计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
2025,34(2):183−194 [doi: 10.15888/j.cnki.csa.009772] [CSTR: 32024.14.csa.009772]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于增强残差特征的孪生网络图像隐写分析①

刘佳梅, 苏　海

(华南师范大学 人工智能学院, 佛山 528225)
通信作者: 苏　海, E-mail: suhai@m.scnu.edu.cn

摘　要: 图像隐写分析旨在检测图像是否经过隐写术处理从而携带了秘密信息. 基于孪生网络的隐写分析算法通过计算待检测图像左右分区的不相似性以此判断图像是否携带秘密信息, 是目前深度学习图像隐写分析算法里面准确度较高的网络. 然而, 基于孪生网络的图像隐写分析算法仍然存在一些局限性. 首先, 孪生网络在预处理层和特征提取层中叠加的卷积块, 忽略了隐写信号从浅层传递到深层过程中容易丢失的问题. 其次, 现有的孪生网络使用的SRM 滤波器仍然沿用其他网络使用的高通滤波器来抑制图像内容, 忽略了生成的残差图大小单一的问题. 为了解决以上问题, 本文提出了基于增强残差特征的孪生网络图像隐写分析方法. 本文方法设计了一种基于注意力的倒残差模块, 通过在预处理层和特征提取层的卷积块后添加基于注意力的倒残差模块, 重用图像特征, 引入注意力机制, 增强网络对图像纹理复杂区域的特征图赋予更多权重. 同时为了更好地抑制图像内容, 提出多尺度滤波器, 将残差类型调整为多个尺寸不同的卷积核进行操作, 丰富残差特征. 实验结果表明, 本文提出的基于注意力的倒残差模块和多尺度滤波器相较于现有方法分类效果更佳.

关键词: 隐写分析; 孪生网络; 多尺度滤波器; 基于注意力的倒残差

引用格式: 刘佳梅,苏海.基于增强残差特征的孪生网络图像隐写分析.计算机系统应用,2025,34(2):183–194. http://www.c-s-a.org.cn/1003-3254/9772.html

Siamese Network Image Steganalysis Based on Enhanced Residual Features

LIU Jia-Mei, SU Hai

(School of Artifical Intelligence, South China Normal University, Foshan 528225, China)

Abstract: Image steganalysis aims to detect whether an image undergoes steganography processing and thus carries secret information. Steganalysis algorithm based on Siamese networks determines whether an image carries secret information by calculating the dissimilarity between the left and right partitions of the image to be detected. This approach currently boasts relatively high accuracy among deep learning image steganalysis algorithms. However, Siamese network-based image steganalysis algorithms still have certain limitations. First, the convolutional blocks stacked in the preprocessing and feature extraction layers of the Siamese network overlook the issue of steganographic signals easily being lost as they are transmitted from shallow to deep layers. Second, SRM filters used in existing Siamese networks still employ high-pass filters from other networks to suppress image content, ignoring single-sized generated residual maps. To address the above problems, this study proposes a Siamese network image steganalysis method based on enhanced residual features. The proposed method designs an attention-based inverted residual module. By adding the attention-based inverted residual module after the convolutional blocks in the preprocessing and feature extraction layers, it reuses image features, introduces an attention mechanism, and enables the network to assign more weights to feature maps of complex-textured image regions. Meanwhile, to better suppress image content, a multi-scale filter is proposed, adjusting the residual types

① 基金项目: 广东省基础与应用基础研究基金 (2021A1515110673)
收稿时间: 2024-07-18; 修改时间: 2024-09-03; 采用时间: 2024-09-10; csa 在线出版时间: 2024-11-28
CNKI 网络首发时间: 2024-11-29

to operate with convolutional kernels of different sizes, thereby enriching residual features. Experimental results show that the proposed attention-based inverted residual module and multi-scale filter provide better classification performance compared to existing methods.

Key words: steganalysis; Siamese network; multi-scale filter; attention-based inverted residual module

在信息安全领域, 图像隐写分析作为图像隐写术的对抗方法, 一直是研究的重点和热点. 图像隐写术通过将隐写信息嵌入载体图像中, 生成在视觉上与原始图像基本一致的载密图像, 从而实现隐秘通信的目的[1]. 然而, 图像隐写分析技术旨在检测图像中是否包含隐写信息, 以防止隐秘通信的发生和传播. 现有研究表明, 尽管深度学习在图像隐写术中具有较高的应用潜力, 但其安全性仍然较差, 易被隐写分析器检测. 目前, 大多数隐写分析器主要针对传统的自适应隐写算法设计. 图像隐写分析在实际应用中能够有效防范隐写术的滥用, 维护社会的稳定与安全. 当前, 图像隐写分析的研究主要集中在二分类问题上, 即判别待检测图像是否包含隐写信息.

由于加载隐写信号前后的图像相似性极高, 隐写分析任务与常规视觉任务存在显著差异. 在图像隐写分析领域, 秘密信息的嵌入通常被视为在载体图像上添加低幅值的隐写信号[2], 使其转化为一个低信噪比的分类问题. 由于隐写信号含量较少, 增强隐写信号并减少图像内容对隐写分析的干扰, 一直是该领域的重点研究方向. 目前, 大多数深度学习隐写分析方法基于主流的分类模型[3−6], 通过提取隐写信号特征, 抑制图像内容, 从而判断这些特征是否为真实的隐写信号. 然而, 大部分深度学习分类网络在分类层难以区分隐写信号与其他高频信号. 尽管许多学者尝试通过特征分类进行隐写分析, 但这些方法的精度仍有进一步提升的空间.

You 等人[7]于 2021 年提出了一种基于孪生网络的隐写分析算法, 为隐写分析提供了新思路. 该算法将输入图像分为左右两半区域, 通过计算两半区域特征向量的相似性, 根据不相似性来区分载密图像. 基于孪生网络的隐写分析算法在当年的检测精度达到了最优. 其主要优势在于, 将传统方法中对高频特征的二分类判断, 转变为对图像不同区域的隐写信号差异的检测. 孪生网络在预处理层使用 SRM[8]滤波器进行残差图的计算, 特征提取层使用多个卷积块进行隐写特征的提取, 分类层将特征提取层得到的特征向量经过全局平均池化后融合两个图像块的特征, 经过全连接层进行分类. 并且额外添加了一个相似损失函数, 对左右子区域的隐写特征进行相似性计算.

然而孪生网络在进行隐写特征提取的时候, 忽略了微量的隐写信号在下采样和卷积层传递中容易丢失, 并且在预处理层使用的 SRM 滤波器是将不同的残差类型通过填充补 0 扩充成 30 个 5×5 的卷积核, 生成的残差图大小单一. 本文为了弥补卷积层传递和下采样过程中的隐写信号特征的丢失, 引入由逐点卷积和深度可分离卷积的倒残差模块[9], 重用图像特征, 扩展通道, 丰富残差特征. 由于自适应隐写算法更倾向将隐写信号嵌入在图像纹理复杂区域, 我们在倒残差模块中引入 SE[10]注意力, 对隐藏在图像复杂纹理和边缘区域的特征图赋予更多权重, 增强与嵌入区域相对应的特征图的显著性. 同时, 本文为了更好地抑制图像内容, 改进 SRM 滤波器, 提出多尺度滤波器来丰富生成不同的残差图. 多尺度滤波器将残差类型调整为多个尺寸为 5×5 和 3×3 卷积核进行操作, 进行不同感受野大小的提取噪声残差, 丰富残差特征.

本文的主要贡献如下.

(1) 提出了基于注意力的倒残差模块, 通过在卷积块中加入基于注意力的倒残差模块, 重用图像特征, 尽可能减少隐写信号在下采样和深层传递过程的丢失.

(2) 提出了多尺度滤波器, 重新设计了 SRM 滤波器的卷积核, 结合多个尺寸不同大小的卷积核丰富残差图特征.

(3) 在 WOW 和 S-UNIWARD 等不同嵌入率的数据集中, 本方法较现有的隐写分析算法准确率更高.

1 国内外研究现状

1.1 基于传统的图像隐写分析模型

传统的图像隐写分析是基于人工设计的特征来提取隐写信号, 通过不断增加特征维度和多样性来捕捉更加隐蔽的隐写信号. 最初的特征提取是利用对图像

像素或者频域系数这些信息生成各种阶数的矩[11], 表示为图像的特征. 随后, 特征提取方法逐渐转向利用临近多个像素或频域系数之间的复杂相关性. Fridich 等人[8]提出一种基于 SRM (spatial rich model) 模型的富模型特征提取算法, 通过融合残差图上的共生矩阵, 增加高维特征的多样性来捕捉到更多的隐写痕迹.

目前, 富模型隐写分析通过设计多样化的线性和非线性滤波器, 取得了良好的性能, 其特征维数已经达到了数万维以上[12–14]. 富模型已经成为隐写分析器的关键工具, 在现代隐写分析网络中得到广泛应用.

1.2 基于深度学习的图像隐写分析模型

基于深度学习的图像隐写分析模型通过其强大的学习能力, 有效地提取隐写特征, 构建端到端的隐写分析器, 在性能上达到了与富模型隐写分析相当甚至更优的效果. 深度学习图像隐写分析器分为预处理层、特征提取层, 分类层 3 个阶段. 大多数模型在这 3 个层次上进行了不同程度的改进, 以增强对隐写特征的提取能力.

在预处理层, 研究者们主要通过设计和调整滤波器的训练方式, 以抑制图像内容干扰并提高分类准确率. 此处滤波器指网络的第 1 层卷积核, 训练方式分为固定式权重类、启发式权重类、随机式权重类. Xu 等人[3]提出了使用固定式权重 KV 核, 固定式权重不参与反向传播. Ye 等人[2]使用启发式权重即 30 个 5×5 的 SRM 滤波核, 启发式权重参与反向更新. SRNet[5]使用随机化初始化权重并参与反向传播, 但难以收敛, 训练速度较慢, 在实际应用中并未广泛应用. 另外, 其他研究者也提出了一些特别的滤波器, 如 RXGNet[15]使用固定的 GPD 滤波器, Luo 等人[16]使用 MD-CFR 滤波器, 但提升效果不明显.

在特征提取层, 研究者们致力于减少隐写信号的损失, 微调网络结构来增强隐写特征的提取能力. 由于平均池化会抑制隐写信号, J-XuNet[17]和 YangNet[18]通过步长为 2 的卷积层代替平均池化. Weng 等人[19]提出的 LWENet 使用多视点全局池化 (MGP) 来生成多视点特征, 取代了传统的全局平均池化. Zhang 等人[20]针对内容自适应隐写算法倾向修改纹理信息丰富区域的像素值, 微调网络结构, 提出基于空间和通道多重注意力机制的方法增强隐写特征的提取. Liu 等人[21]提出注意力下采样模块 (ADM) 帮助特征从浅层传递到深处. Luo 等人[22]将 Transformer 与卷积结合, 使得网络提取到全局隐写特征. Yu 等人[23]引入知识蒸馏[24]提出 AMSRKD 网络, 解决隐写分析在现实场景下有效载荷不匹配的问题, 增强网络对隐写特征的提取能力.

在分类层, 研究者们主要通过对提取到的隐写特征进行分类, 一般使用全连接层和交叉熵损失函数. You 等人[7]提出了一种基于孪生网络的隐写分析算法. 该算法将输入图像分为左右两半区域, 通过计算两半区域特征向量的相似性, 根据不相似性来区分载密图像. 它将传统方法中对高频特征的二分类判断, 转变为对图像不同区域的隐写信号差异的检测, 有很大的发展前景. 然而, 基于孪生网络的隐写分析算法在预处理层往往采用通用的 SRM 滤波器来生成残差图, 尽管这种方法能够提取一些隐写特征, 但由于 SRM 滤波器生成的残差图大小单一, 无法充分捕获到复杂多样的隐写信号. 此外, 该算法在特征提取层中堆叠多个卷积块来提取深层次的隐写特征, 但忽略了微弱的隐写信号在层层传递中逐渐减弱甚至消失, 最终影响分类器的准确率.

综上所述, 现有的深度学习图像隐写分析模型大多集中在改进预处理层和特征提取层, 以增强隐写信号的提取. 孪生网络通过对图像不同区域的隐写信号进行检测, 将分类问题转化为相似问题, 展现了较大的发展潜力. 然而, 孪生网络的预处理层捕获的残差图大小单一, 无法捕获更加多样化的隐写特征, 导致对复杂隐写信号的探测能力不足. 同时, 特征提取层中堆叠的卷积块使得微弱的隐写信号在传递过程中易于丢失. 针对这些问题, 我们提出了 ERFS-Net, 进一步提高分类准确率.

2 方法

2.1 ERFS-Net

由于 You 等人[7]提出的孪生网络的预处理层使用 SRM 滤波器捕获的残差图大小单一, 导致无法捕获丰富的残差特征. 同时, 孪生网络的特征提取层中堆叠的卷积块使得微弱的隐写信号在传递过程中易于丢失. 针对以上不足, 本文提出基于增强残差特征的孪生网络图像隐写分析方法 ERFS-Net. 它由孪生子网和分类层两个部分组成, 其中孪生子网由预处理层和特征提取层组成. ERFS-Net 模型的总体结构如图 1 所示, 由孪生子网、分类层组成. 输入图片对半分成左右两个图像块, 两个子图像块分别经过同一个共享权重的孪

生子网负责对微弱隐写信号的计算和提取. 孪生子网包括预处理层和特征提取层. 分类层对孪生子网提取出的两个子图像块的隐写特征进行融合并且分类.

ERFS-Net 是在 You 等人[7]的算法基础上改进而来. 训练过程中, 输入一对载体图像和载密图像, 然后将每张图像垂直对半分为左右两半区域. 接着, 将每个对半分后的左右子图像块分别输入到孪生子网中, 以计算和提取隐写信号. 值得注意的是, 两个子图像块使用的孪生子网是同一个网络, 共享相同权重, 即两个子图像块经过统一的预处理层和特征提取层来提取隐写信号特征. 在分类层中, 将提取到的两个图像块的隐写特征分别进行全局平均池化为同一大小尺寸的高级特征向量, 经过两端进行分类. 一端对两个图像块的特征进行融合, 经过全连接层转化为二维特征向量, 随后使用交叉熵分类损失判断待分类图像的类型. 具体地, 用 0 作为载体图像对应的预测标签, 用 1 作为载密图像对应的预测标签; 比较二维特征向量两种分类结果的概率大小, 当分类结果为载体图像的概率较大时, 输出待分类图像的下标索引为 0, 表示待分类图像为载体图像; 当分类结果为载密图像的概率较大时, 输出待分类图像的下标索引为 1, 表示待分类图像为载密图像; 另一端添加相似损失函数, 对输出的两个子图像块隐写特征进行相似性计算. 隐写信号在区分载体和载密图像中扮演关键角色, 载密图像的高级特征向量包含的隐写信号越多, 就越容易使得类内距离减小, 从而降低类内不同载体图像语义差异的干扰程度. 该相似损失函数引用对比学习, 驱使模型减小对同类载体图像之间的语义差异, 增加对图像不同子区域之间的隐写信号差异的关注. 如果图像没有被嵌入隐写信号, 那么图像子区域之间相似性较高, 反之, 则相似性较低.

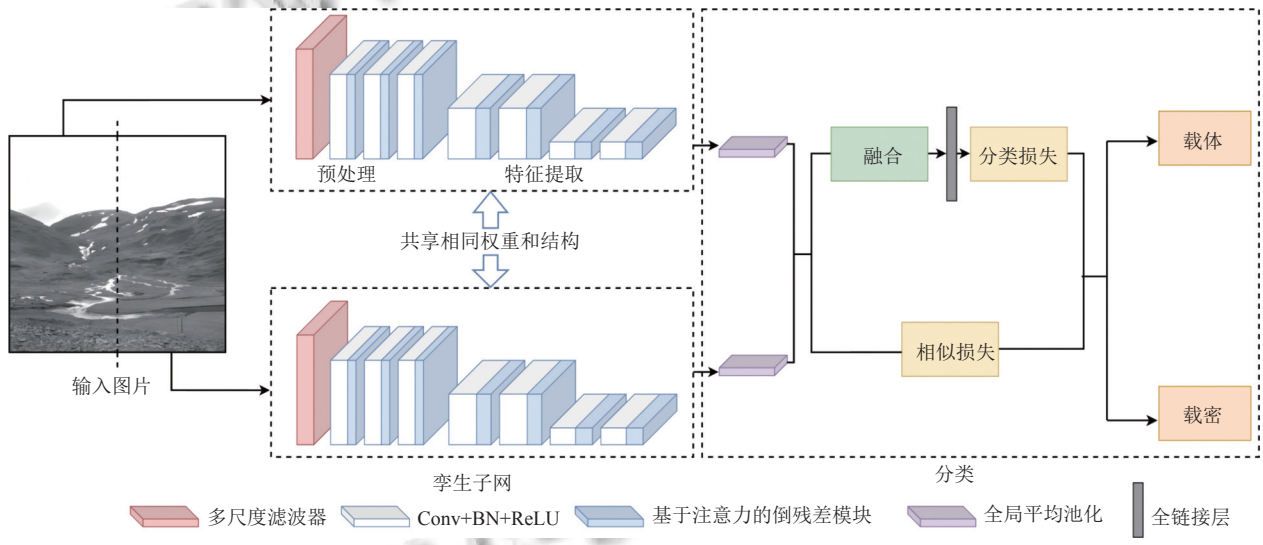


图 1 ERFS-Net 的组成结构

2.1.1 孪生子网

孪生子网负责对输入图像左右对半切的两个子图像块进行隐写信号的计算和提取. 它由预处理层和特征提取层组成, 其具体结构如图 2 所示. 预处理层由多尺度滤波器、A 模块和基于注意力的倒残差模块组成. 多尺度滤波器由 17 个 3×3 卷积核和 13 个 5×5 卷积核组成, 它会对输入的每个图像块进行噪声残差计算, 抑制图像内容, 并得到 30 个特征图. 随后依次经过 3 个残差 A 模块, 全面获取每个图像块的隐写特征. 其中每个残差 A 模块后都添加了一个基于注意力的倒残差模块. 基于注意力的倒残差模块用来减少隐写信号深层传递的消失. 图 2 中 Conv3-30 代表 3×3 卷积, 通道数 30, 默认步长为 1; Conv3-64 代表 3×3 卷积, 通道数 64, 步长为 2; SE-IR-30 代表通道数为 30 的基于注意力的倒残差模块; 后续卷积块表示意义也如此.

特征提取层由两个 B 模块和 A 模块相互交替组成, 并且每个 A、B 模块都增加了一个基于注意力的倒残差模块, 本文第 2.2 节详细介绍了基于注意力的倒残差模块的结构. A 模块与 B 模块的具体结构如图 3 所示. A 模块组成如下, 用一个 3×3 的卷积核对残差图的数据进行处理后, 用 BN 算法进行归一化处理, 再用

ReLU 激活函数激活数据; 紧接着用另一个 3×3 的卷积核对激活的数据进行处理, 再用 BN 算法进行归一化, 最后使用残差连接输入数据和最后的输出使用 ReLU 函数后进行输出. B 模块与 A 模块的结构类似, 但第 1 个卷积块使用的是步长为 2 的 3×3 卷积核. 此外, 在 B 模块的残差连接处还增加了一个步长为 2 的 1×1 的卷积块, 缩小图像尺寸并且增加通道数丰富残差特征. A、B 模块同时采用残差连接, 缓解梯度消失. 网络使用 A、B 模块交替深层提取隐写特征, 并且每个 A、B 模块间都添加了一个基于注意力的倒残差模块用来尽可能减少图像特征下采样过程中微量隐写信号的消失, 通过基于注意力的倒残差模块重用图像特征, 扩展通道, 同时引入注意力机制, 增强网络对图像纹理复杂区域的特征图赋予更多权重.

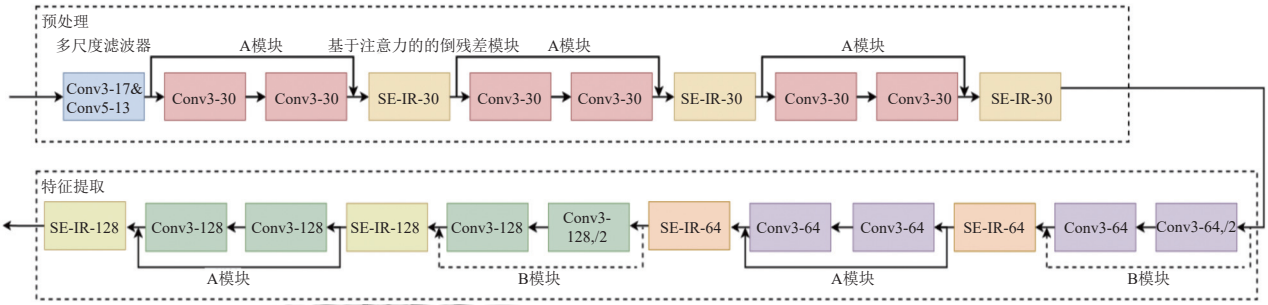


图 2 孪生子网的组成结构

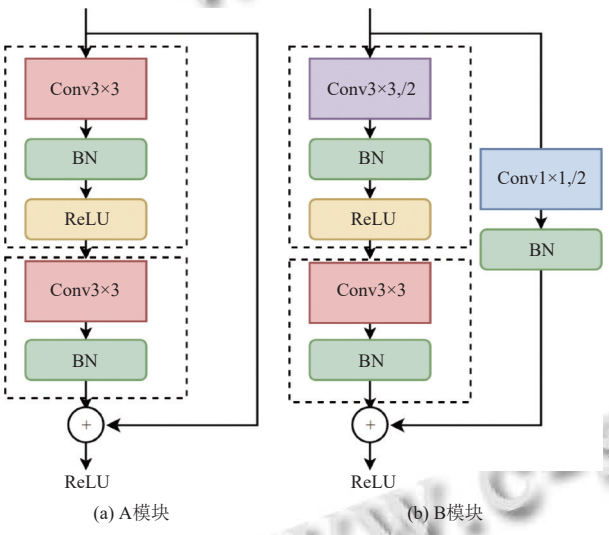


图 3 A 模块和 B 模块的组成结构

2.1.2 分类层

分类层对孪生子网输出的两个子图像块的隐写特征进行分类. 首先, 我们将两个子图像块的特征向量分别经过全局平均池化变为同样尺寸大小, 其中一端将两个子图像块的特征向量按照 SID[25]的方法进行融合, 经过全连接层和交叉熵损失分类. 另一端引入相似损失函数, 进行两个子图像块间隐写特征的相似性计算, 促进网络缩小类内距离, 扩大类间距离, 增强准确率. ERFS-Net 孪生子网的数据传输流程如表 1 所示. 每个图像块经过孪生子网后都会输出一个 128 维度的特征向量. 接下来网络要对这些经过全连接层后的特征向量使用交叉熵分类损失分类, 并且同时使用相似损失计算每个子图像块提取到的隐写特征向量的相似性, 从而提高分类准确率. 交叉熵分类损失函数的定义如式 (1) 所示:

$$L_{\mathrm{CLS}}(p,y)=\begin{cases}-\log(p), & y=1\\ -\log(1-p), & y=0\end{cases} \tag{1}$$

其中, $p\in[0,1]$代表全连接层的输出, $y\in\{0,1\}$代表真实标签值. 载体图像 cover 标签为 0, 载密图像 stego 标签为 1. 交叉熵分类损失通过计算全连接层的输出, 使用概率评估真实标签与预测标签的距离. 然而, 交叉熵分类损失难以在潜在的隐写信号特征中分辨出其他的高频信息和真正的隐写信号. 网络另外添加相似损失函数, 具体定义如式 (2) 所示:

$$\begin{aligned}L_{\mathrm{SML}}=&\frac{1}{2}(1-y)\left\|f_{\mathrm{sub}_i}-f_{\mathrm{sub}_j}\right\|_2^2\\&+\frac{1}{2}y\left[\max\left(0,m-\left\|f_{\mathrm{sub}_i}-f_{\mathrm{sub}_j}\right\|_2\right)\right]^2\end{aligned} \tag{2}$$

其中, $y\in\{0,1\}$代表真实标签值, 0 代表载体图像, 1 代表载密图像. m代表相似损失的边距, f_{sub_i}和f_{sub_j}代表左右子区域的隐写特征向量. 该相似损失函数使用欧氏距离来度量每个特征向量之间的距离, 驱使网络减少对图像不同区域之间内容的差异的关注, 增加对图像不同区域之间高频信号差异的关注. 最终, 我们将相似损失函数与交叉熵分类损失函数进行加权组合, 形

成网络的总损失函数. 如式 (3) 所示, 其中λ代表加权比例.

$$L = L_{\text{cls}}(p,y) + \lambda L_{\text{SML}} \tag{3}$$

表 1 ERFS-Net 孪生子网结构参数

序号	模块	输出大小	重复次数	输出通道
0	输入图像	128×128	—	1
1	多尺度滤波器	128×128	1	30
2	A模块	128×128	3	30
3	基于注意力的倒残差	128×128	3	30
4	B模块	64×64	1	64
5	基于注意力的倒残差	64×64	2	64
6	A模块	64×64	1	64
7	B模块	32×32	1	128
8	基于注意力的倒残差	32×32	2	128
9	A模块	32×32	1	128
10	全局平均池化	1×1	1	128
11	全连接层	—	1	2

2.2 基于注意力的倒残差模块

由于微弱的隐写信号从浅层传递到深层容易消失, 因此本文提出基于注意力的倒残差模块. 它通过在倒残差模块中加入 SE 注意力机制[10], 具体的结构如图 4 所示. 倒残差模块[9]与标准的残差块结构不同, 两者区别如图 5 所示. 在标准残差块结构中, 先使用 1×1 卷积实现降维, 再通过 3×3 标准卷积提取特征, 最后使用 1×1 卷积实现升维. 标准残差是一个两头大, 中间小的沙漏型结构. 但在倒残差结构中, 先使用 1×1 卷积实现升维, t 为扩展因子, 再通过 3×3 的 DW 卷积 (逐通道卷积) 提取特征, 最后使用 1×1 卷积降维恢复原来通道数. 倒残差调换了降维和升维的顺序, 并将 3×3 标准卷积替换为 DW 卷积, 呈现两头小, 中间大的梭形结构. 这种结构有助于弥补隐写信号传递过程中的消失, 扩展通道, 重用图像特征.

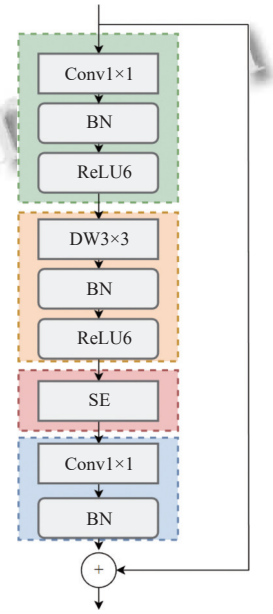
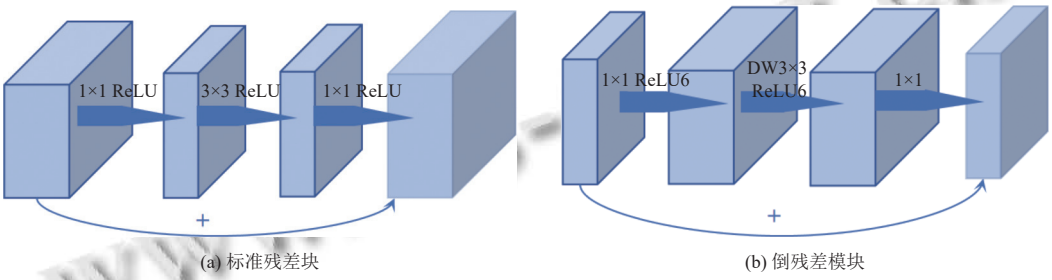


图 4 基于注意力的倒残差模块



(a) 标准残差块 (b) 倒残差模块

图 5 标准残差块和倒残差模块结构

由于孪生网络在提取隐写特征过程中生成的特征图都是给予相同重视, 忽视了自适应隐写算法的特性. 由于自适应隐写算法倾向于将隐写信号嵌入在图像纹理复杂区域, 因此本文在倒残差模块中引入 SE 注意力机制, 使得网络在提取隐写特征的过程中, 对隐藏在图像纹理和边缘区域的特征图赋予更多权重. SE 注意力模块如图 6 所示, 通过压缩、激励两个操作分配不同通道的权重值, 使得网络可以更多关注纹理复杂区域的隐写信号, 丰富残差图特征.

2.3 多尺度滤波器

当前孪生网络沿用绝大部分深度学习隐写分析器使用的 SRM 滤波器[8], 对第 1 层卷积核使用人工设计的权重进行初始化, 使得训练初期网络就具备了基本的抑制图像内容, 捕获隐写信号能力. 然而孪生网络经过 SRM 滤波后的残差图大小单一, 使得捕获的残差特征有限. 为了全面的丰富残差特征, 本文设计了多尺度滤

波器, 结合多个尺寸不同大小的卷积核丰富残差图特征. 滤波器是隐写分析的重要成分, 多样化的高通滤波器, 能够从不同的方向分析图像, 让网络增强残差特征, 提高隐写分析的性能. 现阶段深度学习隐写分析器仍然会使用 SRM 中 30 个不同计算范围、不同方向的高通滤波器对卷积核进行初始化. 这 30 个高通滤波器对应 7 种残差类型, 如图 7 所示. 每个类可通过旋转得到不同方向的滤波器, 其中 (a) 类包含 8 个滤波器, (b) 类包含 4 个滤波器, (c) 类包含 8 个滤波器, (d) 类包含 4 个滤波器, (e) 类包含 1 个滤波器, (f) 类包含 4 个滤波器, (g) 类包含 1 个滤波器. 这 30 个高通滤波器探究了中心元素与周围不同位置, 数量元素之间的关系.

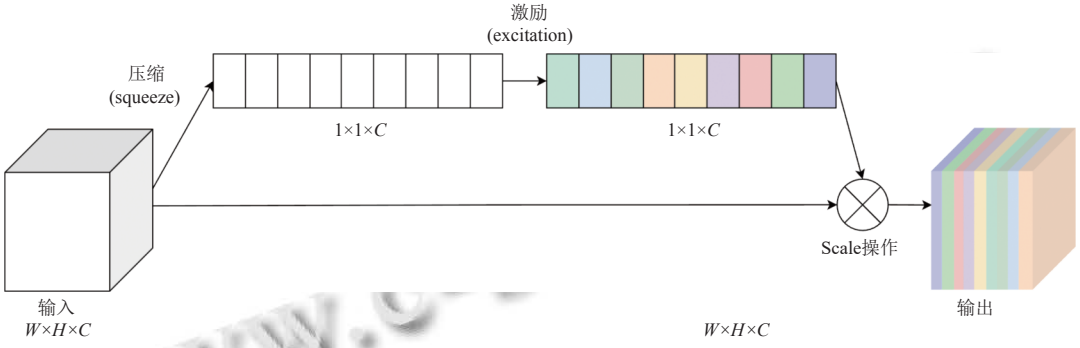


图 6 SE (squeeze-and-excitation) 注意力模块

图 7 SRM 的 30 个高通滤波器的 7 种残差类型

然而经过训练, SRM 中的 30 个滤波器除了有效残差类型值其余权值都是补 0 填充成 5×5 卷积核, 即 30 个滤波器都是探究中心元素与 5×5 区域的元素依赖关系. 残差图像包含的信息单一, 为了丰富残差图的多样性, 我们重新设计了补 0 填充的区域大小. 我们将残差类型中垂直方向的数值个数不大于 3 的填充成 3×3 卷积核, 设计出了多尺度滤波器, 即 17 个 3×3 卷积核与 13 个 5×5 卷积核结合, 如图 8 所示. 多尺度滤波器将不同残差类型调整为多个尺寸为 3×3 卷积核与 5×5 卷积核进行卷积操作, 进行不同深度和感受野大小的提取噪声残差. 具体来说, 5×5 卷积核可以捕获到更大的全局特征, 3×3 卷积核捕获图像的局部特征, 两者结合获得不同尺度的信息进行融合相加, 丰富残差图的类型, 提高分类准确率.

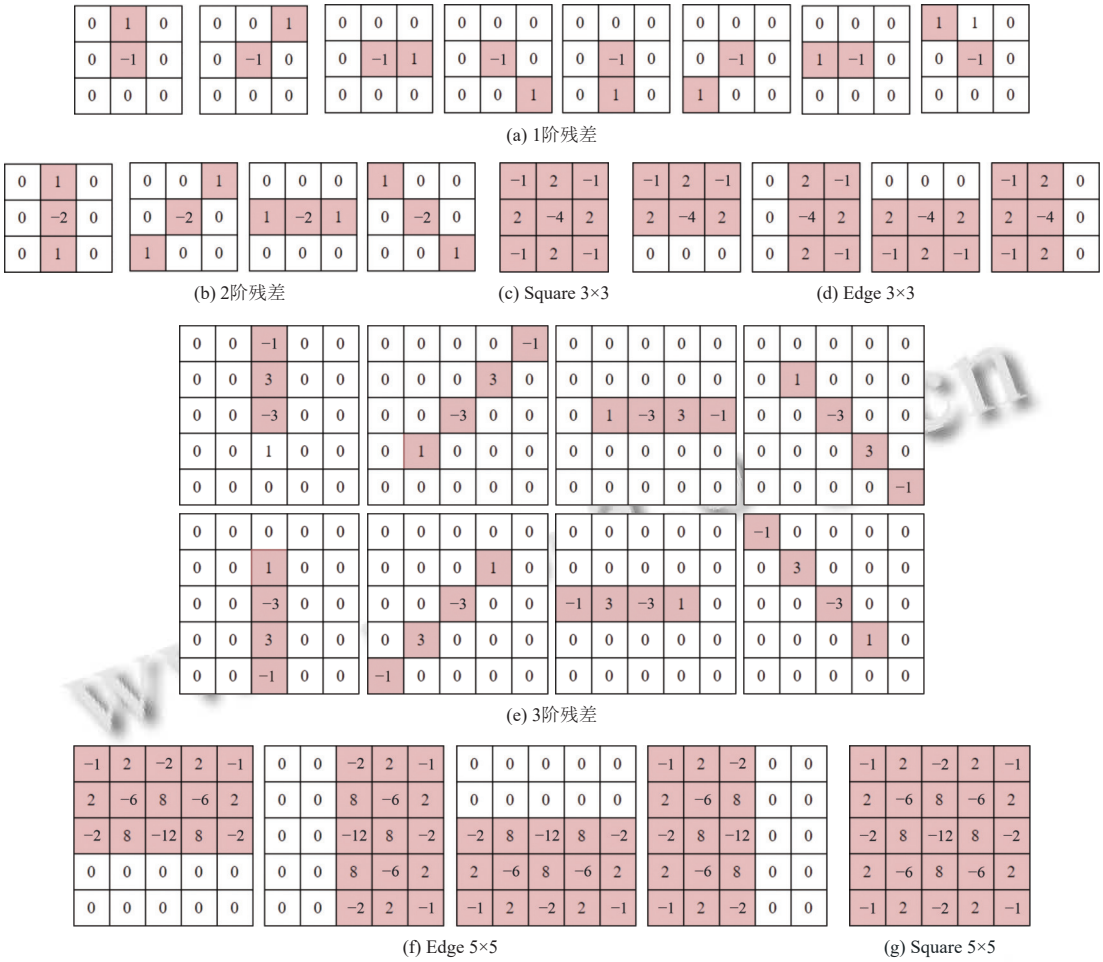


图 8 多尺度滤波器结构

3 实验与分析

3.1 实验设置

ERFS-Net 使用了 Bossbase 1.01[26]数据集进行实验, 该数据集是一个被广泛应用于隐写分析的灰度图像数据集. 它包含 10 000 张 512×512 大小 PGM 格式灰度正常载体图片. 这些图片来源于 7 种型号的相机, 并由相机获取的 RAW 格式图片经过缩放、裁剪等操作以后保存成 PGM 格式得到. 如图 9 所示, 该数据集图像内容涵盖了风景、人物、建筑、动物、生活等不同场景. 图片纹理特性上, 该数据集既含有平滑区域较多的图像也包括了纹理较复杂的图像. 由于处理原始分辨率的计算成本较高, 我们在实验中统一使用 Matlab 中的 *imresize*(·) 函数将图像尺寸统一缩小为原始的一半, 以提高训练速度并减少计算资源的消耗. 目前与绝大部分深度学习隐写分析器一样, 我们没有检测深度学习隐写算法, 因为它仍然处于萌芽阶段. 我们的数据集采用自适应隐写算法 WOW[27]和 S-UNIWARD[28]作为数据集的隐写方法. 对于每种隐写算法, 我们分别设定了不同的有效载荷值, 包括 0.1 bpp、0.2 bpp、0.3 bpp、0.4 bpp. 有效载荷的计算公式如式 (4) 所示:

$$payload = \frac{n_{\text{secret}}}{n_{\text{cover}}} \tag{4}$$

其中, n_{secret} 表示嵌入秘密信息的位数, n_{cover} 表示载体图像的位数. 数据集分为 3 部分, 训练集、验证集和测试集的比例是 6:1:3. 载荷单位为 bpp (bits per pixel).

ERFS-Net 的训练参数设置与孪生网络相同, 使用 $\beta 1 = 0.9$、$\beta 2 = 0.999$ 和 $\varepsilon = 1\times10^{-7}$ 的 Adamax[29]优化器来训练网络, 批量大小设置为 32, 包含 16 个载体-载密对. 损失函数的平衡超参数 λ 设置为 0.1, 超参数 m 为 1.0, 初始学习率设置为 0.001, 采用 MutiStepLR, 学习率在第 300 个 epoch 和第 400 个 epoch 时, 下降 10%. 使用 L2 正则化防止模型过拟合. ERFS-Net 使用 PyTorch 1.7.1, 使用 NVIDIA A30 GPU 训练. 值得注意

的是, 在实验中的训练集、验证集和测试集之间不存在相同的覆盖. 考虑到低嵌入率的数据集收敛困难, 我们采用迁移学习, 网络首先针对 0.4 bpp 有效载荷进行训练, 依次按照“0.4 bpp→0.3 bpp→0.2 bpp→0.1 bpp”的顺序进行训练. 本文使用准确率 (*Acc*), 评估隐写分析器的性能, 如式 (5) 所示:

$$Acc = \frac{P+N}{C+S} \tag{5}$$

其中, C表示载体图像的数量, S 表示载密图像的数量, 预测为载体图像的数量为N, 预测为载密图像的数量为P. 准确率越高, 代表其隐写分析检测器性能越高.

图 9 Bossbase 1.01 数据集图片示例

3.2 对比实验

为了验证本文提出的 ERFS-Net 模型的有效性, 我们选取了几种经典的图像隐写分析模型进行对比. YeNet[2]、SRNet[4]、Zhu-Net[3]、MixNet[30]、FPFNet[31]、Agarwal-Net[32]、Yang-Net[33]、M-CNet[34]等经典模型, 它们都采用了传统的分类网络结构, 并使用交叉熵损失函数进行反馈传播. 此外, 我们还选取了两种在孪生网络 SiaStegNet[7]下进行改进的模型 AGSANet[35]、SiaIRNet[36]进行对比. AGSANet 利用 U-Net 生成残差信息, 并引入注意力机制, 以便网络能够更好地关注图像的纹理区域. 而 SiaIRNet 则采用深度可分离卷积和倒残差结构, 以提取出更丰富有效的残差特征.

为了确保实验结果的公正性, 我们将 AGSANet 和 SiaIRNet 两个网络的训练集、验证集和测试集比例均设置为 6:1:3. 对于 AGSANet, 我们使用了原论文的的实验结果. 对于 SiaIRNet, 鉴于原论文未提供单独使用 Bossbase1.01 数据集进行训练的实验数据, 它的实验结果为本人重新复现得出的. 我们在 Bossbase 1.01 数据集中使用了 S-UNIWARD 和 WOW 两种不同嵌入率的隐写算法, 以测试不同模型的性能. 具体的实验结果如表 2 所示.

根据表 2 的结果, 我们可以明显地看出, 本文提出的隐写分析器 ERFS-Net 在对抗不同嵌入率下的 S-UNIWARD 和 WOW 隐写术中展现出显著的优势. 其分类准确度明显高于其他经典分类网络, 表明隐写分析器需要能够准确地区分载体图像和载密图像, 而仅仅使用经典分类网络中的分类损失是无法实现对细粒度分类的准确性. 相比于基于 SiaStegNet 改进的

AGSANet 和 SiaIRNet, 本文方法的准确性也更高. 特别是在 S-UNIWARD 隐写算法的 0.4 bpp 嵌入率下, 本文方法比 SiaStegNet 基底网络提高了 2.15%. 此外, 对于低嵌入率图像, 由于隐写信号容量较少且更隐蔽, 更难捕获, 本文方法相对于近几年的先进模型也表现出较优的性能.

表 2 在嵌入率为 0.1–0.4 bpp 下, 4 种隐写分析器对抗 S-UNIWARD 和 WOW 隐写术的检测准确率 (%)

隐写分析模型	WOW				S-UNIWARD			
	0.1	0.2	0.3	0.4	0.1	0.2	0.3	0.4
Ye-Net (2017)	56.67	66.90	73.06	76.80	52.84	60.00	65.54	68.30
SRNet (2019)	66.29	75.40	83.46	86.90	56.87	67.40	77.78	81.60
Zhu-Net (2020)	69.87	76.70	85.26	88.20	60.75	71.50	80.37	84.70
MixNet (2022)	70.45	77.55	86.21	88.76	63.95	71.00	80.38	86.33
FPFNet (2023)	74.56	81.53	87.03	89.15	68.75	76.35	84.57	87.55
Agarwal-Net (2024)	71.23	77.18	83.97	88.00	64.36	71.12	79.92	84.64
Yang-Net (2024)	54.90	66.40	75.40	88.10	58.50	69.30	80.30	86.30
M-CNet (2024)	74.45	83.51	88.85	92.41	69.00	79.90	86.25	90.90
SiaStegNet (2021)	76.52	83.69	88.16	90.37	70.44	80.25	86.50	90.13
AGSANet (2022)	77.21	85.64	89.29	90.62	71.51	80.89	87.23	90.22
SiaIRNet (2023)	77.40	85.68	89.32	90.69	71.53	80.92	87.25	90.30
ERFS-Net (Ours)	**80.10**	**87.47**	**91.12**	**92.64**	**75.66**	**84.77**	**89.44**	**92.28**

3.3 消融实验

我们使用 0.4 bpp 的 S-UNIWARD 隐写算法数据集, 比较了在孪生网络 SiaStegNet 基底上添加倒残差模块和多尺度滤波器对本文隐写分析检测器性能的影响, 如表 3 所示.

表 3 0.4 bpp 下, 基于注意力的倒残差模块和多尺度滤波器对模型检测 S-UNIWARD 算法的影响

隐写分析器	准确率 (%)
SiaStegNet	90.13
SiaStegNet＋多尺度滤波器	91.04
SiaStegNet＋倒残差	91.36
SiaStegNet＋基于注意力的倒残差	91.86
ERFS-Net (Ours)	**92.28**

从表 3 的第 2、4 行我们可以看到, 单独在孪生网络基底上分别添加多尺度滤波器和基于注意力的倒残差模网络性能各自都能提高约 1% 的准确率. 因此, 我们能得出结论, 多尺度滤波器确实可以通过利用不同大小的卷积核组合来丰富残差图的多样性, 基于注意力的倒残差模块也可以弥补隐写信号深层传递过程中的消失问题, 两者都能提高网络分类准确率. 从表 3 的第 3、4 行我们可以看到, 若基于注意力的倒残差模块去除 SE 注意力机制, 网络性能反而下降. 综上, 本文联合使用多尺度滤波器和基于注意力的倒残差模块设计的 ERFS-Net 准确率能够达到 92.28% 的优异性能.

3.4 泛化实验

我们针对相同隐写术算法但不同嵌入率的数据集进行了网络训练和测试, 并将结果列于表 4. 同时, 我们还针对使用不同隐写术算法但相同嵌入率的数据集进行了相似实验, 结果见表 5. 需要注意的是, 这两个表中的训练集和测试集之间不存在相同的载体-载密对.

表 4 嵌入率不匹配时, 本文的隐写分析器对抗 S-UNIWARD 不同嵌入率的准确率 (%)

训练集嵌入率 (bpp)	测试集嵌入率 (bpp)			
	0.1	0.2	0.3	0.4
0.4	57.10	73.35	85.95	**92.28**
0.3	64.06	82.51	**89.44**	92.80
0.2	71.23	**84.77**	88.96	91.34
0.1	**75.66**	83.31	86.12	88.58

表 5 隐写算法不匹配时, 本文的隐写分析器对抗 0.4 bpp 数据集的准确率 (%)

隐写术	Test		
	S-UNIWARD	WOW	HUGO
S-UNIWARD	**92.28**	87.05	47.54
WOW	83.31	**92.64**	65.04

在表 4 中, 第 1 列展示了训练集的嵌入率, 第 1 行表示了测试集的嵌入率. 观察实验结果可以发现, 当嵌入率匹配时, 分类器的准确率最高. 此外, 当将高嵌入率模型用于低嵌入率数据集的检测时, 其准确率与嵌入率匹配时相比有明显差距. 相反, 将低嵌入率模型用于高嵌入率数据集的检测时, 其准确率差距较小. 这可能是因为高嵌入率会在低嵌入率的基础上再进行嵌密, 导致两者嵌密区域存在重合. 因此, 低嵌入率模型具有更好的泛化性能.

而在表 5 中, 第 1 列表示训练时所采用的隐写算法, 第 1 行表示测试时所使用的隐写算法. 对于 S-

UNIWARD 与 WOW 隐写算法, 当隐写算法匹配时, 分类器的准确率最高. 然而, 在隐写算法不匹配的情况下, 训练于 S-UNIWARD 和 WOW 数据集的模型相互检测时, 其准确率会在一定范围内下降. 值得注意的是, 在检测 HUGO 隐写算法时, 分类器效果很差. 这可能是由于 HUGO 的失真函数与 S-UNIWARD 完全不同, 因此需要重新训练适应于不同失真函数的隐写算法模型.

4 结语

本文提出了基于增强残差特征的孪生网络图像隐写分析方法 ERFS-Net. 相较于以往的模型, ERFS-Net 模型以孪生网络为基底, 设计了基于注意力的倒残差模块和多尺度滤波器, 以提升隐写分析的效果. 基于注意力的倒残差模块引入倒残差重用图像特征, 减少隐写信号在深层传递中的消失, 并且使用 SE 注意力机制, 使得网络更关注纹理复杂区域的特征图, 增强残差特征. 多尺度滤波器将残差类型调整为多个尺寸不同的卷积核结合进行卷积操作, 丰富了残差特征, 解决了残差图大小单一的问题. 大量的实验证明, 相较于现有方法, ERFS-Net 在 Bossbase1.01 数据集下, 尤其在 S-UNIWARD 和 WOW 隐写算法下, 具有显著的优势.

参考文献

1 Johnson NF, Jajodia S. Exploring steganography: Seeing the unseen. Computer, 1998, 31(2): 26–34. [doi: 10.1109/MC.1998.4655281]

2 Ye J, Ni JQ, Yi Y. Deep learning hierarchical representations for image steganalysis. IEEE Transactions on Information Forensics and Security, 2017, 12(11): 2545–2557. [doi: 10.1109/TIFS.2017.2710946]

3 Xu GS, Wu HZ, Shi YQ. Structural design of convolutional neural networks for steganalysis. IEEE Signal Processing Letters, 2016, 23(5): 708–712. [doi: 10.1109/LSP.2016.2548421]

4 Yedroudj M, Comby F, Chaumont M. Yedroudj-Net: An efficient CNN for spatial steganalysis. Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Calgary: IEEE, 2018. 2092–2096.

5 Boroumand M, Chen M, Fridrich J. Deep residual network for steganalysis of digital images. IEEE Transactions on Information Forensics and Security, 2019, 14(5): 1181–1193. [doi: 10.1109/TIFS.2018.2871749]

6 Zhang R, Zhu F, Liu JY, *et al*. Depth-wise separable convolutions and multi-level pooling for an efficient spatial CNN-based steganalysis. IEEE Transactions on Information Forensics and Security, 2020, 15: 1138–1150. [doi: 10.1109/TIFS.2019.2936913]

7 You WK, Zhang H, Zhao XF. A Siamese CNN for image steganalysis. IEEE Transactions on Information Forensics and Security, 2021, 16: 291–306. [doi: 10.1109/TIFS.2020.3013204]

8 Fridrich J, Kodovsky J. Rich models for steganalysis of digital images. IEEE Transactions on Information Forensics and Security, 2012, 7(3): 868–882. [doi: 10.1109/TIFS.2012.2190402]

9 Sandler M, Howard A, Zhu ML, *et al*. MobileNetV2: Inverted residuals and linear bottlenecks. Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 4510–4520.

10 Hu J, Shen L, Sun G. Squeeze-and-excitation networks. Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 7132–7141.

11 Farid H. Detecting hidden messages using higher-order statistical models. Proceedings of the 2002 International Conference on Image Processing. Rochester: IEEE, 2002. II-905–II-908.

12 Holub V, Fridrich J. Random projections of residuals for digital image steganalysis. IEEE Transactions on Information Forensics and Security, 2013, 8(12): 1996–2006. [doi: 10.1109/TIFS.2013.2286682]

13 Song XF, Liu FL, Yang CF, *et al*. Steganalysis of adaptive JPEG steganography using 2D Gabor filters. Proceedings of the 3rd ACM Workshop on Information Hiding and Multimedia Security. Portland: Association for Computing Machinery, 2015. 15–23.

14 Yu J, Li FY, Cheng H, *et al*. Spatial steganalysis using contrast of residuals. IEEE Signal Processing Letters, 2016, 23(7): 989–992. [doi: 10.1109/LSP.2016.2575100]

15 Su AT, He XL, Zhao XF. JPEG steganalysis based on ResNeXt with Gauss partial derivative filters. Multimedia Tools and Applications, 2021, 80(3): 3349–3366. [doi: 10.1007/s11042-020-09350-2]

16 Luo WW, Dang JW, Wang WR, *et al*. Low-complexity JPEG steganalysis via filters optimation from symmetric property. Multimedia Systems, 2021, 27(3): 371–377. [doi: 10.1007/s00530-021-00780-y]

17 Xu GS. Deep convolutional neural network to detect J-UNIWARD. Proceedings of the 5th ACM Workshop on Information Hiding and Multimedia Security. Philadelphia: Association for Computing Machinery, 2017. 67–73.
18 Yang JH, Kang XG, Wong EK, *et al*. Deep learning with feature reuse for JPEG image steganalysis. Proceedings of the 2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). Honolulu: IEEE, 2018. 533–538.
19 Weng SW, Chen MF, Yu LF, *et al*. Lightweight and effective deep image steganalysis network. IEEE Signal Processing Letters, 2022, 29: 1888–1892. [doi: 10.1109/LSP.2022.3201727]
20 Zhang XB, Zhang XP, Feng GR. Image steganalysis network based on dual-attention mechanism. IEEE Signal Processing Letters, 2023, 30: 1287–1291. [doi: 10.1109/LSP.2023.3313517]
21 Liu JH, Jiao G, Sun XY. Feature passing learning for image steganalysis. IEEE Signal Processing Letters, 2022, 29: 2233–2237. [doi: 10.1109/LSP.2022.3217444]
22 Luo G, Wei P, Zhu SW, *et al*. Image steganalysis with convolutional vision transformer. Proceedings of the 2022 IEEE International Conference on Acoustics, Speech and Signal Processing. Singapore: IEEE, 2022. 3089–3093.
23 Yu LF, Li YW, Weng SW, *et al*. Adaptive multi-teacher softened relational knowledge distillation framework for payload mismatch in image steganalysis. Journal of Visual Communication and Image Representation, 2023, 95: 103900. [doi: 10.1016/j.jvcir.2023.103900]
24 Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network. arXiv:1503.02531, 2015.
25 Tsang CF, Fridrich J. Steganalyzing images of arbitrary size with CNNs. Electronic Imaging, 2018, 30(7): 121-1–121-8.
26 Bas P, Filler T, Pevný T. “Break our steganographic system”: The ins and outs of organizing BOSS. Proceedings of the 13th International Conference on Information Hiding. Prague: Springer, 2011. 59–70.
27 Holub V, Fridrich J. Digital image steganography using universal distortion. Proceedings of the 1st ACM Workshop on Information Hiding and Multimedia Security. Montpellier: ACM, 2013. 59–68.
28 Holub V, Fridrich J. Designing steganographic distortion using directional filters. Proceedings of the 2012 IEEE International Workshop on Information Forensics and Security (WIFS). Costa Adeje: IEEE, 2012. 234–239.
29 Kingma DP, Ba J. Adam: A method for stochastic optimization. arXiv:1412.6980, 2017.
30 Amrutha E, Arivazhagan S, Sylvia Lilly Jebarani W. MixNet: A robust mixture of convolutional neural networks as feature extractors to detect stego images created by content-adaptive steganography. Neural Processing Letters, 2022, 54(2): 853–870. [doi: 10.1007/s11063-021-10661-0]
31 Li JT, Wang XD, Song YF, *et al*. FPFnet: Image steganalysis model based on adaptive residual extraction and feature pyramid fusion. Multimedia Tools and Applications, 2024, 83(16): 48539–48561.
32 Agarwal S, Jung KH. Digital image steganalysis using entropy driven deep neural network. Journal of Information Security and Applications, 2024, 84: 103799. [doi: 10.1016/j.jisa.2024.103799]
33 Yang SC, Jia XX, Zou FH, *et al*. A novel hybrid network model for image steganalysis. Journal of Visual Communication and Image Representation, 2024, 103: 104251. [doi: 10.1016/j.jvcir.2024.104251]
34 Singh B, Sur A, Mitra P. Multi-contextual design of convolutional neural network for steganalysis. Multimedia Tools and Applications, 2024, 83(32): 77247–77265. [doi: 10.1007/s11042-024-18545-w]
35 Gan ZH, Cheng XH, Pang ZL, *et al*. Highly accurate end-to-end image steganalysis based on auxiliary information and attention mechanism. Journal of Electronic Imaging, 2022, 31(6): 063003.
36 Li H, Wang JW, Xiong N, *et al*. A Siamese inverted residuals network image steganalysis scheme based on deep learning. ACM Transactions on Multimedia Computing, Communications and Applications, 2023, 19(6): 214.

(校对责编: 王欣欣)