中图法分类号:TP37　　**文献标识码**: A　　**文章编号**: 1006-8961(XXXX)XX-0001-15

论文引用格式:Si Ruotong, Tang Yichao, Zhang Xinpeng, Li Sheng, Qian Zhenxing. An Image Watermarking Method for Camera Imaging Style Protection[J/OL]. Journal of Image and Graphics, XXXX:1-15. DOI: 10.11834/jig.260043. (司若童, 汤毅超, 张新鹏, 李晟, 钱振兴. 保护相机成像风格的水印方法[J/OL]. 中国图象图形学报,XXXX:1-15. DOI: 10.11834/jig.260043.)[DOI:10.11834/jig.260043]

保护相机成像风格的水印方法

司若童,汤毅超,张新鹏,李晟,钱振兴

复旦大学计算与智能创新学院,上海　200082

摘　要:**目的**　由相机图像信号处理(image signal processing,ISP)流程所决定的成像风格是制造商的核心知识产权。然而,攻击者可通过代理模型攻击来窃取该风格。具体来说,攻击者利用采集的RAW-RGB图像对训练代理ISP模型,生成与目标相机风格高度相似的RGB图像。现有水印方法主要针对常规信号攻击和物理信道攻击设计,难以抵抗此类非线性的代理模型攻击。为此,本文提出一种面向代理模型攻击的相机成像风格保护鲁棒水印方法StyleSign。**方法**　该方法基于端到端设计,通过联合优化水印编码器、内部代理模块和解码器三个模块实现对成像风格的保护。首先,设计多尺度水印编码器,其中采用注意力机制与离散小波变换相结合的模块,以增强水印鲁棒性。然后,设计内部代理模块,用于在训练过程中模拟代理模型攻击。该模块采用双分支网络结构,去马赛克分支基于全局引导色彩映射网络准确模拟图像风格,RAW分支采用基于离散小波变换和通道注意力机制的U-Net结构以在模拟成像风格的同时保留水印信息。最后,利用编码器和内部代理模块的输出对解码器进行联合优化,使其能够从攻击者所采用的代理ISP模型输出的图像中准确提取水印。**结果**　在Zurich RAW to RGB数据集上的实验结果表明,StyleSign对图像质量影响较小,水印图像在PSNR(37.26 dB)、SSIM(0.9893)和LPIPS(0.0425)等指标上均接近原始图像质量。该方法在RAW-to-sRGB、AWNet、MW-ISPNet和Airia CG这四种代理模型攻击下均表现出较好的鲁棒性,水印提取误码率分别低至1.07%、1.19%、0.99%和0.49%,优于对比水印方案。**结论**　所提出的水印框架能够在多种代理模型攻击场景下保持水印的鲁棒性与可提取性,为相机成像风格的知识产权保护提供了一种有效且具备泛化能力的技术方案。

关键词:水印;相机;图像信号处理;版权保护;成像风格

An Image Watermarking Method for Camera Imaging Style Protection

Si Ruotong, Tang Yichao, Zhang Xinpeng, Li Sheng, Qian Zhenxing

College of Computer Science and Artificial Intelligence, Fudan University, Shanghai 200082, China

Abstract: **Objective** The imaging style of a digital camera, determined by its proprietary image signal processing (ISP) pipeline, constitutes a core intellectual property and a critical brand asset for manufacturers. It encompasses distinct visual characteristics including color tendency, tone and atmosphere, spatial sharpness and detail, and noise reduction, which together form brand-identifiable aesthetics, as exemplified by Canon's Picture Style system and Nikon's Vivid mode. However, the widespread availability of open-source deep ISP models and large-scale paired RAW-RGB datasets has made surrogate model attacks a severe threat. In such attacks, an adversary can train a data-driven deep ISP network on paired RAW-RGB datasets, where RAW images are collected by the adversary's device and RGB images are captured by the target camera, to mimic its proprietary imaging style with high fidelity, and can even launch black-box theft without revealing

收稿日期:2026-01-19;**修回日期**:2026-04-30

基金项目:国家自然科学基金(62572125、62502093);上海自然科学基金(25ZR1401019)

Supported by: National Natural Science Foundation of China(62572125、62502093); Natural Science Foundation of Shanghai(25ZR1401019)

the model's structure or parameters. Existing digital watermarking methods are predominantly designed to resist conventional signal processing attacks or physical channel attacks, and they prove inadequate against surrogate model attacks. These attacks involve a highly nonlinear, data-driven transformation that can inadvertently destroy embedded watermarks during the style learning process, which is fundamentally distinct from traditional attack paradigms. To solve this problem, this paper proposes StyleSign, a robust watermarking framework specifically designed to protect camera imaging styles against surrogate model attacks. The framework embeds an invisible watermark into every output RGB image of the protected ISP pipeline, thereby ensuring that the watermark information survives the nonlinear transformations of surrogate attacks and remains reliably extractable from the attacker's generated outputs, providing verifiable evidence of style theft. **Method** StyleSign adopts an end-to-end trainable architecture that jointly optimizes three core modules: a multi-scale watermark encoder, an internal surrogate module, and a decoder. Specifically, the multi-scale watermark encoder is embedded into the protected ISP pipeline to imperceptibly embed a binary watermark into the final RGB image. To enhance robustness against subsequent nonlinear transformations, the encoder employs a squeeze-and-excitation-based discrete wavelet transform (SEDWT) module as its core unit. This module decomposes the fused image and watermark features into multiple frequency sub-bands and applies channel attention module to emphasize style-relevant components, allowing the watermark to be embedded into style-relevant features rather than relying on the semantic content of the image. The key innovation of this framework is the internal surrogate module, which is designed to simulate the behavior of an attacker's surrogate ISP model during training. This module takes the same RAW image as the ISP pipeline as input and learns to reconstruct the watermarked RGB image, effectively mimicking the style transfer process of surrogate attack while preserving the embedded watermark. Architecturally, it adopts a dual-branch design. The demosaicing branch leverages a global guided color mapping (GGCM) network to accurately capture and model the global color and tone characteristics of the target imaging style. Meanwhile, the RAW branch utilizes U-Net structure in which standard pooling operations are replaced with discrete wavelet transforms to retain high-frequency watermark details during downsampling and reconstruction, and an efficient channel attention module is integrated into the skip connections to further enhance watermark-related features. The final output is the pixel-wise average of the two branches' results. Finally, the decoder is jointly optimized on two inputs: the directly watermarked image from the encoder and the mimic image generated by the internal surrogate module. **Result** Experimental results on the Zurich RAW to RGB dataset demonstrate the effectiveness of StyleSign. In terms of fidelity, the embedded watermark introduces negligible visual distortion, with watermarked images attaining a peak signal-to-noise ratio (PSNR) of 37.26 dB, a structural similarity index measure (SSIM) of 0.9893, and a learned perceptual image patch similarity (LPIPS) of 0.0425, all close to the original image quality and outperforming the compared watermarking schemes. In terms of robustness, StyleSign demonstrates strong performance under both conventional image processing attacks and surrogate model attacks. For conventional attacks including color jitter, Gaussian noise, Gaussian blur, JPEG compression, and resizing, the watermark extraction bit error rate (BER) ranges from 0.63% to 0.83%, showing robustness against these distortions. More importantly, under four representative surrogate model attacks with different architectures, namely RAW-to-sRGB, AWNet, MW-ISPNet, and Airia CG, StyleSign achieves BER values as low as 1.07%, 1.19%, 0.99%, and 0.49%, respectively, outperforming the compared watermarking schemes. Ablation studies further verify that the internal surrogate module is indispensable for ensuring robustness against surrogate model attacks, while the discriminator effectively ensures the visual quality of the watermarked images. **Conclusion** The proposed StyleSign framework effectively solves the problem that it is difficult for existing watermarking methods to resist surrogate model attacks for camera imaging style theft. Through the joint optimization of the multi-scale watermark encoder, the dual-branch internal surrogate module, and the decoder, StyleSign maintains excellent watermark robustness and reliable extractability across various surrogate model attack scenarios, while having a minor impact on image quality. This work provides an effective and generalizable technical solution for protecting the intellectual property of camera imaging styles, and also offers a new research idea for the core intellectual property protection of camera manufacturers.

Key words: Watermarking; Camera; Image Signal Processing; Copyright Protection; Camera Imaging Style

Si Ruotong, Tang Yichao, Zhang Xinpeng, Li Sheng, Qian Zhenxing. An Image Watermarking

Method for Camera Imaging Style Protection—SCID[J/OL]. Journal of Image and Graphics. DOI:10.11834/jig.260043.(司若童,汤毅超,张新鹏,李晟,钱振兴.保护相机成像风格的水印方法—SCID[J/OL].中国图象图形学报.DOI:10.11834/jig.260043.)

0 引 言

相机的成像风格是在硬件和软件协同工作下设计和定义的独特视觉特征,通过对色彩倾向、色调与氛围、空间锐度与细节及降噪特性等图像处理要素的调控,形成具有品牌辨识度的视觉特征(Bose等,2003)。这种风格主要源自图像信号处理(image signal processing,ISP)中的RAW-to-RGB转换流程。ISP是相机中的关键处理单元,负责将传感器捕获的RAW图像转换为适合人类视觉感知的RGB图像,该过程通过白平衡、去马赛克、色彩校正等一系列专用模块实现(Ramanath等,2005)。例如佳能的Picture Style系统通过调节饱和度、锐度与色调曲线生成品牌特有的美学风格(Timacheff等,2011),而尼康的Vivid模式则采用基于色相分区的非线性饱和度增强,同时搭载针对单反镜头特性优化的边缘感知锐化算法(Hess等,2013)来形成特有的成像风格。这些独特的成像特征是相机品牌辨识度的关键,是重要知识产权与品牌资产。

随着神经网络在图像信号处理中的广泛应用,针对相机成像风格的代理模型攻击逐渐成为现实威胁,攻击者可利用大量已开源的基于深度学习的ISP模型(Guo等,2025;Li等,2025)训练代理ISP模型,从而窃取目标相机的成像风格。研究表明,局部色调映射、去马赛克和降噪等风格定义操作均可被转化为可学习的函数。这类ISP模型在具备数千组RAW-RGB图像对的条件下,即可训练出性能较好的网络,生成高质量的RGB图像。目前已有一些专为学习相机ISP设计的大规模公开数据集(Lu等,2025),利用此类数据集训练的ISP模型,能够将智能手机传感器捕获的RAW图像转换为接近单反相机成像质量与风格的RGB图像。除利用公开数据集外,攻击者亦可自行构建配对数据集,例如通过固定拍摄场景,使用智能手机采集RAW图像并同步用专业单反相机获取对应的RGB图像,进而通过代理模型攻击窃取相机的成像风格。此类攻击无需公开代理模型的具体结构或参数,即可以黑盒形式提供RAW到RGB的映射服务,从而规避直接检测与取证,这对相机厂商的核心数字资产与品牌价值构成威胁。然而,现有研究多聚焦于高保真地模拟相机ISP成像风格,针对代理模型攻击场景下的相机成像风格保护问题尚缺乏有效的解决方案。

为解决上述问题,本文提出一种端到端水印框架StyleSign,用于保护相机成像风格免受代理模型攻击。该框架的整体结构如图1所示,将水印编码器嵌入受保护的ISP流程中,使得相机拍摄的RGB图像中均含有水印,若攻击者使用这些RGB图像训练代理模型,其输出的模拟RGB图像中仍可能保留水印信息。版权方可从攻击者输出的图像中利用解码器提取水印,若误码率(bit error rate,BER)低于预设阈值,即可为侵权鉴定提供技术支撑。

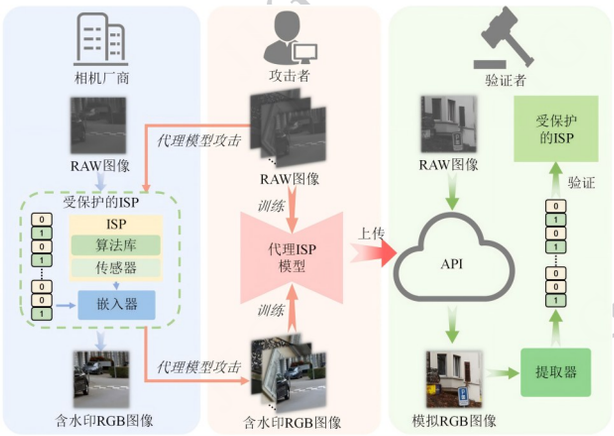

图1 基于水印的相机成像风格版权保护框架

Fig. 1 Framework for watermarking-based copyright protection of camera imaging styles

StyleSign设计了三个联合优化的模块:多尺度水印编码器、内部代理模块与解码器。具体实现路径如下:多尺度水印编码器以通道注意力模块(squeeze-and-excitation networks,SENet)(Hu等,2018)为主体,结合离散小波变换构建SEDWT(squeeze-and-excitation-based discrete wavelet transform)模块,实现多尺度水印嵌入,增强水印鲁棒性。内部代理模块采用双分支结构模拟代理模型攻击行为:去马赛克分支通过全局引导色彩映射网络实现对目标成像风格的高保真模拟;RAW分支则采用基于U-Net(Ronneberger等,2015)的架构,以离散小波变换替代传统池化操作,减少高频信息损失,并引入

高效通道注意力模块增强对水印信息的保持能力。解码器则在训练阶段利用内部代理模块的输出进行优化，使其能够适应代理模型攻击对水印分布的影响，从而从攻击者输出的图像中有效提取水印。

为验证所提出的StyleSign方法的有效性，本文开展如下实验：首先进行图像质量评估，通过峰值信噪比（peak signal-to-noise ratio，PSNR）、结构相似性指数（structural similarity index measure，SSIM）和学习感知图像块相似度（learned perceptual image patch similarity，LPIPS）等指标量化原始图像与水印图像的差异，验证水印嵌入对视觉效果的影响，实验表明，StyleSign对图像质量的影响较小。其次进行鲁棒性评估，测试StyleSign抵抗常规攻击和代理模型攻击的能力，无论面对常规　　攻击还是代理模型攻击，StyleSign均能保持较低的BER，展现出较强的泛化性与鲁棒性。

1　相关工作

1.1　深度ISP模型

随着深度学习在计算机视觉领域的广泛应用，使用数据驱动的神经网络替代或增强传统图像信号处理器已成为重要研究方向（Dos等，2025）。深度ISP模型旨在学习从RAW图像到高质量RGB图像的映射，在提升图像质量、优化处理效率以及适应复杂场景方面展现出较好的效果。主要可概括为以下几种技术路线：

端到端映射。这是早期深度ISP研究的主要方向，核心思想是用神经网络模型直接学习从RAW到RGB的整体映射关系。此类模型通常采用编码器-解码器架构（如U-Net），通过大量成对的RAW-RGB数据进行训练，以像素级重建损失（如L1、L2损失）和感知损失作为优化目标。代表性工作包括PyNET（Ignatov等，2020）及其改进版本PyNET-CA（Kim等，2020），采用倒金字塔结构实现由粗到精的图像生成；AWNet（attentive wavelet network）（Dai等，2020）通过引入小波变换和注意力机制，进一步增强高频细节的保留能力。

分阶段ISP建模。为了提升模型的可解释性和在不同子任务上的性能，部分研究将ISP流程显式划分为多个处理阶段。这类方法通常将成像过程拆分为低级恢复（如去马赛克、降噪、白平衡）与高级增强（如色调映射、色彩风格调整）等子任务，并为不同阶段设计相应的子网络结构。通过任务解耦，可以针对不同性质的处理步骤引入更合适的网络结构与损失函数。代表性工作如CameraNet（Liang等，2021）和RAW-to-sRGB（Zhang等，2021），明确采用双阶段框架进行建模；TENet（Lu等，2021）则通过重新设计并固定任务执行顺序以优化信息流；Uni-ISP（Li等，2025）通过引入设备感知嵌入和特殊的跨相机训练策略，同时学习多个相机的正向与逆向ISP，从而提升网络性能。

性能重建与细节恢复。为追求更高的成像质量，研究者进一步探索了以重建精度和细节恢复能力为核心目标的深度ISP方法。这类工作通常在网络架构中引入更精细的频域建模、注意力机制或多分支融合策略，以缓解传统卷积操作中的信息损失，从而提升色彩还原与细节保真度，代表性工作包括MW-ISPNet（Ignatov等，2020）创新性地将多级小波变换引入网络架构，取得了较好的细节重建效果；Airia CG（Ignatov等，2020）针对图像质量的不同评价维度采取了差异化设计，在PSNR指标上实现了显著提升；Dark-ISP（Guo等，2025）则针对低光条件下的RAW-RGB映射进行研究，实现了较好的重建性能。

上述研究表明，随着深度ISP模型的发展，攻击者已能够利用此类模型作为代理ISP模型，精准模拟相机的成像风格，这对成像风格保护构成了新的挑战。

1.2　数字图像水印

数字图像水印技术是指在图像中不可察觉地嵌入认证信息的过程，该技术是知识产权保护和创作者署名的重要工具。传统水印方案包括空域方法和频域方法（Tang等，2024；Tang等，2025），这些方法对常见传输失真（如JPEG压缩、高斯噪声、维纳滤波及几何变换）展现出较强的鲁棒性。

随着深度学习技术的快速发展，水印技术从传统的人工特征设计转向自动学习数据特征来优化水印的嵌入与提取（Zhang等，2026；Guo等，2026）。根据所能抵抗的攻击类型不同，现有深度水印方案可大致分为以下两类：抵抗常规信号攻击的水印和抵抗物理信道攻击的水印。

抵抗常规信号攻击的水印。这类方案主要针对图像在传输、存储和编辑过程中常见的数字失真，如

JPEG压缩、噪声添加、模糊和几何变换等。HiDDeN（hiding data with deep network）（Zhu等，2018）提出了首个端到端可训练的深度水印模型，通过对抗训练联合优化编码器、失真层和解码器，实现了优异的不可感知性及对常见攻击的鲁棒性。这一框架为后续深度水印技术的发展奠定了重要基础。MBRS（mini-batch of real and simulated JPEG compression）（Jia等，2021）采用模拟真实JPEG压缩的混合失真层，显著增强了对JPEG伪影的鲁棒性；而DWSF（dispersed watermarking with synchronization and fusion）（Guo等，2023）则通过分散嵌入策略，提升了对几何攻击与复合攻击的抵抗能力；RAIMARK（resolution-agnostic implicit watermarking based on implicit neural representation）（Wang等，2024）将图像转化为连续的数学函数后嵌入水印，提升了对裁剪、压缩、缩放等攻击的鲁棒性；BICA（watermarking using bidirection-interactive and context-aware networks）（Yin等，2025）引入双向交叉注意力模块，融合局部与全局特征以提升水印鲁棒性。

抵抗物理信道攻击的水印。这类方案侧重于应对图像在物理世界中传播时所经历的复杂干扰，如打印-扫描、屏幕翻拍、覆膜等。StegaStamp（Steganographic Stamp）（Tancik等，2020）通过物理可实现的噪声建模，提出了可经受打印拍摄过程的物理鲁棒水印；PIMoG（Fang等，2022）通过引入专门模拟屏幕显示特性的失真层，有效抵抗了屏幕翻拍攻击；WRAP（watermarking approach robust against film-coating upon printed photographs）（Liu等，2023）专门针对覆膜图像场景，实现了从覆膜照片中稳定提取水印；Physical Marker（Xue等，2025）针对透明背景商标的打印-拍摄场景，实现了在透明背景商标中稳定提取水印信息；RoPaSS（robust watermarking for partial screen-shooting scenarios）（Ma等，2025）重点研究了抗部分屏幕拍摄的鲁棒性，并对该方案面临的挑战和解决方案进行了深入分析。PEE框架（post-encoding enhancement）（Liu等，2025）针对屏幕拍摄攻击引入透视变形、光学模糊、传感器噪声等混合失真，提高了水印对屏幕拍摄攻击的鲁棒性。

现有方案虽然能有效抵抗常规信号攻击和物理信道攻击，但尚未针对代理模型攻击场景下的相机成像风格保护进行专门研究。常规攻击对图像进行数字处理，引入信号失真；物理攻击则因跨载体传播导致几何畸变、色彩偏移等失真；而代理模型攻击通过训练神经网络窃取图像风格，其自适应学习过程会无意识地破坏水印，攻击机理与前述两类有本质不同，因此现有方案难以有效抵抗此类攻击。

2 方　法

2.1 整体架构

为保护相机成像风格免受代理模型攻击，提出一种端到端的水印方法StyleSign，其整体架构如图2所示。StyleSign的核心思想是将风格窃取攻击（即代理模型的训练过程）本身纳入水印方法的训练中作为失真层，提升水印对代理模型攻击的鲁棒性。为实现这一目标，构建了包含三个模块的端到端框架。三个模块及其功能如下：

多尺度水印编码器 E：内置于受保护的ISP中，将水印信息 $\boldsymbol{M}$ 嵌入ISP输出的RGB图像 $\boldsymbol{I}$ 中，生成水印图像 $\boldsymbol{I}_{\mathrm{w}}$，在视觉上与原始图像 $\boldsymbol{I}$ 保持一致。该模块将水印融合到图像中与风格相关的特征中。

内部代理模块 A：模拟攻击者训练代理ISP模型的行为，以RAW图像 $\boldsymbol{R}$ 为输入，输出模拟攻击后的RGB图像 $\boldsymbol{I}_{\mathrm{A}}$，并以水印图像 $\boldsymbol{I}_{\mathrm{w}}$ 作为学习目标。该模块在学习成像风格的同时，需保留 $\boldsymbol{I}_{\mathrm{w}}$ 中的水印信息。

水印解码器 D：从攻击者生成的模拟RGB图像中提取水印信息。在训练时从两类图像中提取水印：一是原始水印图像 $\boldsymbol{I}_{\mathrm{w}}$，二是内部代理模块输出的模拟RGB图像 $\boldsymbol{I}_{\mathrm{A}}$。通过同时优化对这两类图像的提取性能，解码器能够适应代理模型攻击引入的分布变化，从而在真实攻击场景下仍能可靠提取水印。

在训练阶段，本框架通过单一前向传播与联合优化进行训练，损失函数综合了图像保真度损失、水印提取损失、对抗损失和代理重建损失，实现水印嵌入、代理攻击模拟与水印提取的联合学习。部署时，训练好的多尺度水印编码器 E 内置于受保护ISP中，对目标ISP生成的每张RGB图像进行水印嵌入。因此攻击者在构建数据集时，拍摄得到的RGB图像都包含水印。在检测阶段，收集可疑代理模型 S 的输出 $\boldsymbol{I}_{\mathrm{S}}$，采用解码器 D 提取水印 $\boldsymbol{M}'$，若BER低于预设的阈值，即可作为风格窃取的有效证据。

StyleSign实现了从被动防御到主动适应的策略转变，使得基于受保护ISP输出的图像训练的代理模型输出中均携带可提取的水印，从而为相机成像

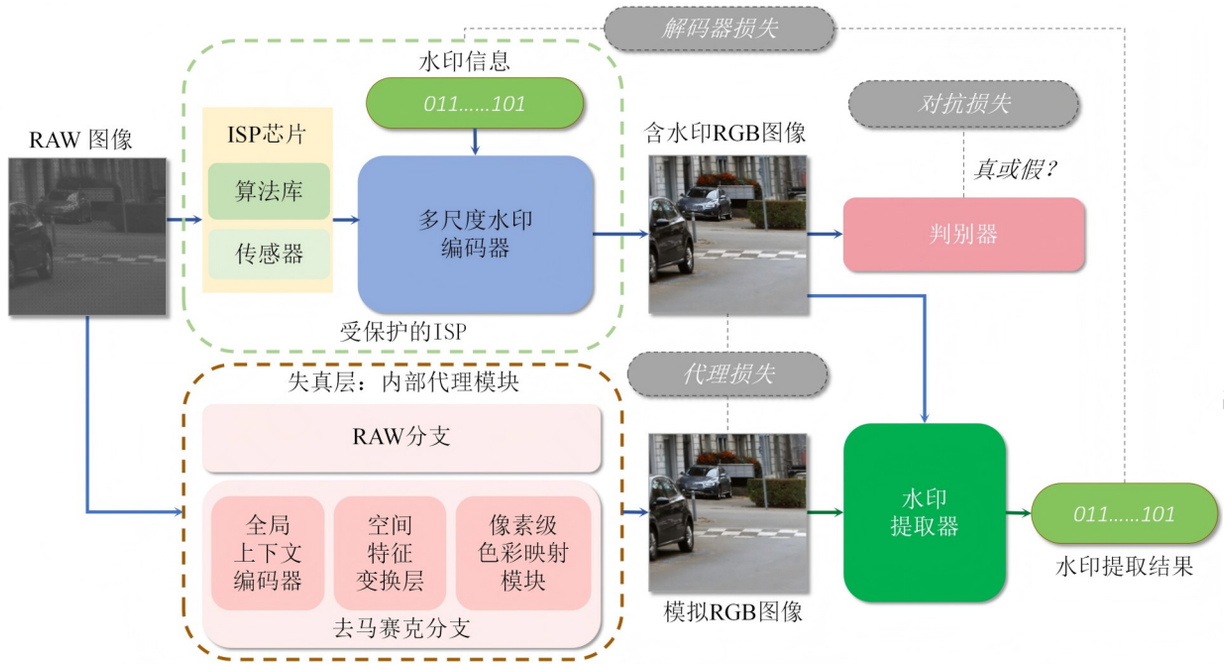

图2　StyleSign整体框架

Fig. 2　The overall framework of StyleSign

风格的版权保护提供了一种可行的解决方案。

2.2　多尺度水印编码器

为了提高水印提抗代理模型攻击的鲁棒性，设计了一种基于SEDWT模块的多尺度水印编码器，将二进制水印信息$\boldsymbol{M}\in\{0,1\}^k$嵌入相机ISP处理后的RGB图像$\boldsymbol{I}\in\mathbf{R}^{C\times H\times W}$中，生成视觉不可感知的水印图像$\boldsymbol{I}_{\mathrm{w}}$，如图3(a)所示，其整体流程可表述为：

$$\boldsymbol{I}_{\mathrm{w}}=E(\boldsymbol{I},\boldsymbol{M}) \tag{2}$$

首先，图像$\boldsymbol{I}$经过一个卷积层提取特征图$\boldsymbol{F}_{\mathrm{img}}$；同时，水印$\boldsymbol{M}$通过全连接层和卷积层被映射为与$\boldsymbol{F}_{\mathrm{img}}$具有相同空间尺寸和通道数的水印特征图$\boldsymbol{F}_{\mathrm{w}}$。随后，将$\boldsymbol{F}_{\mathrm{img}}$与$\boldsymbol{F}_{\mathrm{w}}$沿通道维度拼接，得到融合特征图$\boldsymbol{F}_{\mathrm{fuse}}$，作为SEDWT模块的输入。

SEDWT模块是编码器的核心单元，其内部结构如图3(b)所示。对于输入特征图$\boldsymbol{F}_{\mathrm{fuse}}$，经过一个3×3卷积层后，进行离散小波变换(discrete wavelet transform，DWT)，将其分解为四个子带：低频近似分量$\boldsymbol{F}_{\mathrm{LL}}$以及水平、垂直、对角方向的高频细节分量$\boldsymbol{F}_{\mathrm{LH}}$、$\boldsymbol{F}_{\mathrm{HL}}$、$\boldsymbol{F}_{\mathrm{HH}}$：

$$\{\boldsymbol{F}_{\mathrm{LL}},\boldsymbol{F}_{\mathrm{LH}},\boldsymbol{F}_{\mathrm{HL}},\boldsymbol{F}_{\mathrm{HH}}\}=DWT(\boldsymbol{F}_{\mathrm{fuse}}) \tag{3}$$

为了使水印嵌入与风格有关的特征中，对每个子带分别引入通道注意力机制。以子带$\boldsymbol{F}_{\mathrm{b}}$，$\mathrm{b}\in\{\mathrm{LL},\mathrm{LH},\mathrm{HL},\mathrm{HH}\}$为例，首先通过全局平均池化将通道的空间信息压缩，再经过两个全连接层与Sigmoid激活函数映射为通道权重$\boldsymbol{a}_{\mathrm{b}}$，随后将权重$\boldsymbol{a}_{\mathrm{b}}$与$\boldsymbol{F}_{\mathrm{b}}$逐通道相乘，得到输出$\hat{\boldsymbol{F}}_{\mathrm{b}}$：

$$\hat{\boldsymbol{F}}_{\mathrm{b}}=\boldsymbol{F}_{\mathrm{b}}\cdot\boldsymbol{a}_{\mathrm{b}} \tag{4}$$

该注意力机制使得网络能够更关注与成像风格相关的特征，从而将水印信息嵌入到与风格相关的特征中，而非依赖图像的语义内容。

为提高水印图像的质量，SEDWT模块采用残差连接。将$\boldsymbol{F}_{\mathrm{fuse}}$经过一个步长为2的3×3卷积层下采样至$H/2\times W/2$后，与子带$\{\hat{\boldsymbol{F}}_{\mathrm{LL}},\hat{\boldsymbol{F}}_{\mathrm{LH}},\hat{\boldsymbol{F}}_{\mathrm{HL}},\hat{\boldsymbol{F}}_{\mathrm{HH}}\}$按通道维度进行拼接，作为SEDWT模块的输出$\hat{\boldsymbol{F}}$。

整个编码器包含3个SEDWT模块，实现水印在多尺度频域上的分层嵌入。最后一个SEDWT模块的输出经过小波上采样重建为水印图像$\boldsymbol{I}_{\mathrm{w}}$。

为保障水印的不可感知性，采用均方误差损失约束像素级差异，即图像保真度损失为：

$$L_{MSE}=\left\|\boldsymbol{I}_{\mathrm{w}}-\boldsymbol{I}\right\|_2^2 \tag{5}$$

同时引入对抗性训练机制，通过判别器D促使$\boldsymbol{I}_{\mathrm{w}}$与$\boldsymbol{I}$在分布上尽可能一致，对抗损失定义为：

$$L_{adv}=E[\log D(\boldsymbol{I})]+E[\log(1-D(\boldsymbol{I}_{\mathrm{w}})] \tag{6}$$

编码器的总损失函数由二者加权构成：

$$L_{\mathrm{encoder}}=\lambda_1 L_{\mathrm{MSE}}+\lambda_2 L_{\mathrm{adv}} \tag{7}$$

式中λ_1和λ_2为超参数，分别设置为0.4，0.6，用于平衡视觉质量与对抗约束。通过上述设计，水印信息

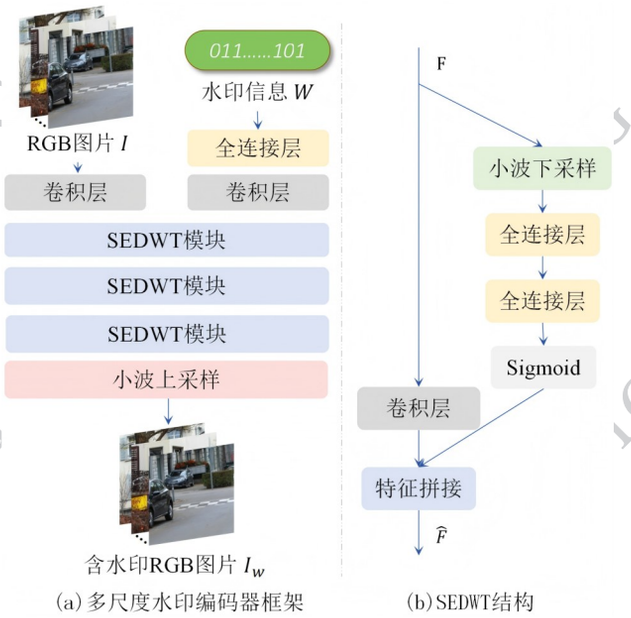

图3　多尺度水印编码器结构

Fig. 3　Structure of multiscale watermark encoder

在频域和通道维度进行嵌入，提高了水印抵抗代理模型攻击的鲁棒性。

2.3　内部代理模块

StyleSign设计了一个端到端的内部代理模块A，如图4所示。该模块本质上作为可学习的失真层，用于模拟攻击者利用水印图像训练代理模型以窃取成像风格的行为，其核心任务学习$\boldsymbol{I}_{\mathrm{w}}$的风格特征和水印分布。该模块以RAW图像$R$作为输入，输出模拟RGB图像$\boldsymbol{I}_{\mathrm{A}}$，采用含水印RGB图像$\boldsymbol{I}_{\mathrm{w}}$作为学习目标，通过学习从RAW到含水印RGB图像的映射，实现高保真的成像风格窃取，同时保持水印信息不被破坏。为兼顾风格模拟的准确性与水印信息的可提取性，StyleSign设计了两个分支：去马

赛克分支与RAW分支，将两者输出RGB图像逐像素取平均，得到内部代理模块输出的模拟RGB图像$\boldsymbol{I}_{\mathrm{A}}$。

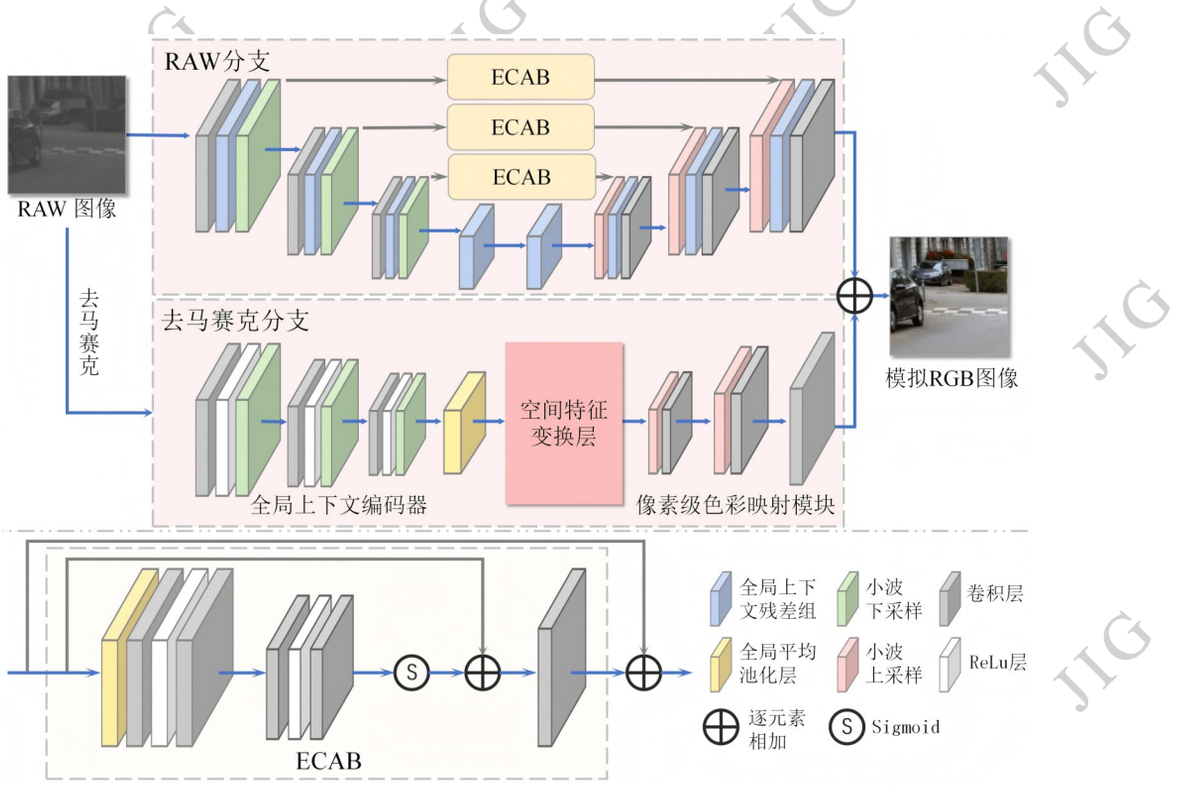

图4　内部代理模块结构

Fig. 4　Structure of internal surrogate module

2.3.1　去马赛克分支

去马赛克分支以RAW图像$\boldsymbol{R}$去马赛克后的图片$\boldsymbol{R}_{\mathrm{D}}$作为输入。虽然$\boldsymbol{R}_{\mathrm{D}}$已具备基本的彩色，但其在色彩还原和纹理细节上仍与目标相机成像风格存在显著差距。为此，设计了全局引导色彩映射网络(global guided color mapping network，GGCM)，通过提取全局风格特征并对局部像素进行精细化调整，实现对目标成像风格的高保真模拟。具体而言，GGCM网络由三个核心部分构成：

全局上下文编码器：该模块旨在提取输入图像

$\boldsymbol{R}_{\mathrm{D}}$的全局色彩和风格特征，由若干步长为2的卷积层和全局平均池化层组成，逐步将特征的空间维度压缩，最终输出一个低维的全局风格向量$\boldsymbol{s}$。该向量包含图像的整体色调、对比度分布和色彩倾向等与风格相关的信息，提供全局引导；

空间特征变换层（spatial feature transform layer，**SFT**）：SFT层（Wang等，2018）将全局风格向量$\boldsymbol{s}$映射为用于调整中间特征的仿射参数。$\boldsymbol{s}$分别输入到两个全连接网络f_{scale}和f_{shift}中，生成缩放因子$\gamma \in \mathbf{R}^{C}$和偏移量$\beta \in \mathbf{R}^{C}$，$C$为特征通道数：

$$\gamma = f_{\mathrm{scale}}(\boldsymbol{s}), \beta = f_{\mathrm{shift}}(\boldsymbol{s}) \tag{8}$$

通过这种方式，全局风格信息被转化为调制信号；

像素级色彩映射模块：该模块利用仿射参数γ和β，对$\boldsymbol{R}_{\mathrm{D}}$经过卷积层提取的中间特征图$\boldsymbol{F}$进行逐通道、逐像素的仿射变换：

$$\boldsymbol{F}_{\mathrm{out}} = \gamma \cdot \boldsymbol{F} + \beta \tag{9}$$

该模块在全局风格向量s的指导下，实现了对局部像素值的精细调整，能够在保持图像空间结构的同时，有效学习全局图像风格。变换后的特征图$\boldsymbol{F}_{\mathrm{out}}$最终通过卷积层重建为具有目标成像风格的RGB图像$\boldsymbol{I}_{\mathrm{D}}$。

GGCM网络以端到端的方式进行训练，学习从$\boldsymbol{R}_{\mathrm{D}}$到$\boldsymbol{I}_{\mathrm{D}}$的映射，在实现较为精准的色彩还原的同时保留图像的纹理细节，避免局部颜色失真，提升了风格模仿的视觉一致性。

2.3.2　RAW分支

RAW分支以RAW图像$\boldsymbol{R}$作为输入，通过端到端学习直接模拟从传感器RAW图像到RGB图像的映射过程。该分支采用U-Net架构作为基础框架，并以离散小波变换替代传统的池化与上采样操作，从而在特征下采样与重建过程中更好地保留高频细节，降低分辨率变化对水印特征的影响。

编码阶段：由4个下采样层构成。RAW图像$\boldsymbol{R}$经过卷积层提取浅层特征$\boldsymbol{F}_{\mathrm{enc}}$，随后采用小波下采样分解为低频子带$\boldsymbol{F}_{\mathrm{l}}$和高频子带$\boldsymbol{F}_{\mathrm{h}}$。低频子带主要包含图像结构信息，作为下一个下采样层的输入；高频子带则通过跳跃连接传至对应的解码层，为后续图像重建补充高频细节。为进一步保留水印信息，本文在跳跃连接处引入高效通道注意力模块（efficient channel attention block，ECAB）（Wang等，2020）。ECAB以$\boldsymbol{F}_{\mathrm{h}}$为输入，首先通过全局平均池化得到通道统计向量$\boldsymbol{z}$；随后对$\boldsymbol{z}$应用卷积核大小为3的一维卷积，得到局部跨通道交互后的特征向量$\boldsymbol{w}$；最后经过Sigmoid函数归一化，生成注意力权重a，并对原始特征图$\boldsymbol{F}_{\mathrm{h}}$进行逐通道重校准得到$\bar{\boldsymbol{F}}_{\mathrm{h}}$：

$$\bar{\boldsymbol{F}}_{\mathrm{h}} = a \cdot \boldsymbol{F}_{\mathrm{h}} \tag{10}$$

ECAB的输出$\bar{\boldsymbol{F}}_{\mathrm{h}}$通过跳跃连接输入解码阶段对应的上采样层中。该操作使网络能够自适应地增强与水印相关的特征。

解码阶段：由4个上采样层构成，每层通过小波上采样恢复图像分辨率，以上一层的低频特征和对应编码层的高频特征$\bar{\boldsymbol{F}}_{\mathrm{h}}$作为输入，将其按通道拼接，然后通过逆小波变换进行上采样。经过4次逆小波变换，最终得到与输入RAW图像同分辨率的RGB图像$\boldsymbol{I}_{\mathrm{R}}$。

通过离散小波变换与高效通道注意力模块的协同作用，RAW分支提升了水印特征在RAW图像到RGB图像映射过程中的保持能力。

2.3.3　loss函数

内部代理模块通过最小化其输出$\boldsymbol{I}_{\mathrm{A}} = A(R, I_{\mathrm{w}})$与水印图像$\boldsymbol{I}_{\mathrm{w}}$之间的重构误差进行训练，代理重建损失函数定义为：

$$L_{\mathrm{surrogate}} = \left\| \boldsymbol{I}_{\mathrm{A}} - \boldsymbol{I}_{\mathrm{w}} \right\|_{2}^{2} \tag{11}$$

该优化过程引导内部代理模块同时学习ISP的成像风格与嵌入的水印信息，即使水印在非线性映射过程中经历变换，其关键特征仍得以在生成的风格化图像中保持可提取性。通过端到端联合优化，水印特征得以与成像风格表征协同建模，使解码器能够从模拟攻击后的图像中更稳定地提取水印。

2.4　水印解码器

StyleSign中，水印解码器的设计目标是从含水印RGB图像$\boldsymbol{I}_{\mathrm{w}}$和模拟RGB图像$\boldsymbol{I}_{\mathrm{A}}$中，同时稳定提取水印信息。该解码器由三个卷积块、全局平均池化层和全连接层构成，最终通过Sigmoid函数输出与原始水印长度一致的预测比特序列$\boldsymbol{M}_{\mathrm{w}}$和$\boldsymbol{M}_{\mathrm{A}}$。

作为端到端训练系统的一部分，解码器与编码器及内部代理攻击模块同步优化，同时接收来自编码器的水印嵌入图像与代理攻击模块生成的模拟图像作为输入，通过最小化如下水印提取损失函数实现对分布偏移的自适应：

$$L_{\mathrm{decoder}} = L_{\mathrm{BCE}}(\boldsymbol{M}_{\mathrm{w}}, \boldsymbol{M}) + L_{\mathrm{BCE}}(\boldsymbol{M}_{\mathrm{A}}, \boldsymbol{M}) \tag{12}$$

式中L_{BCE}为二进制交叉熵损失。该联合优化使解码器能够同时学习从原始水印图像和攻击后图像中提取水印，从而建模攻击引入的分布变化。训练完成后，解码器能够从使用受保护数据训练的代理模型S输出的图像$\boldsymbol{I}_S = S(\boldsymbol{R}, \boldsymbol{I}_w)$中可靠地提取水印：

$$\boldsymbol{M} = D(\boldsymbol{I}_S) \tag{12}$$

提取得到的水印序列$\boldsymbol{M}'$与原始水印$\boldsymbol{M}$之间的BER为成像风格窃取行为提供了可验证的证据。

3 实　验

3.1 实验设置

3.1.1 数据集

在实验中，主要采用Zurich RAW to RGB数据集（ZRR）（Ignatov 等，2020）。ZRR数据集包含48403组配对的RAW-RGB图像，RAW图像的尺寸为448 × 448 × 1，由华为P20智能手机拍摄；对应的RGB图像尺寸为448 × 448 × 3，由佳能5D Mark IV单反相机拍摄，所有图像对均在同一视角下对同一场景采集，使得RAW图像与RGB图像一一对应。该数据集被划分为46839组训练图像对和1204组测试图像对。

选择ZRR数据集主要基于以下考量：首先，ZRR数据集是目前深度ISP领域广泛采用的公开数据集，多数深度ISP网络基于该数据集进行训练和评估，使用ZRR便于与现有工作进行对比。其次，数据集采用的拍摄设备具有一定的代表性：华为P20的RAW数据代表了常见的移动CMOS传感器输出，佳能5D Mark IV的RGB图像则体现了复杂的专业成像风格，若能在此数据集上成功实现成像风格保护，可在一定程度上反映该方法对其它类似映射的泛化性。同时，ZRR数据集包含大量不同场景的图像，图像内容较为丰富，能够验证水印方法对图像内容变化的鲁棒性。综上，采用ZRR数据集能够为本研究的实验验证提供有效支撑。

3.1.2 对比方法

将StyleSign与五种水印方法进行比较：HiDDeN（Zhu 等，2018）、DWSF（Guo 等，2023）、RAIMARK（Wang 等，2024）、Stegastamp（Tancik 等，2020）和WRAP（Liu 等，2023）。选取这些水印方案基于以下考量：它们均为该领域代表性方法，且在技术路线上具有差异性，能够覆盖不同的水印攻击场景与挑战，便于全面、客观地评估StyleSign在处理代理模型攻击方面的性能。

3.1.3 实施细节

本文使用Python3.8和Pytorch1.10.1构建StyleSign模型。为确保公平性，所有实验均在RTX3090 GPU上完成，所有方法均嵌入30位随机比特作为水印信息，并采用相同的训练测试划分。模型采用Adam优化器，初始学习率设为10^{-5}，批量大小（batch size）设置为4，共训练5个轮次。

3.2 水印图像的图像质量评估

为评估水印嵌入对图像视觉质量的影响，对原始RGB图像与对应的水印图像进行可视化对比分析。图5展示了原始图像、水印图像以及二者的残差图。为更清晰地观察细微差异，残差图进行了十倍强度放大。

从可视化结果可以看出，HiDDeN嵌入水印的图像存在偏色，残差图有模糊的边缘伪影；DWSF的水印图像和残差图中均存在较为明显的块状；RAIMARK的残差图中，纹理复杂的部分存在线状和点状伪影；Stegastamp的水印图像中部存在偏黄的伪影，且残差图中存在较为明显的物体轮廓；WRAP的水印图像在路面、墙壁等平摊区域存在伪影，且残差图中明显存在分布均匀的纹理痕迹；而StyleSign方法的水印图像中没有明显的失真或伪影，图像的边缘结构、纹理细节及整体色彩分布均与原始图像保持一致。即使在残差放大十倍的条件下，视觉感知并不明显，表明所提出方法能够在保证水印嵌入的同时维持较高的视觉质量。

本文进一步使用PSNR、SSIM和LPIPS对水印嵌入前后的图像质量进行评估，结果汇总于表1。在PSNR方面，StyleSign水印图像的PSNR达到37.26dB，略优于HiDDeN（34.87dB）、DWSF（37.03dB）、RAIMARK（33.71dB）和WRAP（36.65dB），明显优于Stegastamp（28.68dB），说明StyleSign对原始图像的像素值扰动最小。在SSIM方面，HiDDeN、DWSF、RAIMARK和WRAP的SSIM分别为0.9671、0.9743、0.9123、0.9587，而Stegastamp的SSIM仅为0.8939，这是由于水印嵌入产生了明显伪影，降低结构相似性；StyleSign的SSIM达到0.9893，高于其他对比方法。在LPIPS方面，HiDDeN和Stegastamp高达0.0819和0.0735，在人类视觉下能明显感知图像质量的变化；DWSF、

图5　不同水印方案的水印图像视觉质量

Fig. 5　Visual quality of the watermarked images using different watermarking schemes

RAIMARK 和 WRAP 的 LPIPS 分别为 0.0667、0.0610、0.0694，依然高于本文方法；StyleSign 取得了最低值0.0425，说明在人类视觉感知层面引入的失真最小。

综上所述，StyleSign 设计多尺度水印编码器，通过注意力机制使得水印嵌入在与图像风格相关的特征中，对图像质量造成的影响最小，水印图像能够保持较高的视觉质量。

3.3　鲁棒性评估

3.3.1　抵抗常规攻击

为全面评估StyleSign对常见图像处理操作的鲁棒性，设计了五类典型的常规攻击测试，包括：色彩抖动（color jitter，CJ）、高斯噪声（gaussian noise，GN）、高斯模糊（gaussian blur，GB）、JPEG 压缩（JPEG）以及缩放操作（Resize）。这些攻击模拟了图像在传输、存储或后处理过程中可能遭遇的真实失真，是评估水印系统鲁棒性的基础测试。

表1　不同水印方案的水印图像视觉质量评估

Table 1　Visual evaluation of watermarked images using different watermarking schemes

水印方案	PSNR	SSIM	LPIPS
HiDDeN	34.87	0.9671	0.0819
DWSF	37.03	0.9743	0.0667
RAIMARK	33.71	0.9123	0.0610
WRAP	36.65	0.9587	0.0694
Stegastamp	28.68	0.8939	0.0735
StyleSign	**37.26**	**0.9893**	**0.0425**

注：加粗字体为图像质量最优值。

攻击的具体参数设置如下：1）色彩抖动：亮度、

对比度、饱和度和色调的扰动范围为原始值的±20；2）高斯噪声：标准差为0.02；3）高斯模糊：高斯核大小11 × 11，标准差σ = 1.5；4）JPEG压缩：质量因子为50；5）缩放攻击：下采样至原尺寸的75%，再上采样回原始分辨率。实验中，将StyleSign与HiDDeN、WRAP和DWSF三种代表性水印方法进行对比，所有方法的编解码器均使用相同设置进行训练，嵌入长度为30比特的水印信息。

测试流程如下：首先使用各水印方案的编码器对ZRR数据集中的RGB图像嵌入水印，随后对水印图像依次施加上述五类常规攻击，最后使用各方案的解码器提取水印并计算BER，实验结果如表2所示。

表2　不同水印方案在常规攻击下的BER

Table 2　BER of different watermarking schemes under conventional attacks

水印方案	CJ	GN	GB	JPEG	Resize
HiDDeN	1.63	1.53	1.35	1.48	1.47
DWSF	1.37	1.22	1.02	1.20	1.14
RAIMARK	1.39	0.95	0.93	0.95	0.97
WRAP	1.27	1.14	1.00	1.12	1.10
Stegastamp	1.06	0.96	0.85	0.90	0.91
StyleSign	**0.83**	**0.64**	**0.63**	**0.64**	**0.63**

注：加粗字体为BER(%)最优值。

从表中可看出，在所有五类常规攻击下，StyleSign均表现出最低的BER，优于HiDDeN、DWSF、RAIMARK、Stegastamp和WRAP。具体而言，StyleSign在色彩抖动攻击下的BER为0.83%，在高斯噪声和高斯模糊攻击中，StyleSign的BER分别低至0.64%和0.63%，优于其他对比方案，说明其对常见的图像增强或平滑处理具有鲁棒性；在JPEG压缩和缩放攻击下，StyleSign的BER仍稳定在0.64%和0.63%，优于其他方案，表明其通过多尺度小波嵌入与通道注意力机制，能够有效抵抗压缩与几何形变带来的高频信息损失。StyleSign设计内部代理模块作为失真层，由此引入更复杂的失真模式，提高了水印的泛化性和鲁棒性。

综上所述，尽管StyleSign是针对代理模型攻击设计的专用水印框架，但在传统常规攻击测试中同样展现出较好的稳定性和鲁棒性，进一步验证了其作为端到端成像风格保护水印方案的实用价值与泛化能力。

3.3.2　抵抗代理模型攻击

为了评估水印在代理模型攻击下的鲁棒性，选取了四种技术路线不同且较为有代表性的代理ISP模型作为攻击者：AWNet（Dai等，2020）、RAW-to-sRGB（RTS）（Zhang等，2021）、MW-ISPNet（MWNet）（Ignatov等，2020）和Airia CG（Ignatov等，2020）。这些模型均属于深度ISP网络，常被用于学习并模拟特定相机的成像风格，其网络结构与优化目标存在差异，能够模拟代理模型攻击的多种可能性，从而较为全面地验证StyleSign的泛化能力与鲁棒性。

在攻击设置上，对四个代理模型采用统一的训练策略，以模拟现实场景中攻击者的行为。为了确保对比实验的公平性，对于所有待评估的水印方法均使用相同的数据集、相同的流程来训练四种代理ISP模型。具体而言，对于每一种水印方法，首先将其生成的水印图像与对应的RAW图像配对，构成训练数据集；随后，分别训练四个代理ISP模型，使其学习从RAW到含水印RGB图像的映射。整个训练过程中，所有代理ISP模型均使用一致的损失函数、优化器及训练迭代次数。训练完成后，采用两种设置提取水印：1）直接使用各方法的解码器从代理模型生成的RGB图像中提取水印，并计算BER；2）将基线水印方法的编码器与解码器替换StyleSign中的对应模块，保持内部代理模块与训练流程不变，进行联合优化，结果如表3中带星号（*）的条目所示。

实验结果验证了StyleSign在抵御代理模型攻击方面的优势。在设置1）下，DWSF、RAIMARK和WRAP无法有效抵抗代理模型攻击，其BER接近随机猜测水平（约50%），表明嵌入的水印在攻击过程中已基本失效；HiDDeN和Stegastamp虽然在RTS、AWNet和MWNet上表现出一定的提取能力（BER为10.46%-36.19%），但在结构更为复杂的Airia CG攻击下BER分别为44.15%和49.99%，说明其泛化能力有限。在设置2）下，对比方法的BER有所改善，HiDDeN*的BER降至1.74%-6.99%，DWSF*的BER降至2.33%-6.53%，RAIMARk*的BER降至1.82%-3.71%，WRAP*的BER降至1.38%-1.75%，Stegastamp*的BER降至3.35%-10.54%，均低于其设置1）下的水平，但仍高于StyleSign。StyleSign在全部四种代理ISP模型下均表现出优异的鲁棒性，其

BER 分别低至 1.07%（RTS）、1.19%（AWNet）、0.99%（MWNet）及 0.49%（Airia CG）。这主要得益于以下三方面机制：

表 3　不同水印方案在代理模型攻击下的 BER

Table 3　BER of different watermarking schemes under surrogate attacks

水印方案	代理ISP模型			
	RTS	AWNet	MWNet	Airia CG
HiDDeN	24.01	20.08	24.07	44.15
HiDDeN*	1.74	6.99	1.98	4.76
DWSF	44.61	49.99	50.10	50.28
DWSF*	6.53	2.33	4.37	2.87
RAIMARK	45.23	44.31	44.85	46.64
RAIMARK*	3.71	2.66	3.11	1.82
WRAP	50.02	50.30	49.84	50.05
WRAP*	1.63	1.48	1.75	1.38
Stegastamp	10.46	12.09	36.19	49.99
Stegastamp*	9.22	10.54	3.35	6.48
StyleSign	**1.07**	**1.19**	**0.99**	**0.49**

注：加粗字体为BER（%）最优值。

内部代理模块设计：StyleSign 在训练阶段设计内部代理模块，模拟代理模型攻击的过程，使水印信号能够抵抗代理模型攻击的非线性映射，从而在攻击后仍保持可提取性；

联合优化的水印嵌入与风格学习：通过水印编码器与内部代理模块在同一训练框架内联合优化，促使水印信息能够抵抗代理模型攻击；

面对未知攻击的泛化能力：内部代理模块采用双分支结构，更充分地模拟相机 ISP 的映射关系，使得训练过程覆盖了更广泛的风格变换空间，有助于提升对未知攻击模型的适应性。

综上所述，StyleSign 通过其结构设计应对了代理攻击带来的分布偏移问题，为 ISP 成像风格的保护提供了一种兼具鲁棒性与泛化性的实用解决方案。

3.4　消融实验

本节在包含 1204 组 RAW-RGB 图像对的 ZRR 测试集上进行消融实验，以验证内部代理模块与判别器对 StyleSign 性能的贡献。

为验证内部代理模块的作用，分别比较了包含与不包含该模块的 StyleSign 在四种代理模型攻击下的 BER。表 4 显示，当移除内部代理模块时，BER 从原有的约 1% 上升至约 46%，接近随机猜测水平，说明水印在代理模型攻击过程中几乎完全失效。该现象表明：在不使用内部代理模块作为失真层的情况下训练，编码器与解码器无法适应代理模型攻击引入的非线性映射，导致水印无法提取，说明了采用内部代理模块的必要性。

表 4　内部代理模块的消融实验

Table 4　Ablation studies on the internal surrogate module

测试模型	RTS	AWNet	MWNet	Airia CG
StyleSign	**1.07**	**1.19**	**0.99**	**0.49**
w/o 内部代理模块	45.89	46.31	46.38	45.66

注：加粗字体为BER（%）最优值。

为了验证判别器在保持图像视觉质量方面的作用，对判别器进行了消融实验，采用 PSNR、SSIM 和 LPIPS 作为评估指标。表 5 显示：移除判别器后，

表 5　判别器的消融实验

Table 5　Ablation studies on the discriminator

水印方案	PSNR	SSIM	LPIPS
StyleSign	**37.26**	**0.9893**	**0.0425**
w/o 判别器	20.32	0.8588	0.2280

注：加粗字体为图像质量最优值。

图 6　判别器消融实验中水印图像视觉质量对比

Fig. 6　Visual quality of the watermarked images in ablation studies on the discriminator

水印图像的 PSNR 从 37.26dB 下降至 20.32dB，SSIM 从 0.9893 下降至 0.8588，同时 LPIPS 显著上升至 0.2280；图 6 显示，移除判别器后，水印图像上存在分布广泛的点状伪影。该结果表明，缺少判别器

会导致嵌入过程产生明显的可见失真，使水印图像在结构与视觉质量上受到影响。因此说明了判别器的有效性。

综上所述，消融实验验证了内部代理模块与判别器对StyleSign性能的关键作用：前者提升了对代理模型攻击的鲁棒性，后者有效提高了水印嵌入后图像的视觉质量，从而共同提高了StyleSign在真实攻击场景中的整体性能。

4 结 论

StyleSign是一种专为保护相机ISP成像风格免受代理模型攻击的水印方案，通过将代理模型攻击模拟为可学习的失真层，在多尺度水印编码器、内部代理模块与解码器的端到端联合优化下，实现了对代理模型攻击的有效防御。实验结果表明，该方法在ZRR数据集上表现良好，水印图像的PSNR为37.26dB，SSIM为0.9893，LPIPS为0.0425；对未知代理模型攻击具有更强的鲁棒性，BER在多种代理模型攻击下均低于1.2%，优于HiDDeN、DWSF、RAIMARK、Stegastamp和WRAP等现有方法。

本研究的主要贡献在于提出并验证了面向相机成像风格保护的水印方案。通过引入内部代理模块作为噪声层，模拟代理模型攻击，水印在训练阶段即能够适应代理模型的非线性映射，从而对代理模型攻击具有鲁棒性。多尺度水印编码器结合离散小波变换与通道注意力机制，有助于在不同频率上实现更稳定的水印嵌入，保持更好的视觉一致性。解码器通过联合优化，能够从攻击者所采用的代理ISP模型输出的图像中准确提取水印。

尽管StyleSign在实验中表现出色，但仍存在一定局限性：实验主要基于公开数据集与深度ISP模型，与真实攻击者行为存在差异可能影响实际部署效果。此外，RAW-RGB数据的获取仍面临挑战，未来需进一步与相机厂商合作，构建更贴近实际的攻击模拟与验证环境。

综上所述，StyleSign为相机成像风格保护提供了一种有效、鲁棒且可泛化的技术方案，不仅提升了水印抵抗代理模型攻击的鲁棒性，也为相机制造商的核心知识产权保护提供了新的思路与方法。未来工作将侧重于真实场景下的性能验证与动态攻击防御机制的进一步优化。

参考文献(References)

Bose T and Meyer F. 2003. Digital Signal and Image Processing. New York: John Wiley & Sons, Inc.

Dai L, Liu X, Li C and Chen J. 2020. AWNet: Attentive wavelet network for image ISP//European Conference on Computer Vision. Cham: Springer International Publishing: 185-201 [DOI: 10.1007/978-3-030-67070-2_11]

Dos Santos C F G, Arrais R R, Da Silva J V S, Da Silva M H M, De Araujo Neto W B G, Lopes L T, et al. 2025. ISP meets deep learning: A survey on deep learning methods for image signal processing. ACM Computing Surveys, 57(5): 127:1-127:44 [DOI: 10.1145/3708516]

Fang H, Jia Z, Ma Z, Chang E C and Zhang W. 2022. PIMoG: An effective screen-shooting noise-layer simulation for deep-learning-based watermarking network//Proceedings of the 30th ACM International Conference on Multimedia. Lisboa: ACM: 2267-2275 [DOI: 10.1145/3503161.3548049]

Guo H, Zhang Q, Luo J, Guo F, Zhang W, Su X, et al. 2023. Practical deep dispersed watermarking with synchronization and fusion//Proceedings of the 31st ACM International Conference on Multimedia. Ottawa: ACM: 7922-7932 [DOI: 10.1145/3581783.3612015]

Guo J, Gao X, Yan Y, Li G and Pu J. 2025. Dark-ISP: Enhancing RAW image processing for low-light object detection//Proceedings of the IEEE/CVF International Conference on Computer Vision. Hawaii: IEEE: 9583-9593

Guo N, Huang Y, Niu B N, Guan H, Lan F P and Zhang S W. 2025. Concurrent watermark embedding and feature point enhancement with perceptual constraint and guidance. Journal of Image and Graphics, 30(4):1072-1083 (郭娜, 黄樱, 牛保宁, 关虎, 兰方鹏, 张树武. 2025. 感知约束和引导下的特征点增强局部水印算法. 中国图象图形学报, 30(4):1072-1083) [DOI: 10.11834/jig.240348]

Hess A. 2013. An Introduction to the Nikon Creative Lighting System. Peachpit Press

Hu J, Shen L and Sun G. 2018. Squeeze-and-excitation networks//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE: 7132-7141 [DOI: 10.1109/CVPR.2018.00745]

Ignatov A, Timofte R, Zhang Z, Liu M, Wang H, Zuo W, et al. 2020. AIM 2020 challenge on learned image signal processing pipeline//European Conference on Computer Vision. Cham: Springer International Publishing: 152-170 [DOI: 10.1007/978-3-030-67070-2_9]

Ignatov A, Van Gool L and Timofte R. 2020. Replacing mobile camera ISP with a single deep learning model//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Seattle: IEEE: 2275-2285 [DOI: 10.1109/

CVPRW50498.2020.00276]

Jia Z, Fang H and Zhang W. 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. Chengdu: ACM: 41-49 [DOI: 10.1145/3474085.3475324]

Kim B H, Song J, Ye J C and Baek J. 2020. PyNet-CA: Enhanced PyNet with channel attention for end-to-end mobile image signal processing//European Conference on Computer Vision. Cham: Springer International Publishing: 202-212 [DOI: 10.1007/978-3-030-67070-2_12]

Li L, Yao M, Meng X, Yu M, Xue T and Gu J. 2025. Uni-ISP: Towards unifying the learning of ISPs from multiple mobile cameras. IEEE Transactions on Image Processing, 34: 6126-6137 [DOI: 10.1109/TIP.2025.3607617]

Liang Z, Cai J, Cao Z and Zhang L. 2021. CameraNet: A two-stage framework for effective camera ISP learning. IEEE Transactions on Image Processing, 30: 2248-2262 [DOI: 10.1109/TIP. 2021.3051486]

Liu G, Si Y, Qian Z, Zhang X, Li S and Peng W. 2023. WRAP: Watermarking approach robust against film-coating upon printed photographs//Proceedings of the 31st ACM International Conference on Multimedia. Ottawa: ACM: 7274-7282 [DOI: 10.1145/3581783.3612529]

Liu Z, Qin J, Xiang X, Luo Y and Tan Y. 2025. Post-encoding enhancement: A screen-shooting resistant watermarking scheme with feature enhancement and hybrid distortion simulation. Knowledge-Based Systems, 322: 113673 [DOI: 10.1016/j. knosys. 2025.113673]

Lu L, Guan N, Wang Y, Jia L, Luo Z and Yin J. 2021. Tenet: A framework for modeling tensor dataflow based on relation-centric notation//ACM/IEEE 48th Annual International Symposium on Computer Architecture. Valencia: IEEE: 720-733 [DOI: 10.1109/ISCA52012.2021.00062]

Lu Y, Qian Y, Rao Z, Xiao J, Chen L and Xiong H. 2025. RGB-event ISP: The dataset and benchmark//The Thirteenth International Conference on Learning Representations.

Ma Z, Fang H, Yang X, Chen K and Zhang W. 2025. RoPaSS: Robust watermarking for partial screen-shooting scenarios//Proceedings of the AAAI Conference on Artificial Intelligence, Philadelphia: AAAI: 19332-19339 [DOI: 10.1609/aaai.v39i18.34128]

Ramanath R, Snyder W E, Yoo Y and Drew M S. 2005. Color image processing pipeline. IEEE Signal Processing Magazine, 22(1): 34-43 [DOI:10.1109/MSP.2005.1407713]

Ronneberger O, Fischer P and Brox T. 2015. U-Net: Convolutional networks for biomedical image segmentation//International Conference on Medical Image Computing and Computer-Assisted Intervention. Cham: Springer International Publishing: 234-241 [DOI: 10.1007/978-3-319-24574-4_28]

Tancik M, Mildenhall B and Ng R. 2020. StegaStamp: Invisible hyperlinks in physical photographs//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE: 2117-2126 [DOI: 10.1109/CVPR42600.2020.00219]

Tang Y, Wang C, Xiang S 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]

Tang Y, Wang S, Han R and Wang C. 2025. A novel robust reversible watermarking scheme using fractional-order polar complex exponential transform. IEEE Transactions on Multimedia, 28: 241-255 [DOI: 10.1109/TMM.2025.3623522]

Timacheff S. 2011. Canon EOS Digital Photography Photo Workshop. 8th ed. Indianapolis: John Wiley & Sons

Wang Q, Wu B, Zhu P, Li P, Zuo W and Hu Q. 2020. ECA-Net: Efficient channel attention for deep convolutional neural networks//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE: 11531-11539 [DOI: 10.1109/CVPR42600.2020.01155]

Wang X, Yu K, Dong C and Loy C C. 2018. Recovering realistic texture in image super-resolution by deep spatial feature transform//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE: 606-615 [DOI: 10.1109/CVPR.2018.00070]

Wang Y, Zhu X, Ye G, Zhang S and Wei X. 2024. Achieving resolution-agnostic DNN-based image watermarking: A novel perspective of implicit neural representation//ACM International Conference on Multimedia. Melbourne: ACM: 10354-10362 [DOI: 10.1145/3664647.3681138]

Xue Y, Tan L, Li G, Qian Z, Li S and Zhang X. 2025. Physical marker: Revealing invisible hyperlinks hidden in printed trademarks//Proceedings of the AAAI Conference on Artificial Intelligence, Philadelphia: AAAI: 9068-9075 [DOI: 10.1609/aaai.v39i9.32981]

Yin B and Yin K. 2025. Robust image watermarking using bidirection-interactive and context-aware networks. IEEE Transactions on Circuits and Systems for Video Technology, s35(8): 7683-7696 [DOI: 10.1109/TCSVT.2025.3543969]

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]

Zhang Z, Wang H, Liu M, Wang R, Zhang J and Zuo W. 2021. Learning RAW-to-sRGB mappings with inaccurately aligned supervision//Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal: IEEE: 4328-4338 [DOI: 10.1109/ICCV48922.2021.00431]

Zhu J, Kaplan R, Johnson J and Fei-Fei L. 2018. HiDDeN: Hiding data with deep networks//Proceedings of the European Conference on Computer Vision. Munich: Springer: 657-672 [DOI: 10.1007/978-3-030-01267-0_40]

作者简介

司若童，女，硕士研究生，主要研究方向为多媒体安全。E-mail：rtsi23@m.fudan.edu.cn

钱振兴，通信作者，男，教授，主要研究方向为多媒体安全、人工智能应用与安全。E-mail：zxqian@fudan.edu.cn

汤毅超，男，博士后研究员，主要研究方向为多媒体信息安全。E-mail：yichao_tang@fudan.edu.cn

张新鹏，男，教授，主要研究方向为人工智能安全、多媒体信息安全。E-mail：zhangxinpeng@fudan.edu.cn

李晟，男，副教授，主要研究方向为多媒体安全、人工智能安全。E-mail：lisheng@fudan.edu.cn