Beijing Forest Studio

北京理工大学信息系统及安全对抗实验中心

模型水印攻击方法

硕士研究生 邢凤桐

2024年08月18日

问题回溯

- 总结反思
  - 讲解语速较快，内容详略不当
  - 未考虑听众的观感和接受程度
  - 部分图表缺少关键性说明和解释

- 相关内容
  - **2024.05.19 李嘉玮《深度学习模型后门攻击检测》**
  - **2023.11.12 邢凤桐《DNN模型水印及其鲁棒性评估》**
  - **2023.03.12 邢凤桐《深度神经网络模型水印保护方法》**

内容提要

- 预期收获
- 题目内涵解析
- 研究背景与意义
- 研究历史与现状
- 知识基础
- 算法原理
  - RD-IWAN
  - HIWANet
- 特点总结与工作展望
- 参考文献

背景简介

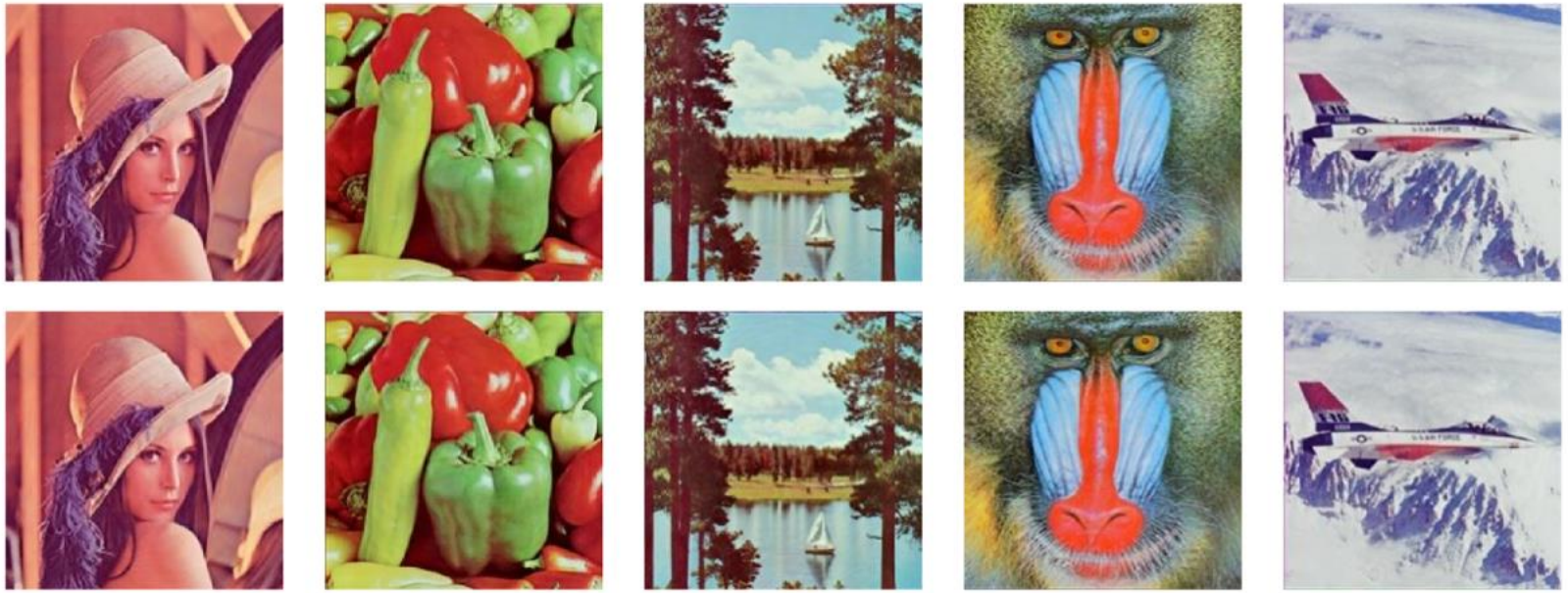
- 预期收获
  - 掌握模型水印攻击的基本概念
  - 了解模型水印攻击的历史现状及应用场景
  - 理解模型水印攻击的技术原理
  - 明确模型水印攻击的发展趋势和未来前景

内涵解析与研究目标

- 题目内涵解析（模型水印攻击方法）
  - 水印：一种在数字媒体中嵌入可见或不可见标记的技术
  - 模型水印：一种隐藏在模型中且不影响模型本身功能的特定信息
  - 模型水印攻击：一种针对机器学习模型中水印的攻击方式，旨在干扰、破坏或绕过嵌入在模型中的水印信息
  - 模型水印攻击方法：参数修改、对抗样本生成、模型剪枝等方法或手段，以破坏或绕过模型中的水印信息
- 研究目标
  - 面向人工智能领域模型知识产权保护
  - 研究模型水印信息保护、模型水印检测、模型水印鲁棒性等关键问题
  - 结合深度学习、模型水印嵌入、模型窃取攻击与防御等理论
  - 改进模型水印技术，实现模型水印准确性和鲁棒性的显著提升

研究背景

- 研究背景
  - 机器学习模型在商业和研究领域广泛应用，其训练过程繁琐、花销昂贵
  - 模型水印是一种常用的保护模型知识产权的方法
  - 对抗攻击、隐私泄露、非法侵权、数据滥用
  - 模型水印攻击破坏水印信息的完整性，所有权验证过程受到影响



研究意义

- 研究意义
  - 知识产权保护：提高对机器学习模型知识产权的保护水平，防止模型被盗用或未经授权使用
  - 模型安全性：提高对机器学习模型的安全性认识，加强对潜在攻击的防范和对抗能力
  - 隐私保护：强化模型隐私保护机制，确保模型中的敏感信息不会被泄露或滥用

  - 技术挑战：为解决复杂的技术挑战提供实践机会，推动了模型水印技术和对抗技术的发展和进步
  - 商业应用：保护商业利益，确保模型的合法使用和收益

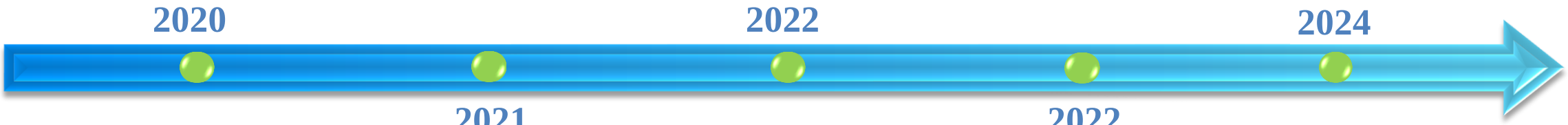
模型水印攻击研究致力于优化模型水印算法！

研究历史　模型水印攻击

Geng等人根据水印的知识量，构建相应的水印图像数据集，训练CNN模型以使用这些数据集去除水印

Wang等人结合渐进式预处理方法，设计了水印攻击残差密集网络WARDN，结合了感知损失和MSE损失进行优化，有效去除水印图像中低频特征中的水印信息

Wang等人结合特征提取模块FEM和水印攻击模块WAM学习图像的高级抽象特征，设计非对称损失函数保持受攻击的水印图像的质量

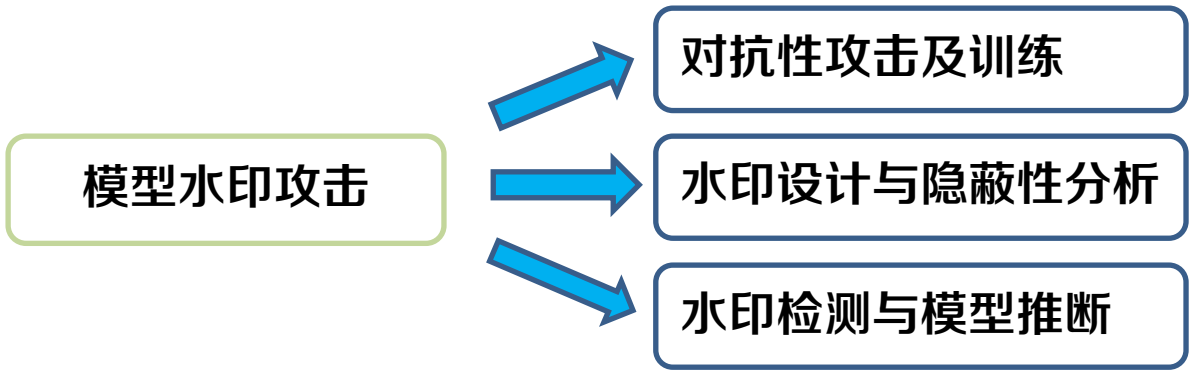


2021

Hatoum等人研究了全卷积神经网络FCNN作为去噪攻击对水印图像的影响，改善了训练过程和去噪性能，保留图像详细结构去除了噪声

2022

Li等人利用具有跳跃连接的编码器-解码器构成生成器，确保生成图像的不可感知性，引入基于特征提取的感知损失和判别网络，使生成的图像的外观和分布与原始图像相似，有效去除水印信息



研究现状

- 模型水印算法改善
  - 模型水印的鲁棒性不足，易受到对抗攻击的破坏，需改进算法以增强水印的鲁棒性
    - 设计更加鲁棒的模型水印嵌入算法、提高水印检测的准确性等，以增强模型水印的安全性
  - 模型水印的隐蔽性不足，易被攻击者检测和破坏水印，需设计更具隐蔽性的水印方案
    - 设计更加隐蔽的模型水印嵌入算法、使得水印信息更难以被攻击者发现或破坏
  - 模型水印检测的准确性不足，存在漏检或误判的情况，需提高检测算法的准确性
- 对抗攻击及防御
  - 对抗攻击手段多样化
    - 对抗样本攻击、模型剪枝攻击、参数修改攻击等，旨在破坏模型中的水印信息
  - 对抗攻击防御全面化
    - 模型水印攻击研究促进了防御方法的改进和完善

知识基础　模型水印攻击

- 模型水印
  - 基本概念：一种隐藏在模型中且不影响模型本身功能的特定信息
  - 原理：通过修改模型参数（内部结构、输入输出等）让模型过拟合到只有模型所有者知道的异常输入输出关系，用来宣称模型的所有权
  - 目的：防止模型被窃取，保护模型的知识产权
  - 分类
    - 可见水印、不可见水印
    - 白盒水印、黑盒水印、灰盒水印、无盒水印
  - 模型水印攻击
    - 基本概念：一种针对机器学习模型中水印的攻击方式
    - 目的：破坏或篡改模型水印，损害水印的完整性或可信度，使其验证失效

模型水印攻击旨在破坏模型中的水印，模糊模型的所有权验证

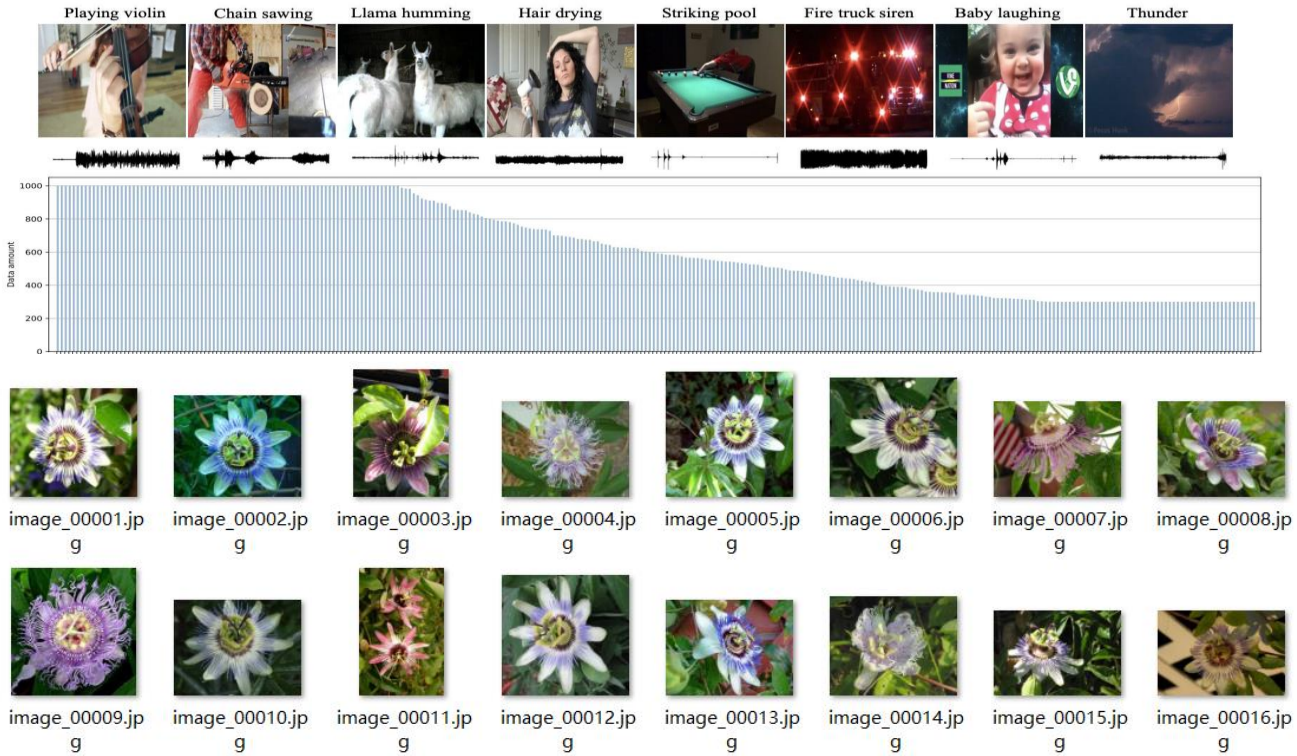
知识基础　模型水印攻击分类

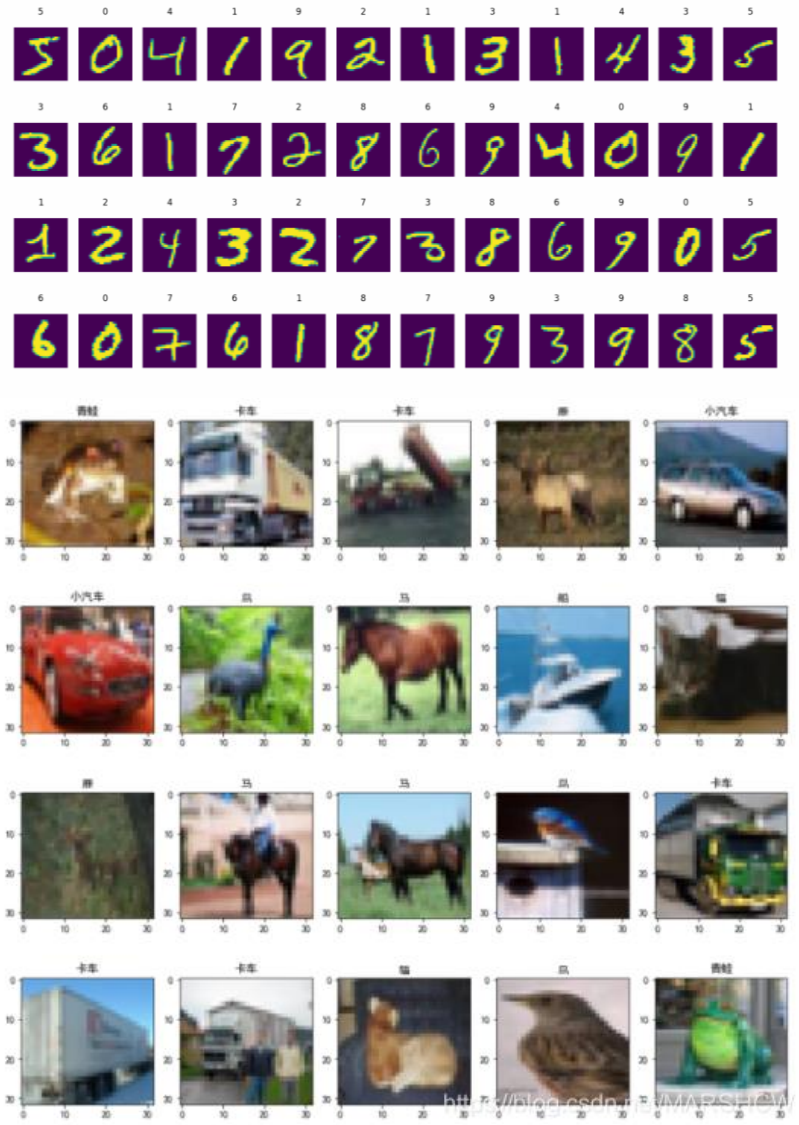
- 模型水印攻击分类（从攻击方法角度）
  - 移除攻击（Removal Attacks）：完全删除水印信息
    - 利用各种技术手段和算法来从数字内容中去除水印，使水印不可见或无法识别
  - 修改攻击（Modification Attacks）：对水印进行篡改或修改
    - 通过修改水印信息改变水印的内容或位置，破坏水印的完整性和可靠性
  - 伪造攻击（Forgery Attacks）：伪造一个合法的水印
    - 误导接收方，使其接受伪造的水印信息
  - 干扰攻击（Interference Attacks）：向数字内容中添加干扰或噪声来破坏水印信息
    - 使水印信息变得不可靠或无法识别
  - 定位攻击（Localization Attacks）：识别和定位数字水印的位置或特征
    - 更容易对水印进行移除、修改或伪造

模型水印攻击可通过多种方式干扰或破坏模型水印验证过程

知识基础 数据源

- 数据源说明
  - 图片、音频、视频、文本等
  - MNIST、CIFAR10、ImageNet 、SST-2…
  - 自定义数据集





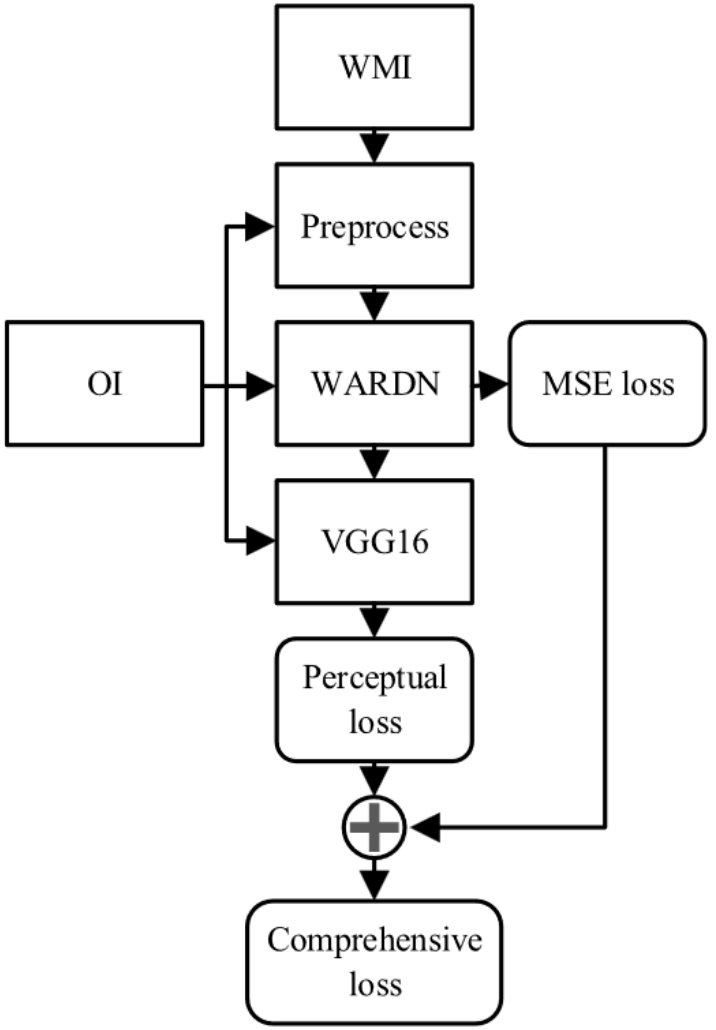
【IEEE Transactions on Circuits and Systems for Video Technology】

RD-IWAN: Residual Dense Based Imperceptible Watermark Attack Network

T	目标	提高水印攻击方法的不可感知性
I	输入	1个原始模型、2个图像数据集（2000个图像）、4个水印算法
P	处理	1. 利用渐进式预处理实现水印增强，使网络准确学习水印信息 2. 基于水印攻击残差密集网络去除水印图像的中低频特征的水印信息 3. 利用MSE损失和VGG感知损失对模型进行优化
O	输出	2000个攻击后的水印图像

P	问题	攻击带水印的图像(WMI)时，易造成图像质量的严重下降
C	条件	黑盒水印情景，模型相关信息和架构不可见
D	难点	攻击方法对原始图像(OI)的视觉质量损害降低
L	水平	TCSVT 2022（CCF B类）

- 算法核心
  - 原理：以原始图像为优化目标，放大加水印图像和原始图像间的差异，通过神经网络将水印图像恢复为原始图像，从而攻击水印信息
    - 不向图像中添加噪声，而处理水印所在的图像的低频
  - 渐进式预处理（水印增强）
  - 水印攻击网络（WARDN）
    - 残余密集网络RDN
  - 损失函数优化（综合损失）
    - MSE损失（逐像素）
    - 感知损失（语义特征）
      - 比较图像特征，重建更多的细节和边缘信息



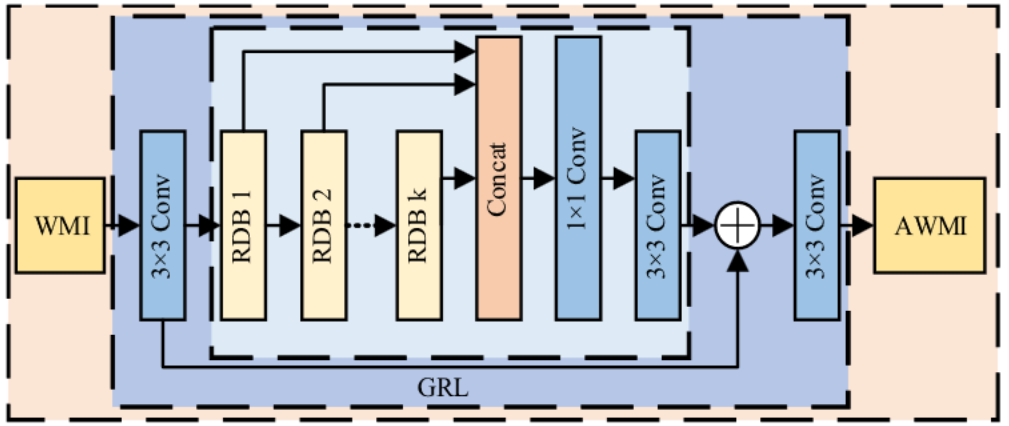
创新分析　渐进式预处理

- 预处理（水印增强）
  - 目的：使得模型更好学习水印信息，提高WARDN的水印攻击强度
  - 增强公式：$I_{WE} = (I_W - I) \times E + I_W$
    - I_W：单个水印图片
    - I：原始图片
    - I_{WE}：增强后的图片
    - E：增强倍数

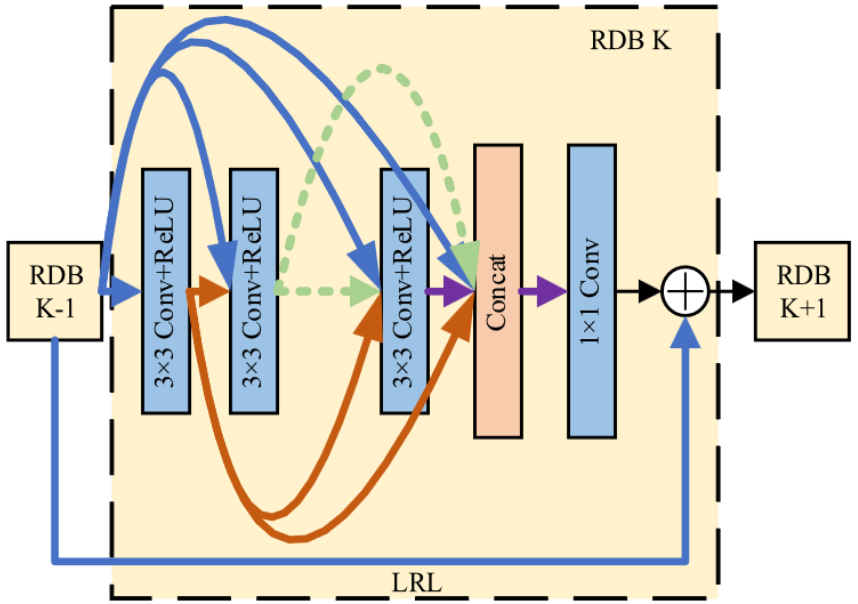
  - 存在问题：使用未经处理的测试集测试时，原始图片会在一定程度上被破坏，攻击的不可察觉性降低
  - 渐进式预处理：根据迭代次数改变增强水印信息的倍数
    - 前100次迭代$E = 5$，每隔100次倍数E减小1
    - 第401次迭代时，水印信息不会被增强
    - 优点：网络准确学习水印信息，保证了强大的攻击能力和不可感知性，不破坏原始图片

创新分析 WARDN

- 水印攻击残差密集网络（WARDN）
  - 浅层特征提取网络（SFEN）
    - 目的：提取图像低频信息
    - 设置卷积层
      - 提取浅层特征并进行全局残差学习
  - 残差密集块（RDB）
    - 密集连接的卷积层
    - 局部特征融合（LFF）（1×1 Conv）
      - 使网络训练更加稳定
      - 使模型能够更准确地学习水印特征
    - 局部残差学习（LRL）
      - 改善信息流
      - 使模型能够更准确地学习水印特征



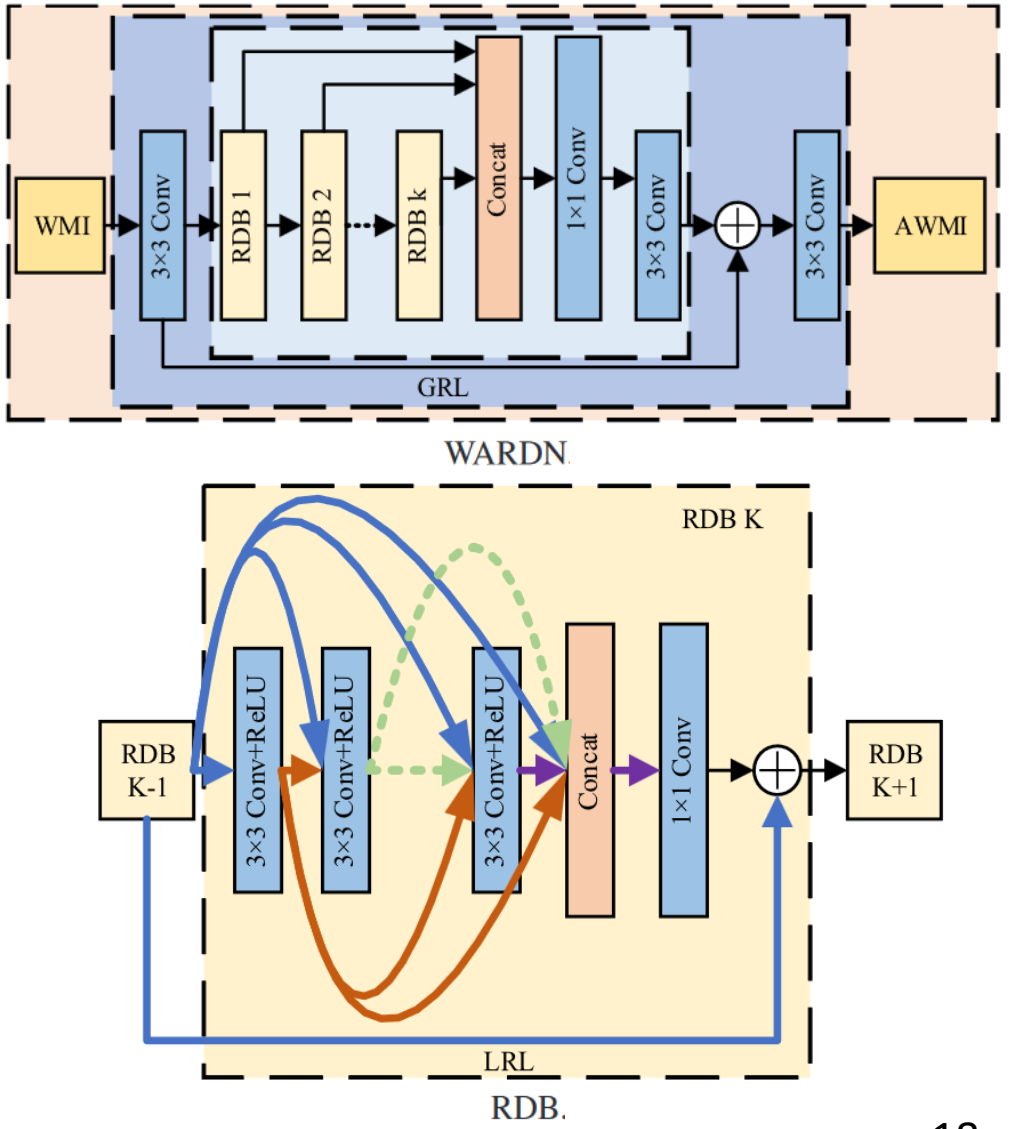
WARDN



RDB.

创新分析 WARDN

- 水印攻击残差密集网络（WARDN）
  - 密集特征融合（DFF）
    - 全局特征融合（GFF）
      - 融合从残差密集块RDB中提取的所有特征，得到全局特征
      - 用于具有全局特征的联合自适应学习
    - 全局残差学习（GRL）
      - 将浅层特征提取网络SFEN提取的特征添加到全局特征融合GFF得到的特征中
        - » WARDN有效地攻击水印
  - 局限性
    - 一些水印信息完全隐藏在水印图像中，无法提取



创新分析　综合损失

- 损失函数优化
  - 单一损失函数
    - 逐像素损失易产生模糊和缺乏边缘的结果
    - 基于特征比较的感知损失其像素空间的覆盖不均匀，可能会产生细微的视觉伪影
  - 综合损失函数：$L_C = \varphi L_{MSE} + \lambda L_{VGG}$
    - L_{MSE}为MSE损失函数（逐像素比较），L_{VGG}为VGG损失函数（语义特征比较）

$$L_{MSE} = \frac{1}{C}\sum_{m=1}^{C}(I - I_w^*)^2$$

$$L_{VGG} = \frac{1}{C_j H_j W_j}||V_j(I_w^*) - V_j(I)||_2^2$$

  - I_w^*为水印攻击网络后的图像
  - $V_j(x)$是预训练VGG16 V处理图像时j层的特征
  - C_j，H_j和W_j指特征图大小

- 数据资源
  - 数据集：PASCAL VOC2012、USC-SIPI
    - 训练集： PASCAL VOC2012中1000张大小为256×256的彩色图像
    - 测试集： PASCAL VOC2012和USC-SIPI中1000张大小为256×256的彩色图像

  - 水印算法：基于四元数极谐傅里叶矩（QPHFMs）的水印算法
    - 对常见的图像处理攻击和几何攻击具有更强的鲁棒性

实验设计　对比方法与评价指标

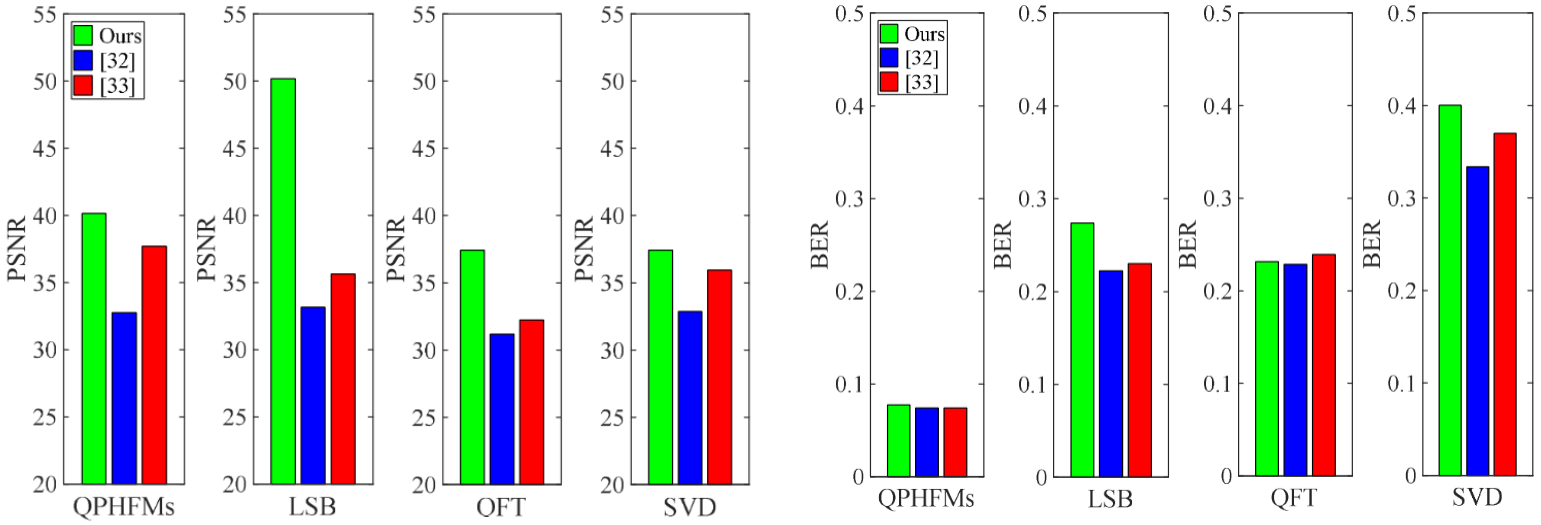
- 对比方法
  - 水印算法
    - 基于最低有效位（LSB）的水印算法：空间域水印算法，对图像质量的影响最小，难以察觉
    - 基于四元数傅里叶变换（QFT）的水印算法：变换域水印算法，对图像处理攻击抵抗能力较强
    - 基于奇异值分解（SVD）的水印算法：特征值分解水印算法，对传统攻击具有鲁棒性
  - 水印攻击方法（降噪攻击）
    - [32]L. Geng et al. 2020：Real-time attacks on robust watermarking tools in the wild by CNN
    - [33]M. W. Hatoum et al. 2021：Using deep learning for image watermarking attack
- 评价指标
  - 峰值信噪比（Peak Signal-to-Noise Ratio，PSNR）

$$PSNR = 10log_{10}\frac{MN\ max(I^2)}{\sum_{i=1}^{M}\sum_{j=1}^{N}[I(i,j)-I^*(i,j)]^2}$$

    - I为水印图像，I^*为攻击后的水印图像，M和N分别是图像的高度和宽度
  - 误码率（Bit Error Rate，BER）

实验设计　对比实验

- 实验结果
  - 基于LSB的水印算法的PSNR远高于其他方法
  - RD-IWAN攻击时间最短，表明RD-IWAN在攻击效率上具有优势



	Ours	[32]	[33]
Attack time (s)	0.3013	0.5420	0.4994

COMPARISON OF SIMILAR METHODS

	QPHFMs	LSB	QFT	SVD
WMI				
WPSNR	40.7173	55.8889	35.6334	39.3639
Ours				
PSNR	40.1280	50.1766	37.4140	37.4193
BER	0.0774	0.2737	0.2316	0.4004
[32]				
PSNR	32.7485	33.1471	31.1706	32.8811
BER	0.0742	0.2224	0.2285	0.3338
[33]				
PSNR	37.6848	35.6430	32.2358	35.9570
BER	0.0742	0.2300	0.2393	0.3701

实验设计　消融实验

- 实验设置：7种不同的预处理模式和训练模式
  - 实验结果模式1-5：具有不同程度信息增强预处理（增强倍数E为0.5,0.4,0.3,0.2,0.1）的训练模式
  - 模式6：无信息增强的训练模式
  - 模式7：具有渐进预处理的训练模式
  - 使用MSE损失训练的模型
    - 具有更强的攻击能力
    - 生成的AWMI较为模糊，不可感知性低

(a) WMI (b) MSE loss (PSNR=32.5363, BER=0.1913) (c) Perceptual loss (PSNR=34.8531, BER=0.0510) (d) Comprehensive loss (PSNR=34.4012, BER=0.1903)

  - 使用感知损失训练的模型
    - 保留更多边缘信息和纹理细节，生成的AWMI更清晰，视觉质量高，不可感知性强
    - 不具有较强的水印攻击能力
- 实验结论
  - 使用综合损失训练模型，既能保证强大的攻击能力，又能保证高的不可感知性

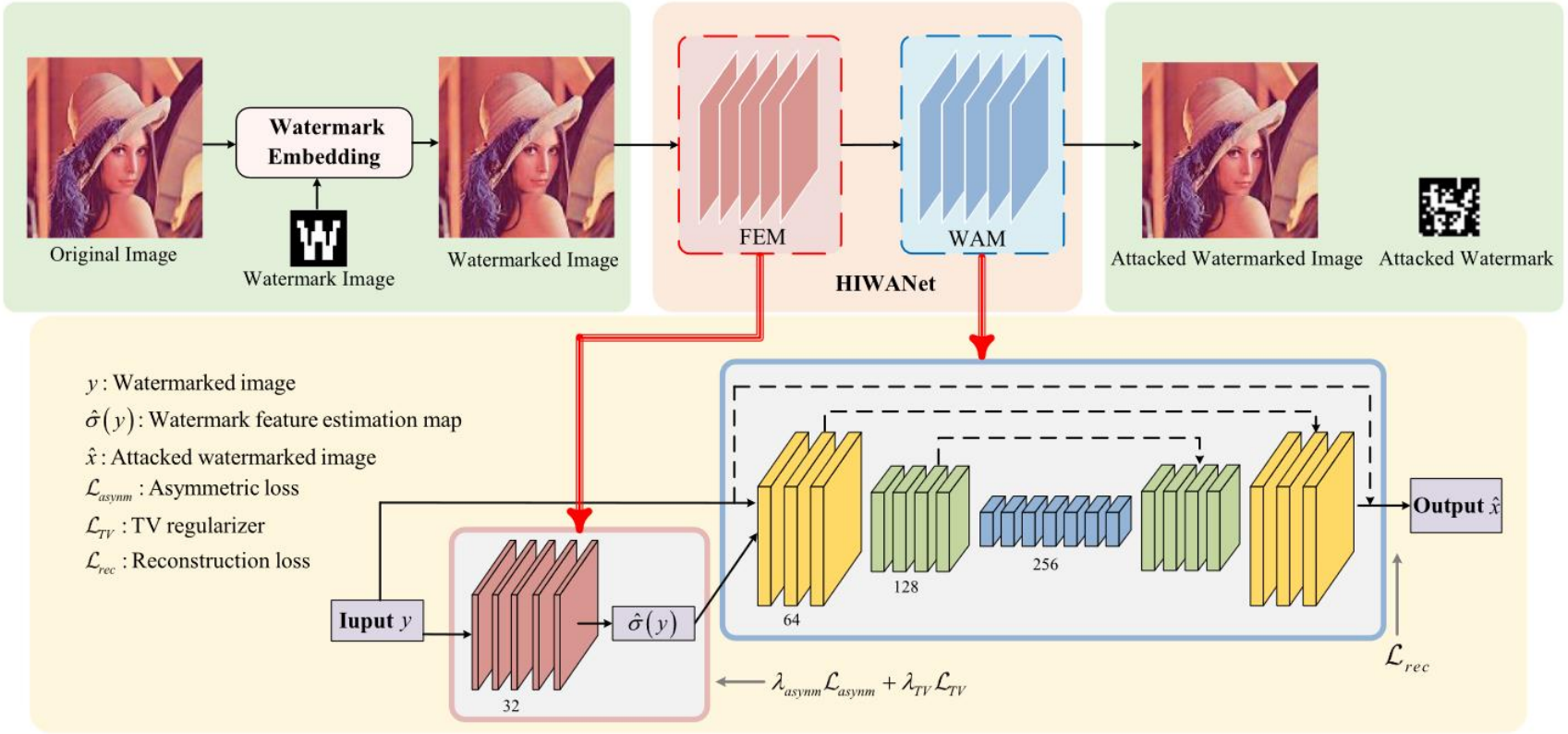
【Engineering Applications of Artificial Intelligence】
HIWANet: A high imperceptibility watermarking attack network

T	目标	提高水印攻击方法的不可感知性
I	输入	1000个原始图像
P	处理	1. 利用QPHFMs水印算法将大小为16×16的二进制水印嵌入到大小为256×256的原始图像中 2. 利用FEM捕获水印信息特征，构建WAM学习图像的高级抽象特征 3. 结合非对称损失函数、全变分损失函数、重构损失函数生成目标损失函数，对模型进行优化
O	输出	1000个攻击后的水印图像

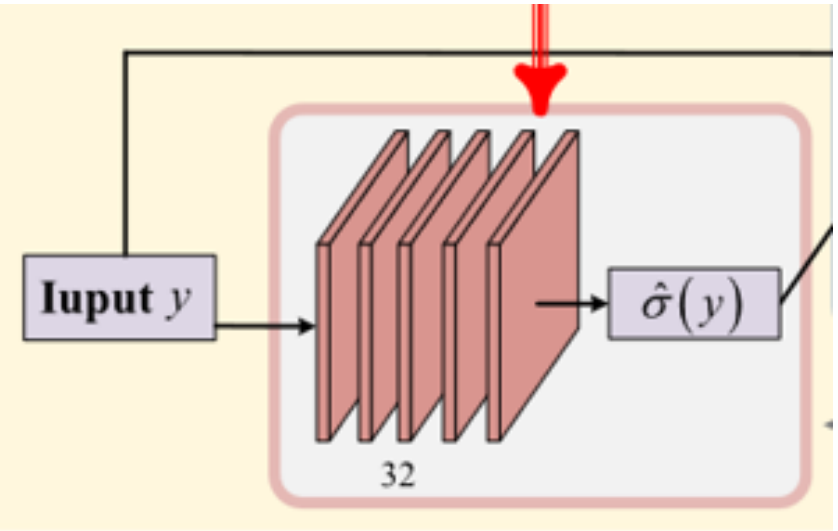
P	问题	现有水印攻击方法使得图像视觉质量恶化
C	条件	图像领域黑盒水印
D	难点	捕获水印信息后高级抽象特征的提取
L	水平	2024中科院 2区，CiteScore:9.60

HIWANet 算法核心

- 算法原理图
  - 特征提取模块（FEM）
  - 水印攻击模块（WAM）



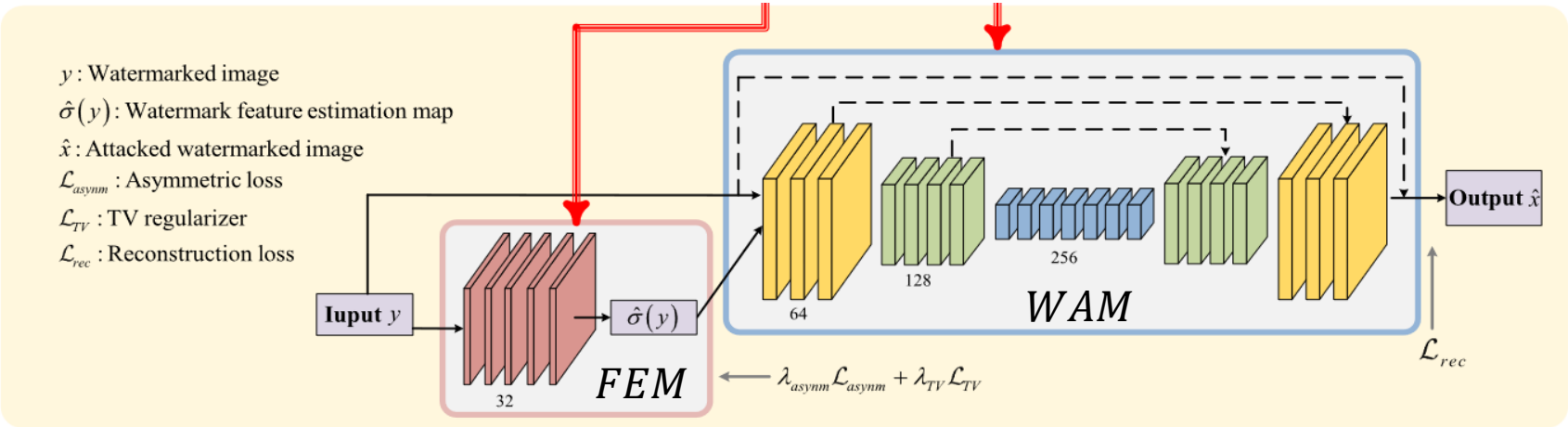
- 特征提取模块（FEM）
  - 5层全卷积网络
    - 每个卷积层的通道数设置为32，卷积核为3×3
  - 激活函数：ReLU
  - 输入：水印图像y
  - 输出：水印特征估计图$\hat{\sigma}(y) = F_m(y, W_m)$
    - W_m为FEM的权重参数
    - $F_m(\cdot)$为网络映射函数
  - 有限元法
    - 捕获水印图像中的低级和高级特征，以预测水印的位置和强度
  - 不存在池化层和批量归一化
    - 池化层和批量归一化会导致一些细节的丢失，从而影响水印攻击的性能

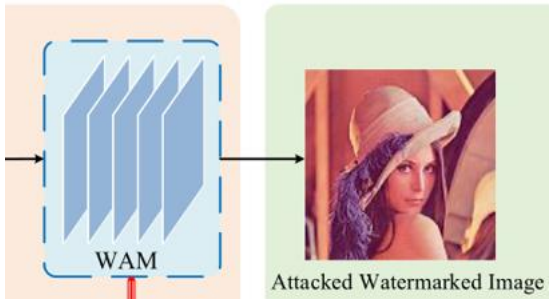


- 水印攻击模块（WAM）
  - 16层的U-Net结构
    - 具有编码器-解码器架构，卷积核为3×3
  - 输入：水印特征估计图$\hat{\sigma}(y) = F_m(y, W_m)$
  - 输出：预测结果$\hat{x} = F_a(y, \hat{\sigma}(y); W_a)$
    - 最终的攻击图像$\hat{x} = y + R(y, \hat{\sigma}(y); W_a)$
      - $R(y, \hat{\sigma}(y); W_a)$为WAM输入和输出间的残差

Component	Input Channels	Output Channels	Number of Layers	Kernels
FEM	3	3	4	3 × 3
WAM Encoder	6	64	2	3 × 3
WAM Downsample1	64	128	3	3 × 3
WAM Downsample2	128	256	6	3 × 3
WAM Upsample1	256	128	3	3 × 3
WAM Upsample2	128	64	2	3 × 3
WAM Output	64	3	1	1 × 1

HIWANet相关参数





创新分析　目标损失函数

- 损失函数优化
  - 非对称损失函数：为水印样本和无水印样本分配权重，使网络更关注学习水印区域

$$L_{asynm} = \sum_i |\alpha - I_{\hat{\sigma}(y_i)-\sigma(y_i)<0}| \cdot (\hat{\sigma}(y_i) - \sigma(y_i))^2 , I_e = \begin{cases} 1, e < 0 \\ 0, e \geq 0 \end{cases}$$

    - $\sigma(y_i)$为像素i处的水印特征真实值，$\hat{\sigma}(y_i)$为像素i处的水印特征估计图
  - 全变分损失函数：限制水印特征估计图的平滑度

$$L_{TV} = ||\nabla_h \hat{\sigma}(y)^2||_2^2 + ||\nabla_v \hat{\sigma}(y)^2||_2^2$$

    - $\nabla_h(\nabla_v)$表示沿水平（垂直）方向的梯度运算
  - 重构损失函数

$$L_{rec} = ||\hat{x} - x||_2^2$$

    - $\hat{x}$ 为FEM中的输出
  - 目标损失函数$L = L_{rec} + \lambda_{asynm} L_{asynm} + \lambda_{TV} L_{TV}$

实验设计　数据资源

- 实验设置
  - 数据集：PASCAL VOC2012
    - 原始图像：PASCAL VOC2012中1000张大小为256×256的彩色图像
    - 水印图像：1个16×16的二进制图像
  - 水印算法：基于四元数极谐傅里叶矩（QPHFMs）的水印算法
    - 对常见的图像处理攻击和几何攻击具有更强的鲁棒性
  - 学习率：0.0001
  - batch size：16
  - patch size：32
  - Epoch：200
  - PyTorch 1.10
  - Python 3.9

Fig. 4. A binary watermark image of size 16 × 16.

实验设计　对比方法与评价指标

- 对比攻击方法
  - L. Geng et al. 2020：Real-time attacks on robust watermarking tools in the wild by CNN
  - M. W. Hatoum et al. 2021：Using deep learning for image watermarking attack
  - Li et al. 2022：Concealed attack for robust watermarking based on generative model and perceptual loss
- 对比水印算法
  - 基于最低有效位（LSB）的水印算法
  - 基于四元数傅里叶变换（QFT）的水印算法
- 评价指标
  - 峰值信噪比（Peak Signal-to-Noise Ratio，PSNR）
  - 误码率（Bit Error Rate，BER）

实验设计　对比实验

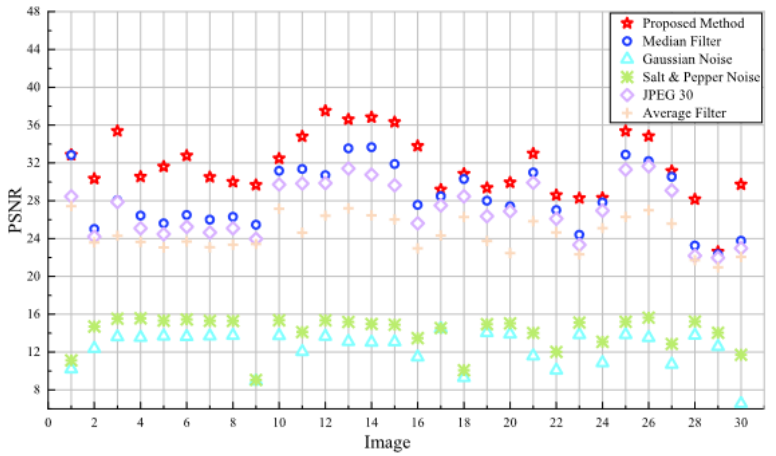
- 攻击方法对比
  - 实验结果
    - 与（Geng et al., 2020）和（Hatoum et al., 2021）相比，PSNR分别提高了21%和29%
    - 与（Geng et al., 2020）和（Hatoum et al., 2021）相比，BER分别提高了39%和75%
      - 基于降噪的方法进行水印攻击，在消除水印信息的同时消除噪声
      - 图像中一些高频信息丢失，损害了被攻击图像的质量
    - 与（Li et al., 2022）相比，PSNR提高了13%，BER提高了27%

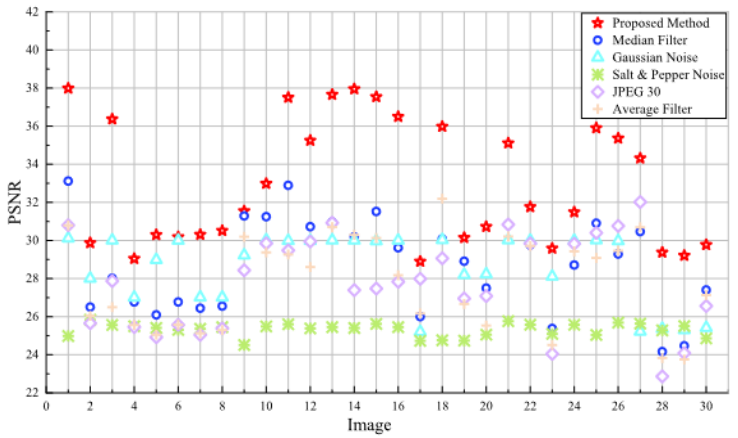
Methods	PSNR	BER
Geng et al. (Geng et al., 2020)	26.13	0.1176
Hatoum et al. (Hatoum et al., 2021)	24.56	0.0935
Li et al. (Li et al., 2022)	27.99	0.1286
Our HIWANet	31.66	0.1637

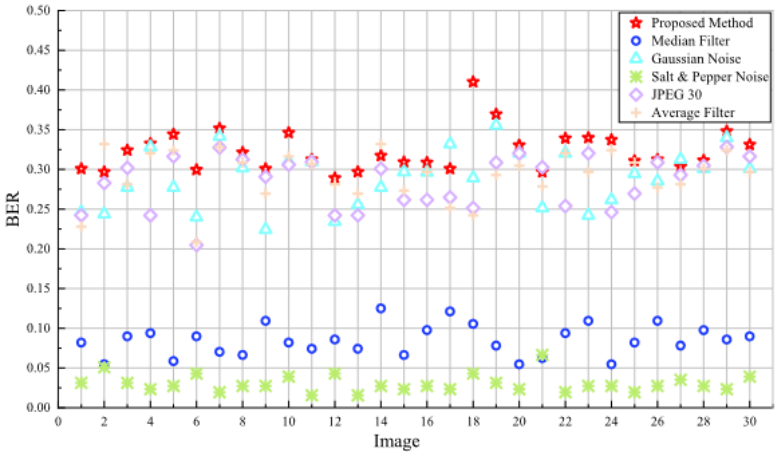
  - 实验结论
    - HIWANet具有优越的攻击能力，实现了高水平的隐蔽性

实验设计　对比实验

- 水印算法对比
  - 实验设置
    - 1000张256×256的彩色图像作为原始图像
    - 1张16×16的二进值图像作为水印图像
    - 随机选择输出的30张图像进行测试
  - 实验结论
    - HIWANet的PSNR和BER远高于传统攻击
    - 攻击能力优越







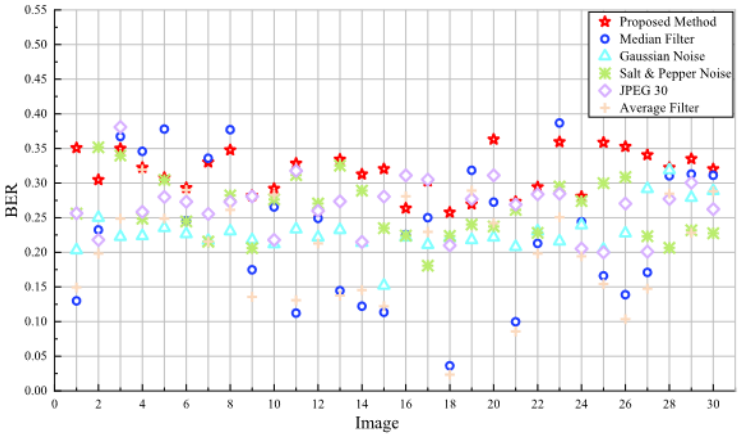


Fig. 6. The PSNRs and BERs using the LSB-based watermarking algorithm. **Fig. 7.** The PSNRs and BERs using the QFT-based watermarking algorithm.

实验设计　消融实验

- 实验设置
  - 从目标损失函数中删除非对称损失建立基线损失
    - 目标损失函数 = 全变分损失函数 + 重构损失函数
  - 用对称损失替换非对称损失构造目标损失函数
    - 目标损失函数 = 对称损失函数 + 全变分损失函数 + 重构损失函数
- 实验结果
  - 使用非对称损失训练的模型表现出更高的PSNR和BER值
    - 不对称损失为加水印和未加水印区域分配不同权重，引导网络专注于学习带水印区域
    - 对称损失会导致网络过度关注非水印区域，无法准确去除水印信息

	Baseline	Sym + Baseline	Asym + Baseline (Ours)
PSNR	28.8172	29.1765	30.4726
BER	0.0928	0.1138	0.1472

特点总结与未来展望

特点总结与未来展望

- 特点总结

算法	RD-IWAN	HIWANet
优势	1. 不向图像中添加任何噪声，只处理水印所在的图像的低频 2. 具有高度的不易察觉性 3. 攻击效率高	1. 在保持水印图像质量的同时，有效地破坏了水印信息 2. 实现了误码率的显著提升 3. 隐蔽性高
劣势	1. 一些水印信息已经完全隐藏在水印图像中，无法提取 2. 无法销毁作为优化目标的原始图像	攻击效率难以保证

- 未来展望
  - 对抗性攻击技术的进步：利用更复杂的对抗性攻击方法有效地绕过模型水印保护
  - 对抗性防御机制的加强：加强模型水印的设计和部署，确保模型的安全性
  - 隐蔽性和鲁棒性的提升：未来的模型水印技术会更注重在模型中嵌入水印的隐蔽性和鲁棒性，使得攻击者更难以检测和去除水印

参考文献

[1] Wang C, Hao Q, Xu S, et al. RD-IWAN: Residual dense based imperceptible watermark attack network[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 32(11): 7460-7472.

[2] Wang C, Li X, Xia Z, et al. HIWANet: A high imperceptibility watermarking attack network[J]. Engineering Applications of Artificial Intelligence, 2024, 133: 108039.

[3] Geng L, Zhang W, Chen H, et al. Real-time attacks on robust watermarking tools in the wild by CNN[J]. Journal of Real-Time Image Processing, 2020, 17: 631-641.

[4] Hatoum M W, Couchot J F, Couturier R, et al. Using deep learning for image watermarking attack[J]. Signal Processing: Image Communication, 2021, 90: 116019.

道德经

知人者智，自知者明。胜人者有力，自胜者强。知足者富。强行者有志。不失其所者久。死而不亡者，寿。

谢谢！