中图法分类号:TP18;TP37　　**文献标识码**: A　　**文章编号**: 1006-8961(2026)01-0197-15

论文引用格式:Zhang G F, Li X, Su Z P, Fang H and Lian C S. 2026. Deep robust image watermarking driven by frequency awareness. Journal of Image and Graphics, 31(1):0197-0211(张国富, 李鑫, 苏兆品, 方涵, 廉晨思. 2026. 频率感知驱动的深度鲁棒图像水印. 中国图象图形学报, 31(1):0197-0211)[DOI:10.11834/jig.250094]

频率感知驱动的深度鲁棒图像水印

张国富[1,2,3],李鑫[1],苏兆品[1,2,3*],方涵[4],廉晨思[3,5]

1. 合肥工业大学计算机与信息学院,合肥　230601; 2. 智能互联系统安徽省实验室(合肥工业大学),合肥　230009;
3. 音视频智能防识联合实验室,合肥　230000; 4. 新加坡国立大学计算机学院,新加坡　117417,新加坡;
5. 安徽省公安厅物证鉴定管理处,合肥　230000

摘　要: **目的**　近年来,基于深度学习的水印方法得到了广泛研究。现有方法通常对特征图的低频和高频部分同等对待,忽视了不同频率成分之间的重要差异,导致模型在处理多样化攻击时缺乏灵活性,难以同时实现水印的高保真性和强鲁棒性。为此,本文提出一种频率感知驱动的深度鲁棒图像水印技术(deep robust image watermarking driven by frequency awareness, RIWFP)。**方法**　通过差异化机制处理低频和高频成分,提升水印性能。具体而言,低频成分通过小波卷积神经网络进行建模,利用宽感受野卷积在粗粒度层面高效学习全局结构和上下文信息;高频成分则采用深度可分离卷积和注意力机制组成的特征蒸馏块进行精炼,强化图像细节,在细粒度层面高效捕捉高频信息。此外,本文使用多频率小波损失函数,引导模型聚焦于不同频带的特征分布,进一步提升生成图像的质量。**结果**　实验结果表明,提出的频率感知驱动的深度鲁棒图像水印技术在多个数据集上均表现出优越性能。在COCO(common objects in context)数据集上,RIWFP在随机丢弃攻击下的准确率达到91.4%;在椒盐噪声和中值滤波攻击下,RIWFP分别以100%和99.5%的准确率达到了最高水平,展现了其对高频信息的高效学习能力。在ImageNet数据集上,RIWFP在裁剪攻击下的准确率为93.4%;在JPEG压缩攻击下的准确率为99.6%,均显著优于其他对比方法。综合来看,RIWFP在COCO和ImageNet数据集上的平均准确率分别为96.7%和96.9%,均高于其他对比方法。**结论**　本文所提方法通过频率感知的粗到细处理策略,显著增强了水印的不可见性和鲁棒性,在处理多种攻击时表现出优越性能。

关键词: 鲁棒图像水印;小波卷积神经网络;深度可分离卷积;注意力机制;多频率小波损失

Deep robust image watermarking driven by frequency awareness

Zhang Guofu[1,2,3], Li Xin[1], Su Zhaopin[1,2,3*], Fang Han[4], Lian Chensi[3,5]

1. School of Computer and Information Technology, Hefei University of Technology, Hefei　230601, China; 2. Intelligent Interconnected Systems Laboratory of Anhui Province (Hefei University of Technology), Hefei　230009, China; 3. Joint Laboratory of Intelligent Prevention and Recognition of Audio and Video, Hefei　230000, China; 4. School of Computing, National University of Singapore, Singapore　117417, Singapore; 5. Department of Physical Evidence Identification, Anhui Public Security Department, Hefei　230000, China

Abstract: **Objective** With the rapid advancement of digital technology and the extensive use of the internet, the creation

收稿日期:2025-03-21;**修回日期**:2025-04-15;**预印本日期**:2025-04-22

*通信作者:苏兆品　szp@hfut.edu.cn

基金项目:教育部人文社会科学研究规划基金项目(24YJA870011);国家自然科学基金项目(62302146);安徽省重点研究与开发计划项目(202104d07020001);

Supported by: MOE (Ministry of Education in China) Project of Humanities and Social Sciences(24YJA870011); National Natural Science Foundation of China(62302146); Anhui Provincial Key R&D Program, China(202104d07020001)

and dissemination of image content have undergone remarkable changes. Social media platforms, such as Douyin and WeChat Moments, have provided users with vast stages for showcasing and sharing visual works. These platforms have not only motivated the flourishing development of image creation but have also injected strong momentum into the "visual economy". However, with the widespread distribution of image works, copyright issues have become increasingly prominent. Numerous malicious actors use technical means to steal, alter, or even commercialize the works of others without authorization, severely damaging the rights of original creators. This phenomenon not only disturbs the creative ecosystem but also threatens the sustainable development of digital content. Therefore, effectively protecting the copyright of image works in an open network environment has become a crucial issue in technical research and industrial practice. Digital image watermarking technology is one of the effective methods to address this problem. The earliest studies in digital watermarking encoded secret information in the least significant bit of image pixels; however, this method is easily detected by statistical measures. Researchers then shifted their attention to frequency domains, determining that encoding information in the discrete cosine transform and discrete wavelet transform domains was highly robust. However, all these traditional methods heavily relied on shallow, manually crafted image features, indicating that they did not fully use the carrier image, resulting in substantial limitations in robustness. In recent years, with the development of deep learning, numerous deep learning-based models have been applied to digital image watermarking to enhance robustness. However, most of these approaches treat the low-frequency and high-frequency components of feature maps equally, overlooking the considerable differences between different frequencies. This uniform treatment limits the model's adaptability to diverse attacks, making it difficult to achieve high fidelity and strong robustness in watermarking. Various attacks affect image frequency bands differently. For example, median filtering primarily impacts the distortion of high-frequency regions, whereas JPEG compression has a greater impact on low-frequency distortion. The low-frequency part of an image typically conveys the smooth appearance of the global area, while the high-frequency part captures the rich details of local regions, such as edges, textures, and other complex features. Ignoring the differences between frequency components hinders effective feature learning, making it difficult to achieve robustness and invisibility in watermarking, which also limits the flexibility of the model in handling diverse information. In robust image watermarking models, fidelity and robustness are crucial evaluation metrics. Effectively capturing image features plays a key role in enhancing the fidelity and robustness of the watermarking model. **Method** The proposed method enhances watermarking performance by processing low-frequency and high-frequency components through differentiated mechanisms. Specifically, low-frequency components are modeled using a wavelet convolutional neural network of cascaded wavelet decomposition, deep convolution, and inverse wavelet reconstruction. This approach leverages wide receptive field convolutions to efficiently capture global structures and contextual information at a coarse granularity, enhancing the capability of the model to learn low-frequency information while maintaining computational efficiency. In contrast, high-frequency components are refined using a feature distillation block comprising depthwise separable convolutions and attention mechanisms. This design strengthens image details and enables efficient extraction of high-frequency information at a fine granularity. The Haar wavelet is used for decomposition due to its computational efficiency and simplicity, increasing its suitability for integration into convolutional neural networks. Additionally, a multifrequency wavelet loss function is employed to guide the model's focus on feature distribution across different frequency bands, further improving generated image quality. Multifrequency wavelet transforms allow image representation at different scales. By decomposing the image into multiple levels, details from low-frequency to high-frequency are captured, which is crucial for image reconstruction, preserving additional details and reducing distortion. Subsequent wavelet decompositions are performed only on the low-frequency components after applying the first-level wavelet transform to the image. **Result** Experimental results demonstrate that the proposed deep robust image watermarking driven by frequency awareness (RIWFP) achieves superior performance across multiple datasets. On the COCO dataset, RIWFP attains an accuracy of 91.4% under dropout attacks. Under salt and pepper noise and median blur attacks, RIWFP achieves a maximum accuracy of 100% and 99.5%, respectively, highlighting its effective learning of high-frequency information. On the ImageNet dataset, RIWFP records an accuracy of 93.4% and 99.6% under crop and JPEG attacks, respectively, notably outperforming other methods. Overall, RIWFP achieves average accuracies of 96.7% and 96.9% on the COCO and ImageNet datasets, respectively, surpassing existing methods. Ablation experiments further confirm that RIWFP enhances watermark

invisibility and robustness by integrating wavelet convolution with a distillation mechanism, showing particular resilience to high-frequency-related noise and filtering attacks. **Conclusion** The proposed method substantially improves watermark invisibility and robustness through a frequency-aware, coarse-to-fine processing strategy. This method demonstrates superior performance against various attacks, providing an effective solution for digital image watermarking.

Key words: robust image watermarking; wavelet convolutional neural network; depthwise separable convolution; attention mechanism; multi-frequency wavelet loss

0 引　言

鲁棒数字水印技术是确保数字图像的版权得到保护的有效手段(Hu 等,2024;Kadian 等,2021),通过在载体图像中嵌入不可见的特定信息,实现了对图像的版权保护。当出现版权争议时,可以通过提取水印信息验证版权归属(Zhang 等,2024;Evsutin 和 Dzhanashia,2022)。为实现版权保护,需要满足鲁棒性与不可见性这两个属性(Wan 等,2022;Tang 等,2024)。鲁棒性是其最关键的属性,表示水印能够抵抗各种失真,即使载体图像受到攻击时,也能够准确提取水印。不可见性则是其最基本的属性,即在不干扰正常视觉感知的前提下,在图像中嵌入信息。

数字水印的早期研究将秘密信息编码在图像像素的最低有效位(van Schyndel 等,1994),但该方法易被统计检测手段破解(Dumitrescu 等,2003)。为此,研究人员转向频域编码,发现离散傅里叶变换域(刘西林 等,2022)和离散小波变换域(Guo 和 Georganas,2003)的信息嵌入具有更强的鲁棒性。然而,这些传统方法依赖于人工设计的浅层图像特征,未能充分挖掘载体图像的潜在信息,导致其鲁棒性存在显著局限。近年来,随着深度学习技术的突破,基于深度神经网络的模型(Tancik 等,2020;Zhong 等,2021;Ding 等,2022;Wang 等,2023;Luo 等,2024b)被引入数字图像水印领域,显著提升了水印对抗攻击的能力。

但在大多数基于深度学习的鲁棒图像水印方案中,往往倾向于对特征图的低频和高频部分进行同等对待,忽视了不同频率之间的重要区别。不同的攻击对图像频带的影响不同,如中值滤波主要影响了图像的高频区域失真,而 JPEG(joint photographic experts group)压缩对低频的失真影响更大。通常,图像的低频部分传达了全局区域的平滑外观,而高频部分则捕捉了局部区域的丰富细节,如边缘、纹理和其他复杂特征。忽视不同频率差别不利于对图像特征的学习,缺乏处理多样化信息的灵活性,在实现水印的鲁棒性和不可见性方面带来了挑战。对于鲁棒图像水印模型而言,保真性与鲁棒性是其重要的衡量指标。对图像特征图进行有效学习有利于提升水印模型的保真性与鲁棒性。

因此,本文提出频率感知驱动的深度鲁棒图像水印技术(deep robust image watermarking driven by frequency awareness, RIWFP),通过频率感知的方式用不同的机制处理不同的成分。本文的工作如下:1)设计基于高低频双流的频率感知模块,对图像特征进行学习,以提高模型的鲁棒性;2)设计多频率小波损失函数,以提高水印的不可见性;3)对比实验表明,RIWFP 相较于最先进的技术,在水印的鲁棒性与不可见性上表现更佳。

1　相关工作

2018 年,Zhu 等人(2018)利用生成式对抗网络,提出首个端到端的水印模型 HiDDeN(hiding data with deep network)。此模型主要包含编码器、解码器、噪声层和鉴别器。编码器负责将水印嵌入载体图像中,噪声层提供被攻击后的载体图像,解码器则需要从被攻击后的载体图像中提取水印信息,鉴别器负责评判图像含水印的概率。该模型打破了之前的水印研究思路,标志着基于深度学习的水印算法迈入全新阶段。Liu 等人(2019)提出一个两阶段可分离深度学习水印框架用于盲水印技术。该框架包含一个编码器—解码器网络和一个含噪声的解码器网络。该框架中的两个网络单独训练,使其能够抵抗未知的噪声攻击。Ahmadi 等人(2020)提出支持在不同的变换域进行操作的水印框架。它由两个具有残差结构的全卷积神经网络组成,用于实时处理嵌入和提取操作,同时用一个强度因子控制图像中

水印的强度。Jia等人(2021)在前人工作的基础上提出一种新型的训练方法MBRS(mini-batch of real and simulated JPEG compression),对每个小批次随机选择模拟JPEG压缩、真实JPEG压缩和无噪声层中的任一种,增强了对JPEG压缩的鲁棒性。Wu等人(2023)在Jia等人(2021)的基础上,利用改进的信息处理器提升对裁剪的鲁棒性,并采用双鉴别器解决生成图像产生伪影的问题,即IMP方法(enhancing robustness and imperceptibility of blind watermarking with improved message processor)。Huang等人(2023)提出一种基于生成对抗网络的注意力的鲁棒图像水印模型,通过引入注意机制模块挖掘原始图像的全局特征。同时为了更有效地获取图像特征,还设计了一个特征融合模块,增强了水印的鲁棒性。Luo等人(2024a)提出一种基于Transformer的模型WFormer(Transformer-based soft fusion model for robust image watermarking),利用自注意力和交叉注意力机制有效提取水印特征并挖掘其与图像的长程相关性,解决了传统方法中水印冗余问题。同时引入特征增强模块捕捉载体图像与水印的跨特征依赖关系,进一步优化融合效果,显著提升了水印的不可见性、鲁棒性和嵌入容量。Fang等人(2023)提出基于流模型的鲁棒水印(flow-based robust watermarking,FBRIW)框架,保证了编码器和解码器的紧密耦合,从而显著提高水印的鲁棒性与不可见性。He等人(2024)提出一种基于流模型与Transformer的模型IWFormer(Transformer-based invertible neural network for robust image watermarking),通过减少冗余特征的嵌入,提高水印提取的精确性,从而增强了水印的鲁棒性。此外,其设计的仿射变换模块能够挖掘图像的全局相关性,进一步增强水印的鲁棒性。

上述方法虽然可以抵抗一定强度的常见图像处理攻击,但它们忽略了水印攻击对图像高低频信息影响的差异性,当遇到高强度攻击或复杂的组合攻击时仍然存在难以提取完整水印的问题。

2 频率感知驱动的深度鲁棒图像水印

在本节中,首先介绍FBRIW(Fang等,2023)基本框架,分析其鲁棒性,并提出改进思路。

2.1 FBRIW及鲁棒性分析

FBRIW水印框架主要包含3个部分:编码器、噪声池和解码器,如图1所示(Fang等,2023)。

编码器由8个相同的可逆神经网络块组成。它将大小为$3 \times W \times H$的载体图像$\boldsymbol{x}_{\mathrm{co}}$和长度为$L$的二进制水印信息$\boldsymbol{m}$作为输入。对于第$i$个可逆神经网络块而言,以$\boldsymbol{m}^i$和$\boldsymbol{x}_{\mathrm{co}}^i$作为输入,其输出为

$$\boldsymbol{x}_{\mathrm{co}}^{i+1} = \boldsymbol{x}_{\mathrm{co}}^i + \phi(\boldsymbol{m}^i) \tag{1}$$

$$\boldsymbol{m}^{i+1} = \boldsymbol{m}^i \cdot \exp[\rho(\boldsymbol{x}_{\mathrm{co}}^{i+1})] + \rho(\boldsymbol{x}_{\mathrm{co}}^{i+1}) \tag{2}$$

式中,ϕ和ρ是需要设计的网络模块。在经过最后一个可逆神经网络块后,可以获得大小为$3 \times W \times H$的编码图像$\boldsymbol{x}_{\mathrm{en}}$和损失信息$\boldsymbol{r}$。

噪声池包含各种常见的图像水印攻击,其中每一种攻击作为一种噪声层。在训练过程中,噪声池随机选择一个攻击,为模型提供一个噪声层。噪声层输入已编码的图像$\boldsymbol{x}_{\mathrm{en}}$,并输出相同大小的被攻击的图像$\boldsymbol{x}_{\mathrm{at}}$。

解码器与编码器共享相同的结构和参数,但信息流方向相反。具体而言,解码器以攻击后的图像$\boldsymbol{x}_{\mathrm{at}}$和长度为$L$的全零辅助量$\boldsymbol{z}$作为输入,输出提取的水印信息$\boldsymbol{m}_{\mathrm{ex}}$和恢复的图像$\boldsymbol{x}_{\mathrm{re}}$。对于第$i$个可逆神经网络块而言,它以$\boldsymbol{x}_{\mathrm{at}}^{i+1}$和$\boldsymbol{z}^{i+1}$作为输入,其输出为

$$\boldsymbol{z}^i = [\boldsymbol{z}^{i+1} - \rho(\boldsymbol{x}_{\mathrm{at}}^{i+1})] \cdot \exp[-\rho(\boldsymbol{x}_{\mathrm{at}}^{i+1})] \tag{3}$$

$$\boldsymbol{x}_{\mathrm{at}}^i = \boldsymbol{x}_{\mathrm{at}}^{i+1} - \phi(\boldsymbol{z}^i) \tag{4}$$

在FBRIW中,ϕ模块和ρ模块采用相同的网络架构,不同的是,ϕ模块将水印信息上采样到与图像相同的维度,而ρ模块负责将图像降采样到与水印信息相同的维度。如图2所示,ρ模块使用简单的6个二维卷积操作,图像的低频特征和高频特征进行等价处理。

需要注意的是,图像的低频区域代表图像中变化缓慢的部分,包含图像的主要结构和整体亮度信息,而图像的高频区域代表图像中变化剧烈的部分,包含图像的纹理和细节信息。不同的攻击对图像的高低频影响不同,例如JPEG压缩主要影响图像低频区域,而中值滤波主要影响图像高频区域。ρ模块的等价处理忽视了不同频率的区别,缺乏处理不同信息的灵活性,不利于模型更好地学习图像特征以应对多样化的攻击。为此,本文设计基于高低频双流的ρ模块,以提高模型抗攻击的鲁棒性。

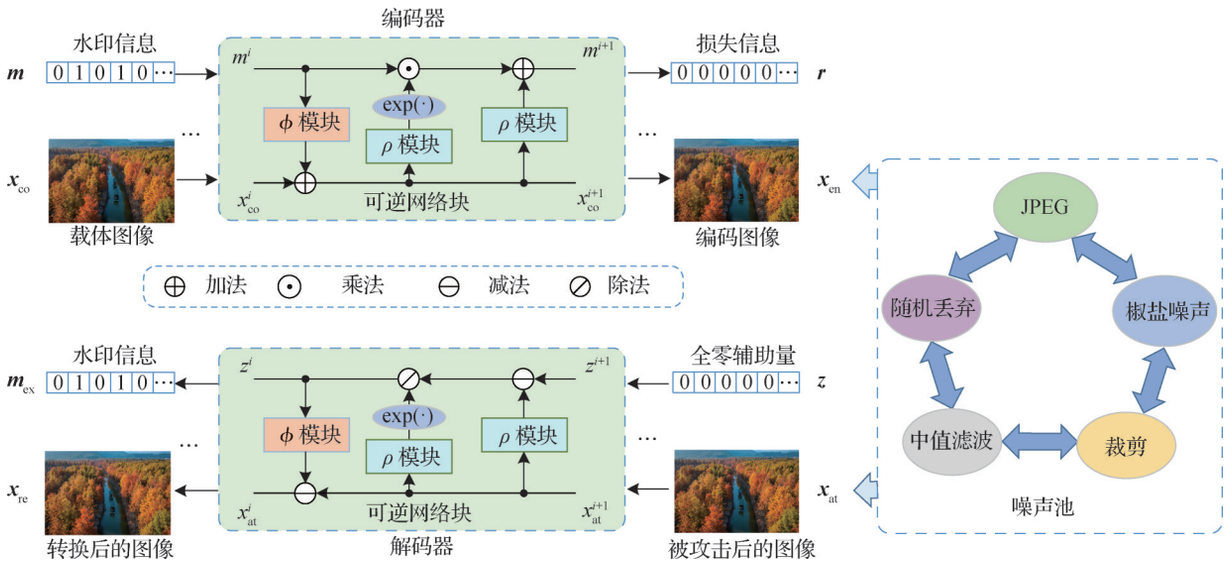


图 1　基于流模型的水印框架

Fig. 1　The flow-based watermarking framework

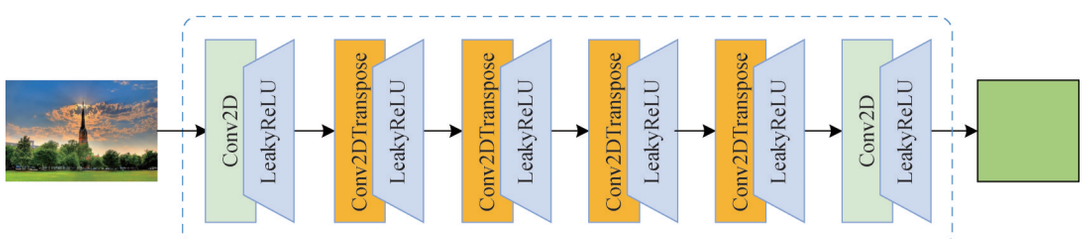


图 2　ρ 模块

Fig. 2　ρ module

2.2　基于高低频双流的 ρ 模块

基于高低频双流的 ρ 模块如图 3 所示，主要包含特征扩展模块、小波卷积神经网络、特征蒸馏块以及下采样子网络块。其中，小波卷积神经网络利用低频子带在粗粒度水平上学习图像的全局结构和上下文信息，有助于模型在处理低频失真（如 JPEG 压缩）时表现出更强的鲁棒性。特征蒸馏模块可以对高频信息进行提炼，使得模型在面对高频失真（如中值滤波和椒盐噪声）时能够更好地保留细节特征。

1）特征扩展块。特征扩展块由 Conv2DTranspose + LeakyReLU（leaky rectified linear unit）和 Conv2D + LeakyReLU 和组成，其核心目标是逐步提取更深层次的特征，并将输入图像的信息映射到更高维度的特征空间。在此过程中，每个 Conv2D 层通过不同大小的感受野学习局部与全局信息，而 LeakyReLU 作为激活函数，有助于缓解梯度消失问题，并在负值区域保留少量信息流动，以增强特征表达能力。

该模块的逐层卷积操作提升了模型对复杂图像的识别能力，从而在更高维度的特征空间中编码更加丰富的图像信息。最终，经过一系列非线性变换，为后续任务提供高质量的特征表示。

2）小波卷积神经网络。小波卷积神经网络通过级联小波变换（wavelet transform，WT）分解层，结合深度可分离卷积（depthwise separable convolution，DSConv）操作，有效捕捉输入信号的多频带特征，在提取全局信息的同时增强了对低频信息的提取能力。图 4 展示了小波卷积神经网络在单个通道上操作的一个例子。其核心设计如下：（1）级联小波分解。输入信号首先经过小波变换，分解为 4 个频率分量，包含 1 种低频分量和 3 种高频分量。在级联过程中，仅对低频分量进行进一步的小波分解，而高频分量保留不变。这种分层分解结构逐步生成新的分量，但每次分解仅作用于上一层的低频部分，形成多层次频率分解。（2）深度卷积操作。对每一层分解得到的 4 个频率分量，分别使用一个小卷积核进行深度卷积（逐通道卷积），以提取各频带的特征信息。这种设计允

许网络在不同频率分量上独立处理，增强了对低频和高频特征的区分能力。(3)逆小波重构与特征融合。每一层的4个卷积结果通过逆小波变换(inverse wavelet transform，IWT)重构为特征图，并与下一层的低频分量相加(最后一层的下一层输入为零)。通过逐层回溯，最终将第一层的卷积结果与下一层的输出融合，生成网络的输出。本文采用Haar小波进行分解，因其计算高效且实现简单，适合在卷积神经网络中集成。

总体而言，小波卷积网络通过级联小波分解、深度卷积和逆小波重构，实现了在粗粒度级别上对主要结构信息的有效处理，强化了模型对低频信息的学习能力，同时保持了计算的高效性。

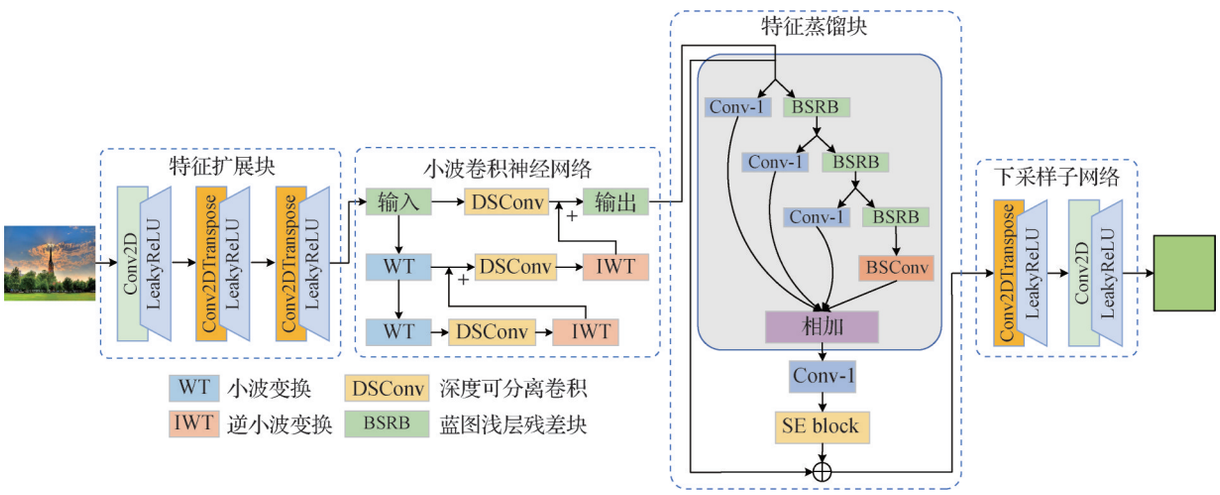


图3　基于高低频双流的ρ模块

Fig. 3　ρ module based on high-low frequency dual-stream

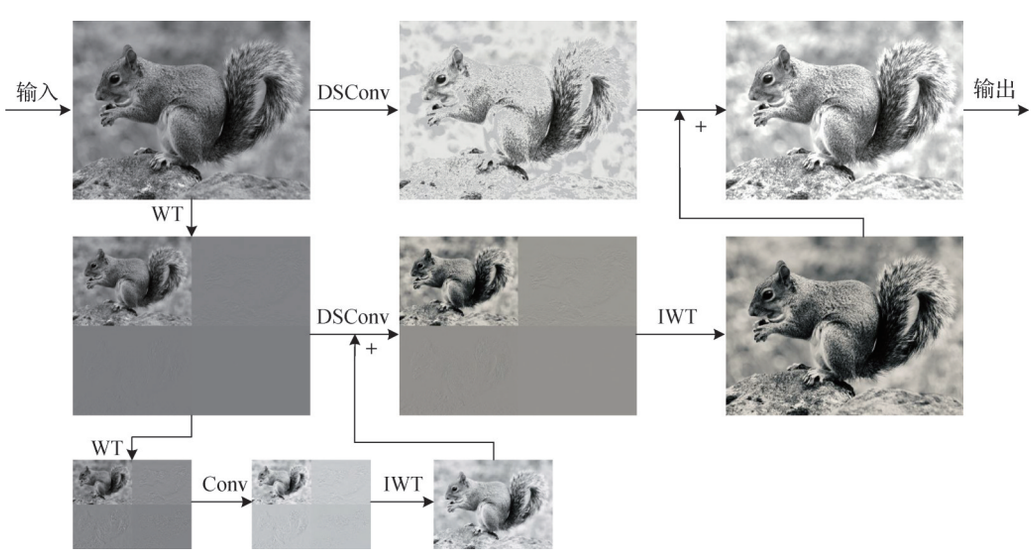


图4　小波卷积神经网络

Fig. 4　Wavelet convolutional neural network

3)特征蒸馏块。ρ模块利用小波卷积神经网络输出经过低频子带信息增强后的特征图。包含丰富细节信息(例如图像纹理细节)的高频分量也需要在精细粒度层面上对特征进行细化。本文使用特征蒸馏块增强高频细节，使模型能够更有效地保留和提取高频信息。在设计上，该模块由蓝图浅层残差块(blueprint shallow residual block，BSRB)(Li等，2022)与通道注意力机制相结合，以充分利用高频特征信息并提高模型的表达能力。

具体而言，首先，输入特征图经过一系列1 × 1卷积层进行通道间的信息交互，以调整和融合不同通道的特征表达。同时，特征送入BSRB与蓝图浅

层卷积(blueprint shallow convolutional，BSConv)(Li等，2022)进行特征蒸馏，以强化高频信息，如边缘、纹理和细节，从而提高水印嵌入和提取过程的鲁棒性。这一过程如图3所示。最后，通道注意力机制(对应图3中的SE block)通过感知动态强化图像细节，建模了不同通道之间的相互依赖关系，以便逐通道地调整特征。这使得网络能够有选择地强调高频信息特征，并通过空间信息学习抑制较不重要的特征。通道注意力机制包含3个阶段，如图5所示。

第1阶段是挤压操作，使用通道空间特征作为表示，通过平均池化层(对应图5中的AvgPool2D)创建一个通道描述符。此操作的目的是生成统计信息并捕获每个通道的全局信息。第2阶段是激励操作，包括两个全连接层(fully connected layer，FC)，一个SiLU(sigmoid linear unit)激活函数和一个sigmoid激活函数。SiLU保持了sigmoid函数的平滑性，即使输入为负值时，仍然能够提供一定的输出。这一特性有助于保持基于流的架构中的信息流动，加速水印模型的收敛并提高训练效率。此阶段的任务是学习每个通道的依赖程度。它以灵活且非互斥的方式建模每个通道的依赖关系，动态调整每个通道的权重，并输出调整后的通道权重。第3阶段是缩放操作，负责将激励操作获得的权重应用于原始特征图。

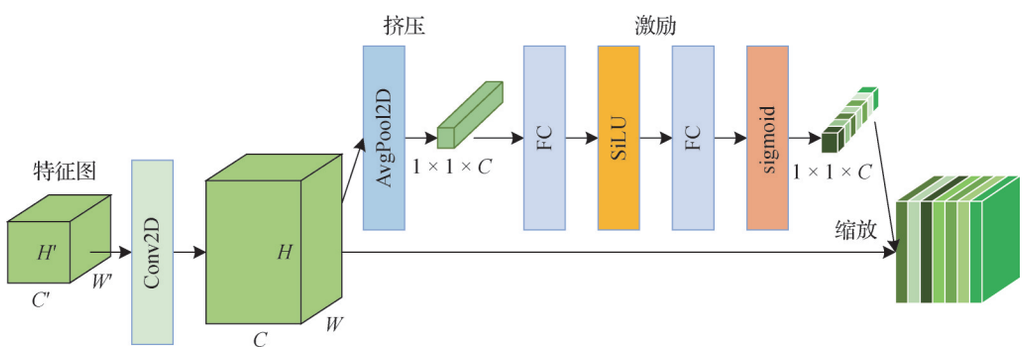


图5　通道注意力机制

Fig. 5　Channel attention mechanism

4)下采样子网络。其作用是在逐步压缩特征图的空间尺寸的同时保留关键信息，使其最终匹配水印信息的维度。首先，Conv2DTranspose层用于对特征图进行局部重构，调整特征图的尺度和分布，使其适应后续的精细降维过程。接着，Conv2D + Leaky-ReLU进一步对特征图进行处理，采用合适的步长和卷积核进行下采样，使得最终的特征图在尺寸上与水印信息保持一致。此步骤不仅能够去除冗余信息，还能保留最具判别力的特征，确保水印嵌入或提取时具备高质量的输入。

下采样子网络的组合策略兼顾了空间变换和特征提取，通过Conv2DTranspose提供更好的信息调整能力，由标准卷积实现有效的降维，使得模型在不同输入条件下依然能够保持稳定的特征表达能力。

由于改进的ρ模块基于高低频双流，采用以粗到细的方式自适应地处理特征，实现了多尺度特征提取，有利于模型对不同频率特征的学习，进而提升水印的性能。

2.3　多频率小波损失函数

1)图像多频率小波损失。多频率小波变换能够在不同的尺度下表示图像。通过将图像分解成多个层次，能够捕捉到图像从低频到高频的不同细节，这对于图像的重建尤为重要，能保留更多的细节，减少失真。为此本文采用了多频率小波变换损失。具体而言，对图像使用小波变换，后续只对低频分量进行进一步的小波分解。图像一层小波损失可以表示为

$$L_{\mathrm{mfl}} = MSE\left[WT(\boldsymbol{x}_{\mathrm{co}}), WT(\boldsymbol{x}_{\mathrm{en}})\right] \tag{5}$$

$$\boldsymbol{x}_{\mathrm{en}} = E(\boldsymbol{x}_{\mathrm{co}}, \boldsymbol{m}, \theta_{\mathrm{E}}) \tag{6}$$

式中，MSE(mean squared error)为均方误差，WT(wavelet transform)为小波变换，$\boldsymbol{x}_{\mathrm{co}}$为载体图像，$\boldsymbol{x}_{\mathrm{en}}$为编码图像，$\boldsymbol{m}$为水印信息，$E$为图1所示的编码器，$\theta_{\mathrm{E}}$为$E$的参数。

2)空间全局图像损失。为了确保水印图像保持不可见，它必须与载体图像非常相似。为了实现这一点，纳入空间全局图像损失，以保证生成图像的质量。此损失函数可以表示为

$$L_{\mathrm{st}} = MSE(\boldsymbol{x}_{\mathrm{co}}, \boldsymbol{x}_{\mathrm{en}}) \tag{7}$$

3)编码器的整体损失。编码器的训练目标是更新参数θ_E,以确保载体图像$\boldsymbol{x}_{co}$与编码图像$\boldsymbol{x}_{en}$尽可能相似,使其难以区分。因此,编码器的损失可以表示为

$$L_E = \omega_1 L_{mf} + \omega_2 L_{st} \tag{8}$$

式中,$\omega_1, \omega_2 \in \mathbf{R}^+$负责平衡这两个损失的权重,$L_{mf}$表示多频率小波损失。

解码器D训练目的是更新参数θ_D,以使水印信息$\boldsymbol{m}$与提取的水印信息$\boldsymbol{m}_{ex}$之间的差异最小,最终使它们相同。因此,解码器损失可以表示为

$$L_D = MSE(\boldsymbol{m}, \boldsymbol{m}_{ex}) \tag{9}$$

$$\boldsymbol{m}_{ex} = D(\boldsymbol{z}, \boldsymbol{x}_{at}, \theta_D) \tag{10}$$

式中,$\boldsymbol{z}$为全零辅助量,$\boldsymbol{x}_{at}$为被攻击后的图像。因此,最终的损失为编码器损失和解码器损失的加权组合,具体为

$$L_t = \lambda_1 L_E + \lambda_2 L_D \tag{11}$$

式中,$\lambda_1, \lambda_2 \in \mathbf{R}^+$负责平衡这两个损失的权重。

3 实验结果与分析

本节首先从基线、实施细节、数据集和评价指标这4个方面介绍实验设置,然后从不可见性和鲁棒性两个方面展开实验验证。

3.1 实验设置

1)基线。为了验证本文所提方案RIWFP的不可见性和鲁棒性,将其与几种先进的水印方法进行比较,包括FBRIW(Fang等,2023)、MBRS(Jia等,2021)、WFormer(Luo等,2024a)、IWFormer(He等,2024)、IMP(Wu等,2023)以及实离散分数Krawtchouk变换(real discrete fractional Krawtchouk transform, RDFrKT)(刘西林等,2022)。如前所述,FBRIW、IWFormer以及本文的RIWFP均属于基于流模型的水印网络,而IMP、MBRS和WFormer是3种基于生成式对抗网络的深度水印方法。RDFrKT是基于传统方法的水印框架。此外,在消融实验中,对RIWFP中的不同改进组合进行了研究,全面验证每个改进部分的效果。

2)实施细节。对于参数的设置,参考了相关文献(Jia等,2021;Luo等,2024a;Fang等,2023;He等,2024;Wu等,2023;刘西林等,2022),分别用于FBRIW、MBRS、WFormer和IWFormer。在MBRS中,编码器损失、解码器损失和鉴别器损失的权重分别配置为1、10和0.000 1。FBRIW将空间损失和编码器损失的权重分配为1和10。WFormer将图像质量损失、鉴别器损失和解码器损失权重分别设置为3、0.000 1、10。在IWFormer中,图像损失、低频损失和水印损失的权重分别设置为3、10和0.1。在RDFrKT中,其分数阶阶数选取为0.6和0.4。IMP将编码器损失和解码器损失的权重设置为1和10,同时鉴别器损失的两个权重分别设置为0.000 1和0.000 3。对于RIWFP,多频率小波变换损失和空间损失的权重分别分配为1和0.5,同时编码器损失和解码器损失的权重分别为2.5和9。本文使用了两层小波神经网络和小波损失函数。

此外,所有深度学习的比较方法都使用Adam优化器;IMP、MBRS和WFormer中的学习率设置为0.001,而FBRIW、IWFormer和RIWFP中的学习率为0.000 1。所有深度学习的比较方法均通过PyTorch(Paszke等,2019)实现。所有实验均在NVIDIA RTX 3060 GPU上进行训练和验证。RDFrKT在MATLAB中编写,采用Windows 10操作系统。

噪声层包含5种常见攻击包括:椒盐噪声、中值滤波、JPEG压缩、随机丢弃和裁剪。

3)数据集。本文选择DIV2K数据集(Agustsson和Timofte,2017)作为训练集,使用从ImageNet数据集(Deng等,2009)和COCO(common objects in context)数据集(Lin等,2014)中各随机选择的2 000幅图像作为测试集评估各个模型的性能。

4)评价指标与参数值。本文使用峰值信噪比(peak signal-to-noise ratio, PSNR)(Huynh-Thu和Ghanbari,2008)和结构相似性(structural similarity, SSIM)(Wang等,2004)评价水印的不可见性。通常,PSNR和SSIM的值越大,水印的不可见性越好。与此同时,为了测试水印的鲁棒性,使用比特准确率(bit accuracy, ACC)衡量水印信息的提取准确性。通常情况下,ACC值越大,说明水印的鲁棒性越好。除了应用表1中常见图像处理来攻击图像,还使用它们的组合攻击分别对带水印的图像进行处理,并计算平均比特准确率,以此作为综合评估水印鲁棒性的依据。本文将所有方法的载体图像尺寸都设置为128 × 128像素,嵌入水印信息长度统一设置为64比特。

3.2 不可见性

如图6所示,本文分别计算了来自ImageNet和

表 1　各种常见攻击方法

Table 1　Various common attack methods

攻击类型	描述
AS-a	调整图像的饱和度，强度因子为 a
AH-b	调整图像的色调，强度因子为 b
AC-c	调整图像的对比度，强度因子为 c
MF-d	窗口大小为 d 的中值滤波
DP-e	将离散像素以 e 的比例随机替换为零
SP-f	添加椒盐噪声，比例为 f
QF-g	质量因子为 g 的 JPEG 压缩
GB-h	标准差为 h 的高斯滤波
CP-i	将连续像素以 $1-i$ 的比例随机替换为零
GN-j	标准差为 j 的高斯噪声
BA-k	调整图像亮度，强度因子为 k

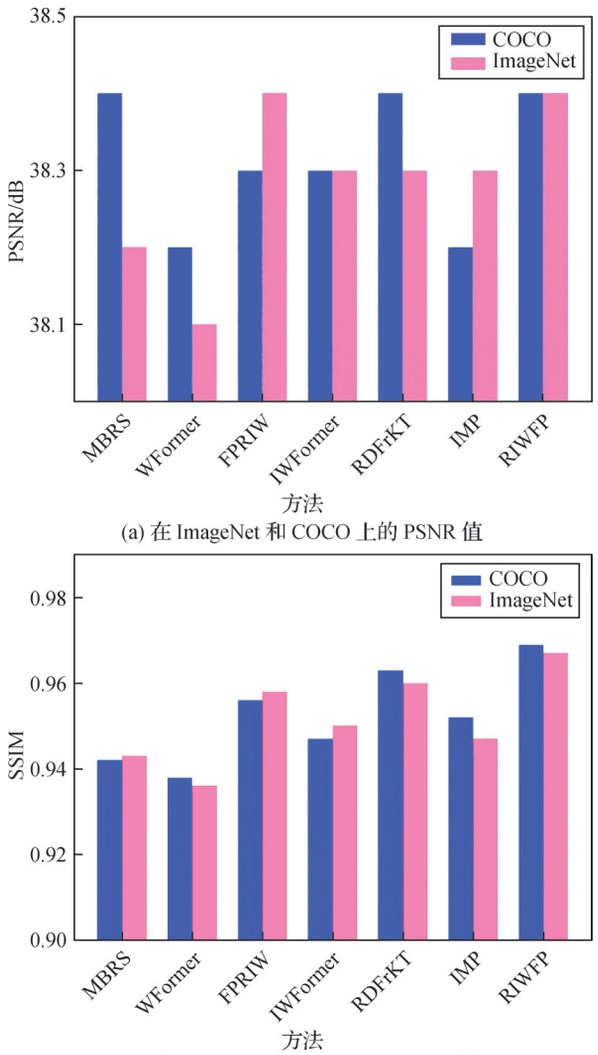
COCO 中各 2 000 幅含水印载体图像的 PSNR 以及 SSIM。由此看出，RIWFP 模型生成水印不可见性更佳，在 ImageNet 和 COCO 数据集上的 SSIM 分别达到 0.967 和 0.969，处于领先地位，其主要原因可能是因为多频率小波损失函数有利于模型关注不同的频带，生成更高质量的编码图像。

3.3　鲁棒性

1）常见图像处理攻击。图 7 展示了在 5 种常见图像处理攻击下，比较方法在 ImageNet 和 COCO 数据集上的鲁棒性表现。实验结果表明，RIWFP 在不同攻击条件下的水印提取准确度均显著优于其他对比方法，展现了其嵌入水印较强的鲁棒性。

在 COCO 数据集上的实验中，RIWFP 展现了更好的性能。在 DP-0.6 攻击下，RIWFP 的准确率达到 91.4%，超过了 IMP 的 87.5% 以及 FBRIW 的 87.6%，这可能是由于所比较方法未能充分关注图像的频域信息。此外，在 SP-0.09 和 MF-9 攻击下，RIWFP 分别以 100% 和 99.5% 的准确率达到了最高水平，这可能是由于频率感知机制有效学习水印提取所需的关键成分，从而在高强度攻击下表现出极强的鲁棒性。从方法的角度来看，流模型展现出其更优的性能，而传统方法表现不佳。综合来看，RIWFP 在 COCO 数据集上的平均准确率为 96.7%，优于其他方法。

对于 ImageNet 数据集，在 CP-0.7 攻击下，RIWFP 以 93.4% 的 ACC 进一步证明了其在学习图像特征方面的能力，使得水印在图像部分丢失的情况下依



(a) 在 ImageNet 和 COCO 上的 PSNR 值

(b) 在 ImageNet 和 COCO 上的 SSIM 值

图 6　对比方法不可见性测试结果

Fig. 6　The objective invisibility results obtained by the compared methods ((a) PSNR values on ImageNet and COCO; (b) SSIM values on ImageNet and COCO)

然保持高提取准确度。对于 QF-50 攻击，RIWFP 的 ACC 达到 99.6%，略高于 FBRIW 的 99.2%，比 WFormer 高 3.8%，显示出其对低频信息较强的学习能力，小波卷积神经网络和多频率小波损失函数的结合有效抵御了噪声对低频信息的干扰，确保了水印的稳定提取。总体来看，RIWFP 平均 ACC 为 96.9%，显著高于其他方法，比 RDFrKT 高 17.5%，这充分体现了小波卷积神经网络、特征蒸馏机制以及多频率小波损失函数的协同作用。

2）组合攻击。由于 ρ 模块是基于高低频分流的机制，选择了对高低频影响各异的攻击进行组合。

表 2 展示了在 ImageNet 数据集上，比较方法在组合攻击（包括中值滤波、DP 和 JPEG 压缩）下的鲁

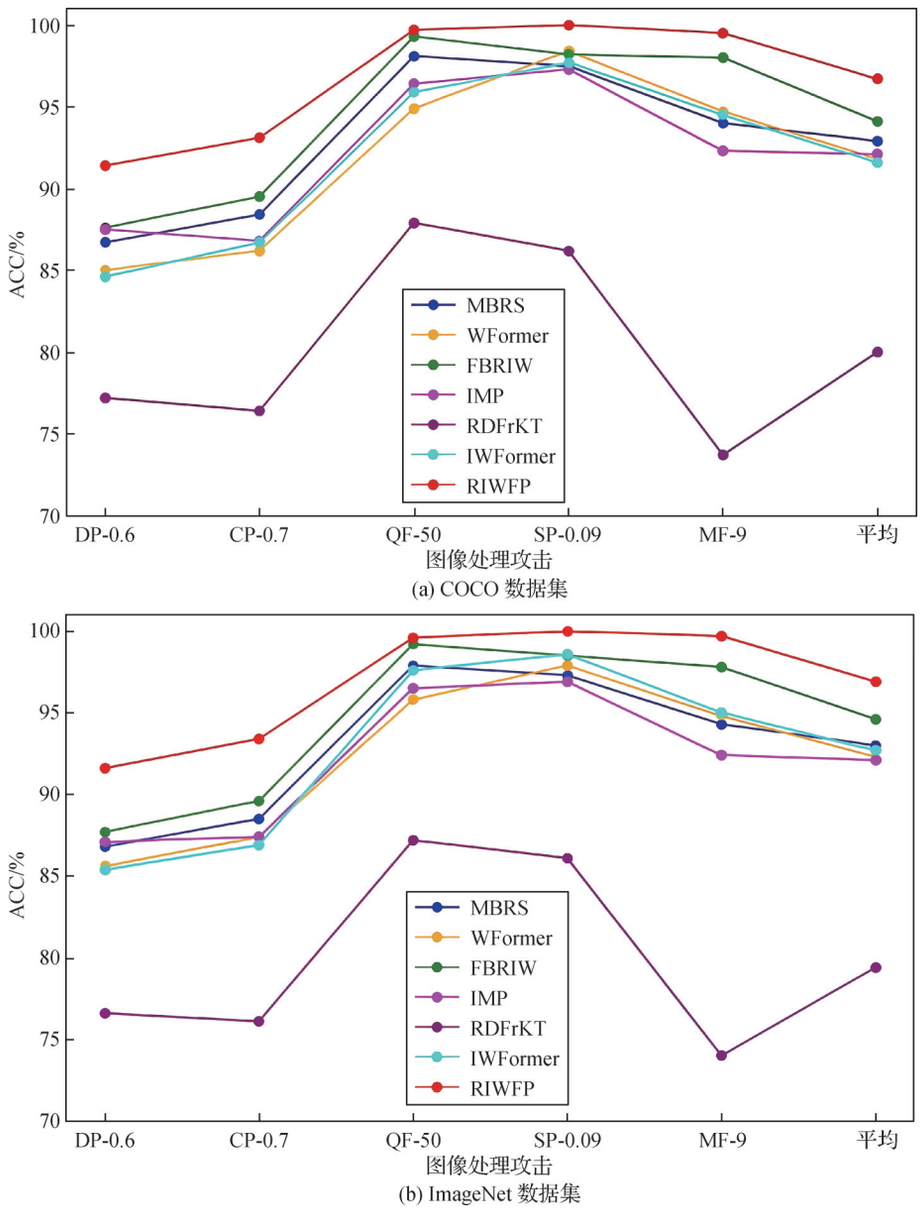


图7　对比方法在COCO和ImageNet数据集上针对5种常见图像处理攻击所获得的ACC值

Fig. 7　The ACC values obtained by the compared methods against the five common image processing attacks on the COCO dataset and the ImageNet dataset ((a) ACC values on the COCO dataset; (b) ACC values on the ImageNet dataset)

棒性。针对MF-3 + DP-0. 9 + QF-90这种组合攻击，RIWFP的准确率达到98. 2%，保持了较强的稳定性和较高的准确率，超越IMP的95. 2%，这可能是由于ρ模块有效的高低频分流机制。然而，随着DP强度的降低(MF-5 + DP-0. 8 + QF-80)，RIWFP的准确率略微下降至93. 3%，但依然保持了领先地位。

通过对ImageNet数据集上RIWFP模型的实验结果进行分析，可以看出RIWFP在应对中值滤波、DP和JPEG压缩等复杂组合攻击时表现出显著的鲁棒性，而其他方法可能由于未充分利用频率信息导致其鲁棒性不足。尽管随着攻击强度的增加，RIWFP的准确率有所下降，但其在面对较强的攻击时，依然保持了较高的性能，远超其他方法。如IMP在MF-5 + DP-0. 8 + QF-80攻击下仅达到89. 3%的准确率；WFormer在MF-5 + DP-0. 8 + QF-80攻击下仅达到84. 3%的准确率，比RIWFP低9%。这表明RIWFP具有强大的适应能力，能够有效应对不同类型的复杂攻击，而其他方法未能充分利用图像信息，导致其在面对复杂攻击时鲁棒性不足。

表3展示了不同方法在ImageNet数据集上面对两种组合攻击的表现情况。RIWFP在面对不同攻击组合时表现出较强的鲁棒性，尤其是在高强度的

攻击下。对于SP-0.03 + CP-0.9这种攻击组合，RIWFP的准确率达到了98.3%，在所有方法中表现最为优秀，显示出其在椒盐噪声和裁剪攻击的复杂组合下的鲁棒性。其他方法，如IMP和RDFrKT的准确率分别为94.2%和80.4%，明显低于RIWFP，表明RIWFP在这些干扰下的性能表现突出。

对于SP-0.04 + CP-0.8攻击，RIWFP的准确率仍然处于领先地位，为95.4%。尽管攻击强度有所提升，但RIWFP依然能够保持较高的性能，而其他方法由于未能充分利用高低频信息，导致在面对高强度组合攻击时性能欠佳。JPEG压缩攻击也是一个重要的干扰因素，对于JPEG-90 + CP-0.9攻击，RIWFP的准确率为97.8%，依然优于其他模型。JPEG压缩带来的图像质量下降对所有模型的影响较大，但RIWFP凭借小波卷积网络高效地学习图像低频信息，能够有效应对这一挑战。

表2　对比方法在3种攻击组合下获得的ACC值

Table 2　The ACC values obtained by the compared methods under three combinations of attacks

/%

攻击	MBRS	RDFrKT	IMP	FBRIW	WFormer	IWFormer	RIWFP
MF-3 + DP-0.9 + QF-90	96.3	65.9	95.2	96.8	94.7	95.7	**98.2**
MF-5 + DP-0.9 + QF-85	92.5	63.2	92.8	94.6	91.7	93.8	**96.2**
MF-5 + DP-0.9 + QF-80	93.0	62.8	91.6	94.1	88.1	90.0	**95.5**
MF-5 + DP-0.8 + QF-80	90.9	59.4	89.3	92.0	84.3	87.4	**93.3**

注：加粗字体表示各行最优结果。

表3　对比方法在两种攻击组合下获得的ACC值

Table 3　The ACC values obtained by the compared methods under two combinations of attacks

/%

攻击	MBRS	RDFrKT	IMP	FBRIW	WFormer	IWFormer	RIWFP
SP-0.03 + CP-0.9	95.5	80.4	94.2	96.2	90.4	94.7	**98.3**
SP-0.04 + CP-0.8	90.6	78.1	88.5	91.8	89.2	90.3	**95.4**
JPEG-90 + CP-0.9	94.7	76.3	95.6	95.9	93.0	94.0	**97.8**
JPEG-80 + CP_0.8	91.1	74.9	92.3	92.9	88.6	90.5	**96.3**
MF-3 + CP-0.9	95.9	67.8	95.8	96.5	93.2	94.4	**98.5**
MF-5 + CP-0.8	91.4	60.2	90.6	92.3	88.9	91.1	**96.9**
SP-0.03 + DP-0.9	99.1	82.1	98.6	98.4	97.6	98.5	**99.9**
SP-0.05 + DP-0.8	98.3	76.7	98.5	97.7	94.2	96.0	**98.7**

注：加粗字体表示各行最优结果。

3）泛化性实验。在本实验中，为了进一步评估模型在未训练攻击下的鲁棒性，引入了几种新的攻击类型，包括饱和度调、色调调整、亮度调整和高斯噪声。通过在ImageNet数据集上的实验，展示了不同攻击类型对模型性能的影响，如表4所示。在ImageNet数据集上，RIWFP在高斯噪声攻击下表现出色。在GN-0.01时，RIWFP的准确率为99.9%，几乎不受影响；在GN-0.05时，RIWFP的准确率仍然高达98.4%，明显优于其他方法。其他方法在GN-0.05的攻击下表现较差，表现出较大的准确率下降，如RDFrKT比特准确率为50.6%。WFormer比特准确率为94.9%，这可能是由于高低频双流机制使模型学习到更鲁棒的特征。

对于饱和度调整攻击，RIWFP在AS-5时表现最佳，准确率为98.5%。在亮度调整攻击下，RIWFP表现稳定。在BA-1.1时，RIWFP的准确率为98.2%，在BA-1.2时，准确率略降至94.3%，但仍然领先于其他方法，显示出较强的亮度适应能力。

色调调整攻击对RIWFP的影响较为显著。在AH-0.44时，RIWFP的准确率为98.4%，仍保持较高性能；但是随着色调因子增大（AH-0.46和AH-0.48），RIWFP的准确率有所下降，尤其在AH-0.48时，下降较为明显，表明该类型攻击对RIWFP的影响较为强烈。但对比于其他方法，RIWFP下降速度更为平稳，显示出ρ模块更强大的特征提取能力。

在对比度调整攻击下，RIWFP在AC-0.9时表现最佳，准确率为98.7%。当对比度强度加大，RIWFP的性能虽然略有下降，但下降速度缓慢，比之前的准确度仅下降1.3%；而FBRIW在对比度调整攻击强度加大时，性能大幅下降，准确度比之前降低了3.1%，由此充分显示出RIWFP在不同对比度变化下的强大适应能力。

对比于其他深度学习水印模型，RIWFP模型在各种未训练攻击下均表现出较强的鲁棒性。对比实验结果可以看出，RIWFP在面对高斯噪声、饱和度调整、亮度调整、色调调整和对比度调整等攻击时，都能保持较高的准确率，在高斯噪声和亮度调整等攻击下，RIWFP表现尤其优秀，展示了其在图像扰动下的抗干扰能力。尽管在某些强度较大的攻击（如AS-15、AH-0.48等）下，RIWFP的准确率有所下降，但整体上RIWFP的表现仍优于大多数对比方法。这表明，RIWFP具有较强的适应性和鲁棒性，其在面对未知攻击时具有显著的优势。

表4　对比方法在各种未训练攻击下获得的ACC值

Table 4　The ACC values obtained by the compared methods against various untrained attacks

/%

攻击	MBRS	RDFrKT	IMP	FBRIW	WFormer	IWFormer	RIWFP
GN-0.01	98.8	67.8	98.2	99.6	97.4	98.5	**99.9**
GN-0.05	95.5	50.6	93.1	96.2	94.9	95.3	**98.4**
AS-5	98.4	64.2	98.7	92.8	92.4	**99.3**	98.5
AS-10	**96.7**	61.7	96.2	87.6	92.0	95.7	96.3
AS-15	94.1	58.3	93.4	84.7	86.0	92.0	**94.2**
BA-1.1	95.3	70.4	95.9	95.7	94.5	94.3	**98.2**
BA-1.2	92.4	67.9	91.8	93.3	90.4	89.9	**94.3**
AH-0.44	90.2	67.5	90.1	96.4	95.1	97.6	**98.4**
AH-0.46	84.3	64.8	83.6	94.1	93.2	95.8	**97.1**
AH-0.48	80.6	61.6	80.3	89.9	91.0	93.9	**94.7**
AC-0.9	94.3	63.4	93.7	98.0	95.6	96.5	**98.7**
AC-1.1	95.9	61.1	94.5	96.3	96.1	96.7	**96.8**
AC-1.2	93.1	58.3	92.6	93.2	93.7	94.3	**95.5**

注：加粗字体表示各行最优结果。

4）消融实验。为了衡量注意力机制的有效性，本文对比了几种基于流模型水印框架的热力图，使用不同颜色展示了不同的注意程度，如图8所示。

从结果中可以明显观察到，FBRIW方法为图像的各个区域分配了相同的权重，这种均匀分配的方式虽然简单，但忽略了图像中不同频率成分对水印性能的差异性影响。具体而言，低频区域通常包含图像的主要结构信息，而高频区域则包含更多的细节和边缘信息。FBRIW的均匀权重分配可能导致水印在高频区域的鲁棒性不足，从而影响整体性能。而IWFormer仅为一部分高频信息分配了较高的注意力，未能完整学习高频信息。相比之下，RIWFP方法通过引入注意力机制，能够更有效地识别并关注图像中的高频区域。这种机制使得模型能够更好地学习不同频率特征的重要性。实验结果表明，RIWFP方法在高频区域的表现显著优于FBRIW，这不仅提高了水印的不可见性，还增强了其抗攻击能力。

表5展示了在ImageNet数据集上，RIWFP在不

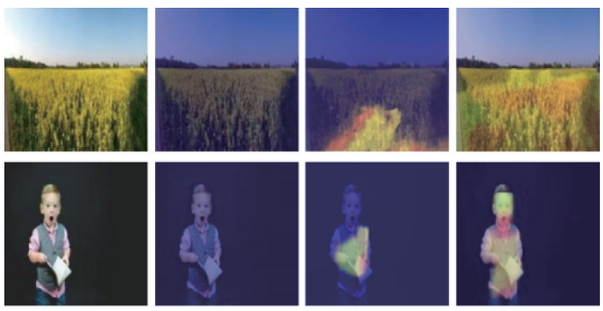


图 8　视觉热力图

Fig. 8　The visual attention heat map

同组件组合下的不可见性和鲁棒性变化。在消融实验中，RIWFP 结合小波卷积和蒸馏机制，展现了全面的性能优势。从具体数据来看，RIWFP 在 CP 攻击下的准确度为 93.4%，比无小波卷积和无蒸馏机制分别高出 1.9% 和 0.6%，表明其对图像整体信息的学习能力更强。在 DP 和 JPEG 压缩攻击下，RIWFP 的准确度分别为 91.6% 和 99.6%，比无小波卷积分别高出 1.5% 和 1.3%，体现了小波卷积对低频信息学习的重要性。而在椒盐噪声和中值滤波攻击下，RIWFP 的准确度分别为 100% 和 99.7%，比无蒸馏机制分别高出 1.3 和 1.8 个百分点，凸显了蒸馏机制对高频信息学习的贡献。从平均值来看，RIWFP 的整体准确度为 96.9%，高于无小波卷积 95.5% 和无蒸馏机制 95.8%，进一步验证了小波卷积和蒸馏机制结合的有效性。低频和高频信息的协同学习使 RIWFP 在面对多种攻击时均表现出色，尤其是在高频相关的噪声和滤波攻击下表现更为突出。

表 5　RIWFP 的消融实验结果

Table 5　The ablation experimental results obtained by RIWFP

组件	SSIM	PSNR/dB	DP-0.4/%	CP-0.7/%	JPEG-50/%	SP-0.09/%	MF-9/%	平均/%
RIWFP-小波卷积神经网络	0.966	**38.4**	90.1	91.5	98.3	99.3	98.5	95.5
RIWFP-特征蒸馏机制	0.965	38.3	91.2	92.8	98.6	98.7	97.9	95.8
RIWFP	**0.967**	**38.4**	**91.6**	**93.4**	**99.6**	**100**	**99.7**	**96.9**

注：加粗字体表示各列最优结果。

4　结　论

1）主要内容。本文提出一种频率感知驱动的深度鲁棒图像水印框架（RIWFP），通过分别采用不同的学习机制处理低频和高频信息，实现了图像特征的精细化处理。不同于大多数深度学习方法将低频和高频信息等同对待，RIWFP 在水印生成过程中，通过差异化的机制提升水印的鲁棒性与不可见性。具体而言，低频信息通过小波卷积神经网络建模，高频信息则采用深度可分离卷积和注意力机制强化图像细节。同时设计了多频率小波损失函数，进一步提升生成图像质量。

2）结果与分析。实验结果表明，所提出的 RIWFP 方法在多种攻击场景下均表现出色。RIWFP 在 COCO 和 ImageNet 数据集上的平均准确率分别为 96.7% 和 96.9%，高于其他现有方法。在 ImageNet 数据集上，对 MF-5 + DP-0.8 + QF-80 3 种攻击组合的准确率达到 93.3%；对 SP-0.04 + CP-80 两种组合攻击的准确率达到 95.4%，超出次佳方法 3.6%。通过消融实验，验证了 RIWFP 结合小波卷积和蒸馏机制在提升水印鲁棒性方面的显著效果，特别是在高频噪声和滤波攻击的情况下表现突出。与经典以及深度的图像水印方法相比，RIWFP 通过频率感知的方式处理低频和高频信息，避免了对图像频带特征差异的忽视。实验结果显示，RIWFP 在应对不同类型的攻击时，表现出更高的鲁棒性。这一优势来源于 RIWFP 对于低频部分的全局结构学习以及对高频部分的细节强化机制的有效结合。

3）不足及原因。尽管 RIWFP 在对抗多种攻击方面表现出强大的鲁棒性，但期望它在所有潜在的对抗性场景下都能具备普遍的鲁棒性是不现实的，特别是在涉及复杂失真或特定领域噪声的情况下。例如，复杂的光学失真、纸张纹理噪声和颜色偏移可能会对水印的鲁棒性造成一定影响。此外，RIWFP 中的特征蒸馏模块和小波卷积神经网络的集成可能会增加计算开销，这对于实时应用或资源受限的环境可能是一个限制。

4)展望。未来的研究将专注于定量分析在打印—相机中出现的失真(如光学失真、纸张纹理噪声和颜色偏移),计划设计针对噪声层的联合训练,以更有效地模拟现实世界中的失真。这种方法将提高水印在打印—相机噪声和其他领域特定挑战中的鲁棒性。为了解决计算复杂性问题,我们将探索轻量级架构和优化技术,以减少RIWFP的计算开销,使其更适合实时应用。

参考文献(References)

Agustsson E and Timofte R. 2017. NTIRE 2017 challenge on single image super-resolution: dataset and study//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops. Honolulu, USA: IEEE: 1122-1131 [DOI: 10.1109/cvprw.2017.150]

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

Deng J, Dong W, Socher R, Li L J, Li K and Li F F. 2009. ImageNet: a large-scale hierarchical image database//Proceedings of 2009 IEEE Conference on Computer Vision and Pattern Recognition. Miami, USA: IEEE: 248-255 [DOI: 10.1109/cvprw. 2009.5206848]

Ding W P, Ming Y R, Cao Z H and Lin C T. 2022. A generalized deep neural network approach for digital watermarking analysis. IEEE Transactions on Emerging Topics in Computational Intelligence, 6(3): 613-627 [DOI: 10.1109/tetci.2021.3055520]

Dumitrescu S, Wu X L and Wang Z. 2003. Detection of LSB steganography via sample pair analysis. IEEE Transactions on Signal Processing, 51(7): 1995-2007 [DOI: 10.1109/tsp.2003.812753]

Evsutin O and Dzhanashia K. 2022. Watermarking schemes for digital images: robustness overview. Signal Processing: Image Communication, 100: #116523 [DOI: 10.1016/j.image.2021.116523]

Fang H, Qiu Y P, Chen K J, Zhang J Y, Zhang W M and Chang E C. 2023. Flow-based robust watermarking with invertible noise layer for black-box distortions//Proceedings of the 37th AAAI Conference on Artificial Intelligence. Washington, USA: AAAI: 5054-5061 [DOI: 10.1609/aaai.v37i4.25633]

Guo H and Georganas N D. 2003. Digital image watermarking for joint ownership verification without a trusted dealer//Proceedings of 2003 International Conference on Multimedia and Expo. Baltimore, USA: IEEE: 497-500 [DOI: 10.1109/ICME.2003.1221662]

He Z Y, Hu R Z, Wu J, Luo T and Xu H Y. 2024. A transformer-based invertible neural network for robust image watermarking. Journal of Visual Communication and Image Representation, 104: #104317 [DOI: 10.1016/j.jvcir.2024.104317]

Hu K, Wang M P, Ma X H, Chen J, Wang X C and Wang X J. 2024. Learning-based image steganography and watermarking: a survey. Expert Systems with Applications, 249: #123715 [DOI: 10.1016/j.eswa.2024.123715]

Huang J T, Luo T, Li L, Yang G B, Xu H Y and Chang C C. 2023. ARWGAN: attention-guided robust image watermarking model based on GAN. IEEE Transactions on Instrumentation and Measurement, 72: #5018417 [DOI: 10.1109/tim.2023.3285981]

Huynh-Thu Q and Ghanbari M. 2008. Scope of validity of PSNR in image/video quality assessment. Electronics Letters, 44(13): 800-801 [DOI: 10.1049/el:20080522]

Jia Z Y, Fang H and Zhang W M. 2021. 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. Virtual Event: ACM: 41-49 [DOI: 10.1145/3474085.3475324]

Kadian P, Arora S M and Arora N. 2021. Robust digital watermarking techniques for copyright protection of digital data: a survey. Wireless Personal Communications, 118(4): 3225-3249 [DOI: 10.1007/s11277-021-08177-w]

Li Z, Liu Y, Chen X, Cai H, Gu J, Qiao Y, et al. 2022. Blueprint separable residual network for efficient image super-resolution//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. New Orleans, USA: IEEE: 832-842 [DOI: 10.1109/CVPRW56347.2022.00099]

Lin T Y, Maire M, Belongie S, Hays J, Perona P, Ramanan D, et al. 2014. Microsoft COCO: common objects in context//Proceedings of the 13th European Conference on Computer Vision——ECCV 2014. Zurich, Switzerland: Springer: 740-755 [DOI: 10.1007/978-3-319-10602-1_48]

Liu X L, Wu Y F, Xiao X Y, Xiao J L and Wang H J. 2022. Real discrete fractional Krawtchouk transform with an application to image watermarking. Journal of Image and Graphics, 27(1): 252-261 (刘西林, 吴永飞, 肖翔宇, 肖嘉龙, 王和锦. 2022. 实离散分数Krawtchouk变换及其在数字图像水印中的应用. 中国图象图形学报, 27(1): 252-261) [DOI: 10.11834/jig.200829]

Liu Y, Guo M X, Zhang J, Zhu Y S and Xie X D. 2019. A novel two-stage separable deep learning framework for practical blind watermarking//Proceedings of the 27th ACM International Conference on Multimedia. Nice, France: ACM: 1509-1517 [DOI: 10.1145/3343031.3351025]

Luo T, Wu J, He Z Y, Xu H Y, Jiang G Y and Chang C C. 2024a. WFormer: a transformer-based soft fusion model for robust image watermarking. IEEE Transactions on Emerging Topics in Computational Intelligence, 8(6): 4179-4196 [DOI: 10.1109/tetci.2024.3386916]

Luo Y J, Zhou T Q, Cui S L, Ye Y F, Liu F and Cai Z P. 2024b. Fixing the double agent vulnerability of deep watermarking: a patch-level

solution against artwork plagiarism. IEEE Transactions on Circuits and Systems for Video Technology, 34(3): 1670-1683 [DOI: 10.1109/tcsvt.2023.3295895]

Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. 2019. PyTorch: An imperative style, high-performance deep learning library//Proceedings of the 33rd Conference on Neural Information Processing Systems. Vancouver, Canada : NeurIPS: 8024-8035 [DOI: 10.5555/3454287.3455008]

Tancik M, Mildenhall B and Ng R. 2020. StegaStamp: invisible hyperlinks in physical photographs//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 2114-2123 [DOI: 10.1109/cvpr42600.2020.00219]

Tang Y C, Wang C T, Xiang S J and Cheung Y M. 2024. A robust reversible watermarking scheme using attack-simulation-based adaptive normalization and embedding. IEEE Transactions on Information Forensics and Security, 19: 4114-4129 [DOI: 10.1109/tifs.2024.3372811]

van Schyndel R G, Tirkel A Z and Osborne C F. 1994. A digital watermark//Proceedings of the 1st International Conference on Image Processing. Austin, USA: IEEE: 86-90 [DOI: 10.1109/icip.1994.413536]

Wan W B, Wang J, Zhang Y M, Li J, Yu H and Sun J D. 2022. A comprehensive survey on robust image watermarking. Neurocomputing, 488: 226-247 [DOI: 10.1016/j.neucom.2022.02.083]

Wang B W, Wu Y F and Wang G L. 2023. Adaptor: improving the robustness and imperceptibility of watermarking by the adaptive strength factor. IEEE Transactions on Circuits and Systems for Video Technology, 33(11): 6260-6272 [DOI: 10.1109/tcsvt.2023.3265970]

Wang Z, Bovik A C, Sheikh H R and Simoncelli E P. 2004. Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing, 13(4): 600-612 [DOI: 10.1109/tip.2003.819861]

Wu Y F, Wang B W, Dai C Y, Yuan Y, Li B, Zheng W Q, et al. 2023. Enhancing robustness and imperceptibility of blind watermarking with improved message processor//Proceedings of 2023 ICASSP IEEE International Conference on Acoustics, Speech and Signal Processing. Rhodes Island, Greece: IEEE: 1-5 [DOI: 10.1109/icassp49357.2023.10097063]

Zhang X Y, Li R Y, Yu J W, Xu Y M, Li W Q and Zhang J. 2024. EditGuard: versatile image watermarking for tamper localization and copyright protection//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 11964-11974 [DOI: 10.1109/cvpr52733.2024.01137]

Zhong X, Huang P C, Mastorakis S and Shih F Y. 2021. An automated and robust image watermarking scheme based on deep neural networks. IEEE Transactions on Multimedia, 23: 1951-1961 [DOI: 10.1109/tmm.2020.3006415]

Zhu J R, Kaplan R, Johnson J and Li F F. 2018. HiDDeN: hiding data with deep networks//Proceedings of the 15th European Conference on Computer Vision——ECCV 2018. Munich, Germany: Springer: 682-697 [DOI: 10.1007/978-3-030-01267-0_40]

作者简介

张国富，男，教授，硕士生导师，主要研究方向为图像安全。
E-mail: zgf@hfut.edu.cn
苏兆品，通信作者，女，副教授，硕士生导师，主要研究方向为图像多媒体安全。E-mail: szp@hfut.edu.cn
李鑫，女，硕士研究生，主要研究方向为图像水印技术。
E-mail: 2022171197@mail.hfut.edu.cn
方涵，男，研究员，主要研究方向为图像水印和信息隐藏。
E-mail: fanghan@nus.edu.sg
廉晨思，女，高级工程师，主要研究方向为多媒体安全。
E-mail: lchsi324@163.com