基于孪生网络和交叉注意力机制的空域和JPEG图像隐写分析

张倩倩[1),2),3)] 李 浩[1)] 张 祎[1)] 马媛媛[2),3)] 罗向阳[1)]

[1)](信息工程大学河南省网络空间态势感知重点实验室 郑州 450001)

[2)](河南师范大学计算机与信息工程学院 河南 新乡 453007)

[3)](河南师范大学河南省教育人工智能与个性化学习重点实验室 河南 新乡 453007)

摘 要 近年来,深度学习在图像隐写分析任务中表现出了优越的性能。然而,此类方法在捕获图像中微弱的隐写噪声时,往往会因下采样过程中大量关键细节信息的丢失,导致在检测空域和JPEG隐写图像时难以同时实现高检测准确率。为此,本文基于孪生神经网络对图像进行分区域细粒度学习,同时利用交叉注意力机制进一步增强模型全局信息感知能力,提出一种跨通道交叉注意力增强的隐写分析方法(CES-Net)。首先,采用孪生神经网络作为主干网对图像进行分区域学习,以细致地感知空域和JPEG图像的像素信息和微弱的隐写噪声,同时,设计了多样化的高通滤波器和多层卷积作为网络预处理层来获取丰富且高质量的隐写噪声残差;接着,改进了特征提取部分,提出了跨通道交叉注意力网络,使模型提取到更多因隐写嵌入对图像像素相关性造成扰动的隐写特征,用于基于秘密噪声残差等弱信息的隐写图像分类任务;最后,融合子网络学习到的不同区域图像的分类特征,并输入全连接层组成的分类模块对载体和载密图像进行分类,提升检测效果。在隐写和隐写分析领域常用的图像数据集BOSSBase-1.01和BOWs2上进行了大量实验,结果表明,CES-Net方法与现有方法相比,对于空域和JPEG图像的多种主流隐写算法均能达到目前最优的检测准确率,其中,对多种空域隐写算法(WOW、S-UNIWARD和HILL)在不同嵌入比率下生成的载密图像,检测准确率最高分别提升1.27%~25.61%、2.1%~21.73%和1.69%~23.46%;对JPEG图像自适应隐写算法J-UNIWARD在不同嵌入比率下生成的载密图像,CES-Net方法对两种质量因子(QF = 75和QF = 85)的JPEG图像隐写检测准确率最高分别提升2.34%和2.06%。

关键词 隐写分析;隐写;孪生网络;交叉注意力机制;信息隐藏

中图法分类号 TP391 **DOI号** 10.11897/SP.J.1016.2025.01305

Siamese Network and Cross-Attention for Spatial and JPEG Image Steganalysis

ZHANG Qian-Qian[1),2),3)] LI Hao[1)] ZHANG Yi[1)] MA Yuan-Yuan[2),3)] LUO Xiang-Yang[1)]

[1)](Key Laboratory of Cyberspace Situation Awareness of Henan Province, Information Engineering University, Zhengzhou 450001)

[2)](Department of Computer and Information Engineering, Henan Normal University, Xinxiang, Henan 453007)

[3)](Engineering Lab of Intelligence Business & Internet of Things, Henan Normal University, Xinxiang, Henan 453007)

Abstract In recent years, deep learning has demonstrated excellent performance in image steganalysis. The deep learning-based steganalysis method constructs an end-to-end network through a data-driven approach for classification tasks, thereby significantly reducing human intervention and achieving outstanding detection performance. However, these methods

收稿日期:2024-10-30;在线发布日期:2025-04-15。本课题得到河南省优秀青年科学基金(252300421233, 222300420058)、国家自然科学基金(U23A20305,62172435,62202495)、国家重点研发计划(2022YFB3102900)、中原学者项目(254000510007)、河南省重点研发专项基金(No. 221111321200)资助。 **张倩倩**,博士研究生,实验师,主要研究领域为信息隐藏与检测、粒计算。E-mail:zhangqianqian@htu.edu.cn。**李 浩**,博士,讲师,主要研究领域为隐写分析技术。**张 祎**,博士,讲师,主要研究领域为图像隐写与分析技术。**马媛媛**,博士,副教授,中国计算机学会(CCF)会员,主要研究领域为图像隐写分析、粒计算。**罗向阳**(通信作者),博士,教授,主要研究领域为图像隐写与隐写分析。E-mail:luoxy_ieu@sina.com。

frequently encounter challenges in achieving high detection accuracy when capturing subtle steganography noise in images. This is primarily due to the substantial loss of critical details during the down-sampling process, which significantly hinders the detection of both spatial-domain and JPEG image steganography. For this reason, this paper presents a comprehensive analysis of the mechanism through which different modules of the steganalysis network effectively capture steganography noise residuals, and proposes CES-Net, a novel steganalysis method. CES-Net leverages Siamese neural networks for fine-grained regional learning of images and incorporates a cross-attention mechanism to further enhance the model′s ability to perceive global information. Firstly, we employ a Siamese neural network as the backbone for regional learning on images, enabling precise perception of pixel information and subtle steganography noise in spatial-domain and JPEG-compressed images. This not only aids in uncovering hidden information within the image but also markedly enhances the model′s sensitivity to specific patterns or features, thereby leading to more accurate and reliable results. In the pre-processing module, we design diverse high-pass filters and multi-layer convolutions to enhance the model′s ability in extracting steganography noise residuals from different regions of the image. This provides abundant and high-quality steganography noise residuals for the subsequent network′s learning process. Secondly, we propose incorporating a cross-channel cross-attention network into the feature extraction module. This enhances the model′s ability to extract steganography features that have been disrupted by embedding, thereby enabling accurate classification of stego images based on subtle information such as secret noise residuals. In this module, CES-Net introduces a novel block that leverages the cross-attention mechanism. This block can effectively capture both intra-channel and inter-channel correlations in residual maps, thereby enhancing the representational capacity of steganography noise features and making the extracted features more discriminative. Finally, at the end of the network, the fine-grained classification features of images in different regions, learned by the sub-networks, are fused through a well-designed aggregation mechanism. These resulting fused features are subsequently fed into a classification module composed of fully connected layers, which is responsible for classifying both cover and stego images. This process effectively improves the overall detection performance. We carry out extensive experiments on BOSSBase-1.01 and BOWs2, two datasets that are widely utilized in the fields of steganography and steganalysis. The results show that the CES-Net achieves state-of-the-art detection accuracy for various mainstream steganography algorithms in both spatial and JPEG images, outperforming existing methods. Specifically, CES-Net improves the detection accuracy by 1.27% to 25.61%, 2.1% to 21.73%, and 1.69% to 23.46%, respectively, for spatial domain algorithms (WOW, S-UNIWARD, and HILL) across different payload conditions. Furthermore, CES-Net achieves a maximum improvement of 2.34% and 2.06% in detection accuracy for two quality factors (QF = 75 and QF = 85) respectively when detecting J-UNIWARD adaptive steganography in JPEG images.

Keywords steganalysis; steganography; siamese neural network; cross-attention; information hiding

1 引 言

隐写是将秘密信息嵌入图像、音频、视频等数字载体中来实现隐蔽通信的技术[1-2]。隐写术的发展，对保障国家特殊部门和关键人群的通信安全具有重要作用。然而，隐写又是一把“双刃剑”，在保护通信安全的同时，若被某些恶意组织利用，实施非法的隐蔽通信，将会给网络空间安全带来严重威胁[3]。隐写分析是隐写的对抗技术，其通过检测由于隐写引

入的嵌入痕迹，来识别带有秘密信息的载体，达到发现、阻止或破坏基于隐写的恶意隐蔽通信行为的目的。

随着互联网及数字图像处理技术的飞速发展，图像隐写分析技术受到了国内外学者的广泛关注。图像隐写分析包括图像隐写检测、隐写定位及隐写信息提取，其中，图像隐写检测主要是设计并提取有效的图像统计特征，来判断其中是否嵌入秘密信息。传统的隐写检测方法是利用手工提取的隐写分析特征，结合集成分类器来识别载密图像。典型的特征有针对空域图像设计的SRM（Spatial Rich Model）[4]及针对JPEG（Joint Photographic Experts Group）图像设计的DCTR（Discrete Cosine Transform Residual）[5]、GFR（Gabor Filter Residual）[6]等富模型特征，可实现图像内容自适应隐写的可靠检测。然而，传统隐写分析方法的特征提取通常需要大量的先验知识。随着深度学习的发展，研究人员将深度学习技术应用到隐写分析领域，提出了多种深度学习隐写分析网络。深度学习隐写分析方法是以数据驱动的方式构建一个端到端的网络来实现分类，这类方法大大减少了人为参与，取得了显著的检测效果。

针对空域图像，Qian等人提出了基于深度学习的隐写分析网络GNCNN[7]（也称Qian-Net）。虽然Qian-Net的性能略逊于SRM，但作为深度学习在隐写分析领域的探索，Qian-Net的提出为后续基于卷积神经网络的隐写分析方法提供了思路。与Qian-Net几乎同一时期提出的Xu-Net[8]在隐写分析领域知识的基础上，设计了一种用于检测自适应隐写的新型CNN架构，其通过改进不同网络层上的激活函数，使深度学习方法的性能首次超过了SRM。之后，基于深度学习的隐写分析方法蓬勃发展，检测性能不断得到提升。比如针对基于CNNs隐写检测方法中存在的载体失配问题设计的J-Net[9]、RCDD[10]方法；Chen等人为平衡训练时间和检测精度提出的基于特征融合的深度学习隐写分析方法[11]等。与空域图像相比，深度学习JPEG图像隐写检测方法起步相对较晚。2017年，Xu等人在Xu-Net基础上提出了一种针对JPEG图像的隐写分析网络Xu-Net-JPEG[12]，证明深度学习网络也可以在复杂的频域隐写检测上胜于基于特征的隐写分析方法。Chen等人也在Xu-Net的基础上提出了一种带有JPEG相位感知的频域隐写分析网络Vnet[13]，Vnet除了借鉴了DCTR等传统频域隐写分析先验知识外，其在网络框架中添加了JPEG相位感知模块来提升隐写检测精度。Yassine提出使用ImageNet预训练的CNNs来改进JPEG图像隐写分析的性能[14]。

可以看出，大多数基于深度学习的隐写分析网络往往仅针对特定图像格式，检测性能通常难以兼顾空域和JPEG图像隐写。为此，Boroumand等人利用残差网络模拟SRM特征筛选过程，提出了一种更深层的深度学习隐写分析器SRNet[15]。SRNet依赖残差网络本身对于信息的跳跃利用较高的优势，既实现了空域隐写分析，也可有效检测JPEG图像隐写。然而，SRNet纯粹依赖深度学习网络方向拟合的方式，导致模型在训练过程中所需要耗费的时间更长，更容易在训练过程中出现损失不动点的情况。2023年，Li等提出了一种基于多重残差卷积和Transformer结构的混合深度网络隐写分析框架ResFormer[16]，进一步提高了空域和JPEG图像隐写检测的性能。然而，该网络采用的CNN结合Transformer的混合架构，导致参数量的显著增加，过高的计算复杂度可能限制其应用。

针对以上问题，本文提出一种跨通道交叉注意力增强的深度学习隐写分析框架，旨在提升兼顾空域和JPEG图像隐写的通用隐写检测效果。与现有深度学习隐写分析模型不同，所提方法通过扩大高通滤波器的规模，来提升模型提取不同域图像隐写噪声残差的能力，并为后续网络的学习提供更丰富的隐写噪声残差。同时，采用孪生神经网络作为主干网对图像进行分区域学习，使模型更细致地感知不同格式图像的像素信息和隐写噪声，增大载体图像和载密图像的类间差异；另外，网络结构中的交叉注意力模块可进一步学习通道间相关性，增强网络对图像复杂纹理区域等更深层次隐写分析特征的捕捉能力，放大提取到的隐写噪声。本文的主要贡献如下：

（1）首次引入交叉注意力机制到隐写分析任务以增强模型跨通道全局隐写特征捕获能力。为使网络学习到不同格式图像中与嵌入更改相关的噪声残差，构建交叉注意力增强模块来计算通道间的相关性，捕捉残差通道的全局依赖关系，进而增强网络对隐写信号的学习能力。

（2）设计多样性的高通滤波器和多层卷积作为网络预处理层来获得丰富且高质量的噪声残差。鉴于高通滤波器对模型在捕获隐写嵌入上的影响，本文利用更多的高通滤波器来全面地描述图像，获取丰富的高通残差，在此基础上，对残差特征进行多层

卷积进一步扩大获得的噪声残差信号，提高残差表示的质量。

(3)提出通道间特征融合的孪生网络隐写检测方法，可同时检测空域和JPEG图像的多种主流隐写算法。为高效感知空域和JPEG图像中微弱的隐写噪声，采用孪生神经网络作为主干网对图像进行分区域学习，同时利用交叉注意力机制进一步增强模型全局信息感知能力，提高图像隐写检测性能。

本文其余部分组织如下：第2节介绍相关方法和研究现状。第3节详细介绍本文构建的基于孪生网络和交叉注意力机制的隐写分析模型。第4节介绍了所用的数据集和应用细节以及通过对比实验验证了所提模型的有效性。最后，第5节总结全文并给出未来的工作展望。

2 相关工作

隐写检测技术的主要手段之一是提取有效的图像统计特征，以发现与识别隐写噪声带来的扰动，进而判断隐写的存在。如前所述，现有的图像隐写检测方法可分为人工特征模型和深度学习模型两类。在人工特征模型中，特征提取和分类器训练是两个分开的步骤，研究的重点是隐写分析特征的构建。比如，SRM[4]是利用多种线性和非线性高通滤波核对图像进行建模得到相应的残差图像，通过残差图像共生矩阵的组合来获得的高维特征。

随着深度学习技术的发展，研究人员将深度学习应用到图像隐写分析任务并提出了多种深度学习隐写分析模型。深度学习模型以数据驱动的方式构建一个端到端的网络来实现分类，包括预处理、特征提取和分类器训练三个阶段，如图1所示。其中，分类器训练和手工模型基本一致，依赖的都是机器学习方法，即通过训练一个二元分类器来输出结果。因此，深度学习隐写分析方法的研究主要集中在预处理和特征提取两个阶段。

需要说明的是，与其它图像分类任务不同，图像

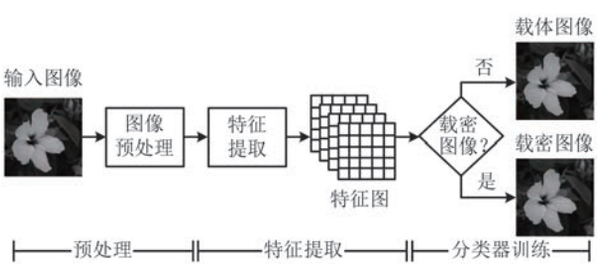


图1 深度学习隐写分析框架

隐写分析任务提取的特征应专注于理解图像中较小的隐写噪声或特征变化，同时需要压制图像内容的影响。因此，深度学习隐写分析模型往往需要对图像进行预处理来获取多样化的残差特征，同时抑制内容对隐写信号的影响，增大载体图像与隐写图像的差异。

2.1 图像预处理方法

在预处理阶段，卷积神经网络大多是在SRM特征设计方法的启发下，设计各种高通滤波器来获取图像残差。可见，传统的图像隐写检测特征设计方法在提升深度学习隐写检测性能方面，仍具有参考借鉴价值。表1列出了现有典型的隐写分析网络中的残差特征滤波核设计。

表1 典型深度学习隐写分析网络预处理阶段滤波核对比

图像类型	Year	Method	残差计算阶段
空域	2015	GNCNN(Qian-Net)[7]	1个滤波核
	2016	Xu-Net[8]	1个滤波核
	2017	Ye-Net[17]	30个滤波核
	2018	Yedroudi-Net[14]	30个滤波核
	2021	SiaStegNet[18]	30个滤波核
	2023	SNRCN2[9]	30个滤波核
频域	2017	Xu-Net-JPEG[12]	16个滤波核
	2017	VNet[12]	4个滤波核
	2018	Zeng-model[19]	25个滤波核
	2022	EWNet[20]	24个滤波核
空域/频域	2019	SRNet[15]	-
	2023	ResFormer[16]	30个滤波核

具体地，针对空域图像，Qian-Net[7]是将1个普通的卷积核替换成固定的高通滤波核，以获取图像的高维残差信息；Xu-Net[8]采用了与Qian-Net相同的预处理滤波核；Ye-Net[17]使用30个5×5大小的滤波核，之后的深度学习隐写分析网络大多采纳最多30个滤波核作为预处理层，比如Yedroudi-Net[14]、SiaStegNet[18]、SNRCN2[9]以及可同时检测空域和JPEG图像隐写的ResFormer[16]等方法。针对JPEG图像，Xu-Net-JPEG[12]使用了16个4×4大小的DCT滤波器进行卷积；VNet[12]沿用了Xu-Net中的预处理层，同时在其基础上额外添加了3个滤波核作为固定的预处理层，因此其滤波核个数为4；Zeng-model[19]使用25个5×5大小的DCT核对图像进行卷积来输出25个残差特征。

上述网络预处理阶段的设计方法有效提高了输入图像的信噪比，使隐写信号在后续的特征提取过程中更容易被检测到。然而，这类方法主要针对特

定格式图像，在检测其他格式图像隐写时，经过预处理后的残差特征图中保留的隐写嵌入更改相对有限，不利于更深层隐写分析特征的提取。不同于上述方法，本文将更多样的高通滤波器和多层卷积块同时作为网络的预处理层，以提高空域和JPEG图像预处理后隐写信号的信噪比，减少信号的损失，提升后续网络提取特征的准确性和可靠性。

2.2　特征提取结构设计

基于CNN的隐写检测方法通常使用各种网络架构自动学习有效的特征，特征的构建主要是依靠卷积网络的反馈学习自动完成。因此，在特征提取阶段，深度学习隐写分析方法主要通过设计不同的网络结构来实现并提升隐写检测的性能。

早期的GNCNN[7]网络结构包括5个卷积层和3个全连接层，模型设计较为简单；Xu-Net[8]的网络框架设计沿用了GNCNN的网络架构；Yedroudj-Net[14]在网络结构上采用了Alex-Net[21]的设计理念；Xu-Net-JPEG[12]采用了20层的全卷积网络，证明了深度网络比宽度网络更容易提取隐写噪声，同时，为防止过深的卷积层造成网络训练过程中出现梯度弥散或者梯度爆炸的情况，采用了与ResNet[22]相同的跳接结构，这在后续的SRNet[15]中也有相应的考虑；VNet[12]是将经过相位分离模块后得到的特征图放入一个线性网络中进行训练；Zeng-model[19]是将预处理后的25个通道残差送入与Xu-Net相似的子网络中分别进行运算；EWNet[20]采用全卷积架构用于任意大小图像的隐写检测；LWENet[23]通过多视图全局池化操作提取多尺度残差特征，同时使用深度可分离卷积减少参数量。

与以一张图像为整体来提取隐写特征的方法不同，SiaStegNet方法[18]采用孪生神经网络架构对图像进行分块独立学习，进一步提升了隐写检测准确率。孪生网络是一种基于两个神经网络建立的耦合构架，基本思想是将数据同时输入到两个完全相同的子网络中，这两个网络共享相同的权重和参数，子网络通常具有深度结构，权重可以通过能量函数或分类损失进行优化，其框架结构如图2所示。

为进一步提高隐写检测性能，近年来，研究者将注意力机制引入隐写分析领域，提出了基于注意力机制的隐写检测方法。比如，2022年，Liu等人提出了基于通道注意力机制与金字塔池化的JPEG图像隐写检测方法CSANet[24]，提升了深度学习隐写检测的性能。2023年，Xie等人[25]提出了基于自注意增强深度残差网络的空域隐写检测方法。同年，

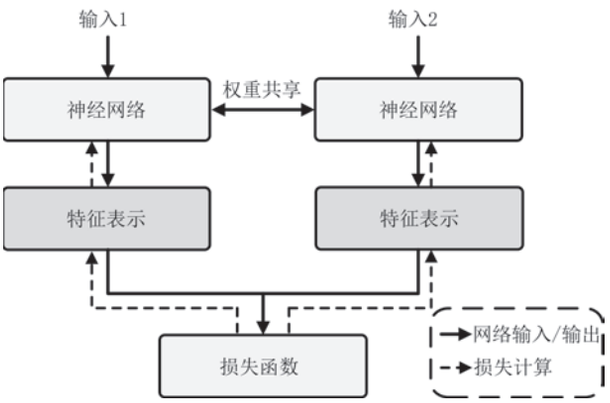


图2　孪生神经网络架构

Wei等人[26]在现有选择信道感知（Selection Channel Aware，SCA）隐写分析方法基础上，引入坐标注意力机制捕获嵌入概率图的关键信息，用以优化现有空域和JPEG图像隐写检测网络，提升其检测性能。

以上方法推动了图像隐写检测技术的发展，其中，基于孪生网络的方法提供了一种细粒度图像隐写检测框架。然而，该方法的侧重点是图像尺寸的多样性，其中设计的捕捉多尺寸空域图像隐写信号的特征提取模块，可能导致对JPEG图像的隐写噪声关注不足。现有基于注意力机制的方法在一定程度上提高了检测性能，然而，坐标注意力机制捕获的是嵌入概率图的关键信息，这通常需要隐写的先验知识；而基于自注意力机制的方法则更多关注的是共同信息，忽视了差异信息的提取和利用，仍会造成特征丢失。

本文借鉴了孪生网络方法的设计架构，但与现有方法不同的是，所提方法一方面扩大了高通滤波器的规模和多样性，以提升模型提取不同域图像隐写噪声残差的能力，为后续网络的学习提供更丰富的隐写噪声残差。另一方面，利用交叉注意力进一步学习通道间相关性，增强网络对图像复杂纹理区域等更深层次隐写分析特征的捕捉能力，放大提取到的隐写噪声，最终共同实现空域和JPEG图像的高精度检测。

3　提出的隐写分析方法

深度学习隐写分析的任务是利用深度神经网络强大的表征学习能力自主地提取隐写图像的异常特征，从而识别载密图像。通常情况下，人类往往会将复杂的事物进行分解，来实现更细致的观察与分析。受这一认知过程的启发，在图像隐写分析领域，

我们采用了类似的策略，即对图像进行划分来深入理解各个局部区域的特性。这种方法不仅有助于揭示隐藏于图像中的信息，还能显著增强模型对特定模式或特征的敏感度，获得更为精确的结果。

基于此，本文以孪生网络为主干网对图像进行分区域学习，同时利用交叉注意力机制增强模型全局信息感知能力，提出一种跨通道交叉注意力增强的深度学习隐写分析方法CES-Net，可同时检测空域和JPEG图像隐写。模型结构如图3所示，包括图像预处理、特征提取及分类识别三个阶段。

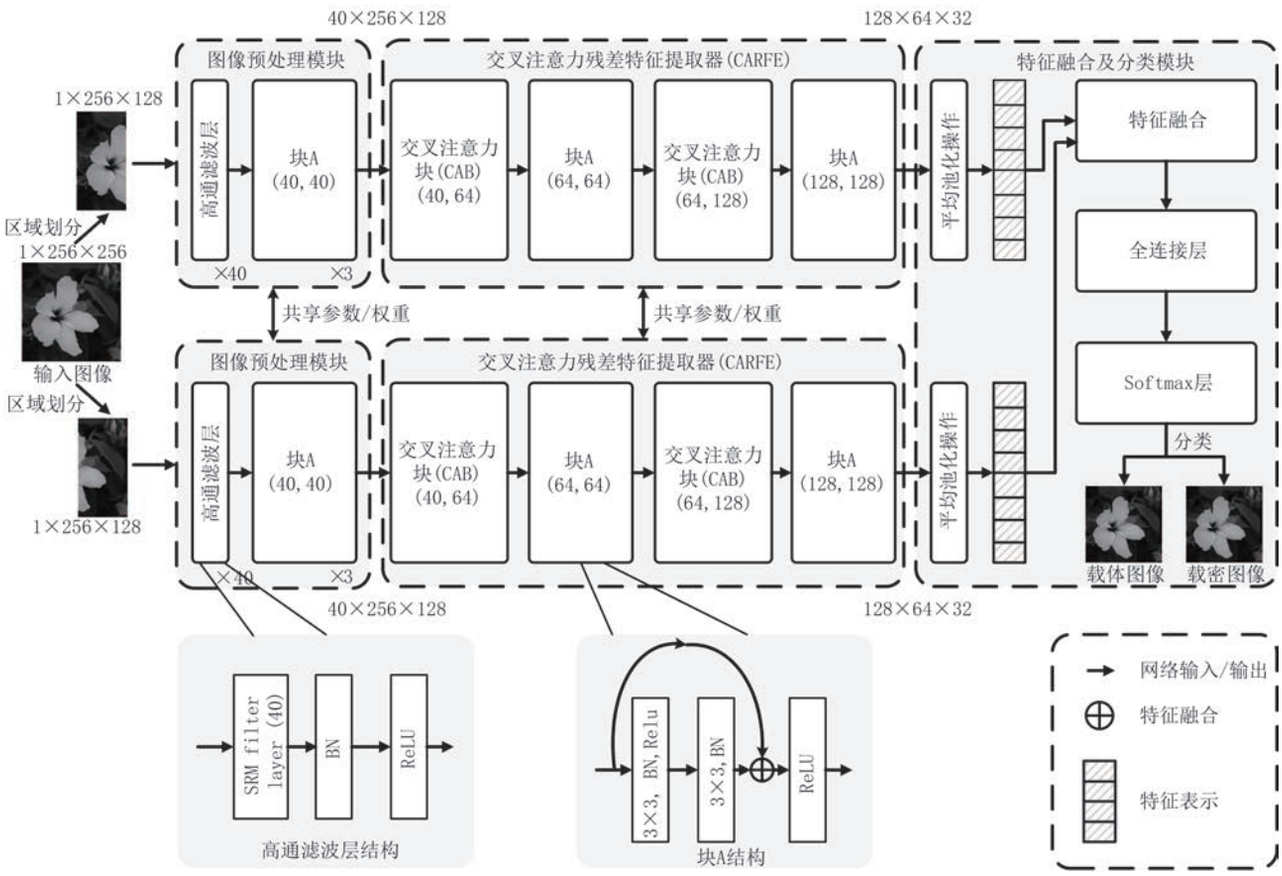


图3 提出的CES-Net方法网络结构图

具体的，给定一张图像$x\in X$，我们按宽度将其分为左、右两个子图x_1和x_2，分别作为孪生网络分支子网络的输入。假设输入图像尺寸为$C\times H\times W$，其中，C表示图像的通道，H和W分别表示图像的长和宽，经过区域划分后子网络输入图像尺寸分别为$C\times H\times(W/2)$。将左、右子图分别输入到2个完全相同的子网络得到子图的表示向量$z(x_i)\in\mathbb{R}^C$，$i\in\{1, 2\}$，这2个子网络结构完全相同并共享权重和参数。每个子网络用于学习子图的特征表示，分别包括图像预处理和残差特征提取2个阶段。

具体的结构和实现细节分别如下。

(1) 图像预处理阶段。预处理的目的是获取图像各个通道的高通残差并保留更多隐写嵌入的修改。在该阶段，本文使用了多样的高通滤波器来获得更丰富的噪声残差，这一操作是在压制图像内容噪声对模型影响的同时，使模型捕获图像中更加多样性的隐写信号，并且有利于后续卷积神经网络学习更丰富的残差特征。

(2) 残差特征提取阶段。该阶段通过构造多层卷积块来学习残差图像中更深层次的隐写分析特征。在该阶段，本文设计了一个交叉注意力块CAB (Cross-Attentional Block)，该Block可同时学习残差通道内和通道间的相关性，丰富残差图中隐写噪声特征的表示能力，使提取的特征更具判别性。

在网络的最后，拼接2个子图向量得到原图的表征$z(x)$，并输入全连接层组成的分类模块，得到图像的预测结果。

下面对所提模型的各部分进行详细介绍。

3.1 图像预处理模块

不同于自然图像分类，隐写分析是一项极具挑

战性的任务，主要原因在于，一方面，图像中嵌入的隐写信号相对于图像内容而言比例极小，如果直接从图像中提取特征，往往会学习到与内容相关的信息；另一方面，隐写痕迹通常是嵌入在图像高频位置上微弱的残差特征，且极易与图像自身的噪声混淆。通常情况下，深度学习隐写分析网络首先对图像进行预处理来提取输入图像的噪声分量，同时压制图像内容的影响。

因此，在本文提出的CES-Net方法前端设置的图像预处理模块，主要用来捕获并增强图像中的隐写噪声残差。该模块由高通滤波层（High-Pass Filtering Layer，HPF）和设计的卷积层块A组成，具体结构如图3所示。其中，高通滤波层用于提取图像中微弱的隐写噪声，增强隐写信号和载体图像间的信噪比；块A用于增强滤波后的噪声残差信号，提高噪声残差表示的质量，其中的BN操作可在一定程度上减少过拟合，加速网络的收敛，ReLU操作有助于缓解梯度消失现象。另外，高通滤波层也可在深度网络中充当正则化项，通过减小可行参数（Feasible Parameter，FP）来实现网络的收敛。

与之前的隐写分析网络类似，在提出网络的HPF层，我们使用固定的SRM高通滤波器对图像进行无填充高通滤波卷积，以使网络学习到不同格式图像中与隐写嵌入更改相关的“噪声残差”，滤波核为5×5的SQUARE核，即

$$K_{SQUARE}=\begin{bmatrix}-1 & 2 & -2 & 2 & -1\\ 2 & -6 & 8 & -6 & 2\\ -2 & 8 & -12 & 8 & -2\\ 2 & -6 & 8 & -6 & 2\\ -1 & 2 & -2 & 2 & -1\end{bmatrix} \tag{1}$$

高通滤波核是一种中心对称结构，在隐写分析模型中，可以利用高通滤波提取图像中某像素点与该像素点周围像素之间的信息差距，使模型可以有效地获取像素之间的共生矩阵。

本文对不同尺寸滤波核进行了相关实验，结果如图4所示。具体为，我们使用J-UNIWARD隐写算法将秘密信息嵌入到图4(a)所示的JPEG格式载体图像中，隐写嵌入率为0.4 bpnzAC。图4(b)～图4(e)分别是不同操作下载体图像和对应载密图像的差异直方图。其中，图4(b)为不执行高通滤波的原始图像差异，图4(c)为使用1阶滤波核进行高通滤波，图4(d)和4(e)的高通滤波核尺寸分别为3×3和5×5。可以看出，5×5滤波可以有效放大载体载密图像间的差异。高通滤波层的结构设计如图3所示，其中包括了40个5×5的SRM高通滤波器，1个批标准化层（Batch Normalization，BN）和1个ReLU层。使用固定的SRM高通滤波器对图像进行无填充高通滤波卷积，可使网络学习到不同格式图像中与隐写嵌入更改相关的“噪声残差”，以从输入图像中提取高频隐写信号，生成可学习的特征图。

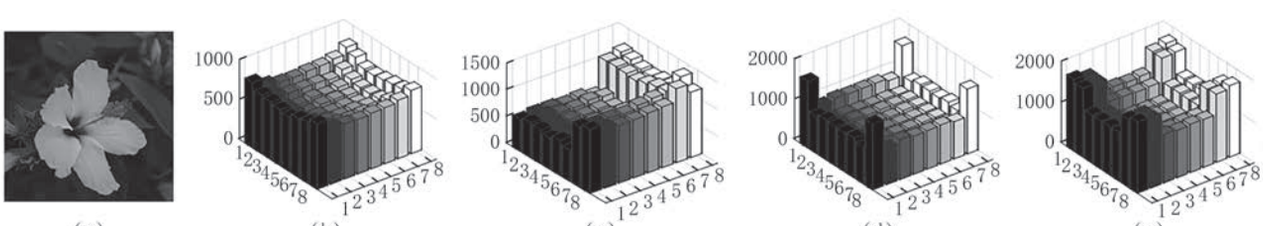


图4　使用J-UNIWARD隐写算法将秘密信息嵌入到载体图像(a)中，(b)为原始载体载密图像差异，(c)～(e)为高通滤波后载体载密图像差异，滤波核分别为1×1、3×3和5×5

然而，与现有隐写分析网络设置不同的是，我们预处理层中的滤波器数量达到了40个，使用多样性的高通滤波器来生成残差图像，以捕获更丰富的隐写分析特征。事实证明，增加滤波器的个数并结合我们设计的网络结构，的确有助于提升检测效果。

此时，对给定图像x使用高通滤波核K_j，$j\in\{1,2,\cdots,40\}$进行卷积可以形式化表示为

$$F_j=x\odot K_j, j\in\{1,2,\cdots,40\} \tag{2}$$

其中，K_j表示第j个滤波核，F_j是第j张噪声残差特征图，$\odot$表示逐元素相乘。

图像经高通滤波的噪声残差图生成过程如图5所示，其本质上是利用高通滤波器对图像像素进行加权求和。高通滤波器从图像的左上角开始计算得到第一个残差，然后根据步长逐行逐列地在图像上滑动，通过将像素值与对应的权重相乘求和获得残差，残差按中心像素的位置依次排列获得残差图。

图像x经过高通滤波层后的噪声残差图$I(x)$可表示为

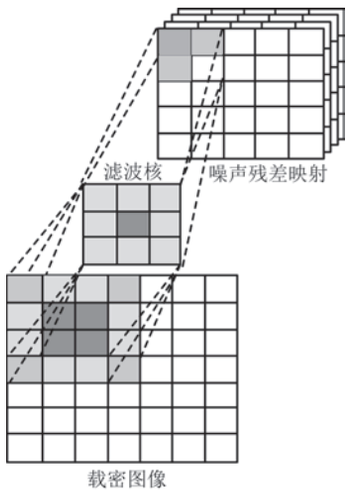


图5 噪声残差图生成过程

$$\boldsymbol{I}(x)=[F_1,F_2,\cdots,F_j,\cdots,F_{39},F_{40}] \tag{3}$$

其中，$\boldsymbol{I}(x)\in\mathbb{R}^{C\times H\times W}$，$C$是通道数，$H$是高度，$W$是宽度；$F_j,j\in\{1,2,\cdots,40\}$表示第$j$张噪声残差特征图，40为特征图的维度。

单纯地提取图像噪声残差并不能作用于所有像素，在残差特征获取过程中往往会掺杂其他特征。为使模型在学习较小的隐写信息残差特征的同时，降低图像中其他特征的影响，本文使用截断操作来限制输入残差特征图的动态范围，这一操作还有利于网络前期的收敛。在我们的方法中，对提取的噪声残差特征使用ReLU激活函数进行限制，即

$$I'=ReLU(I) \tag{4}$$

其中，ReLU函数表示如下：

$$f(x)=\begin{cases}0, & for\ x<0\\ x, & for\ x\geqslant 0\end{cases} \tag{5}$$

由于从卷积操作中学习的特征是多样且互补的，多层卷积操作可以进一步利用通道间的关系来显著扩大噪声残差计算，提高噪声残差表示的质量。因此，除了高通滤波层外，CES-Net方法设计了块A作为图像预处理模块的一部分，用于进一步增强经高通滤波后残差图像中的隐写信号。更多样的高通滤波器和多层卷积块A同时作为网络的预处理层，可显著提高预处理后图像隐写信号的信噪比，提升后续网络提取特征的准确性和可靠性。

卷积块A结构设计如图3所示，其是一个包含2层卷积的残差结构，卷积核尺寸均为3×3，其中，第1层附带BN和ReLU层，第2层仅附带BN层。BN层的作用是使训练数据符合正态分布，同时提升训练时的收敛速度，也可以避免训练时出现梯度弥散或梯度爆炸现象，导致训练结果陷入局部最小值。

在CES-Net网络的预处理模块中使用了3个块A对残差图像进行卷积，每层卷积操作分为两个步骤，首先是在输入的特征图I中采样一个局部感受野R，然后将权重与对应的特征值相乘再求和。经过卷积输出的特征图I'可表示如下：

$$I'(p_0)=\sum_{p_n\in R}w(p_n)\cdot I(p_0+p_n) \tag{6}$$

其中，p_0为输入特征图I和输出特征图I'当前的位置，p_n为局部感受野R上的位置，w为对应的可学习的权重参数。

一般来说，图像的隐写嵌入是对图像像素值的修改，这种修改会破坏图像相邻像素间的相关性，包括局部相关性和全局相关性。从公式(6)可以看出，卷积通常获得的是图像局部特征，除此之外，若网络能够捕获更多隐写破坏的像素间相关性，将对后续特征提取更有利。

在基于富模型的隐写分析方法中，JRM(JPEG Rich Model)特征[27]通过考虑多种JPEG系数块内和块间相关性，尽可能从多个方面获取隐写对系数相关性的扰动，从而提高了JPEG图像隐写的检测正确率。受此启发，本文在特征提取模块中，通过引入交叉注意力机制，来增强模型跨通道全局特征的捕获能力，缓解因卷积操作导致的层与层之间隐写特征信息丢失的问题，提高隐写检测正确率。

3.2 交叉注意力残差特征提取器

特征提取模块用于从训练数据中提取预测信息到目标任务的分布，即学习图像x的残差特征图$I(x)$与标签相关的特征表示。空域图像隐写主要通过修改图像像素的最低有效位来嵌入秘密消息，这一过程会在图像高频区域引入隐写噪声；而JPEG图像隐写则是通过更改图像的DCT系数来嵌入消息，影响的是解压缩后的空域像素值。鉴于此，从图像像素值中提取有效特征能够同时检测空域和JPEG图像隐写。

由于自适应隐写算法的嵌入修改主要集中在图像纹理复杂区域，为了增强网络对深层次隐写特征的捕捉能力，在CES-Net方法的特征提取模块中设计了交叉注意力残差特征提取器(Cross-Attention Residual Feature Extractor，CARFE)。CARFE的核心是交叉注意力块(Cross-Attentional Block，CAB)，利用该Block可进一步捕获残差通道间的相关性，增强预处理模块获得的隐写噪声残差特征。

CAB结构设计如图6所示，具体为：首先是两个卷积层，第1层卷积的卷积核为7×7，步长为2，

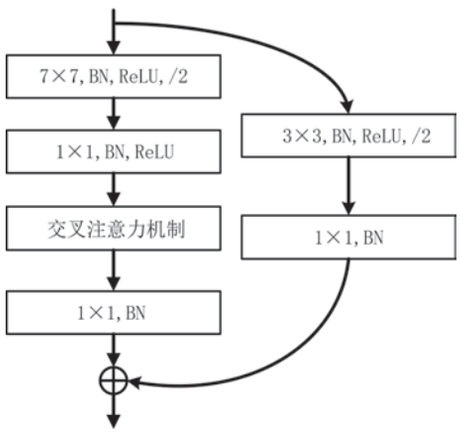


图6　CAB结构图

并附带BN层和ReLU层，卷积方式为无填充卷积，第2层卷积的卷积核为1×1，步长为1，并附带BN层和ReLU层；紧接着是交叉注意力（Cross-Attention，CA）层，该层用来计算残差特征通道间注意力权重，增强网络对隐写信号的捕获能力；最后是1×1的卷积并附带BN层，用于将特征信息集聚，同时防止模型过拟合。此外，为了进一步增强模型的表达能力，CAB中使用了一个包含2层卷积的残差连接，第一层卷积层的卷积核是3×3，步长为2，并附带BN层和ReLU层，第2层是附带BN层的1×1卷积。

下面对CAB的核心组件进行详述。

交叉注意力是自注意力机制的一个扩展，它允许模型在两个不同的序列之间传递信息，从而提升模型对复杂输入的理解和生成能力。交叉注意力机制可表示如下：

$$\boldsymbol{O}=\boldsymbol{W}_o(\mathrm{softmax}(\frac{\boldsymbol{QK}^T}{\sqrt{d_k}})\boldsymbol{V})+\boldsymbol{b}_o \tag{7}$$

其中，$\boldsymbol{O}$是注意力权重的输出，$\boldsymbol{QK}^T$计算的是查询$\boldsymbol{Q}$与键$\boldsymbol{K}\in\mathbb{R}^{n\times d_k}$之间的点积，值$\boldsymbol{V}\in\mathbb{R}^{n\times d_v}$用于计算最终输出时被加权求和，以提取与查询$\boldsymbol{Q}$最相关的信息，$n$为键值序列的长度，$d_q$、$d_k$、$d_v$分别为查询、键和值的维度。$\boldsymbol{W}_o$和$\boldsymbol{b}_o$分别是可学习的权重矩阵和偏置向量，$1/\sqrt{d_k}$用于缩小点积的范围，确保softmax梯度的稳定性。最后的softmax函数对每个查询位置分配一个权重分布。

注意力机制的主要思想是从大量信息中有选择地筛选出少量重要信息并将注意力聚焦于这些重要信息，从而忽略大多不重要的信息。聚焦的过程体现在信息权重系数的计算上，权重越大越聚焦于其对应的value值上，即权重代表了信息的重要性，而value是其对应的信息。

基于CNNs传统卷积学习到的权重是通道间局部的空间规律，CNN在捕捉局部特征方面具有出色的表现，但它们在捕捉跨通道全局相关性方面具有局限性。鉴于隐写信号的特殊性，易导致网络学习到的特征并不具有很强的区分性，因此，我们希望进一步学习不同通道间的相关性，使每个像素可以捕捉残差图间的全局依赖关系，并与卷积神经网络相结合，学习更丰富的隐写特征。然而，自注意力机制更多关注的是共同信息，忽视了差异信息的提取和利用。而在交叉注意力机制中，一个序列中的某个位置可与另一个序列中的所有位置进行注意力权重系数的计算，基于此，本文设计了基于交叉注意力机制的跨通道信息学习块，以搜集残差通道间交叉路径上所有通道的上下文信息，并通过进一步的循环操作，最终捕捉到隐写分析全局特征的依赖关系。图7展示了利用交叉注意力机制捕获通道间相关性的原理过程。

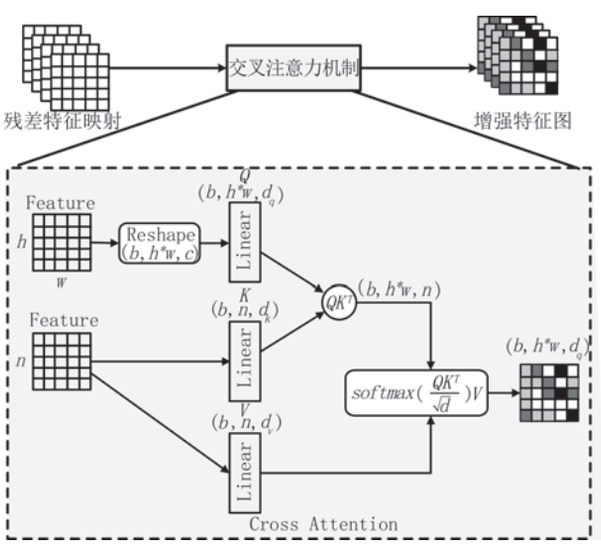


图7　交叉注意力示意图

通常情况下，网络主干后期深层特征的每个通道可以看作是代表某一特定类别的信息，同时这些通道之间存在内在的关联，传统的卷积神经网络往往难以充分捕捉这些通道间的复杂关系。注意力机制可以有效地模拟并建模各通道间的相互依赖关系，通过动态调整各通道的权重，来强化对关键信息的提取并抑制不相关特征的干扰，从而改善特征表示的质量。这种增强机制进一步提升了特征的区分度，使后续的分类过程更加精准。因此，在我们模型的后期利用交叉注意力强模块，以优化通道间的依

赖性和特征表示能力。

具体地，给定图像x的噪声残差图$I(x)$交叉注意力增强形式化描述为：

$$z(x)=\sum_{w,h}Attn(I(x_i))\odot I(x_j)_{w,h} \tag{8}$$

其中，$\odot$表示逐元素相乘，求和是在每个通道的宽度和高度上。

3.3 特征融合及分类模块

特征提取模块为每个子图x_1和x_2生成相应的特征图$z(x_1)$和$z(x_2)$。本文使用全局平均池化(Global Average Pooling，GAP)将特征图中的空间信息压缩成一个固定长度的向量，以从$z(x_1)$和$z(x_2)$中获得更紧凑的特征表示，这不仅有助于减少计算复杂度，而且可以减少过拟合。全局平均池化的计算方式如下：

$$\mathrm{GAP}(z(x_i))=\frac{1}{H\times W}\sum_{h=1}^{H}\sum_{w=1}^{W}z(x_i)_{h,w} \tag{9}$$

其中，$z(x_i)$表示特征图，H和W分别是特征图的高度和宽度，$z(x_i)_{h,w}$表示特征图在(h,w)的值。

将经过全局平均池化后的两个向量GAP$(z(x_1))$和GAP$(z(x_2))$进行融合(融合方式为横向拼接)，作为图像最终的特征表示$z(x)$，即

$$z(x)=Fusion(\mathrm{GAP}(z(x_1)),\mathrm{GAP}(z(x_2))) \tag{10}$$

其中，GAP$(\cdot)$表示全局平局池化，$Fusion$表示特征融合。

网络学习到的图像特征$z(x)$通过一个全连接层得到最终的分类结果，即

$$y=Wz(x)+b \tag{11}$$

其中，W是可学习的权重矩阵，b是偏置向量，y是经过全连接层之后的输出向量。

全连接层之后使用softmax函数得到图像类别概率分布，即

$$p_\phi(i|x)=\mathrm{softmax}(y)=\frac{exp(y_i)}{\sum_j exp(y_j)},\ i,j\in\{0,1\} \tag{12}$$

其中，$p_\phi(i|x)$是深度学习模型ϕ的概率分布，$p_\phi(0|x)$和$p_\phi(1|x)$分别是图像x的载体的概率和载密的概率，i和j遍历所有类别。

最终的输出p_ϕ用于载体载密图像分类，通常情况下，隐写分析的图像预测标签为

$$\hat{y}=\begin{cases}1, p_\phi(0|x)\geqslant 0.5\\ 0, p_\phi(0|x)<0.5\end{cases} \tag{13}$$

其中，0.5是二分类问题默认的分类阈值。

最后，使用梯度下降法对模型进行端到端的参数更新，直至模型收敛。

4 实验结果和性能分析

本节将进行一系列实验来验证所提CES-Net方法的性能，包括与不同深度学习隐写分析模型在典型空域和JPEG图像隐写检测方面的对比、载体和算法失配检测实验、鲁棒图像隐写检测实验、模型参数量和计算量实验以及网络模块消融实验对比和相应的实验分析。

4.1 数据集和评估指标

实验数据集使用的是隐写和隐写分析领域常用的图像数据库BOSSBase-1.01和BOWs2，这两个数据库分别包含10,000张512×512的灰度未压缩自然图像。实验设置如表2所示。

表2 实验设置

名称	值
图像来源	BOSSBase-1.01和BOWs2
图像尺寸	256像素×256像素
图像颜色	灰度
图像格式	PNG，JPEG
JPEG图像压缩质量因子	75；85
空域图像隐写算法	HILL，WOW，S-UNIWARD
空域图像隐写嵌入率	0.1 bpp，0.2 bpp，0.3 bpp，0.4 bpp
JPEG图像隐写算法	J-UNIWARD
JPEG图像隐写嵌入率	0.1 bpnzAC，0.2 bpnzAC，0.3 bpnzAC，0.4 bpnzAC
空域载体图像数量	10 000+10 000 = 20 000幅
空域载密图像数量	20 000×3×4 = 240 000幅
JPEG载体图像数量	20 000×2 = 40 000幅
JPEG载密图像数量	40 000×1×4 = 160 000幅
载体图像数量总计	20 000+40 000 = 60 000幅
载密图像数量总计	240 000+160 000 = 400 000幅

具体实验设置为：

(1)空域图像数据集。使用MATLAB的“imresize”函数将这两个数据库图像缩放成256×256大小，然后将两个数据库合并，组成新的图像数据集作为本文实验数据集，命名为BossBow。由于空域图像自适应隐写算法WOW[28]、S-UNIWARD[29]和HILL[30]的广泛使用，本文利用这3种隐写算法在空域载体图像数据集上分别构建4种嵌入率(0.1 bpp(bits per pixel)、0.2 bpp、0.3 bpp和0.4 bpp)的空域载密图像。

（2）JPEG图像数据集。由于提出的方法同时适用于检测JPEG图像隐写，为获得JPEG图像数据集，使用MATLAB的“imwrite”函数将载体图像数据集压缩成两种质量因子（75和85）的JPEG图像作为JPEG图像隐写检测载体数据库，并利用当前抗检测性能较好的经典JPEG图像自适应隐写术J-UNIWARD[31]（JPEG Universal Wavelet Relative Distortion）在JPEG图像数据集上分别构建嵌入率为0.1 bpnzAC（bits per nonzero AC DCT coefficient）、0.2 bpnzAC、0.3 bpnzAC和0.4 bpnzAC的JPEG载密图像。

本文采用隐写分析常用的评价指标检测准确率（Acc）来评估所提模型的性能，该评价指标广泛应用于现有的隐写分析工作[21-22]。检测准确率表示为

$$\mathrm{Acc}=1-\frac{1}{2}(P_{FA}+P_{MD}) \tag{14}$$

其中，P_{FA}是指载体图像被错误分类为载密图像的比例，P_{MD}是载密图像被错误分类为载体图像的比例。P_{FA}和P_{MD}的计算公式分别如下：

$$P_{FA}=\frac{FP}{TN+FP} \tag{15}$$

$$P_{MD}=\frac{FN}{TP+FN} \tag{16}$$

其中，FP是载体图像被错误分类为载密图像的数量，TN是载体图像被正确分类为载体图像的数量，FN表示载密图像被错误分类为载体图像的数量，TP表示载密图像被正确分类为载密图像的数量。

除了Acc之外，本文也采用了另一个常用的评价指标AUC来衡量模型的表现。AUC值越大表示性能越好。

现有隐写分析算法通常使用成对训练方法来提取隐写分析特征，以学习载体图像和载密图像间微弱的分类特征。和已有的方法一样，本文将载体图像和对应嵌入率的载密图像组成载体载密图像对用于训练和测试。其中，训练集、验证集和测试集按照6∶1∶3的比例随机抽取，即针对不同嵌入率的隐写术，随机从20,000对图像中随机选取12,000对图像作为训练集，2,000对作为验证集，剩余的6,000对作为测试集。实验主要包括以下7个方面：

（1）模型参数设置实验；

（2）空域图像隐写检测实验；

（3）JPEG图像隐写检测实验；

（4）载体和算法失配检测实验；

（5）鲁棒图像隐写检测实验；

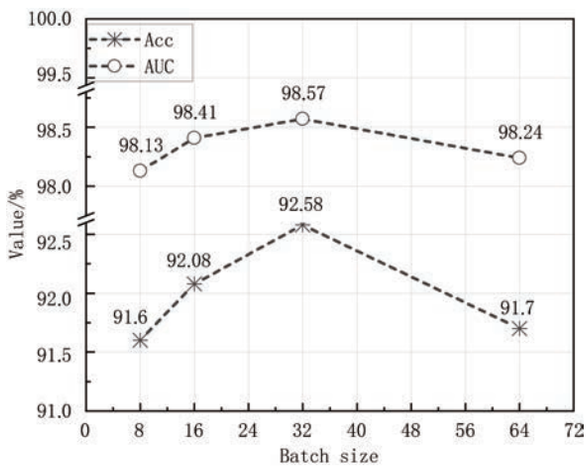
（6）参数量和计算量实验；

（7）消融实验及分析。

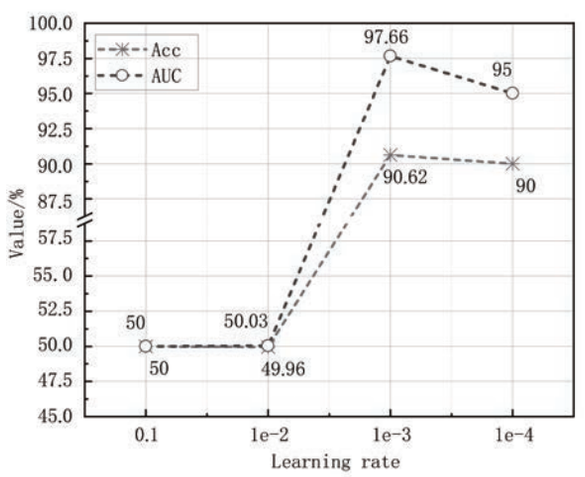
4.2　模型参数设置实验

提出的模型是基于PyTorch深度学习框架来实现的，实验的硬件平台是32 GB显存的NVIDIA TeslaV100显卡。在训练阶段，使用随机梯度下降Adamax优化器来优化网络参数，训练500个epoch。为获得提出方法的最优超参数值，我们重点考察了模型中的两个关键超参数：批大小（Batch Size）和学习率（Learning Rate）。

具体地，我们采用了多组不同大小的Batch size和Learning rate来训练提出的CES-Net模型，其中，Batch size分别取8、16、32和64进行实验，Learning rate分别取1e-1、1e-2、1e-3、1e-4进行实验。实验数据集选择的是空域载体图像数据集BossBow和相应0.4 bpp嵌入率的WOW隐写算法生成的载密图像来对模型进行训练和测试，实验结果如图8所示，



（a）批大小对模型准确率的影响



（b）学习率对模型准确率的影响

图8　网络参数对模型准确率的影响

其中，图8(a)是Batch size对模型准确率的影响，图8(b)是Learning rate对模型准确率的影响。

根据图8的实验结果，我们确定最佳的Batch size为32，即每次训练迭代中处理32对载体-载密图像对。此外，在训练过程中，为了增强模型的泛化能力，在每个迭代周期内对数据进行随机打乱。同时，将初始学习率设定为1e-3，并在设定的Epoch后将其调整为原来的1/10，即1e-4。通过逐步降低学习率来逐渐减小更新步长，以在训练的后期阶段更精细地调整模型参数，最终提高模型的性能和稳定性。

4.3 空域图像隐写检测实验

为了验证本文方法在空域图像隐写检测方面的有效性，本节将提出的隐写分析模型与现有的一些典型隐写分析方法进行了性能对比，对比方法包括6种，即1种传统的空域图像隐写检测方法SRM+EC(Ensemble Classifier)[4](SRM空域富模型特征结合集成分类器)；3种基于深度学习的空域图像隐写检测方法SiaStegNet[18]、CSANet[24]和LWENet[23]；2种同时适用于空域和JPEG图像的深度学习隐写检测方法SRNet[15]和ResFormer[16]。我们使用0.1 bpp、0.2 bpp、0.3 bpp、0.4 bpp共4种嵌入率的3种空域自适应隐写算法WOW、S-UNIWARD和HILL分别进行了实验对比，以全面评估不同方法的性能。

方法的对比实验结果如表3所示。从表3可以看出，图像嵌入率越高，隐写检测方法的准确率越高。同时可以看出，在检测多种嵌入率的不同隐写算法时，深度学习隐写分析方法的检测性能均超过了传统方法。而与现有深度学习隐写分析方法相比，在测试的4种嵌入率下WOW、S-UNIWARD和HILL隐写算法的数据集中，本文方法的检测准确率比SRM+EC、SRNet、SiaStegNet、CSANet、

表3 本文方法与典型空域图像隐写检测方法在WOW、S-UNIWARD和HILL隐写算法上的检测结果对比

隐写检测方法	隐写算法	0.1 bpp		0.2 bpp		0.3 bpp		0.4 bpp	
		Acc(%)	AUC(%)	Acc(%)	AUC(%)	Acc(%)	AUC(%)	Acc(%)	AUC(%)
SRM+EC[4]	WOW	55.20	61.30	67.26	78.70	70.12	83.19	72.26	85.55
	S-UNIWARD	54.47	59.17	60.79	70.86	67.22	78.66	73.03	85.88
	HILL	53.13	56.27	57.38	64.22	62.32	71.61	67.44	78.10
SRNet[15]	WOW	70.58	78.06	86.98	95.46	88.88	96.65	90.85	97.87
	S-UNIWARD	67.80	77.20	79.23	89.76	85.18	94.86	89.10	97.45
	HILL	65.77	74.68	74.30	85.76	81.25	91.86	83.75	94.88
SiaStegNet[18]	WOW	69.87	79.93	86.17	94.14	87.49	96.12	91.04	97.85
	S-UNIWARD	66.64	75.44	78.63	88.60	85.13	94.11	88.89	96.68
	HILL	64.22	72.45	74.20	83.95	78.73	88.89	83.35	93.23
CSANet[24]	WOW	72.37	82.93	87.73	96.25	88.87	96.97	91.16	97.98
	S-UNIWARD	66.49	76.11	79.06	90.07	84.64	94.36	89.81	**97.82**
	HILL	66.93	75.75	75.68	86.59	79.62	90.13	84.95	94.55
LWENet[23]	WOW	74.53	85.25	88.25	96.59	89.83	97.32	91.39	98.15
	S-UNIWARD	67.70	77.51	78.64	89.58	84.27	94.09	89.28	96.99
	HILL	67.19	76.71	75.48	86.46	81.03	91.04	84.58	94.25
ResFormer[16]	WOW	74.97	85.70	88.36	96.51	90.27	97.55	92.13	98.35
	S-UNIWARD	68.48	78.15	79.44	89.67	85.96	95.01	89.64	97.28
	HILL	66.91	76.05	75.17	85.89	81.64	91.81	84.55	94.14
本文方法	WOW	**76.07** ↑(1.10 ~20.87)	**86.91** ↑(1.21 ~25.61)	**89.63** ↑(**1.27** ~22.37)	**97.01** (↑0.42 ~18.31)	**90.99** ↑(0.72 ~20.87)	**97.90** ↑(0.35 ~14.71)	**92.58** ↑(0.45 ~20.32)	**98.57** ↑(0.22 ~13.02)
	S-UNIWARD	**70.58** ↑(2.10 ~16.11)	**80.90** ↑(2.75 ~21.73)	**81.17** ↑(1.73 ~20.38)	**91.32** ↑(1.56 ~20.46)	**86.98** ↑(1.02 ~19.17)	**95.92** ↑(0.91 ~17.26)	**90.60** ↑(0.96 ~17.57)	97.67 ↓-0.15
	HILL	**67.70** ↑(0.51 ~14.57)	**77.25** ↑(0.54 ~20.98)	**77.18** ↑(1.50 ~19.8)	**87.68** ↑(1.09 ~23.46)	**82.92** ↑(1.28 ~20.6)	**92.60** ↑(0.74 ~20.99)	**86.64** ↑(**1.69** ~19.2)	**95.47** ↑(0.59 ~17.31)

注：加粗表示相应情况下最好的结果，下划线表示次优结果，括号中为本文方法与对比方法检测准确率的差异，↑表示提高，↓表示降低。

LWENet、ResFormer高，且与目前最优的模型检测性能相比，提出方法在0.2 bpp的WOW算法、0.1 bpp的S-UNIWARD算法和0.4 bpp的HILL算法上的性能提升最显著，分别提升1.27%、2.10%和1.69%。因此，在本节通过充足的实验验证了我们所提出的隐写分析模型在空域图像隐写检测上的有效性。

4.4　JPEG图像隐写检测实验

为了验证本文方法在JPEG图像隐写检测方面的有效性，本节将提出的模型与6种现有一些典型的JPEG图像隐写分析方法进行了性能对比，包括3种传统方法和3种基于深度学习的隐写检测方法。其中，传统方法为经典的JPEG图像隐写检测特征DCTR结合集成分类器的DCTR+EC方法[5]，以及最新提出的低维特征$DCTR_3$和$DCTR_4$分别结合集成分类器的$DCTR_3$+EC[32]、$DCTR_4$+EC方法[32]；对比的基于深度学习方法为JPEG图像隐写分析方法EWNet[20]，以及2种同时适用于空域和JPEG图像的隐写分析方法SRNet[15]和ResFormer[16]。JPEG图像隐写检测实验在2种质量因子的BossBow_75和BossBow_85图像数据库上进行。使用J-UNIWARD隐写算法在0.1 bpnzAC、0.2 bpnzAC、0.3 bpnzAC、0.4 bpnzAC共4种嵌入率的隐写进行了实验，以全面评估不同方法的性能。

方法的对比实验结果如表4所示。从表4可以看出，与空域图像隐写检测结果一样，图像隐写嵌入率越高，隐写检测方法的准确率越高。同时可以看出，JPEG图像的质量因子越高，相应的载密图像越难检测。在检测QF=75的JPEG图像隐写时，提出方法的检测准确率均优于对比方法SRNet、EWNet和ResFormer，且与目前最优的模型相比，本文方法在0.2 bpnzAC的J-UNIWARD算法上的性能提升最显著，准确率提升达到2.34%；在检测QF=85的JPEG图像隐写时，本文方法的检测准确率在0.1 bpnzAC和0.4 bpnzAC的嵌入率略逊于对比方法，分别低0.49%和0.85%，但是在0.2 bpnzAC和0.3 bpnzAC的嵌入率下，本文方法的检测准确率均明显优于其余3种对比方法，且最多可提升1.50%。因此，本节通过充足的实验验证了我们所提出的隐写分析模型在空域图像隐写检测上的有效性。

表4　提出的方法与典型JPEG图像隐写检测方法在两种质量因子的JPEG图像上使用J-UNIWARD隐写算法的检测结果

质量因子	隐写检测方法	0.1 bpnzAC		0.2 bpnzAC		0.3 bpnzAC		0.4 bpnzAC	
		Acc(%)	AUC(%)	Acc(%)	AUC(%)	Acc(%)	AUC(%)	Acc(%)	AUC(%)
75	DCTR+EC[5]	54.69	58.73	61.94	72.53	70.45	84.59	78.78	92.93
	$DCTR_3$+EC[32]	54.81	56.59	62.05	66.65	70.08	76.38	77.84	85.93
	$DCTR_4$+EC[32]	54.68	58.72	62.09	72.74	70.40	84.83	83.62	93.09
	SRNet[15]	56.50	61.01	75.10	83.10	79.04	90.37	89.95	96.26
	EWNet[20]	64.35	68.47	77.37	83.44	78.58	90.23	89.67	95.87
	ResFormer[16]	64.98	70.37	75.88	84.34	84.82	92.72	90.43	96.84
	本文方法	**65.30** ↑(0.32 ~10.62)	**71.95** ↑(1.58 ~15.36)	**77.60** ↑(0.23 ~15.66)	**86.68** ↑(**2.34** ~20.03)	**85.65** ↑(0.83 ~15.57)	**93.67** ↑(0.95 ~17.29)	**91.37** ↑(0.94 12.59)	**97.55** ↑(0.71 ~19.71)
85	DCTR+EC[5]	53.22	56.16	58.86	67.40	65.81	78.76	73.39	88.44
	$DCTR_3$+EC[32]	53.24	54.28	59.10	62.94	65.87	72.10	73.02	80.76
	$DCTR_4$+EC[32]	54.58	57.01	62.12	68.41	70.71	79.80	79.12	89.20
	SRNet[15]	55.55	58.83	71.65	80.96	78.05	89.63	87.60	95.69
	EWNet[20]	61.52	65.74	73.37	78.86	75.05	87.10	88.18	94.33
	ResFormer[16]	**61.86**	66.40	72.97	82.02	80.88	89.89	87.48	95.13
	本文方法	61.37 ↓0.49	**66.89** ↑(0.49 ~12.61)	**73.80** ↑(0.43 ~14.94)	**83.22** ↑(1.20 ~20.28)	**82.38** ↑(**1.50** ~16.57)	**91.95** ↑(2.06 ~19.85)	**88.85** ↑(0.67 ~15.83)	**95.71** ↑(0.02 ~14.95)

注：加粗表示在相应情况下最好的结果，下划线表示次优结果，括号中为提出的方法与对比方法检测准确率的差异，↑表示提高，↓表示降低。

本文提出的方法在空域和JPEG图像隐写检测方面均取得了性能的提升，究其原因，一方面得益于基于孪生网络对图像分区域进行学习，有助于提取更丰富、更有区别性的细粒度特征，提高检测准确率；另一方面是因为交叉注意力机制是在全局范围内对特征进行加权处理，而不仅仅局限于局部区域。

这种全局感知能力在隐写检测中尤为重要，尤其是在隐写噪声信号相对较弱的情况下。同时，4.3节和4.4节的实验充分验证了我们提出的隐写分析模型在典型的空域和JPEG图像隐写检测上的有效性。

4.5 载体和算法失配检测实验

本小节进行开放场景下方法的性能对比实验，包括载体失配和算法失配的隐写检测。

4.5.1 载体失配检测

为验证提出方法的跨数据库性能，除了在常用的BOSSBase-1.01和BOWs2数据集进行实验外，本文增加了提出方法在ALASKA#2图像数据库的实验，以评估方法在不同数据集上的有效性和可靠性。ALASKA#2[33]是图像隐写分析领域极具挑战性的数据集，旨在将隐写分析技术从实验环境推向实际应用场景。该数据集是一个大型异构的摄影图像数据集，包含多种图像库，本文使用其中的512×512尺寸灰度图像库，并利用Matlab的“imresize”函数将其缩放至256×256大小，用于模型的训练和测试。载密图像使用0.4 bpp嵌入率的WOW隐写算法进行信息嵌入。

实验结果如表5所示。从表5可以看出，与BOSSBase-1.01和BOWs2数据库相比，ALASKA#2图像数据库上的隐写更难检测。而在这三个数据库中，BOSSBase-1.01数据集隐写相对来说比较容易检测，即便是在载体失配情况下，模型在检测BOSSBase-1.01数据集上的隐写时，准确率仍相对较高。

从表5还可以看出，当训练数据集和测试数据集均为ALASKA#2图像数据库时，本文方法的检测性能比SRNet和SiaStegNet均有提升，检测准确率比SRNet提高0.78%，比SiaStegNet提高1.91%。当训练数据集和测试数据集不一致时，比如在ALASKA#2上训练的模型，在检测BOSSBase-1.01数据集隐写时，提出方法的检测准确率比SRNet和SiaStegNet分别高1.29%和0.19%，在检测BOWs2数据集隐写时，比SRNet和SiaStegNet分别高1.63%和1.97%。跨数据库的实验结果表明，本文方法在载体失配这一开放场景下仍具有一定的优势。

表5 SRNet、SiaStegNet和本文方法在三种数据集上检测0.4 bpp嵌入率的WOW隐写的性能对比

训练集	检测器	测试集					
		BOSSBase		BOWs		ALASKA	
		Acc（%）	AUC（%）	Acc（%）	AUC（%）	Acc（%）	AUC（%）
BOSSBase	SRNet[15]	91.19	**98.01**	85.85	**95.09**	68.10	**76.09**
	SiaStegNet[18]	90.53	97.44	83.73	92.59	65.95	75.01
	本文方法	**91.48**	97.51	**86.48**	93.11	**68.45**	75.69
BOWs	SRNet[15]	86.85	95.43	**87.74**	**95.95**	**66.55**	75.13
	SiaStegNet[18]	87.88	96.23	84.41	93.64	64.31	73.13
	本文方法	**88.40**	**96.54**	84.56	93.44	65.15	**75.45**
ALASKA	SRNet[15]	82.00	**94.86**	76.05	**91.32**	68.20	77.18
	SiaStegNet[18]	83.10	93.39	75.72	88.57	67.07	77.22
	本文方法	**83.29**	93.42	**77.69**	88.79	**68.98**	**78.14**

注：加粗表示在相应情况下最好的结果。

4.5.2 算法失配检测

为了评估本文提出的方法在算法失配时的性能，本文使用WOW、S-UNIWARD、HILL和MiPOD[34]四种空域图像隐写算法，以及J-UNIWARD和UERD两种JPEG图像隐写算法进行实验。实验是在一种算法上进行训练，然后在同一嵌入率的另一种算法上进行测试，隐写嵌入率为0.4bpp/bpnzAC，数据集为BOSSBase-1.01。实验结果如表6和表7所示。

表6 SRNet和CES-Net在空域图像隐写算法失配下的检测准确率对比（隐写嵌入率为0.4 bpp）

训练算法	检测器	检测算法			
		WOW	S-UNIWARD	HILL	MiPOD
		Acc(%)	Acc(%)	Acc(%)	Acc(%)
WOW	SRNet	91.07	**84.48**	67.72	71.21
	CES-Net	**91.48**	79.42	**68.28**	**71.89**
S-UNIWARD	SRNet	88.98	89.77	75.17	78.84
	CES-Net	**91.65**	**92.53**	74.75	78.17
HILL	SRNet	82.58	72.58	85.86	78.20
	CES-Net	**86.46**	**78.54**	**88.39**	**81.42**
MiPOD	SRNet	85.24	84.04	81.12	85.03
	CES-Net	**86.61**	**84.67**	80.77	**86.35**

注：加粗表示在相应情况下最好的结果。

表7 SRNet和CES-Net在JPEG图像隐写算法失配下的检测准确率对比（隐写嵌入率为0.4 bpnzAC）

质量因子	训练算法	检测器	检测算法	
			J-UNIWARD	UERD
			Acc(%)	Acc(%)
75	J-UNIWARD	SRNet	89.93	92.47
		CES-Net	**91.37**	**94.56**
	UERD	SRNet	76.81	**93.93**
		CES-Net	**78.69**	93.00
85	J-UNIWARD	SRNet	84.38	92.36
		CES-Net	**87.91**	**93.25**
	UERD	SRNet	85.46	92.23
		CES-Net	**87.92**	**93.25**

注：加粗表示在相应情况下最好的结果。

从表6可以看出，空域图像隐写算法HILL仍然是最难检测的一种隐写，而WOW相对更易检测。其中，基于WOW训练的网络，在检测HILL和MiPOD时，提出方法的准确率比SRNet有微弱的提升，分别提高0.56%和0.68%。从整体来看，基于MiPOD算法训练网络的失配检测性能更好。

从表7可以看出，UERD隐写比J-UNIWARD更容易检测。从整体上看，使用J-UNIWARD训练的网络，在检测UERD隐写时，两种方法均呈现出较好的检测效果。尤其是在质量因子为85时，提出方法的算法适配性能比SRNet有微弱的提升，提升约0.89%。表6和表7的实验结果表明，本文方法在算法适配时具有一定的优势。

4.6　鲁棒图像隐写检测实验

虽然目前的隐写检测方法普遍通过检测WOW、S-UNIWARD、HILL以及J-UNIWARD等隐写算法来评估提出方法的性能，但鉴于面向有损网络信道的图像隐写技术——鲁棒隐写术的不断提出，本节我们对提出的方法在检测最新的鲁棒隐写算法方面的性能进行评估。由于鲁棒隐写算法通常更关注面对社交网络攻击时的抵抗能力，因此其隐写嵌入率通常较低。为此，我们使用ROAST[35]和Zhu[36]两种鲁棒隐写算法在0.1 bpnzAC和0.3 bpnzAC两种嵌入率下进行检测实验，以评估方法的性能。实验数据集为BossBase-1.01数据库的10,000张512×512图像，压缩质量因子为85。实验结果如表8所示。

表8　在鲁棒隐写算法上的检测准确率

隐写方法	隐写检测方法	嵌入率/%	
		0.1 bpnzAC	0.3 bpnzAC
ROAST[35]	$DCTR_3$+EC[32]	56.00	76.70
	$DCTR_4$+EC[32]	58.70	81.79
	SRNet[15]	59.85	83.82
	本文方法	**61.38**	**86.46**
Zhu的方法[36]	$DCTR_3$+EC[32]	59.20	73.99
	$DCTR_4$+EC[32]	61.05	75.28
	SRNet[15]	64.43	82.90
	本文方法	**67.18**	**84.59**

从表8可以看出，相比于传统的$DCTR_3$特征并结合集成分类器的方法以及经典的深度学习隐写检测方法SRNet，本文方法取得了最优的检测准确率。如0.3 bpnzAC嵌入率时，方法在检测ROAST隐写时准确率分别提高了2.64%到9.76%；在检测Zhu方法隐写时的准确率分别提高了1.69%到10.06%，验证了本文方法的有效性。

4.7　参数量和计算量实验

深度学习网络的参数量和计算量是衡量模型复杂度和性能的重要指标。本节从参数量和计算量对提出模型与其他模型的时间和空间复杂度进行量化分析，以评估其在实际应用中的性能，对比方法包括SRNet[15]、SiaStegNet[18]、EWNet[20]、CSANet[24]、LWENet[23]、ResFormer[16]和本文提出的CES-Net方法。实验结果如表9所示。

表9　本文提出的模型与其他模型参数量和计算量对比

隐写检测模型	参数量(Params)	计算量(FLOPs)
SRNet	4.77 M	5.95 G
SiaStegNet	0.71 M	7.28 G
EWNet	3.82 M	2.03 G
CSANet	0.47 M	8.18 G
LWENet	0.38 M	4.82 G
ResFormer	0.39 M	1.80 G
本文方法	0.35 M	3.71 G

从表9可以看出，SRNet和EWNet所需参数量远高于其他方法。本文提出方法的参数量比SRNet减少约92.6%，比EWNet减少约90.8%，同时仍低于ResFormer方法。通常情况下，参数量大的模型往往需要更多的迭代才能收敛，这不可避免地会增加训练时间。

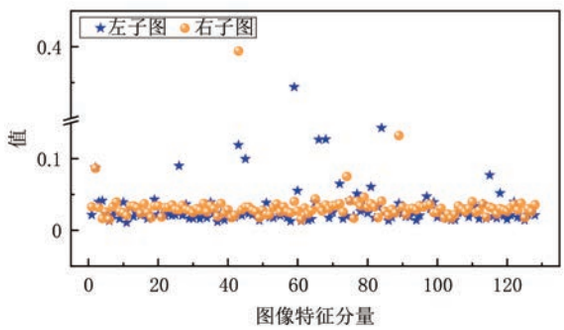
从表9还可以看出，虽然SRNet的参数量最高，但其计算量并不是最高的。在对比方法中，计算量相对较高的是CSANet。SiaStegNet方法和提出方法的计算量比CSANet分别减少约11%和54%。

4.8　消融实验及分析

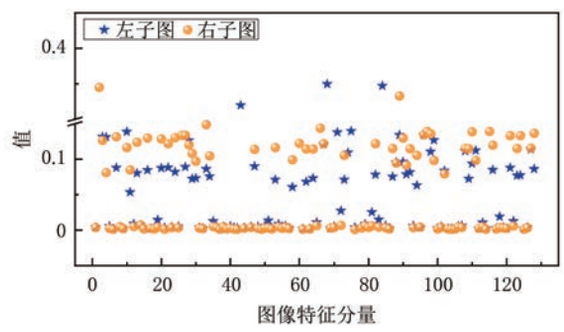
为了验证本文提出的隐写分析网络结构设计的必要性和合理性，本节对提出的模型在不同的配置下进行了消融实验，主要包括基于孪生网络对图像进行分区域学习性能分析、图像预处理模块中滤波核设计和主干网中交叉注意力模块对模型性能的影响。同时，我们对不同模块的消融实验进行了分析，并对CAB模块性能进行了可视化。本文在空域和JPEG图像上均进行了消融实验验证，其中，空域图像隐写检测的消融实验是在0.2 bpp和0.4 bpp嵌入率的WOW、S-UNIWARD和HILL三种隐写算法上进行，JPEG图像隐写检测的消融实验是在0.2 bpnzAC和0.4 bpnzAC嵌入率的J-UNIWARD隐写算法上进行。具体实验结果如下。

4.8.1　孪生网络对图像分区域学习性能分析

图9显示的是本文提出的CES-Net网络输出的



（a）载体图像



（b）载密图像

图9 本文方法输出载体/载密图像子图128-D特征可视化

图像左、右子图128-D（Dimension）特征的分布情况，其中，图9（a）是质量因子为75的BOSSBase-1.01图像数据库中名称为201的载体图像特征分布，图9（b）为使用0.4 bpnzAC嵌入率的J-UNIWARD生成的图9（a）载体图像对应的载密图像特征分布，横坐标表示每一维特征分量，纵坐标表示该维特征分量的值。

从图9可以看出，对于载体图像而言，同一维度不同区域的图像特征分量值相差不大，而对于载密图像，其值差异相对显著。说明将图像进行区域划分并使用孪生网络分区域学习的方法，可有效捕捉载体载密图像差异，提高检测性能。

4.8.2 预处理模块中滤波核设置的消融实验

表10和表11是预处理层中高通滤波核规模对模型性能影响的实验结果，表10是空域图像隐写检测的消融实验结果，表11是JPEG图像隐写检测的消融实验结果，其中，CES-Net是本文提出的网络结构，CES-Net-SRM表示预处理层中高通滤波器使用传统的30个滤波核。

表10 滤波核设置对空域图像隐写检测性能的影响

隐写算法	隐写检测方法	0.2 bpp		0.4 bpp	
		Acc(%)	AUC(%)	Acc(%)	AUC(%)
WOW	CES-Net-SRM	88.47	96.60	91.51	98.19
	CES-Net	**89.63**（↑**1.16**）	**97.01**（↑0.41）	**92.58**（↑1.07）	**98.57**（↑0.38）
S-UNIWARD	CES-Net-SRM	80.62	91.15	90.52	**97.77**
	CES-Net	**81.17**（↑**0.55**）	**91.32**（↑0.17）	**90.60**（↑0.08）	97.67（↓0.10）
HILL	CES-Net-SRM	76.04	86.84	85.96	95.09
	CES-Net	**77.18**（↑**1.14**）	**87.68**（↑0.84）	**86.64**（↑0.68）	**95.47**（↑0.38）

注：加粗表示在相应情况下最好的结果，括号中为本文提出的模型与对比模型中最优检测准确率的差异，↑表示提高，↓表示降低。

表11 滤波核设置对JPEG图像隐写检测性能的影响

质量因子	隐写检测方法	0.2 bpnzAC		0.4 bpnzAC	
		Acc(%)	AUC(%)	Acc(%)	AUC(%)
75	CES-Net-SRM	77.04	86.01	91.36	96.86
	CES-Net	**77.60**（↑**0.56**）	**86.68**（↑0.67）	**91.37**（↑0.01）	**97.55**（↑0.69）
85	CES-Net-SRM	73.37	82.68	88.00	95.79
	CES-Net	**73.80**（↑**0.43**）	**83.22**（↑0.46）	**88.18**（↑0.18）	**95.69**（↓0.10）

注：加粗表示在相应情况下最好的结果，括号中为本文提出的模型与对比模型中最优检测准确率的差异，↑表示提高，↓表示降低。

从表10和表11可以看出，在两种嵌入率的多种空域和JPEG图像隐写检测中，CES-Net模型的检测准确率比CES-Net-SRM均有提升，其中，在WOW、S-UNIWARD和HILL三种空域图像隐写检测的性能分别最高可提升1.16%、0.55%和1.14%；在75和85两种质量因子的JPEG图像隐写检测的性能分别最高可提升0.56%和0.43%。同时，与高嵌入率隐写检测相比，低嵌入率隐写检测的性能提升更为显著，说明CES-Net方法能够有效捕捉空域和JPEG图像微弱的隐写嵌入，提升图像隐

写检测的性能。因此，本节的相关消融验证了本文提出的模型预处理模块中滤波核的设置在图像隐写检测上的有效性。

4.8.3　交叉注意力模块消融实验

表12和表13是模型主干网中交叉注意力模块设计对模型性能影响的实验结果，表12是空域图像隐写检测的消融实验结果，表13是JPEG图像隐写检测的消融实验结果，其中，CES-Net是本文提出的网络结构，CES-Net-Cross表示将主干网中的交叉注意力机制替换为传统卷积。

表12　交叉注意力模块对空域图像隐写检测性能的影响

隐写算法	隐写检测方法	0.2 bpp		0.4 bpp	
		Acc(%)	AUC(%)	Acc(%)	AUC(%)
WOW	CES-Net-Cross	86.08	94.99	89.85	97.18
	CES-Net	**89.63** (↑**3.55**)	**97.01** (↑2.02)	**92.58** (↑2.73)	**98.57** (↑1.39)
S-UNIWARD	CES-Net-Cross	79.97	89.74	90.25	97.28
	CES-Net	**81.17** (↑**1.20**)	**91.32** (↑1.58)	**90.60** (↑0.35)	**97.67** (↑0.39)
HILL	CES-Net-Cross	75.24	84.85	84.82	93.71
	CES-Net	**77.18** (↑**1.94**)	**87.68** (↑2.83)	**86.64** (↑1.82)	**95.47** (↑1.76)

注：加粗表示在相应情况下最好的结果，括号中为本文提出的模型与对比模型中最优检测准确率的差异，↑表示提高。

表13　交叉注意力模块对JPEG图像隐写检测性能的影响

质量因子	隐写检测方法	0.2 bpnzAC		0.4 bpnzAC	
		Acc(%)	AUC(%)	Acc(%)	AUC(%)
75	CES-Net-Cross	76.57	84.39	91.25	96.82
	CES-Net	**77.60** (↑1.03)	**86.68** (↑2.32)	**91.37** (↑0.12)	**97.55** (↑0.73)
85	CES-Net-Cross	73.13	81.23	88.00	95.28
	CES-Net	**73.80** (↑**0.67**)	**83.22** (↑1.99)	**88.18** (↑0.18)	**95.69** (↑0.41)

注：加粗表示在相应情况下最好的结果，括号中为本文提出的模型与对比模型中最优检测准确率的差异，↑表示提高。

从表12和表13可以看出，CES-Net模型相较于CES-Net-Cross在空域和JPEG图像隐写检测任务中均表现出显著的性能提升。其中，CES-Net在空域图像隐写检测的提升幅度较JPEG格式更为显著，这可能是因为空域隐写算法在修改图像时对视觉特征的影响更加显著，而交叉注意力机制能够更有效地捕捉这些微妙的变化。

JPEG图像隐写检测虽然也有性能的提升，但相对于空域的提升幅度稍低，表明交叉注意力在处理不同类型的隐写算法时虽然具有普适性，但在空域算法中表现尤为突出。另外，低隐写嵌入率下CES-Net模型相对于CES-Net-Cross的性能提升更为显著，这表明交叉注意力机制在低嵌入率隐写检测方面具有相对明显优势。这可能是由于低嵌入率隐写的信息变化相对微弱，传统的卷积操作难以有效捕捉这些细微的特征变化，而交叉注意力机制则能够通过跨通道的信息交互来增强模型对隐写信号的感知能力。因此，本节的消融验证了本文提出的模型中交叉注意力机制在图像隐写检测上的有效性。

4.8.4　不同模块消融实验分析

本节将3种不同的结构CES-Net、CES-Net-SRM和CES-Net-Cross进行对比，以分析网络结构模块的设计并验证其合理性。3种结构可视化示意图如图10(a)～(d)所示，横坐标表示隐写嵌入率，纵坐标表示模型检测准确率。图10(a)为3种模型设计方案在检测WOW隐写的性能差异，图10(b)是检测HILL隐写上的性能差异，图10(c)为检测QF=75的J-UNIWARD隐写上的性能差异，图10(d)为检测QF=85的J-UNIWARD隐写上的性能差异。

从图10可以看出，CES-Net的检测准确率均优于其他两种结构，从而可以验证本文提出方法的有效性和合理性。其中，交叉注意力机制比滤波核对

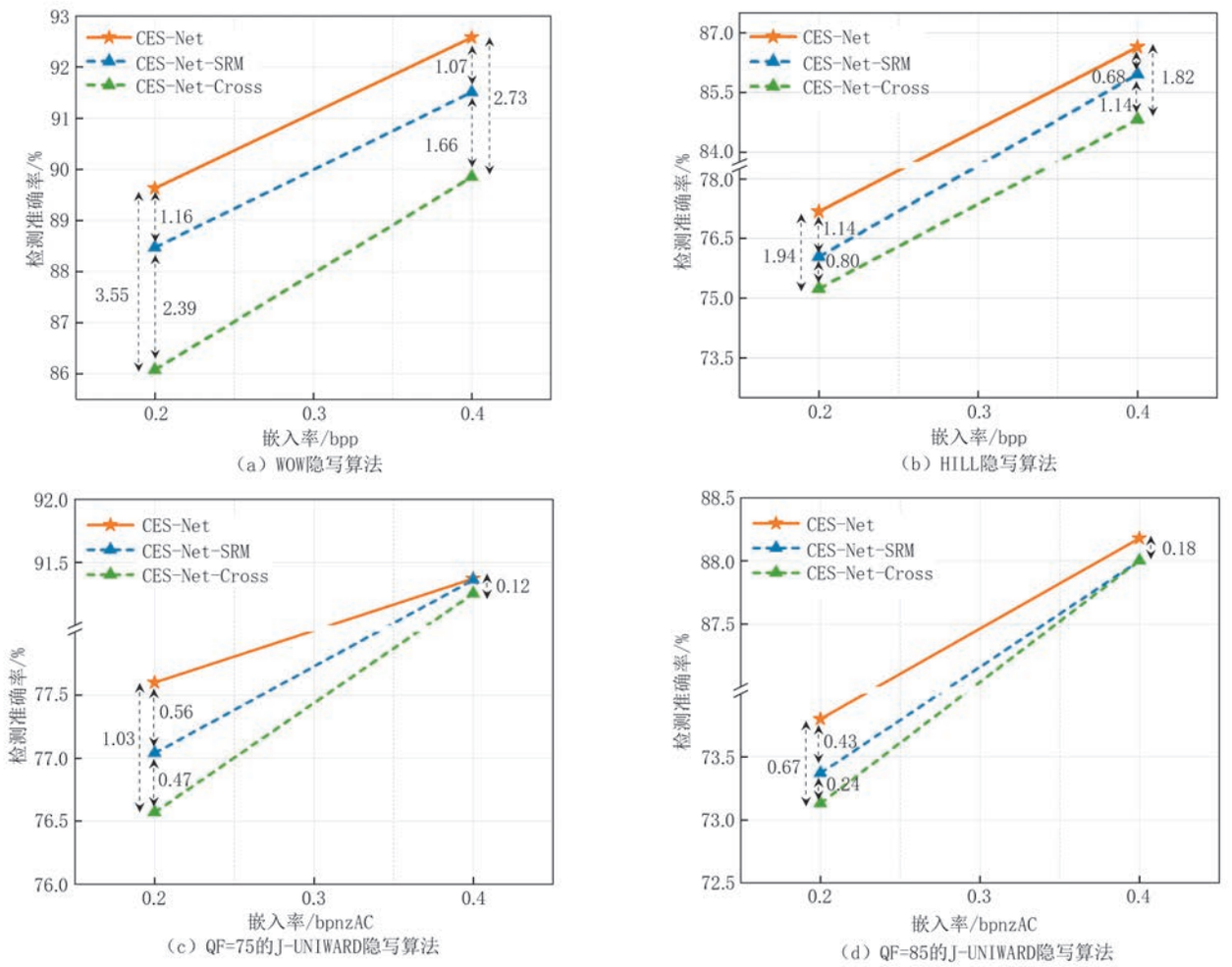


图10 不同模块在空域和JPEG图像隐写检测上的性能表现

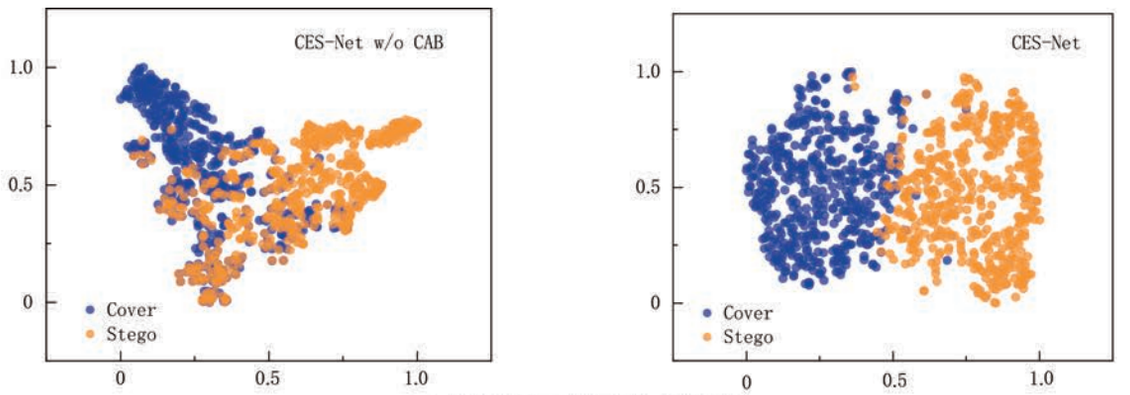
模型性能的影响更为显著，尤其是在低嵌入率隐写检测方面，这一现象更为突出，在0.2 bpp/bpnzAC嵌入率，模型中交叉注意力机制和SRM对模型性能的影响差异分别为2.39%、0.80%、0.47%和0.24%，在0.4 bpp/ bpnzAC嵌入率，模型中交叉注意力机制比SRM对模型性能的提升分别高1.66%、1.14%、0.12%和0.0%。这进一步说明交叉注意力机制可有效捕捉丰富的残差特征，从而提高模型的表达能力。

为更直观地描述交叉注意力模块的分类效果，在下一小节，我们对模型中交叉注意力机制模块对提取特征的分类能力进行可视化。

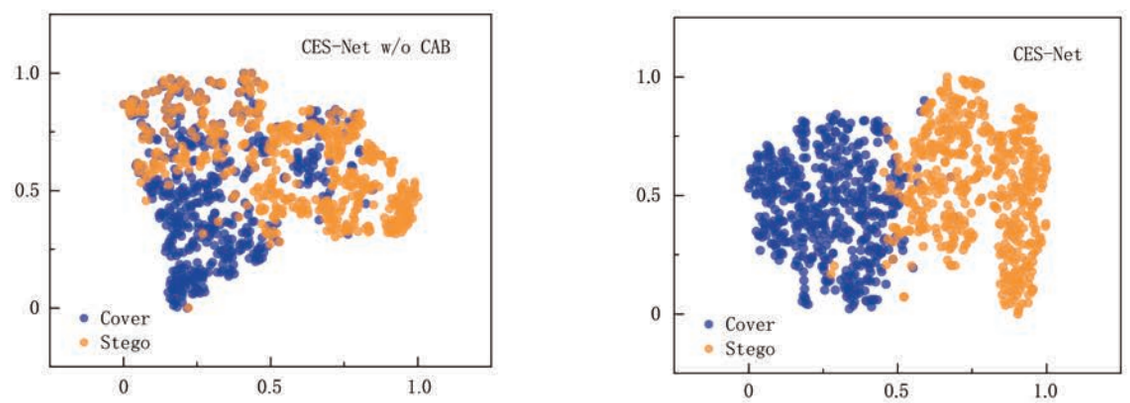
4.8.5 CAB模块分类可视化及分析

本小节是对提出的方法中CAB模块对载体载密图像分类的t-SNE可视化，可视化描述如图11所示，包括实验对比的WOW、S-UNIWARD和HILL三种空域图像隐写及JPEG图像的J-UNIWARD隐写检测可视化，其中，载体载密图像数量分别为500。在图11中，左边图为本文提出的方法中不使用交叉注意力机制（这里表示为：CES-Net w/o CAB）对两类图像的分类结果，右边图为提出的方法（CES-Net）对两类图像的分类结果，蓝色圆点表示载体图像（Cover），黄色圆点表示载密图像（Stego）。

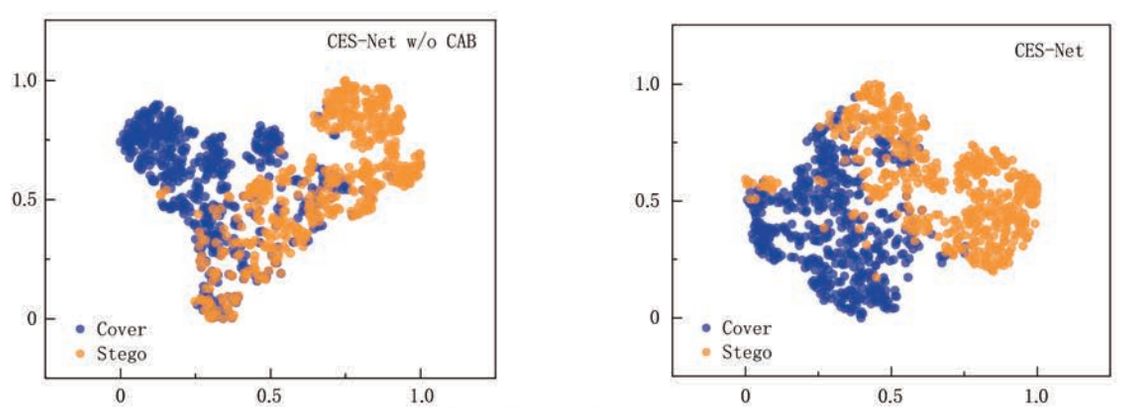
从图11可以看出，提出的模型结构在学习到的特征空间可以很好地分离载体和载密图像，说明我们的模型学习到了一个两类图像的很好的分类面。而CES-Net w/o CAB，由于难以捕获更多通道间残差特征，使得部分载体和载密图像难以区分。可视化结果进一步表明了交叉注意力机制模块设计的合理性和有效性。



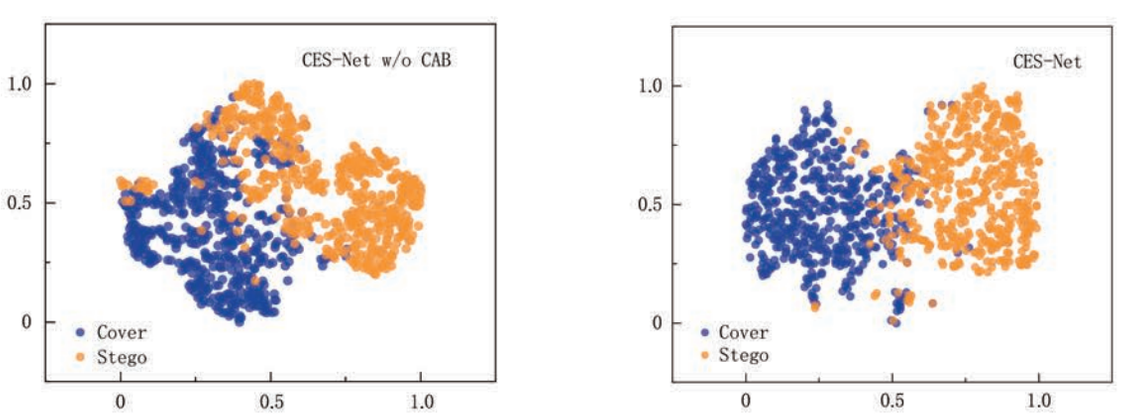
（a）载体图像和WOW隐写图像分类可视化



（b）载体图像和S-UNIWARD隐写图像分类可视化



（c）载体图像和HILL隐写图像分类可视化



（d）载体图像和J-UNIWARD隐写图像分类可视化

图11　网络分类特征t-SNE可视化（与CES-Net w/o CAB方法相比，CES-Net方法生成了更具判别性的特征表示）

5 结束语

本文提出了一种跨通道交叉注意力增强的深度学习隐写分析框架，能够有效检测现有的空域和JPEG图像隐写。该框架的主干网采用孪生神经网络对图像进行分区域细粒度学习，其预处理模块通过多样化的高通滤波器和多层卷积来提取高质量的隐写噪声残差；特征提取模块则利用交叉注意力模块来计算通道间的相关性，从而增强主干网络的特征表达能力；最后，网络学习到的图像分类特征被输入分类模块，以对载体和载密图像进行分类。通过在BOSSbase-1.01和BOWs2图像数据库上进行的一系列实验结果表明，所提模型在多种嵌入率的空域和JPEG图像隐写检测中均表现出显著的性能提升。此外，我们还进行了消融实验，以验证模型结构的合理性。下一步的研究将集中于：(1)设计适用于不同尺寸图像的更高精度隐写分析网络架构，以满足实际应用中的需求；(2)将深度学习隐写分析网络与传统的图像隐写检测特征相结合，进一步提高图像隐写检测的准确率和鲁棒性。

参 考 文 献

[1] Zhong N, Qian Z X, Wang Z C, et al. Batch steganography via generative network. IEEE Transactions on Circuits and Systems for Video Technology, 2021, 31(1): 88-97

[2] Zhou Zhi-Li, Ding Chun, Li Jin, et al. Research on generative steganography. Chinese Journal of Computers, 2023, 46(9): 1855-1887. (in Chinese)
(周志立，丁淳，李进等。生成式隐写研究，计算机学报，2023, 46(9): 1855-1887)

[3] Verma V, Muttoo S K, Singh V B. Detecting stegomalware: Malicious image steganography and its intrusion in windows. Security, Privacy and Data Analytic. Singapore: Springer Singapore, 2022: 103-116

[4] Fridrich J, Kodovsky J. Rich models for steganalysis of digital images. IEEE Transactions on Information Forensics and Security, 2012, 7(3): 868-882

[5] Holub V, Fridrich J. Low-complexity features for JPEG steganalysis using undecimated DCT. IEEE Transactions on Information Forensics and Security, 2014, 10(2): 219-228

[6] Song X F, Liu F L, Zhang Z G, et al. 2D Gabor filters-based steganalysis of content-adaptive JPEG steganography. Multimedia Tools and Applications, 2017, 76: 26391-26419

[7] Qian Y L, Dong J, Wang W, et al. Deep learning for steganalysis via convolutional neural networks//Proceedings of the Media Watermarking, Security, and Forensics. San Francisco, USA, 2015: 94090J

[8] Xu G, Wu H Z, Shi Y Q. Structural design of convolutional neural networks for steganalysis. IEEE Signal Processing Letters, 2016, 23(5): 708-712

[9] Rana K, Singh G, Goyal P. SNRCN2: Steganalysis noise residuals based CNN for source social network identification of digital images. Pattern Recognition Letters, 2023, 171: 124-130

[10] Yu L, Weng S W, Chen M F, et al. RCDD: Contrastive domain discrepancy with reliable steganalysis labeling for cover source mismatch. Expert Systems with Applications, 2024, 237: 121543

[11] Chen Z Q, Yu X Y, Chen R Z. Image block regression based on feature fusion for CNN-based spatial steganalysis//Proceedings of the International Workshop on Digital Watermarking. Beijing, China, 2022: 258-272

[12] Xu G. Deep convolutional neural network to detect J-UNIWARD//Proceedings of the ACM Workshop on Information Hiding and Multimedia Security. Philadelphia, USA, 2017: 67 - 73

[13] Chen M, Sedighi V, Boroumand M, et al. JPEG-phase-aware convolutional neural network for steganalysis of JPEG images//Proceedings of the ACM Workshop on Information Hiding and Multimedia Security. Philadelphia, USA, 2017: 7584

[14] Yousfi Y, Butora J, Khvedchenya E, et al. ImageNet pre-trained CNNs for JPEG steganalysis//Proceedings of the IEEE International Workshop on Information Forensics Security. New York, USA, 2020: 1-6

[15] Boroumand M, Chen M, Fridrich J. Deep residual network for steganalysis of digital images. IEEE Transactions on Information Forensics and Security, 2018, 14(5): 1181-1193

[16] Li H, Luo X Y, Zhang Y. Improving CoatNet for spatial and JPEG domain steganalysis//Proceedings of the IEEE International Conference on Multimedia and Expo. Brisbane, Australia, 2023: 1241-1246

[17] Ye J, Ni J Q, Yi Y. Deep learning hierarchical representations for image steganalysis. IEEE Transactions on Information Forensics and Security, 2017, 12(11): 2545-2557

[18] You W, Zhang H, Zhao X. A Siamese CNN for image steganalysis. IEEE Transactions on Information Forensics and Security, 2020, 16: 291-306

[19] Zeng J S, Tan S Q, Li B, et al. Pre-training via fitting deep neural network to rich-model features extraction procedure and its effect on deep learning for steganalysis. Electronic Imaging, 2017(7): 44-49

[20] Su A T, Zhao X F, He X L. Arbitrary-sized JPEG steganalysis based on fully convolutional network//Proceedings of the International Workshop on Digital Watermarking. Beijing, China, 2021: 197-211

[21] Zhang R, Zhu F, Liu J Y. 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

[22] Shen Jun, Liao Xin, Qin Zheng, et al. Spatial steganalysis of low embedding rate based on convolutional neural network.

Journal of Software, 2021, 32(9): 2901-2915 (in Chinese)
(沈军，廖鑫，秦拯等。 基于卷积神经网络的低嵌入率空域隐写分析。 软件学报，2021，32(9): 2901-2915)

[23] Weng S W, Chen M F, Yu L F, et al. Lightweight and effective deep image steganalysis network. IEEE Signal Processing Letters, 2022, 29: 1888-1892

[24] Liu Q L, Ni J Q, Jian M X. Effective JPEG steganalysis using non-linear pre-processing and residual channel-spatial attention//Proceedings of the IEEE International Conference on Multimedia and Expo. Taipei, China, 2022: 1-6

[25] Xie G, Ren J, Marshall S, et al. Self-attention enhanced deep residual network for spatial image steganalysis. Digital Signal Processing, 2023, 139: 104063

[26] Wei K, Luo W, Liu M, Ye M. Residual guided coordinate attention for selection channel aware image steganalysis. Multimedia Systems, 2023, 29: 2125-2135

[27] Kodovsk J, fridrich J. Steganalysis of JPEG images using rich models//Proceeding of Media Watermarking, Security, and Forensics. Bellingham, USA, 2012: 81-93

[28] Holub V, Fridrich J. Designing steganographic distortion using directional filters//Proceedings of the IEEE International Workshop on Information Forensics and Security. Costa Adeje, Spain, 2012:234-239

[29] Holub V, Fridrich J. Digital image steganography using universal distortion//Proceedings of the ACM Workshop on Information Hiding and Multimedia Security. Montpellier, France, 2013: 59-68

[30] Li B, Tan S Q, Wang M, et al. Investigation on cost assignment in spatial image steganography. IEEE Transactions on Information Forensics and Security, 2014, 9(8): 1264-1277

[31] Holub V, Fridrich J, Denemark T. Universal distortion function for steganography in an arbitrary domain. EURASIP Journal on Information Security, 2014: 1-13

[32] Xia C, Guan Q X, Zhao X F, et al. Improved JPEG phase-aware steganalysis features using multiple filter sizes and difference images. IEEE Transactions on Circuits and Systems for Video Technology, 2019, 30(11): 4100-4113

[33] Cogranne R, Giboulot Q, Bas P. The ALASKA steganalysis challenge: a first step towards steganalysis//Proceedings of the ACM Workshop Information Hiding Multimedia Security. Paris, France, 2019: 125-137

[34] Sedighi V, Cogranne R, Fridrich J. Content-adaptive steganography by minimizing statistical detectability. IEEE Transactions on Information Forensics and Security, 2016, 11(2):221-234

[35] Zeng K, Chen K J, Zhang W M, et al. Upward robust steganography based on overflow alleviation. IEEE Transactions on Multimedia, 2024, 26: 299-312

[36] Zhu Li-Yan, Luo Xiang-Yang, Zhang Yi, et al. Asymmetric distortion steganography method based on superpixel filtering. Chinese Journal of Computers, 2023, 46(7): 1473-1493. (in Chinese)
(朱利妍，罗向阳，张祎等。 基于超像素滤波的非对称失真隐写方法，计算机学报，2023，46(7): 1473-1493)

ZHANG Qian-Qian, Ph. D. candidate, engineer. Her main research interests include image steganography, steganalysis, and granular computing.

LI Hao, Ph. D., lecturer His main research interest is image steganalysis.

ZHANG Yi, Ph. D., lecturer Her main research interests include image steganography and steganalysis.

MA Yuan-Yuan, Ph. D., associate professor. Her main research interests include image steganalysis and granular computing.

LUO Xiang-Yang, Ph. D., professor, Ph. D. supervisor. His main research interests include image steganography and steganalysis.

Background

BackgroundSteganalysis is the counter-steganography domain that aims to detect the existence of steganography within a cover, in order to discover, prevent or disrupt malicious covert communication based on steganography. With the rapid development of Internet and digital image processing technology, image steganalysis technology has received extensive attention from scholars at home and abroad.

Traditional image steganalysis techniques rely on manually designed feature and their performance is not satisfactory. The current state-of-the-art approach, which is based on deep neural network models, solves the problem of manually designing feature and usually achieve superior detection performance. However, the existing image steganalysis models can only detect certain image format steganography, and their detection performance is often difficult to be satisfactory in both spatial and JPEG image steganography.

Therefore, in this paper, a Siamese neural network and cross-attention-based approach, named CES-Net, is proposed for spatial and JPEG image steganalysis to enhance the model's performance in different image formats. Extensive experiments

demonstrate that our model can achieve advanced detection performance in spatial and JPEG image steganography.

This research group has been working on image steganography and steganalysis for a long time, with numerous publications in related fields such as: "A siamese inverted residuals network image steganalysis scheme based on deep learning", "steganographer identification of JPEG image based on feature selection and graph convolutional representation", and "steganalysis feature selection with multidimensional evaluation and dynamic threshold allocation".

This work has been supported by the Natural Science Foundation for Excellent Young Scholars of Henan Province (No. 252300421233, No. 222300420058), the National Natural Science Foundation of China (No. U23A20305, No. 62172435, No. 62202495), the National Key Research and Development Program of China (No. 2022YFB3102900), the Innovation Scientists and Technicians Troop Construction Projects of Henan Province (No. 254000510007), and the Key Research and Development Project of Henan Province (No. 221111321200).