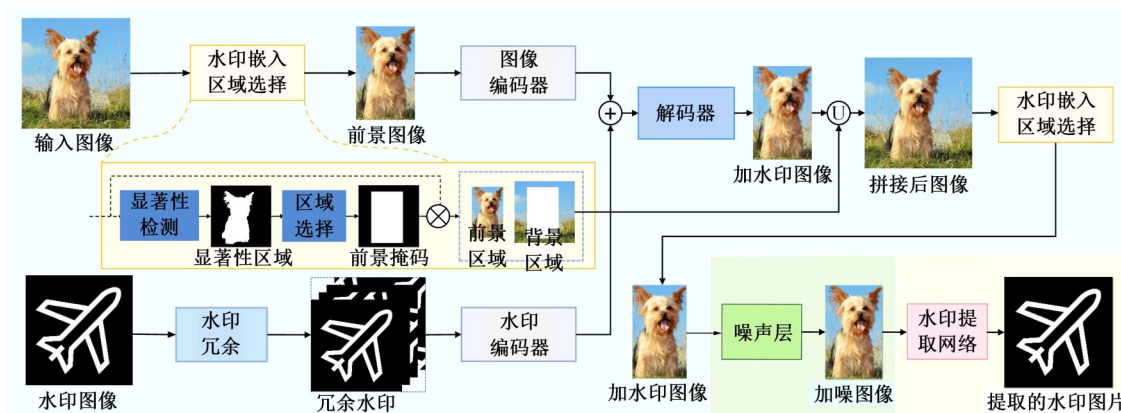


图1 基于显著性检测的抗剪切攻击的图像数字水印方法框架图(⊗代表点乘操作,⊕代表相加操作,Ⓜ代表粘贴操作)

Fig. 1 Framework of anti-cropping digital image watermarking based on saliency detection (⊗represents the operation of dot product, ⊕represents the operation of add, Ⓜrepresents the operation of pasting)

2.1 基于排列重组的水印冗余

剪切图像时,嵌入的水印信息同样被丢弃。基于重复水印信息的传统冗余方法只适用于嵌入水印信息少的情况,当嵌入的水印图像分辨率较高时,简单地重复水印不能有效实现冗余。本文提出一种简单的、适用于图像的水印冗余方法,即将图像分块后再排列,具体方法如图2所示。给定水印图像 $W \in \mathbb{R}^{h \times w}$, 将其分成4块,依次从左上角到右下角编号为 B_1, B_2, B_3, B_4 。重新排列后,

可以得到新的水印图像 W_i 。公式如下:

$$W_i = \mathcal{P}(W, r_i) \quad (1)$$

式中: $\mathcal{P}$ 为排列操作; r_i 为排列顺序。 $r_1 = [1, 2, 3, 4]$, $r_2 = [2, 1, 4, 3]$, $r_3 = [3, 4, 1, 2]$, $r_4 = [4, 3, 2, 1]$ 。

例如采用 r_2 的排列顺序时,水印图像从左上角到右下角依次为 B_2, B_1, B_4, B_3 。当裁剪导致 W_i 的部分信息丢失时,可由其他水印图像 $W_{j|j \neq i}$ 恢复丢失的信息。

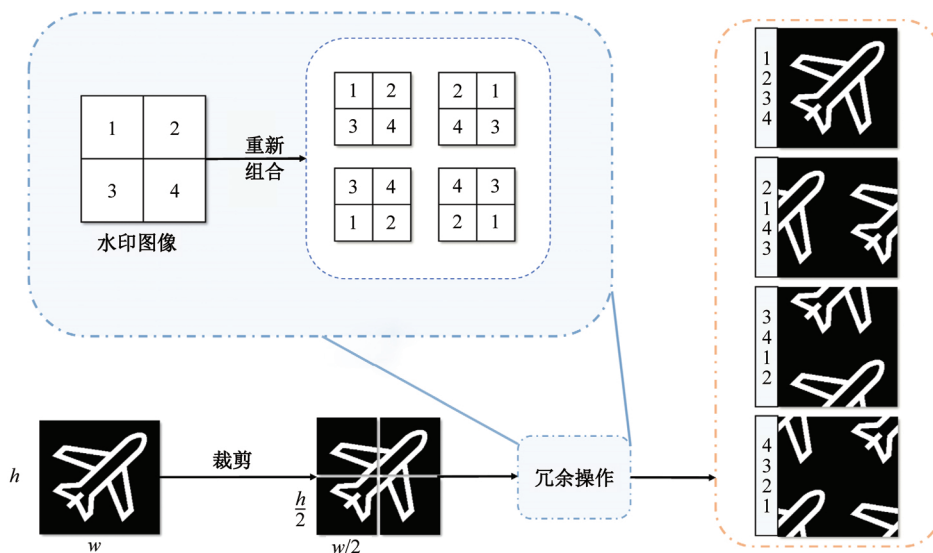


图2 基于排列重组的水印冗余操作示例

Fig. 2 Example of watermarking redundant operation based on permutation and recombination

2.2 基于显著性区域的水印嵌入区域选择

本文使用 PoolNet^[21] 提取图像中的显著性区域作为水印的嵌入区域。图像显著性从语义层面保证了水印被嵌入到图像重要区域,可以有效避免背景剔除或基于显著性的图像剪切导致水印信息丢失。给定显著性网络 $\Phi_s(\cdot; \theta_s)$, 原始图像 I_{co} 的

显著性检测结果记为 S_{co} 。使用阈值 T 得到显著性区域的掩码 M_{co} ($M_{co} = S_{co} \geq T$)。当显著性区域的最小外接矩形 B 的面积 $S \geq 0.4 \times H \times W$ 时,水印的嵌入区域为当前显著性区域的最小外接矩形区域,否则水印的嵌入区域为整张图像。实验发现: 嵌入水印后的图像 I_{em} 经过噪声层处理后的显著性

检测结果 S_{em} 可能与 S_{co} 之间存在较大差异,增加了水印提取的难度。因此,将 S_{co} 作为 S_{em} 的真值,微调显著性网络。损失函数采用二值交叉熵(Binary cross entropy, BCE)损失函数,定义如下:

$$\mathcal{L}_1 = \text{BCE}(S_{em}, S_{co}) \quad (2)$$

2.3 基于 Transfromer 的水印嵌入网络

水印嵌入网络将图像 I_{co} 和冗余的水印图像 $W_{i|i=1}^4$ 作为输入,输出嵌入水印的图像 I_{em} ,其详细结构如图 3 所示。使用多个 Transformer 模块提

取图像特征,并与水印特征相加后送入多层感知机进行融合。融合后的特征经过多个 Transformer 模块恢复出嵌入水印后的图像。具体过程公式化表示如下:

$$I_{em} = \Phi_E(I_{co}, W_{i|i=1}^4; \theta_E) =$$

Decoder(Encoder₁(I_{co}) + Encoder₂($W_{i|i=1}^4$)) (3)
式中:Decoder($\bullet$)为用于恢复图像信息的解码模块;Encoder₁($\bullet$)和 Encoder₂($\bullet$)两个编码模块分别用于编码图像和水印信息。

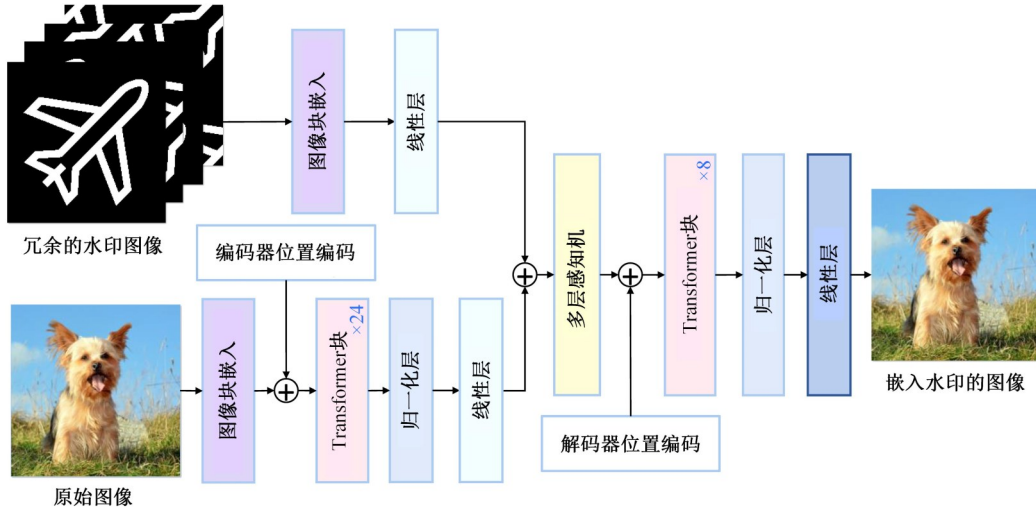


图 3 水印嵌入网络结构图

Fig. 3 Structure of watermark embedding network

图像编码模块 Encoder₁($\bullet$)由 24 个 Transformer 块和一个线性层组成,将图像编码成长度为 1 024 的向量,其公式表示为:

$$v_{im} = FC(N(TB_{24}(P\epsilon(I_{co}) + p_{en}))) \quad (4)$$

式中:FC($\bullet$)为全连接层;N($\bullet$)为层规范化操作;TB₂₄($\bullet$)为由 24 个连续的 Transformer 块组成的网络;P ϵ ($\bullet$)为分块操作;p_{en}为位置编码。

水印编码模块 Encoder₂($\bullet$)由一个全连层将水印转换成长度为 1 024 的向量,其公式化表示为:

$$v_w = FC(P\epsilon(W_{i|i=1}^4)) \quad (5)$$

解码模块 Decoder($\bullet$)由一个多层感知机、8 个 Transformer 块、一个全连层组成,其公式化表示为:

$$I_{em} = FC(N(TB_8(MLP(v_{im} + v_w) + p_{en}))) \quad (6)$$

为了保证嵌入水印的图像 I_{em} 与原图像 I_{co} 的相似性,本文采用均方误差(Mean squared error, MSE)损失函数,定义如下:

$$\mathcal{L}_2 = \text{MSE}(I_{em}, I_{co}) \quad (7)$$

2.4 基于全卷积的水印提取网络

水印提取网络 $\Phi_D(\bullet; \theta_D)$ 如图 4 所示,包括特

征提取部分和水印提取部分。特征提取部分采用 UNet 结构,使用最后一层卷积特征提取水印。水印提取部分采用 4 个独立的提取分支,分别提取冗余的水印图像,然后再使用卷积将冗余的水印图像合并为最终的水印图像。使用 X 表示图像特征,每一个分支使用卷积、批处理规范化、Sigmoid 操作提取水印图像,具体公式如下:

$$W_i' = \sigma(B(C_i(X))) \quad (8)$$

式中: $C_i(\bullet)$ 为第 i 个分支的卷积; $B(\bullet)$ 为批处理规范化; $\sigma(\bullet)$ 为 Sigmoid 函数。

$W_{i|i=1}^4$ 对应水印冗余模块中 4 个重新排列后的水印图像。通过融合重新排列的 $W_{i|i=1}^4$,可以准确地恢复水印图像,其计算公式如下:

$$\hat{W} = C(R(C([P(W_i', r_i')]|_{i=1}^4))) \quad (9)$$

式中: $C(\bullet)$ 为卷积操作; $R(\bullet)$ 为 ReLU 激活函数; $P(\bullet, r_i')$ 为将排列顺序 r_i' 还原成 r_i ; $[\cdot]_{i=1}^4$ 为拼接操作。

本文采用均方误差损失函数训练水印提取网络,其损失函数定义如下:

$$\mathcal{L}_3 = \text{MSE}(\hat{W}, W) + \sum_{i=1}^4 \text{MSE}(W_i', W_i) \quad (10)$$

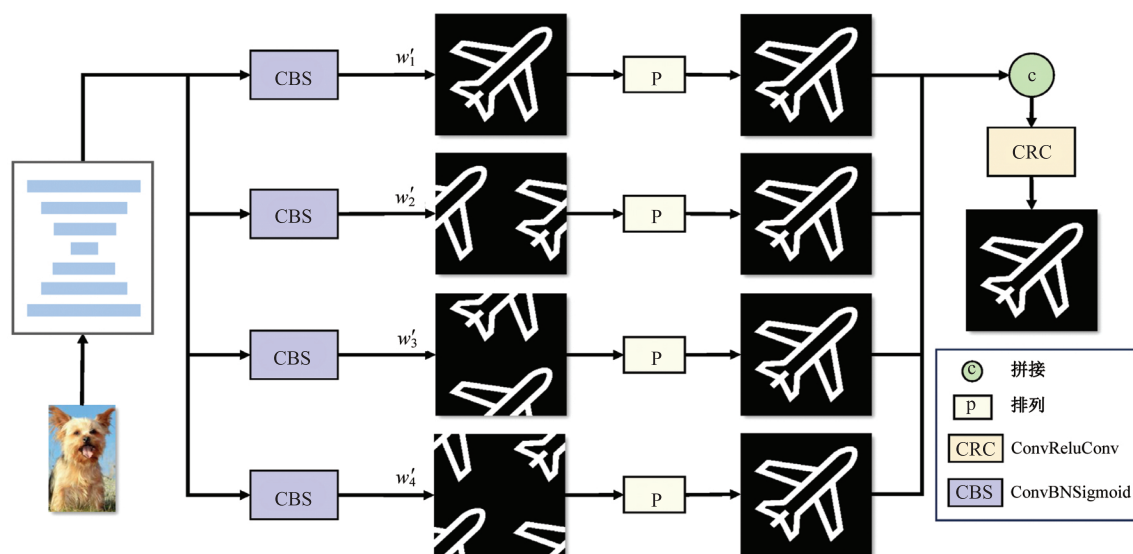


图 4 水印提取网络结构图

Fig. 4 Structure of watermark extraction network

为了防止水印提取网络过拟合,本文使用对抗训练的方式,将无水印图像 I_{co} 作为输入,全零图作为真值,得到额外的损失函数:

$$\mathcal{L}_4 = \text{MSE}(\hat{W}, 0) + \sum_{i=1}^4 \text{MSE}(W'_i, 0) \quad (11)$$

基于式(2)(7)(10)(11),可得整体损失函数定义为:

$$\mathcal{L}_{\text{Total}} = \lambda_1 \mathcal{L}_1 + \lambda_2 \mathcal{L}_2 + \lambda_3 \mathcal{L}_3 + \lambda_4 \mathcal{L}_4 \quad (12)$$

式中: $\lambda_i (i=1, 2, 3, 4)$ 是平衡系数,分别为 1、7、35、20。

本文方法使用 Pytorch 实现,硬件设备使用 NVIDIA 3090,采用 Adam 优化算法,学习率设置为 0.000 01,批处理大小设置为 2,模型训练 30 轮。

3 实验

3.1 实验设置

数据集: 由于没有公开的水印数据集,本文从 ImageNet^[22] 数据集中随机挑选 100 类图像用于训练和测试,将其命名为 ImageNet 100,其中训练集有 128 967 张图像,测试集有 5 000 张图像。水印图像为一张大小为 224×224 像素的黑白图像。

对比方法: 实验选取了传统的水印嵌入方法 DIWm^[8] 和基于深度学习的水印嵌入方法 ReD-Mark^[20] 以及 DAH-Net^[12] 进行对比。DIWm 的原始图像是尺寸为 512×512 像素的灰度图,提取的水印图像为 256×256 像素。ReD-Mark 的原始图像是 512×512 像素的灰度图,提取的水印图像是尺寸为 224×224 像素的灰度图。DAH-Net 的原

始图像是 256×256 像素的三通道 RGB 图像,提取的水印图像是尺寸为 256×256 像素的灰度图。为了公平比较,使用相同的水印图片训练模型,在比较时将水印图像的真值分别缩放到 DIWm、ReD-Mark 和 DAH-Net 提取水印的目标尺寸。

量化指标: 本文采用峰值信噪比 (Peak signal to noise ratio, PSNR) 和结构相似性指数 (Structural similarity index, SSIM) 衡量嵌入水印后图像的质量,使用比特误码率 (Bit error ratio, BER) 和 SSIM 衡量水印图像的质量。

3.2 结果分析

3.2.1 无攻击的图像和水印质量比较

表 1 对比了不同水印方法嵌入水印后的图像质量和提取水印的质量。可以发现,DIWm 嵌入水印后图像的 PSNR 和 SSIM 高于 ReD-Mark 和本文方法,达到了 32.13 和 0.96。然而 DIWm 和 ReD-Mark 只能在灰度图像上嵌入水印。本文方法提取的水印图像质量明显优于 DIWm 和 ReD-Mark, BER 低至 0.002, SSIM 达到了 0.992 4,但相比于 DAH-Net 低了 0.006 7。表 1 中的量化结果表明本文方法具有较好的鲁棒性和不可见性。

图 5 展示了不同水印方法嵌入水印后的图像和提取的水印图像。在没有噪声攻击的情况下,所有水印方法都可以提取出水印图像。ReD-Mark 提取的水印图像质量较差,其中的飞机轮廓不清晰且存在断裂。DIWm^[8] 在灰度图像上嵌入水印,虽然可以提取出水印的轮廓,但水印图像前景区域像素值较低,且背景区域不干净。本文方

表 1 无攻击嵌入水印图像的 PSNR 和 SSIM 及提取的水印图像对应的 BER 和 SSIM

Table 1 PSNR and SSIM of the watermarked image and the BER and SSIM of the extracted watermark without attack

方法	嵌入水印的图像		提取的水印图像	
	PSNR	SSIM	BER	SSIM
DIWm	32.13	0.96	0.124 7	0.817 6
ReDMark	25.03	0.77	0.125 4	0.696 2
DAH-Net	30.17	0.94	0.014 3	0.999 1
本文	31.50	0.92	0.002 0	0.992 4

注:表中加粗字体表示该值的最优解,后同。

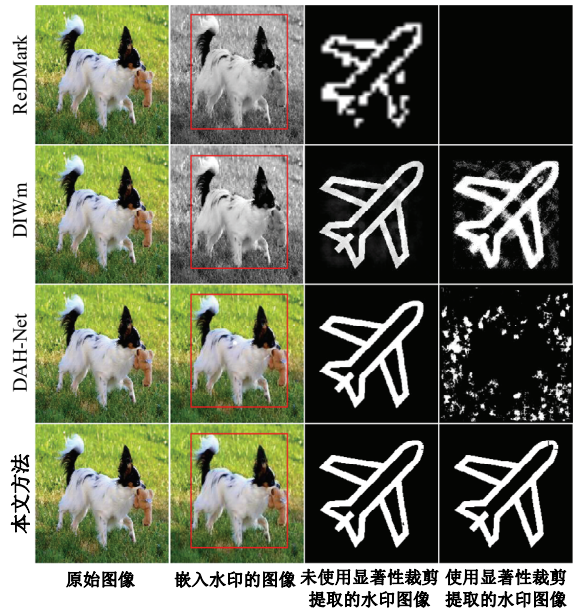


图 5 基于显著性检测的剪切攻击后提取的水印图像

Fig. 5 Extracted watermark under cropping attack based on saliency detection

法和 DAH-Net 不仅可以在彩色图像上嵌入水印,且提取的水印图像质量均较高。

3.2.2 不同攻击方法的水印质量比较

除了对图像进行显著性裁剪的攻击外,本文水印提取方法还可以抵抗其他图像攻击,如高斯滤波、裁剪、中值滤波等。从表 2 中可以看出,使用噪声层本文方法可以有效对抗椒盐噪声、裁剪、像素替换、高斯滤波、中值滤波、高斯噪声、JPEG 压缩和缩放等攻击。在椒盐噪声、裁剪和高斯噪声攻击下,本文方法的 BER 在不同攻击参数下均低于 0.005,并且 SSIM 均高于 0.97。与之相比,DIWm、ReDMark 和 DAH-Net 的 BER 最低为 0.085,SSIM 最高仅为 0.856 6。在椒盐噪声和高斯噪声的攻击下,DIWm 的 SSIM 甚至只有约 0.3,表明其提取的水印图像受到了严重的噪声干扰。在裁剪替换攻击后,本文方法在攻击参数为

0.1 时,SSIM 仅为 0.694 2,这表明在当前参数下基本无法提取水印。但是,当攻击参数增至 0.2 和 0.3 时,SSIM 的数值有所上升,这表明该方法对较大程度的裁剪替换攻击有一定的鲁棒性。然而,DIWm、ReDMark 和 DAH-Net 在这两种参数的攻击下,SSIM 均明显低于本文方法。在 JPEG 压缩的攻击下,DIWm 的鲁棒性最强,在 3 种不同质量参数攻击下的效果均优于 ReDMark、DAH-Net 和本文方法。特别是当质量参数为 50 时,本文方法的 SSIM 仅为 0.687 3,而 DIWm^[8] 的 SSIM 可以达到 0.857 3,但是当质量参数增加时,本文方法也可以在一定程度上提取出水印。可视化结果图如图 6 所示,可以看出在面对大尺度的裁剪攻击时,DIWm、ReDMark 和 DAH-Net 基本提取不出水印,而本文方法虽然在攻击系数为 0.1 的裁剪替换时所提取的水印有所缺失,但在其他参数的攻击下依然能够提取出高质量的水印图像。在高斯滤波、高斯噪声和椒盐噪声的攻击下,DIWm 提取的水印图像有很多噪点,但是本文方法提取出的水印图像背景清晰没有噪声。本文的推理时间为每张图像 0.15 s,虽多于 ReDMark 的每张图像 0.06 s,但是水印的鲁棒性远高于 ReDMark。

3.2.3 基于显著性检测的剪切水印图像质量比较

为了检验不同水印嵌入方法抵抗基于显著性检测的剪切攻击的能力,本文使用 PoolNet^[21] 作为图像显著性区域提取的基础网络。选取图像显著性区域的最小外接矩形作为剪切区域,用于提取水印图像。从表 3 中可以看出,当水印换成了 224×224 像素的飞机图案后,DAH-Net 提取出的水印图像接近嵌入的水印图像,但 ReDMark 和 DIWm 在嵌入水印的图像上提取的水印图像质量都不高,在没有对嵌入水印的图像进行显著性剪切的攻击时,ReDMark 的 SSIM 仅为 0.696 2,DIWm 的 SSIM 仅为 0.817 6。并且经过显著性剪切的攻击后,二者的 BER 都有显著上升。值得注意的是,采用 ReDMark 方法对图像进行显著性裁剪攻击后,水印图像的 SSIM 只下降了 0.014 7,这是因为在很多张裁剪完显著性区域的嵌入水印的图像中,ReDMark 方法是提不出水印图像的,它们的水印图是一张全黑的图像,全黑图像和要嵌入的水印图像的 SSIM 仅有 0.678 4。

表 2 使用不同攻击方法得到的被攻击水印图像对应的 BER、SSIM 以及不同方法的推理时间

Table 2 BER and SSIM of the attacked watermarked images obtained by different attack methods and inference time of different methods

攻击方法	参数	DIWm		ReDMark		DAH-Net		本文方法	
		BER	SSIM	BER	SSIM	BER	SSIM	BER	SSIM
椒盐噪声	0.4	0.367 7	0.314 4	0.135 8	0.690 2	0.226 1	0.378 8	0.004 5	0.973 7
	0.3	0.347 9	0.316 9	0.130 3	0.695 0	0.225 8	0.380 9	0.004 5	0.974 3
	0.2	0.314 0	0.325 6	0.127 4	0.697 1	0.224 8	0.385 4	0.004 5	0.975 3
剪切	0.6	0.194 4	0.596 6	0.163 6	0.682 7	0.085 0	0.856 6	0.003 7	0.982 4
	0.5	0.193 0	0.604 0	0.169 2	0.681 0	0.120 5	0.779 8	0.003 9	0.981 9
	0.4	0.191 6	0.612 7	0.173 3	0.679 9	0.152 5	0.706 7	0.003 9	0.980 8
裁剪替换	0.3	0.169 3	0.583 4	0.176 8	0.679 0	0.204 8	0.518 5	0.097 1	0.769 1
	0.2	0.176 2	0.524 3	0.179 1	0.678 5	0.227 7	0.457 5	0.125 9	0.736 4
	0.1	0.183 7	0.451 8	0.180 1	0.678 4	0.240 2	0.425 4	0.165 6	0.694 2
像素替换	0.4	0.139 9	0.804 7	0.138 0	0.690 4	0.102 2	0.725 6	0.021 9	0.949 2
	0.3	0.146 0	0.783 5	0.151 5	0.683 2	0.141 2	0.626 6	0.037 7	0.918 5
	0.2	0.142 0	0.748 4	0.171 3	0.679 7	0.183 5	0.533 5	0.065 0	0.868 8
高斯滤波	3	0.219 1	0.292 7	0.180 1	0.678 4	0.129 9	0.634 2	0.001 7	0.991 0
	2	0.201 0	0.304 9	0.180 1	0.678 4	0.084 3	0.744 3	0.009 8	0.963 9
	1	0.175 0	0.385 0	0.180 1	0.678 4	0.021 8	0.985 4	0.002 9	0.989 1
中值滤波	7	0.184 5	0.347 1	0.180 1	0.678 4	0.184 3	0.532 9	0.040 4	0.910 2
	5	0.167 3	0.391 3	0.180 1	0.678 4	0.113 6	0.669 4	0.025 4	0.926 2
	3	0.144 9	0.501 9	0.180 1	0.678 4	0.032 2	0.943 7	0.003 3	0.987 7
高斯噪声	0.2	0.348 6	0.318 9	0.180 1	0.678 4	0.224 5	0.380 9	0.004 5	0.975 3
	0.1	0.307 7	0.330 3	0.180 1	0.678 4	0.223 7	0.383 4	0.004 3	0.977 3
	0.05	0.255 0	0.357 8	0.180 1	0.678 4	0.222 3	0.386 8	0.003 7	0.980 3
JPEG 压缩	50	0.099 5	0.857 3	0.134 9	0.687 8	0.224 3	0.446 3	0.173 1	0.687 3
	70	0.101 2	0.868 6	0.129 2	0.692 2	0.226 4	0.443 4	0.175 9	0.683 7
	90	0.091 0	0.932 8	0.126 4	0.697 7	0.230 0	0.438 9	0.085 7	0.835 6
缩放	1.2	0.132 8	0.602 1	0.170 8	0.677 9	0.027 5	0.948 9	0.002 5	0.990 3
	1.1	0.126 1	0.674 4	0.174 0	0.678 8	0.032 8	0.895 8	0.002 3	0.990 5
	0.9	0.135 3	0.593 7	0.179 3	0.678 5	0.033 7	0.891 2	0.002 3	0.990 9
时间/(张·s ⁻¹)		0.29		0.06		0.35		0.15	

3.3 消融实验

为了验证本文方法提出的显著性网络、水印图像冗余模块的有效性,针对以上两个模块进行了以下消融实验。

3.3.1 显著性网络

在本文方法中,对原始图像添加水印之前,首先进行一次显著性检测,以提取图像的显著性区域。在水印提取过程中,采用在添加水印时基于显著性进行微调的显著性网络,以准确定位水印提取的区域。经过显著性网络裁剪后,在这个显著性区域进行水印的提取。实验结果(见表4)显示,在使用显著性网络和未使用显著性网络的情况下,二者嵌入水印的图像质量相差不大。但是在嵌入水印的图像中先提取显著性区域再提取水印时,未使用显著性网络提取的水印图像的SSIM为0.975 9,相比于使用了显著性网络的结果低了0.016 5。同时,未使用显著性网络提取的

水印图像的BER也比使用了显著性网络提取的水印图像高出0.004。从图7可以看出,在未使用显著性网络时,经过显著性网络裁剪后再提取水印时,所提水印会丢失掉边缘的信息。使用显著性网络进行训练后,在嵌入水印的图像中提取显著性区域后再提取水印时可以提取出较高质量的水印图像。实验表明,显著性模块能够有效地将水印集中在图像的显著性区域。

3.3.2 水印冗余模块

由于编码器采用Transformer框架,水印按照编码顺序嵌入原始图像,所以该方法对于剪切攻击的鲁棒性较差。因此,本文方法设计了水印冗余模块,水印在经过水印冗余模块后生成4张重新排列的冗余图像。在水印提取阶段,对提取的4张图像进行重新裁剪,并按照原来的拼接顺序重新组合。通过水印融合模块将4张提取的水印图片融合成一张完整的水印图片。为了验证水印



图 6 不同方法受到噪声攻击后的加噪图像和提取的水印图像

Fig. 6 Noised image and extracted watermark compared with different methods under different attacks

表 3 不同方法受到基于显著性检测的剪切攻击时提取的水印图像质量对比

Table 3 Comparison of the quality of watermark extracted by different methods under cropping attack based on saliency detection

方法	显著性裁剪攻击	BER	SSIM
DIWm	是	0.205 7	0.571 4
	否	0.124 7	0.817 6
ReDMark	是	0.170 4	0.681 5
	否	0.125 4	0.696 2
DAH-Net	是	0.242 9	0.419 6
	否	0.014 3	0.999 1
本文	是	0.002 0	0.992 4
	否	0.000 3	0.998 2

冗余模块的有效性,去掉水印冗余模块,直接使用水印图像作为水印编码器的输入,保持其他设置

表 4 显著性网络的消融实验

Table 4 Ablation experiments of saliency networks

实 验	嵌入水印的图像		提取的水印图像	
	PSNR	SSIM	BER	SSIM
未使用显著性网络	32.210 8	0.916 7	0.006 0	0.975 9
使用显著性网络	31.495 0	0.920 8	0.002 0	0.992 4

不变重新训练水印嵌入网络和提取网络,其结果如表 5 所示。可以发现,移除水印冗余模块前后,本文方法嵌入水印的图像其 PSNR 和 SSIM 基本上没有发生变化,但是使用水印冗余模块后提取的水印图像 SSIM 比未使用水印冗余模块提取的水印图像 SSIM 高了 0.079 5,并且 BER 也降低了 0.03。在测试过程中分别剪切掉嵌入水印的图像,左上角、左下角、右上角和右下角的 1/4 部分

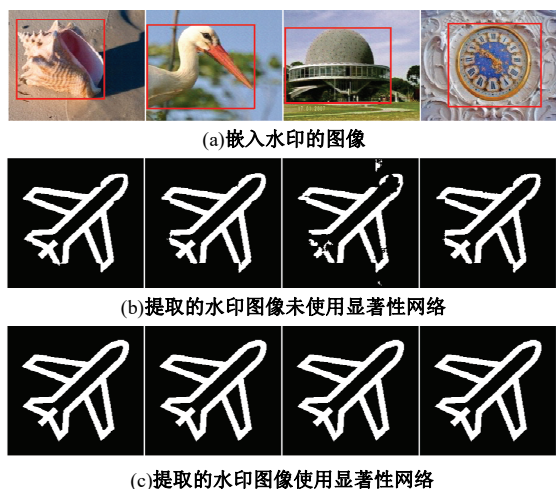


图 7 显著性网络消融实验

Fig. 7 Ablation experiments of saliency network
后,如图 8 所示,未使用水印冗余模块时,分别裁剪掉这 4 个部分会导致其所提取水印对应位置的信息丢失,但是使用水印冗余模块后,丢失的部分可以通过冗余的信息进行补全。

表 5 水印冗余模块的消融实验

Table 5 Ablation experiments of watermark redundant modules

实 验	嵌入水印的图像		提取的水印图像	
	PSNR	SSIM	BER	SSIM
未使用水印冗余模块	31.670 3	0.918 4	0.032 0	0.912 9
使用水印冗余模块	31.495 0	0.920 8	0.002 0	0.992 4

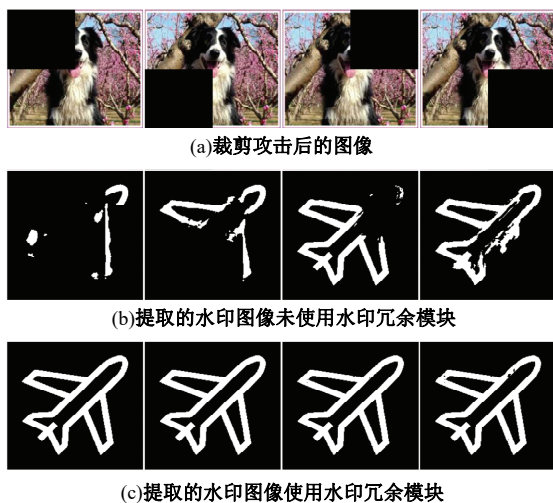


图 8 水印冗余模块的消融实验

Fig. 8 Ablation experiments of watermark redundant modules

4 结束语

本文提出了一种具有较强抗剪切攻击能力的数字图像水印方法。利用显著性检测选取图像中

语义信息丰富的图像区域作为嵌入区域,从语义层面降低图像嵌入水印区域被剪切的可能性。此外,本文设计一种基于排列重组的水印冗余方法。利用 Transformer 网络将冗余水印嵌入图像中,并使用卷积网络提取水印图像。提取时通过冗余水印的融合,进一步提高了水印图像的质量。通过在 ImageNet 上随机选取 100 类图像数据进行训练和测试,结果表明,本文方法可以有效抵抗 9 种常见的攻击方法。在对比方法无法抵抗基于图像内容的剪切攻击时,本文方法依然可以准确地提取水印图像。大量实验表明本文方法的有效性。

参考文献:

- [1] Van Schyndel R G, Tirkel A Z, Osborne C F. A digital watermark[C]//Proceedings of 1st International Conference on Image Processing, Austin, USA, 1994: 86-90.
- [2] Ko H J, Huang C T, Horng G, et al. Robust and blind image watermarking in DCT domain using inter-block coefficient correlation[J]. Information Sciences, 2020, 517: 128-147.
- [3] Cedillo-Hernandez M, Cedillo-Hernandez A, Garcia-Ugalde F J. Improving dft-based image watermarking using particle swarm optimization algorithm[J]. Mathematics, 2021, 9(15): No. 1795.
- [4] Giri K J, Quadri S M K, Bashir R, et al. DWT based color image watermarking: a review[J]. Multimedia Tools and Applications, 2020, 79: 32881-32895.
- [5] Assini I, Badri A, Safi K, et al. A robust hybrid watermarking technique for securing medical image[J]. International Journal of Intelligent Engineering and Systems, 2018, 11(3): 169-176.
- [6] 李喜艳, 周夏冰, 刘征. 基于高容量强鲁棒的水印算法[J]. 吉林大学学报: 工学版, 2022, 52(1): 174-179.
Li Xi-yan, Zhou Xia-bing, Liu Zheng. High-capacity and robust image watermarking algorithm[J]. Journal of Jilin University (Engineering and Technology Edition), 2022, 52(1): 174-179.
- [7] Hsu C S, Tu S F. Enhancing the robustness of image watermarking against cropping attacks with dual watermarks[J]. Multimedia Tools and Applications, 2020, 79(17/18): 11297-11323.
- [8] Liu J, Huang J, Luo Y, et al. An optimized image watermarking method based on HD and SVD in DWT domain[J]. IEEE Access, 2019, 7: 80849-

- 80860.
- [9] Zhu J, Kaplan R, Johnson J, et al. Hidden: hiding data with deep networks[C]//Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 2018: 657-672.
- [10] Lu J, Ni J, Su W, et al. Wavelet-based CNN for robust and high-capacity image watermarking[C] // IEEE International Conference on Multimedia and Expo (ICME), Taipei, China, 2022: 1-6.
- [11] Liu Y, Guo M, 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, France, 2019: 1509-1517.
- [12] Zhang L, Lu Y, Li J, et al. Deep adaptive hiding network for image hiding using attentive frequency extraction and gradual depth extraction[J]. *Neural Computing and Applications*, 2023, 35(15): 10909-10927.
- [13] Marchesotti L, Cifarelli C, Csürka G. A framework for visual saliency detection with applications to image thumbnailing[C] //IEEE 12th International Conference on Computer Vision, Kyoto, Japan, 2009: 2232-2239.
- [14] 刘洋, 毛克明. 基于自适应反馈机制的小差异化图像纹理特征信息数据检索[J]. *江苏大学学报: 自然科学版*, 2025, 46(1): 73-81.
Liu yang, mao Ke-ming. Retrieval of texture feature information data for small differentiated images based on adaptive feedback mechanism[J]. *Journal of Jiangsu University(Natural Science Edition)*, 2025, 46(1): 73-81.
- [15] Tu Y, Niu L, Zhao W, et al. Image cropping with composition and saliency aware aesthetic score map [C]//Proceedings of the AAAI Conference on Artificial Intelligence, New York, USA, 2020: 12104-12111.
- [16] Thahab A T, Shahadi H I. Robust image watermarking against crop and rotation attacks for public communication networks[C]//Third Scientific Conference of Electrical Engineering, Baghdad, Iraq, 2018: 105-110.
- [17] Jia Z, Fang H, Zhang W. Mbrs: enhancing robustness of dnn-based watermarking by mini-batch of real and simulated jpeg compression[C]//Proceedings of the 29th ACM International Conference on Multimedia, Chengdu, China, 2021: 41-49.
- [18] Ding J, Cao Y, Shi C. CropDefender: deep watermark which is more convenient to train and more robust against cropping[J/OL]. [2024-03-05]. <https://doi.org/10.48550/arXiv.2109.06651>
- [19] Tancik M, Mildenhall B, Ng R. Stegastamp: invisible hyperlinks in physical photographs[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, USA, 2020: 2117-2126.
- [20] Ahmadi M, Norouzi A, Karimi N, et al. ReDMark: framework for residual diffusion watermarking based on deep networks[J]. *Expert Systems with Applications*, 2020, 146: No. 113157.
- [21] Liu J J, Hou Q, Cheng M M, et al. A simple pooling-based design for real-time salient object detection [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, USA, 2019: 3912-3921.
- [22] Deng J, Dong W, Socher R, et al. ImageNet: a large-scale hierarchical image database[C] //IEEE Conference on Computer Vision and Pattern Recognition, Miami, USA, 2009: 248-255.