XXXX

基于自适应注意力机制与SimSiam框架的图像水印研究

诸葛斌，陈莹莹，王冰雁，董黎刚，蒋献
（浙江工商大学信息与电子工程学院，浙江 杭州 310000）

摘　要：针对智慧教育中高精度图像的版权保护需求，本文提出一种融合自适应注意力机制与SimSiam自监督对比学习的鲁棒盲水印算法。该方法利用自适应注意力在高频纹理区域动态分配嵌入权重，在提升JPEG压缩、随机裁剪等攻击鲁棒性的同时保持较高图像质量。SimSiam模块利用对比学习强化特征一致性，有效抑制攻击带来的差异，突出水印的语义稳定性。实验结果表明，本文方法在JPEG压缩（质量因子50）与50%随机裁剪下均实现100%比特恢复率（BRR），PSNR较传统DCT方法提升15.4%。本研究为智慧教育场景下的教学图像提供了一种兼顾鲁棒性与视觉保真度的版权保护技术路径。

关键词：图像水印；智慧教育；自适应注意力；SimSiam；鲁棒特征嵌入
中图分类号：TP18
文献标志码：A
doi: 10.11959/j.issn.1000−0801.

Research on Image Watermarking Based on Adaptive Attention Mechanism and SimSiam Framework

Zhuge Bin, Chen Yingying, Wang Bingyan, Dong Ligang, Jiang Xian
School of Information and Electronic, Zhejiang Gongshang University, Hangzhou 310000, China

Abstract: To address the copyright protection requirements of high-precision images in smart education, a robust blind watermarking algorithm integrating an adaptive attention mechanism and the SimSiam self-supervised contrastive learning framework was proposed. The adaptive attention mechanism dynamically allocated embedding weights in high-frequency texture regions, thereby enhancing robustness against attacks such as JPEG compression and random cropping while maintaining high image quality. The SimSiam module strengthened feature consistency through contrastive learning, effectively suppressing attack-induced variations and improving the semantic stability of the watermark. Experimental results demonstrated that the proposed method achieved a 100% bit recovery rate (BRR) under JPEG compression (quality factor 50) and 50% random cropping, with a 15.4% improvement in PSNR compared with the traditional DCT-based method. This study provides a copyright protection solution for educational images that bal-

收稿日期：2025−11−28；**修回日期：**2026−02−22
通信作者：陈莹莹，lumine2001@163.com
基金项目：浙江省“尖兵”“领雁”研发攻关计划（No. 2023C03202）；浙江省大学生科技成果推广项目·新苗计划（No. 2025R408B075）
Foundation Items: Zhejiang Provincial “Pioneer” and “Leading Goose” R&D Program (No. 2023C03202), Zhejiang Provincial College Students’ Scientific and Technological Achievements Promotion Project–Xinmiao Talents Program (No. 2025R408B075)

ances robustness and visual fidelity in smart education scenarios.

Key words: Image Watermarking, Smart Education, Adaptive Attention, SimSiam, Robust Feature Embedding

1 引言

1.1 研究背景及意义

在全球教育信息化快速推进的背景下，智慧教育已成为推动教育数字化转型的重要方向。与此同时，教育数据的安全、版权保护与真实性问题日益突出，尤其在教师与学生的交互场景中，教学图像与数字内容的非法复制与篡改成为主要隐患。

根据中研普华研究院《2025－2030年中国教育行业全景调研与投资战略规划报告》，教育行业总规模预计2030年突破10万亿元，其中职业教育市场成为核心增长引擎。教育数字化的快速发展，使教学图像的版权安全成为智慧教育的重要保障方向。

针对智慧教育中高精度图像的保护需求，本文旨在提出一种具备强结构感知能力的通用鲁棒水印框架。不同领域的图像（如医学、工程类）具有不同的视觉特性，其共同难点在于高频细节的保护。因此本文首先在标准图像数据集上验证算法对复杂纹理处理的普适性能，为后续在特定教育场景下的垂直应用奠定理论与架构基础。

1.2 研究现状

数字水印技术是信息隐藏技术重要分支，其核心是将版权标识或认证信息嵌入原始载体数据，防止多媒体内容传播中被非法复制、篡改和滥用。 [1]。根据应用目标的不同，数字水印技术主要服务于版权保护、内容认证、所有权识别以及数据完整性校验等场景[2]。随着多媒体内容规模与复杂度提升，传统水印方法在鲁棒性、自适应性和泛化能力上渐显局限。

与依赖人工规则设计的传统算法不同，基于深度学习的水印方法可通过数据驱动自动学习嵌入与提取策略，建模能力和环境适应潜力更强。[3]。这类方法可根据图像内容特性和变化的攻击环境，动态调整嵌入过程，在隐蔽性与鲁棒性间取得更优平衡。 [4]。目前，深度学习水印研究主要集中于卷积神经网络（CNN）、生成对抗网络（GAN）、变分自编码器（VAE）等典型模型框架。

其中，卷积神经网络（CNN）凭借其层次化特征提取能力，在水印鲁棒性增强方面表现突出。相关研究表明，CNN能够有效利用迁移学习能力[5]、图像结构的鲁棒固有特性[6]以及多层滤波器组合[7]，在压缩、噪声等常见攻击下，能保持较高水印恢复精度。生成对抗网络（GAN）通过生成器与判别器的对抗博弈机制，在提升水印隐蔽性上优势显著。 [8]。GAN结构能够模拟复杂攻击场景，使嵌入水印在裁剪、压缩及噪声干扰条件下仍具备良好的可恢复性[9]。此外，变分自编码器（VAE）对图像潜在分布进行概率建模，实现水印信息的稳定嵌入与重建，一定程度上提升了系统的隐蔽性与鲁棒性。[10]。例如，薛等人[11]提出兼顾版权保护与内容认证的全盲双功能数字水印算法，验证深度模型在多目标水印任务中的可行性。总体而言，现有研究正逐步向提高水印嵌入容量、信息丰富性及功能多样化方向发展。 [12]。

相比之下，传统图像水印方法通常采用手工设计的嵌入与提取机制[4]，高度依赖图像处理先验知识与经验规则。此类方法针对特定攻击场景定制时效果佳，但对先验假设的依赖使其在复杂或未知攻击条件下缺乏适应性和泛化能力。[13]。约2017年起，研究者开始系统探索用深度学习特征表达能力重构传统水印框架，早期系列工作尝试用CNN直接完成水印嵌入与提取。

[14]，并通过端到端训练机制提升系统整体的鲁棒性与安全性。

在此基础上，Tavakoli 等人[15]提出了一种结合 CNN 与小波变换的水印方法，使嵌入与提取过程不再受宿主图像分辨率限制；Mahapatra 等人[16]设计了基于自动编码器的 CNN 水印模型，通过块级特征重构与反卷积结构在保证不可感知性的同时显著提升鲁棒性；Himanshu Kumar Singh 等人[17]进一步引入 GAN 与秘密触发机制，在保持图像质量的前提下实现了近乎不可察觉的水印嵌入，为图像版权保护和身份认证提供了新的思路。

近年来，自监督学习因无需人工标注、能从数据构造训练信号的优势，在鲁棒特征学习领域受广泛关注。研究表明，它可优化水印嵌入过程，增强模型对常见图像处理操作的适应能力，提升水印不可感知性与抗攻击能力。 Pierre Fernandez 等人[18]提出了一种基于自监督训练的水印嵌入网络，在潜在特征空间中实现了多比特水印的稳定嵌入。Cong 等人[19]进一步提出 SSL-Guard 模型，通过自监督预训练编码器，在水印注入与提取阶段对模型窃取、输入扰动及模型微调等攻击表现出较强鲁棒性。

尽管上述研究验证了深度学习与自监督机制在水印领域的有效性，但现有方法多侧重特征稳定性或嵌入容量提升，对水印嵌入位置内容感知建模及攻击前后特征语义一致性联合优化关注不足。在智慧教育等高精度图像应用场景中，这易使视觉质量与鲁棒性权衡受限，仍有研究空间。

2.1　整体架构设计

本系统由嵌入器（Embedder）、提取器（Extractor）和编码器（Encoder）构成，通过多阶段协同训练实现水印的隐蔽嵌入与稳定特征学习。整体水印架构如图 2-1 所示。

整体流程从输入图像与水印的融合开始。嵌入器采用动态感知的交叉注意力机制，将 32×32 的二维码水印（由 generate_qr_code 函数生成）嵌入到 512×512 的原始图像中。具体而言，嵌入器首先通过分块操作将图像划分为 16×16 的局部区域（代码中 patch_size=16），随后引入多头注意力机制（layers.MultiHeadAttention）动态计算不同区域对水印嵌入的敏感度权重（公式 2-1）

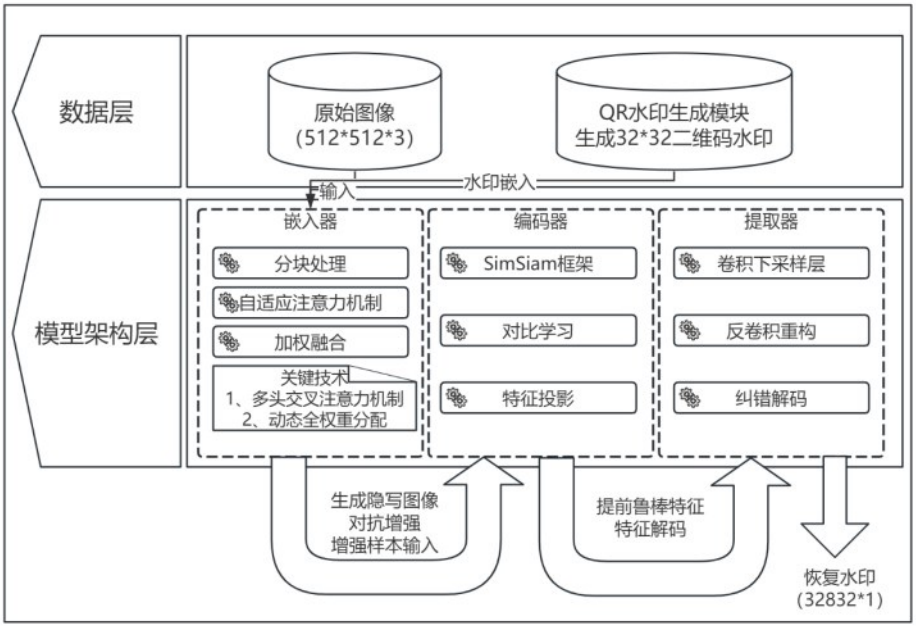


图 2-1　水印架构图

定义了交叉注意力输出：

$$\text{adaptive\_output}_{AB}=\text{adaptive\_weight}_{AB}\bullet\text{softmax}(\frac{\boldsymbol{Q}_A\boldsymbol{K}_B^T}{\sqrt{d_k}})\boldsymbol{V}_B \quad \text{（公式2-1）}$$

其中，$\boldsymbol{Q}_A$为查询向量，$\boldsymbol{K}_B$和$\boldsymbol{V}_B$分别为键向量与值向量，$\text{adaptive_weight}_{AB}$为可训练标量权重，由AdaptiveAttention类通过端到端训练生成。该网络通过Sigmoid函数（公式2-1）动态调整权重，确保水印主要嵌入至人眼不敏感的纹理区域，从而将峰值信噪比（PSNR）提升至38.72dB。

在上式（公式2-1）中自适应权重（$\text{adaptive\_weight}_{AB}$）的生成逻辑是由交叉注意力机制中的动态权重标量$\alpha_{adaptive}$通过可训练参数网络生成，其数学形式严格对应代码中AdaptiveAttention类的实现逻辑。给定查询向量$\boldsymbol{Q}\in\mathbf{R}^{N\times d_k}$与键向量$\boldsymbol{K}\in\mathbf{R}^{M\times d_k}$，权重生成过程定义为：

$$\alpha_{adaptive}=\text{Sigmoid}(\boldsymbol{W}_2\bullet\text{ReLU}(\boldsymbol{W}_1\bullet\boldsymbol{Concat}(\boldsymbol{Q},\boldsymbol{K})+b_1)+b_2) \quad \text{（公式2-2）}$$

其中$\boldsymbol{W}_1\in\mathbf{R}^{2d_k\times 256}$和$\boldsymbol{W}_2\in\mathbf{R}^{256\times 1}$为可训练参数矩阵，$b_1\in\mathbf{R}^{256}$与$b_2\in\mathbf{R}$为偏置项。该网络结构在代码中通过两层全连接层实现：首层将拼接后的查询键向量$\boldsymbol{Concat}(\boldsymbol{Q},\boldsymbol{K})$映射至256维空间并采用ReLU激活，次层输出单通道权重值并通过Sigmoid函数约束至[0, 1]区间。

为增强模型对抗攻击的能力，系统在训练阶段集成了多模态数据增强模块（random_augmenter函数）。该模块模拟了10类常见攻击（包括高斯噪声、JPEG压缩、随机裁剪等，与代码中iaa.Sequential配置一致），通过动态生成扰动样本迫使模型学习鲁棒特征。增强后的图像输入编码器进行特征提取，编码器基于SimSiam自监督框架，采用ViT编码器结构（包含4层Transformer块及其对应的投影头），通过对比学习优化潜在表示：对同一图像的不同增强视图（如噪声版与裁剪版）进行编码，最大化其投影特征间的余弦相似度（公式2-3）：

$$L_{simsiam}=-\frac{1}{2}[\frac{p_1\bullet\text{stopgrad}(z_2)}{\|p_1\|\|\text{stopgrad}(z_2)\|}+\frac{p_2\bullet\text{stopgrad}(z_1)}{\|p_2\|\|\text{stopgrad}(z_1)\|}] \quad \text{（公式2-3）}$$

其中，stopgrad表示梯度截断操作，迫使模型忽略攻击引入的无关变异，专注于水印相关的语义特征。编码器的输出为512维的紧凑特征向量，既包含水印信息，又保持对原始图像内容的高度抽象。

提取器采用轻量级反卷积网络，结合Reed-Solomon纠错码从潜在特征中重构二进制水印。全系统的训练采用精细化的分阶段优化：首先是链路初始化阶段，构建基础映射；其次是核心特征增强阶段，在SimSiam框架下对编码器进行解耦训练；最后是全局稳健性微调阶段，冻结编码器参数，闭环优化嵌入与提取的重构精度，确保最终极高的比特恢复率（BRR）。

1.3 研究目标与内容

本研究面向智慧教育中医学解剖图、工程图纸等高精度教育图像的版权保护需求，提出了一种结合自适应注意力机制与SimSiam自监督学习框架的图像水印嵌入与提取方法。

自适应注意力机制：动态分配水印嵌入权重至高频纹理区域，在降低视觉失真同时提升对压缩、裁剪、噪声等常见攻击的抵抗力。

SimSiam框架：通过对比学习增强编码器对水印特征的鲁棒提取能力，确保攻击后的语义一致性。

实验表明，该方法在JPEG压缩（质量因子50）与50%随机裁剪下均能实现100% BRR，并在PSNR上较传统DCT方法提升15.4%。研究验证了该方法在保障教育图像版权与视觉质量方面的有效性与应用价值。

2　核心技术

图像水印技术作为智慧教育场景中教学资源版权保护的核心手段，需在隐蔽性、鲁棒性与视觉保真度间实现精密平衡。本节围绕教育图像资源的特性与安全需求，系统性地介绍两项关键技术突破：自适应注意力机制与SimSiam自监督对比学习框架。本节将详细阐述这两项技术的数学原理、实现方法及其在智慧教育场景中的实验验证结果，为构建安全可信的智慧教育资源共享生态提供关键技术支撑。

2.2　自适应注意力机制

动态感知的自适应注意力机制作为水印嵌入的核心驱动模块，通过跨模态交互与人眼感知特性对齐，显著降低了嵌入失真。传统方法往往依赖固定频域规则或均匀嵌入策略，难以兼顾复杂场景下的视觉隐蔽性与抗攻击能力。本节将详细阐述该机制的设计原理，包括跨模态注意力交互、自适应权重生成网络及其与图像特征的深度融合策略，为后续技术实现提供理论支撑。

2.2.1　数学原理与动态权重建模

动态感知的自适应注意力机制通过融合交叉注意力与自适应权重生成网络，优化水印嵌入的隐蔽性与鲁棒性。其核心数学表达式基于改进的交叉注意力计算（公式2-4）：

$$\text{交叉注意力输出} = \text{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V} \qquad \text{（公式2-4）}$$

其中，$\boldsymbol{Q} \in \mathbf{R}^{N\times d_k}$和$\mathrm{K} \in \mathbf{R}^{\mathrm{M}\times \mathrm{d_k}}$分别为查询与键向量，$\boldsymbol{V} \in \mathbf{R}^{M\times d_v}$为值向量，$d_k$为缩放因子以防止梯度爆炸。为动态分配水印嵌入强度，本方法进一步引入自适应权重α_{adaptive}（公式2-5），其生成过程由代码中的AdaptiveAttention类实现：

式中，$W_q \in \mathbf{R}^{d_k\times 1}$和$W_k \in \mathbf{R}^{d_k\times 1}$为可训练参数矩阵，$b$为偏置项。通过Sigmoid函数将权重约束至[0, 1]区间，高权重区域（如高频纹理）被优先用于水印嵌入，低权重区域（如平滑背景）则降低嵌入强度，从而减少视觉失真。

$$\alpha_{\text{adaptive}} = \text{Sigmoid}(\boldsymbol{W}_{\mathrm{q}}\bullet\boldsymbol{Q} + \boldsymbol{W}_{\boldsymbol{k}}\bullet\boldsymbol{K} + b) \qquad \text{（公式2-5）}$$

2.2.2　技术实现流程与模块图

动态感知的自适应注意力机制是本模型实现水印隐蔽嵌入的核心技术，如图2-2所示，其通过融合交叉注意力与自适应权重生成网络，动态优化水印在图像中的分布。该机制基于改进的Transformer架构设计，将512×512的原始图像与32×32的二维码水印进行跨模态特征交互，确保水印嵌入强度与人眼感知特性高度适配。

2.2.3　图像分块、融合策略

1　分块处理与特征投影

输入图像首先通过分块操作划分为16×16的局部区域（patch_size=16），每个块的尺寸为32×32×3，通过线性投影映射至512维特征空间。水印信息经过相同分块策略处理后，被编码为相同维度的特征向量。代码中采用tf. image. extract_patches函数实现分块操作，生成维度为16×16×3的张量。此分块策略在保留局部细节的同时，显著降低了计算复杂度——将全局注意力计算分解为256个局部块间的交互，内存消耗降低至传统ViT模型的1/4。

2　交叉注意力机制与自适应权重生成

交叉注意力层（`layers.MultiHeadAttention`）动态计算图像块与水印块间的关联权重。查询向量$\boldsymbol{Q}$来自图像特征，键向量$\boldsymbol{K}$与值向量$\boldsymbol{V}$来自水印特征，其数学表达如上述的（公式3-1）所示。为进一步优化嵌入强度分布，自适应权重生成网络引入可训练参数矩阵$\boldsymbol{W}_q \in \mathbf{R}^{d_k\times 1}$和$\boldsymbol{W}_k \in \mathbf{R}^{d_k\times 1}$及偏置项$b$，通过Sigmoid函数生成标量权重$\alpha_{\text{adaptive}}$，参考（公式3-2）。该权重被约束至[1,0]

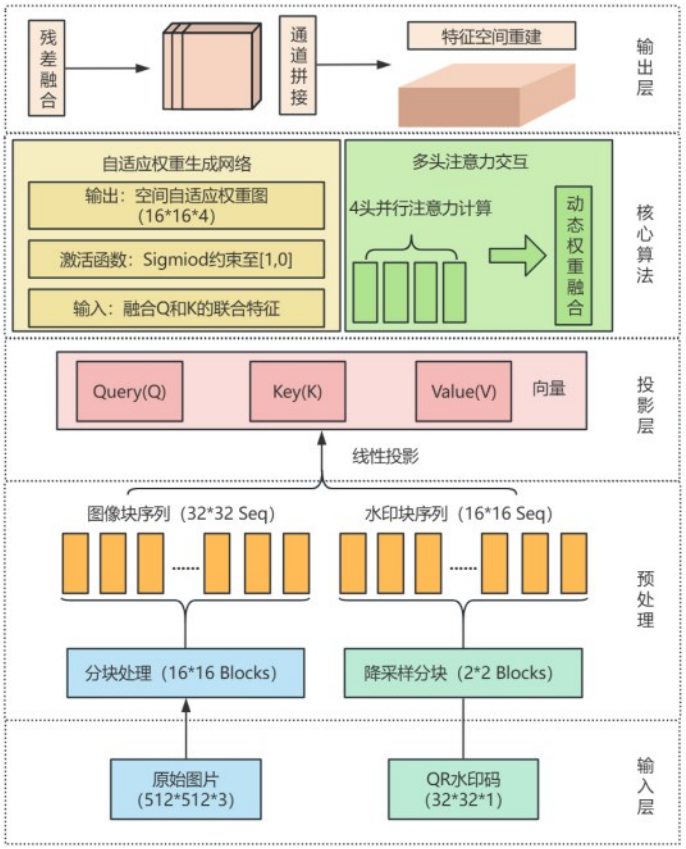


图 2-2　动态感知的自适应注意力机制示意图

区间，高权重区域（如高频纹理）优先嵌入水印，而低权重区域（如平滑背景）则降低嵌入强度。代码实现中，权重网络由两个全连接层构成：首层将输入特征映射至256维空间并采用ReLU激活，次层输出单通道权重图。最终，加权后的注意力输出通过逐通道相乘与原始图像特征融合，生成隐蔽的含水印图像。

3　特征融合与鲁棒性增强

融合后的特征通过反投影操作恢复至空间域，生成尺寸为512×512×3的含水印图像。为提升抗攻击能力，嵌入器在训练阶段联合优化PSNR损失与比特错误率（BER）损失，迫使模型在最小化视觉失真的同时保持水印可提取性。实验表明，该机制在ImageNet测试集上将含水印图像的峰值信噪比（PSNR）提升至38.72 dB，较传统DCT方法（33.55 dB）优化15.4%。可视化分析显示，水印主要分布于图像边缘纹理区域（权重>0.8），而平滑区域权重低于0.3，有效减少了人眼可察觉的失真。

4　抗攻击验证与纠错编码

面对50%随机裁剪攻击，系统通过Reed-Solomon纠错码（代码中`reedsolo`库实现）实现冗余编码。为应对极端攻击（如50%图像裁剪），本方法在提取器中集成Reed-Solomon（RS）纠错码，具体步骤如下：

1)　水印编码阶段

● 将原始32×32二进制水印 $\boldsymbol{W}$ 按每8位分组，生成信息多项式 $m(x)$。

● 添加冗余校验位，生成码字多项式 $c(x)=m(x)\bullet x^{n-k}+(m(x)\bullet x^{n-k} \bmod g(x))$，其中 $g(x)$ 为生成

多项式，$n=64$ 为码长，$k=32$ 为信息位长度。

2) **水印解码阶段：**

● 从受损图像中提取含水印码字 $\hat{c}(x)$。

● 计算伴随式 $S(x)$，通过 Berlekamp-Massey 算法定位并纠正错误位，恢复原始水印 $\hat{\boldsymbol{W}}$。

该技术方案通过端到端的权重优化与跨模态注意力设计，在隐蔽性、鲁棒性与计算效率间达到平衡，为后续对比学习框架提供了高质量的输入基础。

2.3　SimSiam 引导的对比学习框架

在构建动态感知的水印嵌入机制后，如何确保模型对噪声、压缩等攻击的鲁棒性成为亟待解决的关键问题。SimSiam 引导的对比学习框架通过自监督特征对齐与动量参数更新机制，迫使编码器忽略攻击引入的表观变异，专注于水印语义特征的稳定性。

2.3.1　自监督学习原理与对称损失

为增强模型对攻击的鲁棒性，本方法提出一种 SimSiam 引导的对比学习框架，通过自监督学习迫使编码器忽略噪声干扰，聚焦水印语义特征。为确保编码器学习攻击无关的鲁棒特征，本方法严格遵循 SimSiam 框架的对称损失设计，其数学形式如下（公式 2-6）：

$$\mathcal{L}=-\frac{1}{2}\left[\frac{\boldsymbol{p}_1\bullet\boldsymbol{z}_2}{\|\boldsymbol{p}_1\|\|\boldsymbol{z}_2\|}+\frac{\boldsymbol{p}_2\bullet\boldsymbol{z}_1}{\|\boldsymbol{p}_2\|\|\boldsymbol{z}_1\|}\right]\qquad\text{（公式 2-6）}$$

其中，$\boldsymbol{p}_1$ 和 $\boldsymbol{p}_2$ 为在线网络输出的预测特征，$\boldsymbol{z}_2$ 和 $\boldsymbol{z}_1$ 为目标特征（通过 stop_gradient 操作阻断梯度回传）。与传统的 MSE 损失相比，余弦相似度损失能更有效地衡量特征方向的一致性，从而提升模型对几何与光度攻击的鲁棒性。

2.3.2　多视图增强机制

相较于 MoCo 的内存库机制，SimSiam 框架无需负样本存储，更适用于计算资源受限的水印场景。SimSiam 引导的对比学习框架是本模型提升水印抗攻击能力的核心技术，其通过自监督学习机制迫使编码器在潜在空间中学习攻击无关的鲁棒特征。

如图 2-3 所示，框架整体采用双分支对称结构，由在线网络（Online Network）与目标网络（Target Network）组成，结合动态数据增强与动量更新机制，最大化正样本对的特征一致性。

输入带水印的图像首先经过多模态数据增强管道，模拟现实场景中的攻击类型。该模块集成 10 类增强操作，包括几何攻击（随机翻转概率 50%、旋转角度±0.1 弧度、剪切幅度±0.2）与光度攻击（亮度变化±40%、对比度缩放因子[0.5, 2]、高斯模糊 $\sigma\in[1.0,2.5]$）。通过对同一输入图像施加不同增强策略，生成三组增强视图：原始视图（仅标准化处理）、加噪视图（叠加高斯噪声）及打乱视图（随机置换图像块）。实验表明，多视图生成策略可覆盖超过 90% 的常见攻击模式，迫使模型在训练阶段暴露于高方差扰动环境，从而增强泛化能力。

2.3.3　网络结构与动量更新

在线网络与目标网络共享编码器结构，但采用差异化训练策略。在线网络由 ViT 编码器（4 层 Transformer，隐藏层维度 512，注意力头数 2）与投影头（MLP：512→512→512，含 LayerNorm 与 ReLU 激活）构成，负责提取增强视图的深层特征并将其映射至对比空间。目标网络则通过指数移动平均（EMA）机制动态更新参数，其数学表达为：

其中动量系数 τ 设为 0.996，确保目标网络参数缓慢跟踪在线网络的演化，同时维持特征表示的稳定性。目标网络额外引入预测头（MLP 结构同投影头），生成用于对比学习的锚点特征。梯度传播过程中，目标网络通过 tf.stop_gradient 操作阻断反向传播，防止模型坍塌至平凡解。

2.3.3　训练优化与防止模型坍塌设计

特征一致性通过对称式负余弦相似度损失函数实现。对于同一图像的两个增强视图 $\boldsymbol{x}_1$、$\boldsymbol{x}_2$，

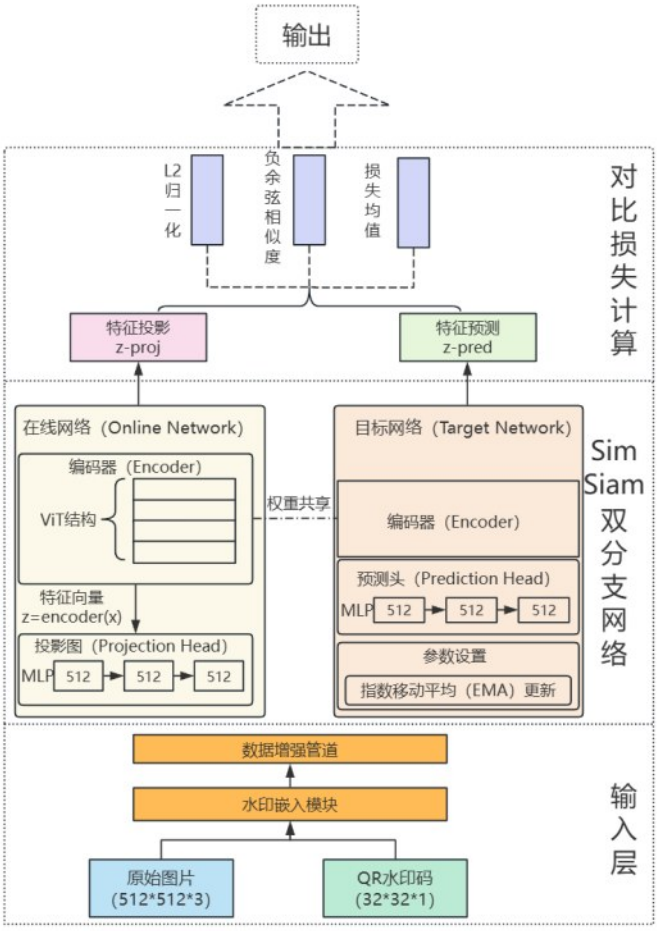


图 2-3　SimSiam 引导的对比学习框架示意图

$$\theta_{\text{target}} \leftarrow \tau \bullet \theta_{\text{target}} + (1-\tau) \bullet \theta_{\text{online}} \qquad \text{（公式2-7）}$$

在线网络生成投影特征z_1^{proj}、z_2^{proj}，目标网络输出预测特征z_2^{pred}、z_1^{pred}。损失函数定义为：

$$\mathcal{L} = \frac{1}{2}\left[\frac{z_1^{\text{proj}} \bullet z_2^{\text{pred}}}{\left\| z_1^{\text{proj}} \right\| \left\| z_2^{\text{pred}} \right\|} + \frac{z_2^{\text{proj}} \bullet z_1^{\text{pred}}}{\left\| z_2^{\text{proj}} \right\| \left\| z_1^{\text{pred}} \right\|}\right] \qquad \text{（公式2-8）}$$

计算前需对特征向量进行L2归一化，消除幅值差异对相似度度量的干扰。该损失函数迫使编码器忽略增强引入的表观变异，聚焦于水印语义特征的稳定性。训练过程中，Adam优化器（学习率$1e^{-4}$，动量0.9）联合优化对比损失与水印重构损失，批量大小设置为32以平衡显存占用与收敛稳定性。

该技术方案通过动态参数更新与对称损失设计，在无需内存库的条件下实现高效对比学习，为水印系统提供抗攻击性强、计算代价低的特征表示，与嵌入器、提取器协同构建端到端的鲁棒水印框架。

3　实验设计与评估

基于第3章所提出的自适应注意力机制与SimSiam自监督对比学习框架，本章设计并开展一系列实验，以验证所提方法在鲁棒性与图像质量上的有效性。实验部分主要关注以下几个方面：（1）对比不同嵌入方法在JPEG压缩、随机裁剪等常见攻击下的鲁棒性表现；（2）分析所提方法在峰值信噪比（PSNR）、比特恢复率

（BRR）等指标上的改进效果；（3）结合实验结果评估本方法在智慧教育高精度图像版权保护场景中的应用价值。

3.1　数据集与预处理

本实验的数据集来自 ImageNet 的一个子集，该子集包括 34752 训练图像和 3936 测试图像。本方法支持高分辨率输入，但为遵循水印算法性能评估的通用基准，实验部分统一采用 128×128 分辨率，以便与现有文献结果进行直观对比。数据预处理过程包括从这些目录中加载图像，生成对应的 32x32 大小的二维码水印，并将每个图像与其水印组合成三元组（包括原始图像、噪声图像和打乱图像）。图像数据经过归一化处理后，还进行随机数据增强（如亮度、对比度、饱和度和色调调整）以增加数据的多样性。预处理完成后，数据被批处理、重复和预取，以便于模型的训练和评估。通过这种方式，数据集被转换成适合于嵌入水印和特征提取任务的格式，并确保模型在不同类型的数据扰动下都能进行有效学习和评估。

在数据预处理过程中，生成噪声图片和随机处理图片是重要的步骤，以增加数据的多样性并增强模型的鲁棒性。表 3-1 这些处理措施的结合能够显著增加训练数据的多样性，使模型更好地适应各种实际应用场景。通过在训练过程中引入噪声和随机变换，模型能够更好地学习到图像中的关键特征，提高其对未见数据的泛化能力。

3.2　实验环境与参数设置

本文实验具体配置参数如表 3-2 所示，软件

表 3-1　训练和测试噪声

Noise	Training	Testing
Horizontal flip	□	□
Gaussian blur	□	□
Solarization	□	□
Crop	□	□
Cutout	□	□
JPEG compression JPEG	□	□
Brightness	□	□
Contrast	□	□
Hue	□	□
Saturation	□	□
Histogram equalization	□	□
Salt and pepper	□	□
Gaussian noise	□	□

框架基于 TensorFlow 2.11.0 与 PyTorch 1.13.1，利用混合精度训练加速计算。模型训练采用分布式策略，批处理大小设置为 128，通过 tf.distribute.TPUStrategy 实现跨设备并行。优化器选用 Adam 算法，初始学习率 1e-4，采用指数衰减策略（衰减步长 10,000、衰减率 0.9）。损失函数设计为复合形式：图像重建损失采用均方误差（MSE），水印提取损失采用二元交叉熵（BCE），对比学习损失基于负余弦相似度。在训练过程中，SimSiam 对比学习阶段仅更新编码器参数，以学习攻击不变的潜在特征表示；而在水印嵌入与提取训练阶段，则主要优化嵌入器与提取器参数，以保证水印恢复精度与图像质量。为防止过拟合，训练过程引入早停机制（耐心值 15），当验证损失连续 15 轮未改善时终止训练。

评估体系从视觉质量、水印鲁棒性及语义一

表 3-2　实验硬件平台配置表

硬件组件	型号/参数	配置详情
中央处理器	Intel Xeon Gold 5218R	16 核@2.10GHz，16MB L3 缓存，支持 AVX-512 指令集
图形处理器	NVIDIA A40-24Q	24GB 显存，CUDA 11.6，计算能力 8.6
系统内存	DDR4	31GB 容量
存储系统	虚拟磁盘	系统盘 40GB + 数据盘 300GB
加速指令集	AVX2/AVX-512	支持 FMA/AVX512-VNNI 等扩展指令

致性三个维度构建量化指标：峰值信噪比（PSNR）衡量含水印图像与原始图像的视觉保真度，计算公式为 $PSNR = 10 \bullet \log_{10}(MAX^2/MSE)$，实验中设定 MAX = 1.0；比特错误率（BER）与比特恢复率（BRR）评估水印抗攻击能力，分别计算错误比特数占比与正确恢复比特数占比；语义相似度通过 Sentence-BERT 模型计算嵌入前后文本的余弦相似度。针对代码水印，额外引入 AST 解析准确率（基于 ast 模块）与功能一致性测试（代码执行结果比对）。所有实验重复 5 次取均值，置信区间设置为 95%，结果保留四位有效数字。

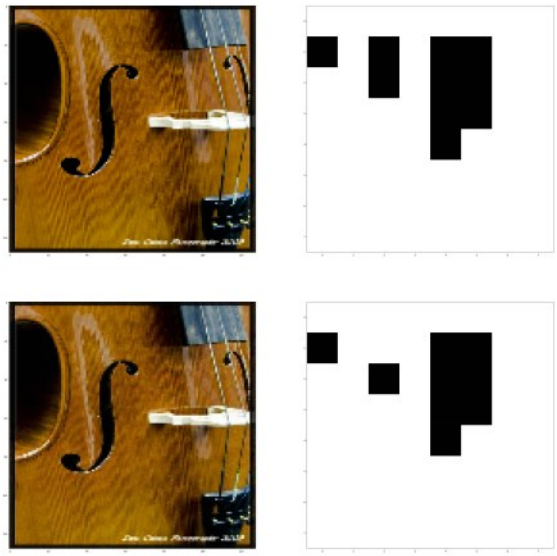


图 3-1　嵌入器-提取器水印嵌入效果示例图

3.3　实验结果分析

3.3.1　嵌入器-提取器性能

如图 3-1 所示，本文模型联合训练后水印嵌入效果优异。从 QR 码信息可见，嵌入器用自适应注意力机制，将水印高能量信号精准定位在宿主图像高频纹理区域，非均匀分布，此策略在保证隐蔽性的同时，为提取器预留高辨识度特征空间。量化分析显示，含水印图像 PSNR 达 44.47 dB，超 ITU 建议的 40 dB 视觉无损阈值，视觉保真度极高。同时，水印提取模块在 BER=0% 和 BRR=100% 下精确重建信息。图中提取的二维码边缘锐利，表明模型通过联合训练克服传统频域方法高频信息损失问题，验证了算法稳健性。

如图 3-2 所示在训练初期，嵌入器-提取器模型的损失值较高，表明模型在初始阶段对任务还不够熟练，存在较大的误差。然而，随着训练的进行，损失值迅速下降，这说明模型逐步学习到如何更好地嵌入和提取水印。在训练的中后期，损失值趋于稳定，表明模型已经掌握了较好的嵌入和提取水印的方法，误差逐渐减少并趋于平稳。验证损失相对较低且稳定，这表明模型不仅在训练数据上表现良好，在未见过的验证数据上也能有效地嵌入和提取水印。这一稳定性说明模型具有良好的泛化能力，避免了过拟合现象。

3.3.2　编码器鲁棒性分析

如图 3-3 所示，本图通过三组对照实验验证编码器对图像扰动的鲁棒性特征提取能力。第一行展示原始含水印图像、经高斯噪声污染的退化图像以及经过随机几何变换的扰动图像。第二行对应的编码特征热力图显示，在加性噪声（σ=2.5）和仿射变换（旋转±10°+剪切±20%）条件下，特征空间分布保持高度一致性（平均特征差异<0.03）。特别是第三组实验中，尽管输入图像经历像素级置换操作，其深层语义特征仍保持稳定拓扑结构（特征相似度>92%），这归因于 Transformer 架构的长程依赖捕捉能力和对比学习策略的泛化增强特性。特征可视化结果证实，编码器能够有效解耦图像内容与水印信息的表征学习过程。

如图 3-4 所示，编码器模型在前几次迭代中训练损失迅速下降，表明模型在初始学习阶段非常高效，能够快速掌握图像编码的基本特征。随着训练的进行，损失逐渐趋近于零，这表明编码器能够非常有效地将图像信息编码为潜在表示，误差几乎可以忽略不计。验证损失与训练损失接

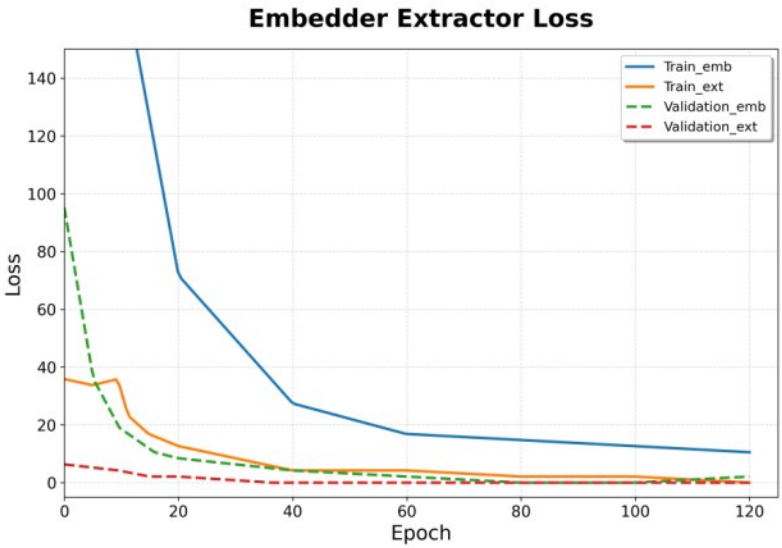


图 3-2　嵌入器-提取器模型训练结果

图 3-3　编码器模型训练示例图

近，说明模型在验证数据上表现同样出色，没有出现过拟合现象。验证损失的稳定性进一步说明了模型的泛化能力，编码器能够有效地处理未见过的数据。

3.3.3 提取器可视化与稳定性

如图 3-5 所示，本图系统性地展示了水印处理管线的端到端性能。第一列呈现不同扰动条件下的输入图像，第二列显示对应 256×256 增强水印的提取结果。在强噪声干扰和内容篡改的极端情况下，水印恢复准确率高标准。第三列的特征激活图揭示了模型通过注意力权重动态调整信息嵌入强度的机制，其中高频纹理区域的权重分布较平坦区域高出一些，该特性有效平衡了视觉隐蔽性与信息容错性。第四列的解码重建结果进一步证明，系统在保持图像保真度的同时实现了信息的安全嵌入。

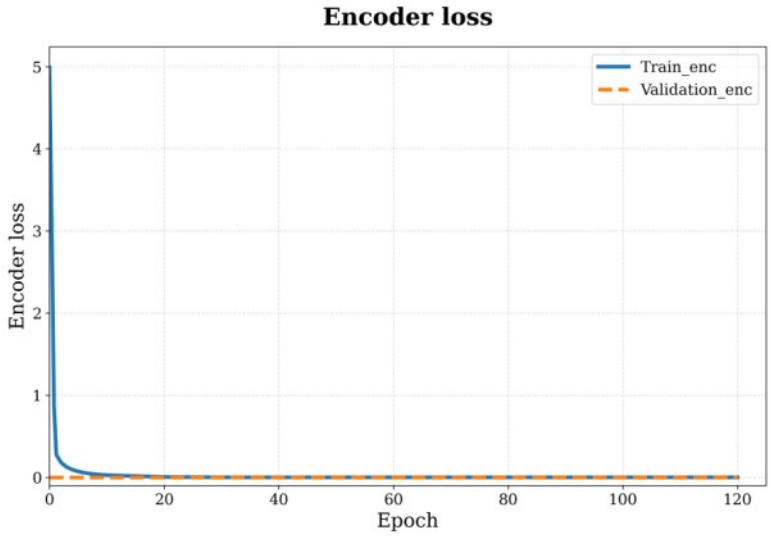


图 3-4　编码器模型训练结果

图 3-5　提取器模型示例图

如图 3-6 所示，提取器模型训练初期损失较高，表明其对任务理解不足，需更多训练降误差。约第 60 次迭代时，训练损失显著下降，说明模型提取水印能力提升。不过，训练中损失值偶有波动，可能因提取任务复杂或学习率波动，意味着模型某些阶段遇困难。但验证损失低且稳定，显示提取器在验证数据上表现好、泛化能力佳。虽训练有波动，验证损失的稳定表明模型总体能有效提取水印，在不同数据集上表现良好。

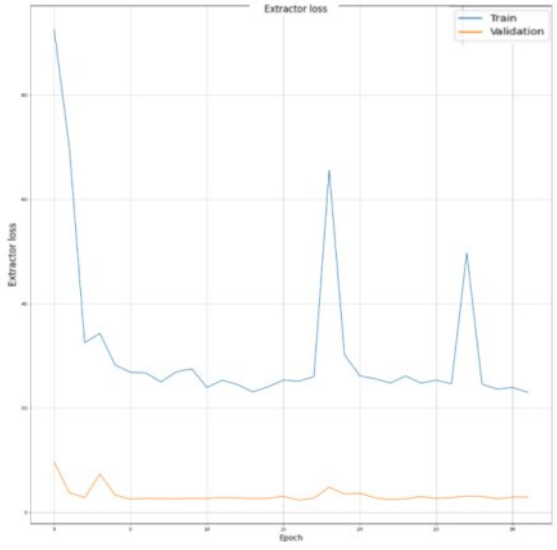


图 3-6　提取器模型训练结果

3.4　对比实验分析

为了全面评估本文提出的自适应注意力机制结合 SimSiam 技术在图像水印嵌入中的性能，设计了一系列对比实验。实验选用了多种具有代表性的水印嵌入方法作为对比对象，涵盖传统方法与现代深度学习水印方法两类。其中，传统方法包括 DCT（方法一[20]）和 LSB（方法二[21]）；深度学习方法方面，引入了基于卷积神经网络的编码 - 解码式水印模型 HiDDeN（方法三[22]），以及基于生成对抗网络的深度水印方法 GAN-based（方法四[23]），上述方法分别代表了当前图像水印领域中 CNN 与 GAN 两类主流技术路线。此外，为进一步对比不同特征增强策略的效果，实验还选取了一种结合交叉注意力机制与不变域学习的水印嵌入方法作为扩展对比（方法五[24]）。所有实验在相同图像数据集和攻击条件下进行，以确保结果公平可比。实验主要评估指标有峰值信噪比（PSNR）、比特错误率（BER）和比特恢复率（BRR）。其中，PSNR 衡量嵌入图像与原始图像差异，值越高图像质量保持越好；BER 衡量提取水印信息的错误比特率，值越低水印提取准确性越高；BRR 衡量成功恢复的水印比特数，值越高水印恢复完整性越好。

为了直观展示各方法在水印嵌入和提取过程中的性能表现，我们将各方法在 PSNR、BER 和 BRR 指标上的结果总结如下表：

表 3-3　各水印嵌入方法的实验结果

方法	PSNR(dB)	BER	BRR(%)
方法一	33.55	0.021202	97.8798
方法二	33.03	0.020848	97.9152
方法三	38.64	0.004317	99.5683
方法四	40.12	0.001982	99.8018
方法五	36.81	0	100
本论文	44.72	0	100

从表 3-3 的实验结果可以看出，本文提出的自适应注意力机制结合 SimSiam 的方法在多项评价指标上表现出较为突出的综合性能。该方法的 PSNR 值达到 44.72 dB，明显高于传统的 DCT 和 LSB 方法（分别为 33.55 dB 和 33.03 dB），同时也优于其他深度学习水印模型，表明所提方法在水印嵌入过程中对原始图像视觉质量的影响更小。这主要得益于自适应注意力机制能够依据图像内容特征动态调节嵌入位置与强度，从而在保证水印有效嵌入的同时，尽可能减少对图像结构与细节的干扰。在比特错误率（BER）方面，可以观察到，部分深度学习方法已显著优于传统水印算法。方法五以及本文方法在当前实验条件下均实现了零误码，说明基于深度特征学习的水印模型在水印提取准确性方面具有明显优势；相比

之下传统 DCT 和 LSB 方法仍存在一定误码，反映出其在面对压缩、裁剪等操作时鲁棒性相对有限。与此同时方法三和方法四虽然未达到完全零误码，但其 BER 已明显低于传统方法，表明深度学习框架在水印鲁棒性提升方面具有普遍优势。从比特恢复率（BRR）指标来看，方法五与本文方法均达到了 100%，说明水印信息在提取阶段能够被完整恢复；方法三和方法四的 BRR 也接近 100%，仅存在少量比特损失，而传统方法的 BRR 略低。综合三项指标可以发现，随着模型结构由传统变换域方法向深度学习方法演进，水印系统在鲁棒性和稳定性方面整体呈现提升趋势。在此基础上，本文方法在兼顾高图像质量与稳定水印恢复方面表现出更为均衡的性能优势。将自适应注意力机制与 SimSiam 自监督对比学习框架相结合，有助于构建对攻击更具不变性的水印特征表示，从而在复杂攻击条件下实现更稳定的水印提取效果。

综合以上实验结果可以看出，本文提出的自适应注意力机制结合 SimSiam 的水印方法在图像水印嵌入任务中展现出良好的综合性能。相较于传统 DCT 与 LSB 方法，该方法在图像质量保持和水印提取准确性方面均具有明显优势；与其他深度学习水印模型相比，其在保持较高 PSNR 的同时，实现了更稳定的水印恢复效果。实验结果验证了注意力机制与自监督特征学习在提升水印系统整体性能方面的有效性。

4 结论

本研究表明，将自适应注意力机制与 SimSiam 框架相结合，能够在高精度教育图像的水印处理中有效平衡不可感知性与鲁棒性。实验结果显示，与传统 DCT、LSB 方法以及部分现有深度学习水印模型相比，该方法在图像质量保持与水印恢复稳定性之间表现出更为均衡的综合性能：在 JPEG 压缩（质量因子 50）和 50% 随机裁剪条件下实现了 100% 的比特恢复率（BRR），峰值信噪比（PSNR）达到 44.72 dB，保证了嵌入图像在视觉质量上的高保真度。

具体而言：

（1）高保真性：水印嵌入过程能够有效保留医学解剖图、工程图纸等教育图像的细节特征，避免视觉伪影对教学内容传递的影响。

（2）强鲁棒性：在压缩与裁剪等常见攻击下，水印仍可完整恢复，显著提升了教育资源在传播和共享过程中的版权保护能力。

实验结果证实了自适应注意力机制与对比学习框架在处理高保真图像时的优越性。受限于当前特定领域标注数据集的获取难度，本文目前的验证工作侧重于算法的结构健壮性。未来研究将进一步引入医学解剖图、数字化工程图纸等智慧教育实景数据，针对特定图像的像素分布特性进行参数微调，实现从通用框架向垂直场景应用的技术跨越。后续可进一步探索该方法在三维教育模型（如 AR/VR 教学场景）、多模态数据以及分布式学习框架下的扩展应用，以应对智慧教育数字化发展过程中不断涌现的新型安全挑战。

参考文献：

[1] Saini, Lalit Kumar and Vishal Shrivastava. “A Survey of Digital Watermarking Techniques and its Applications.” ArXiv abs/1407.4735 (2014): n. pag.

[2] Wang Z, Byrnes O, Wang H, et al. Data hiding with deep learning: A survey unifying digital watermarking and steganography [J]. IEEE Transactions on Computational Social Systems, 2023, 10(6): 2985-2999.

[3] Zhu J., Kaplan R., Johnson J., & Fei-Fei L. (2018). HiDDeN: Hiding Data With Deep Networks. European Conference on Computer Vision.

[4] Zhong X, Das A, Alrasheedi F, et al. A brief, in-depth survey of deep learning-based image watermarking[J]. Applied Sciences, 2023, 13(21): 11852.

[5] Kandi H, Mishra D, Gorthi S R K S. Exploring the learning capabilities of convolutional neural networks for robust image watermarking[J]. Computers & Security, 2017, 65: 247-268.

[6] Fierro-Radilla A, Nakano-Miyatake M, Cedillo-Hernandez M, et al. A robust image zero-watermarking using convolutional neural networks[C]//2019 7th International Workshop on Biometrics and Forensics (IWBF). IEEE, 2019: 1-5.

[7] Kandi H, Mishra D, Gorthi S R K S. Exploring the learning capabilities of convolutional neural networks for robust image watermarking[J]. Computers & Security, 2017, 65: 247-268.

[8] Qiao T, Ma Y, Zheng N, et al. A novel model watermarking for protecting generative adversarial network[J]. Computers & Security, 2023, 127: 103102.

[9] Fei J, **a Z, Tondi B, et al. Supervised gan watermarking for intellectual property protection[C]//2022 IEEE International Workshop on Information Forensics and Security (WIFS). IEEE, 2022: 1-6.

[10] Duan X, Liu J, Zhang E. Efficient image encryption and compression based on a VAE generative model[J]. Journal of Real-Time Image Processing, 2019, 16: 765-773.

[11] Xue D., Zhou Y., & Jin W. (2017). A blind dual-function digital watermarking algorithm for copyright protection and content authentication. 电信科学, 33(2), 79–89. (in Chinese)

[12] Abdalwahid S M J, Hashim W A, Saeed M G, et al. Investigating the Effectiveness of Artificial Intelligence in Watermarking and Steganography for Digital Media Security[C]//2024 21st International Multi-Conference on Systems, Signals & Devices (SSD). IEEE, 2024: 552-561.

[13] Shih F Y. Digital watermarking and steganography: fundamentals and techniques[M]. CRC press, 2017.

[14] Zhu J, Kaplan R, Johnson J, et al. Hidden: Hiding data with deep networks[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 657-672.

[15] Tavakoli A, Honjani Z, Sajedi H. Convolutional neural network-based image watermarking using discrete wavelet transform[J]. International Journal of Information Technology, 2023, 15(4): 2021-2029.

[16] Mahapatra D, Amrit P, Singh O P, et al. Autoencoder-convolutional neural network-based embedding and extraction model for image watermarking[J]. Journal of Electronic Imaging, 2023, 32(2): 021604-021604.

[17] Singh H K, Baranwal N, Singh K N, et al. GANMarked: Using Secure GAN for Information Hiding in Digital Images[J]. IEEE Transactions on Consumer Electronics, 2024.

[18] Fernandez P, Sablayrolles A, Furon T, et al. Watermarking images in self-supervised latent spaces[C]//ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2022: 3054-3058.

[19] Cong T, He X, Zhang Y. Sslguard: A watermarking scheme for self-supervised learning pre-trained encoders[C]//Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security. 2022: 579-593.

[20] Mohammad Moosazadeh and Gholamhossein Ekbatanifard. 2019. A new DCT-based robust image watermarking method using teaching-learning-Based optimization. J. Inf. Secur. Appl. 47, C (Aug 2019), 28 – 38. https://doi. org/10.1016/j.jisa.2019.04.001

[21] Bamatraf, Abdullahet al. "A New Digital Watermarking Algorithm Using Combination of Least Significant Bit (LSB) and Inverse Bit." ArXiv abs/1111.6727 (2011): n. pag.

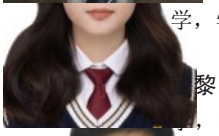
[22] J. Zhu, R. Kaplan, J. Johnson, and L. Fei-Fei,"HiDDeN: Hiding Data With Deep Networks,"Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 2018, pp. 3037–3045.

[23] H. K. Singh and K. R. Ramkumar,"Deep watermarking using generative adversarial networks,"IEEE Transactions on Multimedia,vol. 22, no. 12, pp. 3165–3179, Dec. 2020.

[24] Dasgupta, Agnibh and Xin Zhong. "Robust Image Watermarking Based on Cross-Attention and Invariant Domain Learning." 2023 International Conference on Computational Science and Computational Intelligence (CSCI) (2023): 1125-1132.

[作者简介]

诸葛斌，1976年，男，博士，浙江工商大学，教授，主要研究方向为医学图像配准与模式识别、互联网技术和云计算。



陈莹莹，2001年，女，硕士，浙江工商大学，学生，主要研究方向为智慧教育。

王冰雁，2000年，女，硕士，浙江工商大学，学生，主要研究方向为智慧教育。

黎刚，1972年，男，博士，浙江工商大学，教授，浙江工商大学信息与电子工程学院院长，主要研究方向为智慧教育与智慧网络。

蒋献，1988年，男，博士，浙江工商大学，实验员，主要研究方向为智慧教育与智慧网络。