

【电子与信息科学 / Electronics and Information Science】

基于CNN-Transformer的图像隐写分析

王嘉昊^{1,2}, 阎红灿^{1,2}, 谷建涛^{1,2}

1) 华北理工大学理学院, 河北唐山 063210; 2) 河北省数据科学与应用重点实验室, 河北唐山 063210

摘要: 现有的卷积神经网络(convolutional neural network, CNN)隐写分析模型主要关注隐写图像的局部特征, 尽管CNN通过堆叠更深的卷积层来扩展感受野, 但在全局特征提取方面仍较为薄弱, 而对于大尺寸图像, 关注全局特征有助于提高隐写分析的效果. 提出一种基于CNN-Transformer的图像隐写分析网络模型(CNN-Transformer image steganalysis network, CTS-Net), 能够有效捕捉隐写信号局部特征与全局特征的依赖关系. 在预处理阶段, 采用多尺度残差提取与信息融合, 以提高信噪比; 在特征提取阶段, 结合CNN和Transformer进行局部与全局特征提取, 提高了对大尺寸隐写图像的检测精度; 最后采用全连接层进行分类. 在公开数据集BOSSbase1.01上采用WOW、HILL和S-UNIWARD隐写算法, 使用不同嵌入率进行检测, 结果表明, 在低嵌入率(0.1 bpp)下, CTS-Net模型的检测准确率最优; 在公开数据集ALASKA#2上采用WOW隐写算法, 对16种图像尺寸的数据集进行隐写分析表明, CTS-Net模型在固定大小数据集与异构数据集上能有效利用隐写信号的全局特征, 检测准确率高于SRNet、SiaStegNet、CvT Net和CVTS模型.

关键词: 人工智能; 卷积神经网络; 图像隐写分析; 多尺度残差; 全局特征; 大尺寸图像

中图分类号: TP309

文献标志码: A

DOI: 10.3724/SP.J.1249.2025.02233

Image steganalysis based on CNN-Transformer

WANG Jiahao^{1,2}, YAN Hongcan^{1,2}, and GU Jiantao^{1,2}

1) College of Science, North China University of Science and Technology, Tangshan 063210, Hebei Province, P. R. China

2) Key Laboratory of Data Science and Application of Hebei Province, Tangshan 063210, Hebei Province, P. R. China

Abstract: Current convolutional neural network (CNN) steganalysis models primarily focus on the local features of steganographic images. Although CNNs expand their receptive field by stacking deeper convolutional layers, their ability to extract global features remains limited. For large images, focusing on global features can significantly improve steganalysis performance. We propose a hybrid model named CTS-Net (CNN-Transformer image steganography network) for image steganalysis. This model effectively captures both local and global features dependencies of the steganographic signals. In the preprocessing stage, multi-scale residual extraction and information fusion are applied to improve the signal to noise ratio. In the feature extraction stage, CNN and Transformer are combined to extract both local and global features, enhancing detection accuracy for large size steganographic images. Finally, a fully connected layer is used for classification. Experiments on the public dataset BOSSbase1.01, using the WOW, HILL, and S-UNIWARD steganography algorithms at different embedding rates for detection, show that at a low embedding rate (0.1 bpp), the CTS-Net model achieves the best detection accuracy. On the public dataset ALASKA#2, the WOW steganography algorithm is used for steganalysis across 16 image sets of different sizes. The results demonstrate that the CTS-Net model effectively leverages global features of the

Received: 2024-07-09; **Accepted:** 2024-10-14; **Online (CNKI):** 2024-12-31

Foundation: Social Science Foundation of Hebei Province (HB17GL071); Key Research Projects of North China University of Science and Technology (ZD-YG-202316); Higher Education Teaching Reform Research and Practice Project of Hebei Province (2023GJJG226)

Corresponding author: Professor YAN Hongcan (yanhcan@163.com)

Citation: WANG Jiahao, YAN Hongcan, GU Jiantao. Image steganalysis based on CNN-Transformer [J]. Journal of Shenzhen University Science and Engineering, 2025, 42(2): 233-241. (in Chinese)



steganographic signals on both fixed-size and heterogeneous datasets, achieving superior detection accuracy compared to SRNet, SiaStegNet, CvT Net, and CVTS.

Key words: artificial intelligence; convolutional neural network; image steganalysis; multi-scale residuals; global feature; large size image

图像隐写是一种信息隐藏技术,目的是通过对数字图像进行细微修改,嵌入秘密信息,利用图像中的隐藏通道实现信息的隐蔽传输.图像隐写分析是一种用于检测、分析和识别图像中隐藏信息的方法和工具,它采用统计分析、数字信号处理和机器学习等方法分析隐写图像和载体图像之间的统计差异,并通过增加特征数量和增强分类器性能等方法来提高检测精度^[1].近年来,随着更为复杂的自适应隐写术被陆续提出,图像隐写分析技术也面临着更大的挑战,图像隐写分析的研究与发展变得尤为重要.

传统图像隐写分析方法可分为专用型隐写分析和通用型隐写分析.为应对非自适应隐写算法, KER^[2]提出两种应用直方图特征函数(histogram characteristics function, HCF)的新方法,以分析最低有效位(least significant bit, LSB)隐写算法. LIU等^[3]提出基于特征挖掘和模式识别的隐写分析算法,提取了5种类型的特征,实现了更好地检测最低有效位匹配(LSB matching, LSBM)隐写.为应对更加复杂的自适应隐写术, FRIDRICH等^[4]提出空域富模型(spatial rich model, SRM)方法,使用一组特定设计的滤波器来捕获图像的局部结构和纹理信息,再经过量化与截断提取共生矩阵,使用多个子模型提取出更多类型的特征.该模型能够更好地表征隐写对邻域像素多种相关性的破坏.

传统的隐写分析模型需要人工设计分类特征,这要求具备特别高超的先验知识且效率不高,而深度学习在特征提取方面具有优势^[5],采用端到端的学习方式,能够自主学习丰富的特征,分类结果更准确^[6]. TANG等^[7]首次将图像隐写与卷积神经网络相结合,提出的TanNet模型由卷积层以及全连接层堆叠而成,但分类结果准确率并不高,性能也不如SRM. QIAN等^[8]提出一种基于高斯卷积神经网络(Gaussian-neuron convolutional neural network, GNCNN)的隐写分析模型,使用固定的高通滤波器提取残差信息,使得网络更关注隐写噪声而非图像本身,并在卷积层后引入非线性的高斯激活函数,更好地适配隐写信号的特征. GNCNN模型的分类效率比TanNet高,但不如SRM. XU等^[9]改进了

GNCNN模型,在卷积层中嵌入绝对激活层(absolute activation layer, ABS),以改进后续层中的统计建模,在网络的早期阶段应用双曲正切函数tanh作为激活函数以防止过度拟合,并在每个非线性激活层前先执行批量归一化(batch normalization, BN),从而构建出新的用于图像隐写分析的卷积神经网络Xu-net,性能上首次超过了SRM. YE等^[10]利用SRM在残差映射计算上使用的基本高通滤波集进行初始化,作为正则化器有效地抑制了图像内容,为更好地捕捉嵌入信号的结构,采用截断线性单元的激活函数,更好地适应噪声残差.该模型在精度上明显优于目前最先进的手工特性集SRM. BOROUMAND等^[11]提出深度残差隐写分析网络(steganalysis residual network, SRNet),该模型由12层组成,第1~7层用于提取隐写嵌入时引入的高频噪声,第8~11层用于缩小特征图的维度,因此获得了很高的检测准确率. ZHANG等^[12]提出Zhu-net模型,使用了30个SRM高通滤波器进行残差提取(5个 5×5 的滤波器,25个 3×3 的滤波器),并使用可分离卷积与特征金字塔来提取更具代表性的信息. YOU等^[13]将孪生网络用于图像隐写分析,提出孪生隐写分析网络(siamese steganalysis network, SiaStegNet),在检测大尺寸图像时具有一定效果.

基于纯CNN的隐写分析网络只能捕捉隐写图像的局部特征,无法有效提取全局特征向量.对于大尺寸的隐写图像而言,综合考量全局信息能够更好地区分载体图像(cover)与载密图像(stego).视觉转换器^[14](vision transformer, ViT)由于具有自注意力机制,能够捕获序列元素之间的长距离依赖关系,但无法用于图像隐写分析任务.基于此,本研究提出一种结合CNN和Transformer结构的图像隐写分析模型,旨在突破隐写分析无法有效利用全局信息的瓶颈,高效利用隐写图像局部与全局的依赖关系,从而提高模型的检测准确率.本研究主要贡献包括:①提出一种基于CNN-Transformer的图像隐写分析网络模型(CNN-Transformer image steganalysis network, CTS-Net),从局部与全局角度有效地提取隐写图像的特征向量;②构建一种充分提

取噪声残差的预处理层, 通过从浅层到深层的多尺度残差提取, 强调残差信息的通道相关性与全局性; ③ 在自注意力机制中整合多尺度特征, 实现提高隐写分析准确率的同时, 轻量化自注意力机制; ④ 对不同图像隐写分析算法在固定图像尺寸的数据集上进行对比分析, 验证 CTS-Net 模型的有效性, 同时在异构图像数据集上进行隐写分析. 结果表明, CTS-Net 模型在检测大尺寸图像上表现出优越的准确率.

1 视觉转换器

Transformer 编码器因为其独特的多头自注意力机制(multi-head self-attention, MHSA), 可捕获特征之间的全局依赖, 因此在计算机视觉^[14]和自然语言处理^[15]等领域开展下游任务时的效率均高于 CNN. 与典型的分类问题相比, 图像隐写分析任务需要检测图像中隐藏的噪声信号, 这种隐写变化很难从视觉上直接观察, 且自适应隐写术通过选择性嵌入数据, 使其对图像的统计特征影响极小, 这更加大了图像隐写分析的难度^[16], 这更加大了图像隐写分析的难度. 基于此, 考虑将全局特征的提取引入到图像隐写分析任务中, 充分利用隐写特征来提高检测准确率.

ViT 通过将 Transformer 应用于图像图块序列上来完成图像分类任务^[14]. 其结构如图 1. 其中, 0~9 表示位置嵌入; * 表示额外的类别嵌入.

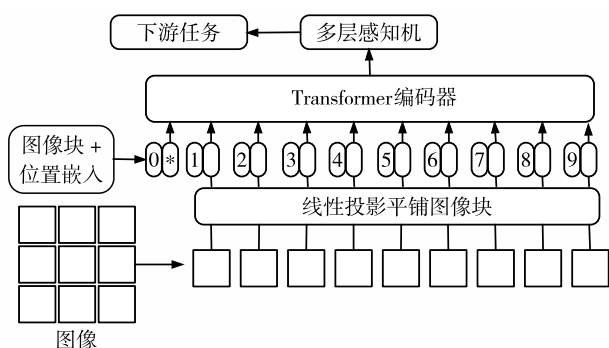


图1 ViT 结构

Fig. 1 ViT structure.

将输入图像数据划分为多个小的图片块, 然后对每个图片块进行线性变换转化为序列特征, 此序列特征平铺后变为可输入到 Transformer 编码器的标记(token), 再经 Transformer 进行特征提取. 因此, Transformer 的多头自注意力机制可表述为

$$\text{MHSA}(X) = \text{concat}(h_1, h_2, \dots, h_m)W \quad (1)$$

$$h_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \quad (2)$$

其中, X 为 Transformer 编码器的输入; concat 为张量拼接函数; W 为可学习的参数矩阵; W_i^Q 、 W_i^K 和 W_i^V 分别为第 i 个注意力头中的可学习的参数矩阵;

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (3)$$

其中, $\sqrt{d_k}$ 为避免内积过大的尺度因子; softmax 为归一化函数.

ViT 使得 Transformer 可更便捷地处理图像数据, 结合 CNN 可有效的进行局部与全局特征的提取.

2 CTS-Net 混合模型

CTS-Net 混合模型用于图像隐写分析时, 利用 Transformer 特有的多头自注意力机制, 进行隐写信号全局特征的提取, 结合卷积层提供的局部特征, 实现对隐写噪声局部与全局的特征提取, 从而提高模型的检测能力. 图 2 为 CNN-Transformer 混合模型的结构框架, 主要由预处理层、特征提取和分类 3 部分组成.

2.1 预处理层

与经典计算机视觉领域中的物体识别和图像分类等任务不同, 隐写分析主要关注细微的隐写噪声而非一般的图像内容^[17]. 隐写图像可看作是在一个图像像素上叠加上一个隐写噪声信号, 在特征提取前对图像进行预处理, 有利于提高信噪比, 因此, 采用噪声残差代替原始像素建模有助于加速网络的收敛.

首先, 采用 SRM 的 30 个滤波器对图像进行初始化, 包含 5 个 5×5 的滤波器(4 个 EDGE 5×5 与 1 个 SQUARE 5×5), 其余 25 个是 3×3 的高通滤波器, 应用小尺寸卷积内核可减少参数数量. 接着使用改进的倒置残差块(inverted residual block, IRB)^[18]进行浅层噪声残差提取, 从 1×1 的点卷积开始, 经轻量级的深度卷积, 将卷积核应用于各个信道, 再经过另一个点卷积混合这些信道信息. 最后, 使用残差连接控制梯度消失.

为强调更为关键的残差信息, 抑制图像内容, 在最后的点卷积前引入高效通道注意(efficient channel attention, ECA)^[19]模块. ECA 结构如图 3,

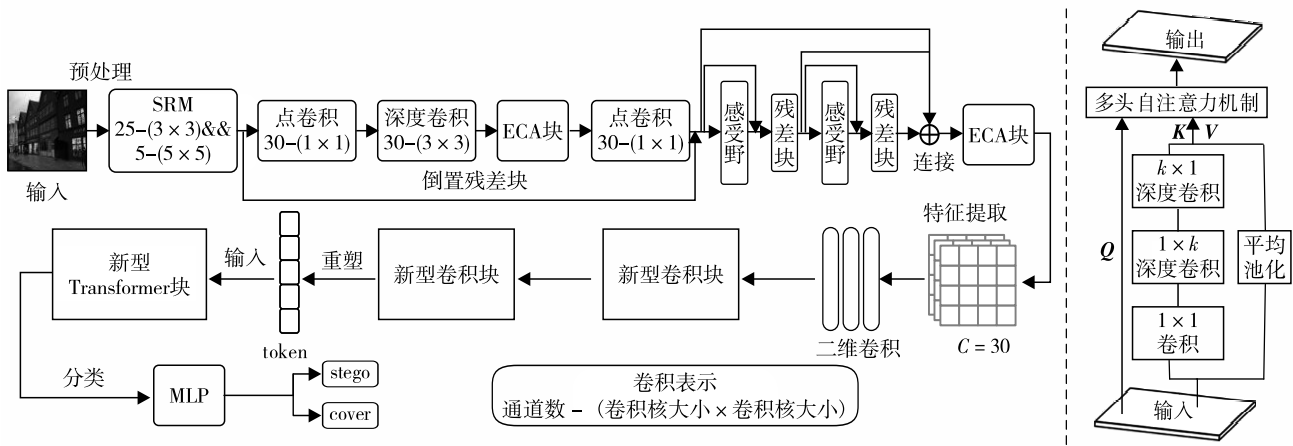
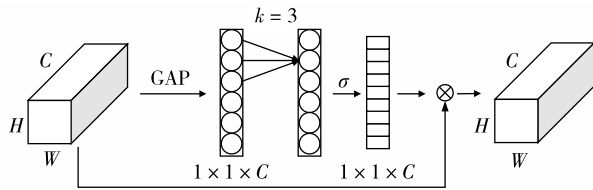


图2 CTS-Net结构

Fig. 2 CTS-Net structure diagram.

它可自适应地调整通道特征的权重, 将输入的高度(H)、宽度(W)和通道数(C)特征映射传递给全局平均池化(global average pooling, GAP), 获得聚合特征 $1 \times 1 \times C$, 代替通道注意力模块SE(squeeze-and-excitation)中使用的全连接层. ECA模块通过执行卷积核大小为 k 的一维卷积来生成通道权重, 大大减少了参数量. 其中, k 由通道数 C 的映射自适应确定. 通过缩放因子 σ 进行归一化, 再将 $1 \times 1 \times C$ 的权重通道作用在原特征映射上. 改进的倒置残差块不仅关注不同通道的全局信息, 也实现了不同通道的加权组合.

图3 ECA结构图^[19]Fig. 3 ECA structure diagram^[19].

最后, 将感受野模块(receptive field block, RFB)^[20]引入残差网络, 设计一个新型多尺度残差提取器, 进行深层噪声残差提取. RFB结构如图4. 其中, $C_{1 \times 1}$ 、 $C_{1 \times 3}$ 、 $C_{3 \times 1}$ 和 $C_{3 \times 3}$ 分别为卷积核为 1×1 、 1×3 、 3×1 和 3×3 的卷积函数; $D_{3 \times 3}$ 分别为卷积核为 3×3 的膨胀卷积函数; r 为膨胀率. RFB使用多分支卷积层来提取多尺度的残差信号, 并使用膨胀卷积模拟更大的感受野, 最终所有分支的残差向量被连接起来, 合并成一个卷积数组. 另外, 在RFB中使用 1×1 与 3×3 的小尺度卷积核, 代替RFB原有的 5×5 卷积核, 以此来减少参数.

2.2 特征提取

与以往的图像隐写分析不同, 本研究在特征提取阶段使用ViT架构来提取隐写信号的全局特征, ViT可以模拟局部和全局噪声残差对隐写图像的依赖性, 这对于自适应隐写术尤为重要, 全局特征的提取有助于增强模型的建模能力, 提高对隐写信息的理解和分析效果.

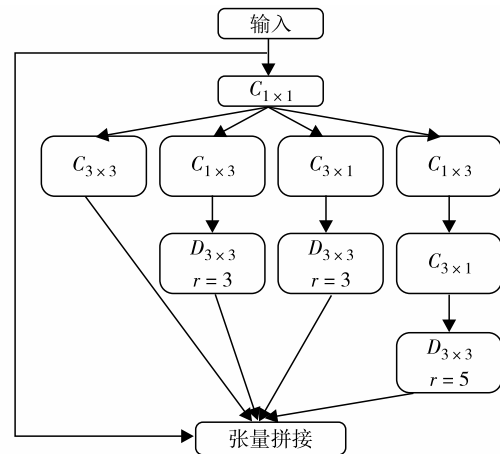


图4 RFB结构图

Fig. 4 RFB structure diagram.

2.2.1 新型卷积块

新型卷积块(new convolution block, NCB)不仅能够保持残差结构的快速推理能力, 还能获得Transformer块高准确率的突出性能. NCB使用多头卷积注意力(multi-head convolution attention, MHCA)^[21]代替MHSA机制作为令牌混合器(token mixer), 并在MetaFormer的范式中使用MHCA和多层感知机(multi-layer perceptron, MLP)层共同构建NCB块. NCB的结构见图5. 其中, $\oplus$ 表示元素级

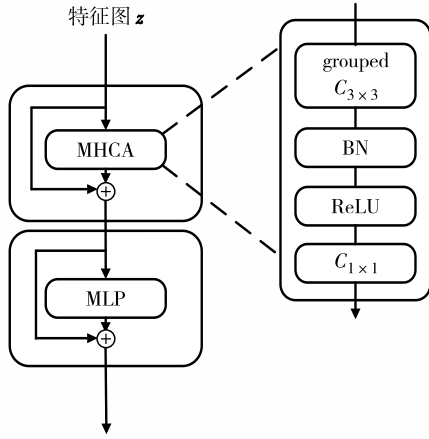


图5 NCB结构图

Fig. 5 NCB structure diagram.

加法. 由图5可见, NCB块的实现主要依靠群卷积与点卷积. MHCA通过多头卷积范式关注不同通道的空间信息, 实现有效提取局部特征. MHCA表达式为

$$\text{MHCA}(z) = \text{concat}(A_1(z_1), A_2(z_2), \dots, A_m(z_m)) \quad (4)$$

其中, 输入特征 $z = [z_1, z_2, \dots, z_m]$, 这 m 个维度构成多头划分; A_i 为第 i 个 ($i = 0, 1, \dots, m$) 单头卷积注意. MHCA的多头形式依靠群卷积进行模拟, 实现局部信道的交互, 使用BN和ReLU激活函数来进一步加快推理速度, 最后, 配备点卷积将所有信道进行融合.

2.2.2 新型Transformer块

虽然NCB已经有效地学习了局部特征, 但仍无法学习噪声信号的全局信息, 因此, 本研究引入了新型Transformer块(new Transformer block, NTB)来提取隐写信号的全局特征. NTB结构如图6. 特征向量在进入NTB之前, 需要先将输入平铺为Transformer可读取的token输入, 接着经过高效多头自注意力(highly efficient multi-head self-attention, HE-MHSA)机制捕获全局信号.

HE-MHSA机制是本研究针对Transformer多头自注意力复杂的计算问题, 提出的一种双分支并行方法, 可简化自注意力机制的计算量. HE-MHSA

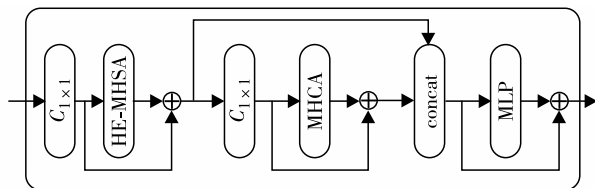


图6 NTB结构图

Fig. 6 NTB structure diagram.

的结构如图2虚线右边部分所示, 对于输入的特征 $X \in \mathbf{R}^{H \times W \times C}$, 在 X 上应用双分支结构生成新的输入形式. 针对第1个分支, 对 X 进行空间可分离卷积操作. 首先进行 1×1 的卷积操作, 以减少乘法的运算量, 再与 $1 \times k$ 的核大小进行深度卷积(depth wise convolution, DWC), 最后与核大小为 $k \times 1$ 的卷积核进行深度卷积. 针对第2个分支, 对 X 进行平均池化操作, 池化窗口大小为 $k \times k$, 将卷积的步幅设置为 k , 因此, 第2分支的输出在空间上表现为 $k \times k$ 的降采样. 最后将两个分支进行合并. 经过双分支操作, X 变为形式 $H/k \times W/k \times C$, 铺平为token序列时变短, 计算量大大降低且具有丰富的特征. 上述过程可表述为式(5)至式(7).

$$C_1 = \text{DWC}_{k \times 1}[\text{DWC}_{1 \times k}(X)] \quad (5)$$

$$C_2 = \text{AvgPool}(X) \quad (6)$$

其中, C_1 与 C_2 为两个分支的输出; $\text{DWC}_{k_1 \times k_2}$ 为与核大小 $k_1 \times k_2$ 的深度卷积; AvgPool 为平均池化操作. 最后, 平铺并连接这些特征图, 可得

$$O = \text{LayerNorm}(\text{concat}(C_1, C_2)) \quad (7)$$

其中, LayerNorm 为层级归一化函数.

如此得到的token序列 O 比原输入 X 铺平时更短, 且 O 还包含了对输入 X 的多尺度上下文信息, 因此, 可代替 X 作为新的输入形式.

在新的输入形式下, 自注意力机制的查询张量(Q)、键值张量(K)和值张量(V)分别为

$$Q = XW^Q \quad (8)$$

$$K = OW^K \quad (9)$$

$$V = OW^V \quad (10)$$

其中, W^Q 、 W^K 和 W^V 为可学习的线性变换权重矩阵. 自注意力机制可表示为

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (11)$$

HE-MHSA机制可有效关注隐写特征的全局性, 在HE-MHSA后, 再接入1个MHCA模块, 则来自HE-MHSA和MHCA的输出特征被连接起来以混合高低频信息, 实现有效提取残差信号的局部-全局特征.

3 实验与分析

3.1 数据集以及实验平台

实验采用的数据集为 BOSSbase1.01^[22] 和

ALASKA#2^[23]. BOSSbase1.01 由 1×10^4 幅灰度图像组成, 使用 Matlab 软件将分辨率为 512×512 像素的原始图片裁剪为 256×256 像素的新数据集, 使用其中的 1.2×10^4 幅作为训练集, 载体图像和载密图像各 6×10^3 幅, 2×10^3 幅作为验证集, 6×10^3 幅作为测试集.

ALASKA#2 是隐写分析挑战赛中提供的数据集, 包括 ALASKA_v2_TIFF_512 和 ALASKA_TIFF_VariousSize 子集. 其中, ALASKA_v2_TIFF_512 包含了 80 005 幅分辨率为 512×512 像素的原生图像. ALASKA_TIFF_VariousSize 包含了 16 种不同分辨率的图像, 从 ALASKA_v2_TIFF_512 中随机选择 1.2×10^4 幅图片作为训练集, 2×10^3 幅图像作为验证集, 然后从 ALASKA_TIFF_VariousSize 的 16 组图像中选择每组 750 幅作为测试集. 所有训练集、验证集和测试集图像互不重叠.

采用 WOW、S-UNIWARD 和 HILL 这 3 种自适应隐写术对图像进行隐写, 嵌入率分别设为 0.4 bpp (每个像素所占用的有效比特数, bits per pixel) 与 0.2 bpp. 在训练 BOSSbase1.01 数据时, 实验环境为 Windows11, 内存为 16 Gbyte, 编程环境为 Python3.9.12. 由于图形处理器 (graphics processing unit, GPU) 的限制, 训练 ALASKA#2 图像数据时实验环境在 Tesla V100 上进行.

3.2 实验参数

由于 GPU 的限制, 实验设置批次为 16, 初始学习率为 1×10^{-4} , 在迭代 100 次后, 学习率降低为 0.1, 最大迭代次数为 500, 且使用了数据增强方法, 包括随机翻转与随机旋转. 图 2 中二维卷积操作、NCB 与 NTB 具体参数如表 1. 其中, C 为通道数; S 为步长; G 为组数; m 为注意力头数量.

3.3 实验模型对比

隐写分析残差网络 (steganalysis residual network, SRNet)^[11] 是一种使用随机初始化卷积核以端

到端方式进行训练的隐写分析模型, 它主要由 4 种结构的卷积块组成, 前 7 层负责噪声残差的提取, 后面部分负责减少特征图的维度, 对空域图像具有较好的检测性能.

孪生隐写分析网络 (siamese steganalysis network, SiaStegNet)^[13] 是一种基于孪生网络的图像隐写分析模型. 该模型将图像划分为左右两部分, 通过融合信号与相似度信号来训练网络, 从而能够高效地进行隐写分析, 且适用于不同分辨率的图像.

CvT Net^[24] 与 CVTStego-Net^[25] (简称 CVTS) 模型均采用 CNN 与 Transformer 相结合的方式进行空域图像隐写分析. 其中, CvT Net 模型使用 Resnet-50 和 Resnet-18 结合通道注意力机制进行预处理, 在特征提取阶段采用卷积 Transformer 提取全局特征. CVTS 采用两种不同的卷积组合块以及通道注意力机制进行特征提取, 在分类阶段使用 ViT 捕获特征的全局依赖关系. 本研究提出的 CTS-Net 在预处理阶段采用了多尺度的噪声残差提取方式, 利用 RFB 膨胀卷积的特性, 模拟全局性的残差提取, 另外 ECA 机制可简单地快速响应卷积层之间的关联, 以获得更高的信噪比. 在特征提取阶段, HE-MHSA 因配备了 MHCA, 能够充分混合高低频信息, 获得较高的检测准确率.

3.4 实验结果与分析

3.4.1 消融实验

为验证 CTS-Net 每个阶段的贡献, 在 0.2 嵌入率下使用 WOW 隐写算法对不同阶段进行实验, 实验结果如表 2. 由表 2 可见, 无预处理阶段, 直接使用特征提取进行隐写分析的检测准确率为 74.77%, 表明多尺度残差提取对后续工作至关重要. 此外, 为突出 Transformer 在隐写分析中的贡献, 不使用 NTB 模块的情况下对网络进行训练, 检测准确率达到 80.15%, 与整体结构的检测结果相比, 这一阶段对模型的贡献率约为 4.5%, 可见, NTB 模块可有效提取并结合隐写信号的局部特征与全局特征, 提高隐写图像的检测准确率, 适用于图

表 1 实验参数设置

Table 1 Experimental parameter settings

层次	参数设置
Conv2D	$C_{3 \times 3}, C = 64, S = 2$
	$C_{3 \times 3}, C = 32, S = 1$
	$C_{3 \times 3}, C = 64, S = 1$
	$C_{3 \times 3}, C = 64, S = 2$
NCB	$C_{3 \times 3}, C = 96, S = 1, G = 3$
NCB	$C_{3 \times 3}, C = 192, S = 1, G = 6$
NTB	$C = 256, k = 8, m = 6$

表 2 不同网络结构的检测准确率、训练时间和推理时间

Table 2 Detection accuracy, training time and inference time of different network structures

模型结构	准确率/%	训练时间/h	推理时间/s
去掉预处理阶段	74.77	2.44	3.54
去掉 NTB 模块	80.15	5.73	3.97
CTS-Net	84.55	6.53	4.33

像隐写分析.

3.4.2 在 BOSSbase1.01 数据集上的比较

为验证 CTS-Net 在固定数据集上的有效性, 与 SRNet 和 SiaStegNet 两种网络结构在不同自适应隐写算法和不同嵌入率下进行对比分析, 嵌入率单位 bpp 表示平均每个像素可以嵌入的有效比特数. 实验结果如表 3.

表3 不同隐写算法的检测准确率

Table 3 Detection accuracy of different steganography algorithms		嵌入率			%
隐写算法	隐写分析模型	0.1 bpp	0.2 bpp	0.4 bpp	
WOW	SRNet	70.65	78.60	86.45	
	SiaStegNet	74.38	83.24	91.06	
	CTS-Net	75.42	84.55	90.95	
HILL	SRNet	62.18	69.76	78.35	
	SiaStegNet	67.56	75.28	84.33	
	CTS-Net	68.87	76.30	85.10	
S-UNIWARD	SRNet	65.12	73.35	82.27	
	SiaStegNet	71.66	81.02	87.00	
	CTS-Net	73.45	80.95	88.58	

注: 灰底数据为该指标下的最优值

由表 3 可见, 在训练集、验证集和测试集数量相同的情况下, CTS-Net 模型在使用 WOW 隐写算法且嵌入率为 0.2 bpp 时的检测准确率为 84.55%, 分别比 SRNet 模型和 SiaStegNet 模型高出 5.95% 和 1.31%; 在使用 S-UNIWARD 隐写算法且嵌入率为 0.4 bpp 的检测准确率为 88.58%, 比 SiaStegNet 模型高出 1.58%; 对于 HILL 隐写算法, 在嵌入率为 0.4 bpp 和 0.2 bpp 时, CTS-Net 模型的准确率分别比 SiaStegNet 模型高出 0.77% 和 1.02%. 在低嵌入率(如 0.1 bpp)下, 针对 3 种不同隐写算法, CTS-Net 的检测准确率均优于 SRNet 模型以及 SiaStegNet 模型, 特别是使用 S-UNIWARD 隐写算法时, CTS-Net 模型的检测准确率相比 SiaStegNet 模型提高了 1.79%. 在较低嵌入率下训练模型时, 可以预先加载高嵌入率下迭代 20 次所得的模型参数, 从而提升检测准确率.

本研究还在模型参数数量和推理速度方面对比了 CTS-Net、SRNet 和 SiaStegNet 的性能. 在嵌入率为 0.2 bpp 的条件下, 使用 WOW 隐写算法进行实验, 结果如表 4. 从表 4 可见, CTS-Net 模型在多个方面优于 SRNet 和 SiaStegNet 模型, CTS-Net 模型的检测准确率比 SiaStegNet 模型高出 1.31%, 且参数量仅

为 0.823×10^6 个, 远低于 SRNet 模型的 4.776×10^6 个, 同时, 本研究方法大幅缩短了模型的训练时间, 仅需要 6.53 h. 在推理时间方面, 本研究方法也表现出更高的效率, 仅需 4.33 s, 明显快于 SRNet 和 SiaStegNet 模型.

表4 不同模型的准确率、参数量、训练时间和推理时间

Table 4 Accuracy, number of parameters, training time and reasoning speed of different models

隐写分析模型	准确率/%	$10^{-6} \times$ 参数量/个	训练时间/h	推理时间/s
SRNet	78.60	4.776	18.55	7.91
SiaStegNet	83.24	0.709	13.65	5.08
CTS-Net	84.55	0.823	6.53	4.33

综上所述, CTS-Net 模型针对自适应隐写术具有一定的泛化能力, 并且针对不同的有效载荷具有较高的检测准确率和较短的推理时间, 适用于图像隐写分析.

固定图像尺寸(256 × 256 像素)的数据集未能使 Transformer 的优越性得以显现, 因此, 本研究拟在 16 种不同大小的数据集上进行对比实验, 以检测本研究方法全局特征的提取能力.

3.4.3 在 ALASKA#2 数据集上的比较

为验证 CTS-Net 模型的预处理方法较 CvT Net 模型和 CVTStego-Net 模型的优越性, 比较了 50 幅载密图像经过各个网络预处理之后隐写信号与噪声之间的信噪比(signal to noise ratio, SNR), 实验结果如图 7.

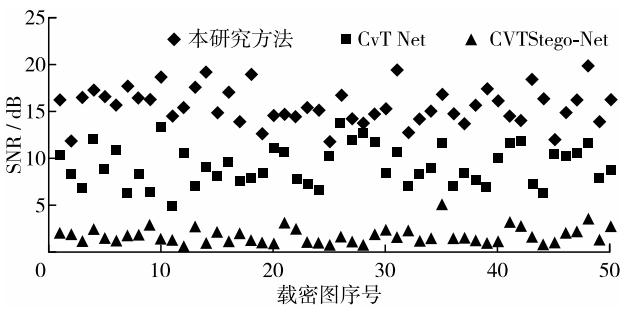


图7 不同预处理结果的 SNR 对比

Fig. 7 SNR comparison of CTS-Net (rhomboid) and CvT Net (square) and CVTStego-Net (triangle) pretreatment results.

从图 7 可见, CTS-Net 模型预处理后的 SNR 值显著高于 CvT Net 模型和 CVTStego-Net 模型, 主要因为 CTS-Net 模型中的多尺度残差提取, 通过浅层至深层逐步提取局部与全局的残差信息, 并结合 ECA 机制放大关键的信道特征, 同时有效抑制了图像原始内容的干扰.

此外, 为验证 CTS-Net 模型在不同分辨率数据集上的有效性, 分别与 SiaStegNet、CvT Net 和 CVTS 隐写分析模型进行实验对比, 这 3 种方法均适用于检测不同分辨率的隐写图像. 实验采用 WOW 自适应隐写算法, 嵌入率为 0.4 bpp, 训练集采用分辨率为 512×512 像素的 1.2×10^4 幅图像数据, 包括载体图像和载密图像各 6×10^3 幅, 其余 15 种分辨率的图像则直接进行测试, 并且根据平方根定律, 调整不同分辨率图像的有效嵌入率, 以获得恒定的统计检测性, 实验结果如表 5.

表 5 不同大小图像的检测准确率

Table 5 Detection accuracy of images of different sizes

图像大小/ 像素	嵌入 率/bpp	SiaSteg- Net/%	CvT Net/%	CVTS/ %	CTS- Net/%
512×512	0.40	73.85	71.25	70.88	72.30
512×640	0.35	71.56	67.55	68.62	69.77
640×512	0.35	71.04	68.09	69.11	70.55
512×720	0.35	73.65	69.30	70.57	71.26
720×512	0.35	73.50	70.42	71.28	73.68
640×640	0.34	74.33	69.10	70.30	70.28
640×720	0.31	74.64	70.68	72.32	74.70
720×640	0.31	71.28	68.90	69.08	72.85
720×720	0.29	71.30	68.92	68.73	71.89
512×1024	0.29	69.78	67.30	68.06	70.55
1024×512	0.29	69.85	68.65	67.34	71.35
640×1024	0.28	70.05	67.28	66.46	70.10
1024×640	0.28	69.23	66.85	66.13	70.85
720×1024	0.25	67.56	65.20	67.04	69.90
1024×720	0.25	66.33	64.72	63.89	69.05
1024×1024	0.22	64.28	63.95	64.37	69.50

注: 灰底数据为该指标下的最优值

由表 5 可见, 在图像较小时 ($512 \times 512 \sim 640 \times 640$ 像素), CTS-Net 模型的检测率不及 SiaStegNet 模型, 主要因为 Transformer 块提取全局信息的能力在小图像上发挥不充分; 随着图像增大, CTS-Net 模型的优越性愈加明显. 得益于 NCB 模块的多头卷积注意力与 NTB 模块的多头自注意力机制相结合, CTS-Net 模型能够有效提取隐写信号的全局特征并充分整合局部信息, 从而在图像尺寸为 640×720 像素至 1024×1024 像素时的检测准确率始终优于其他 3 种网络, 特别在 1024×1024 像素分辨率下, 检测准确率比 SiaStegNet 模型和 CVTS 模型提升了 5%, 比 CvT Net 提升了 5.5%.

结 语

设计了一种基于 CNN-Transformer 的图像隐写分析模型. 首先, 改进的倒置残差块用于残差信号的浅层提取, 配合 ECA 模块关注噪声残差的全局重要性. 其次, RFB 模块结合残差结构, 进行多尺度残差信息提取与融合. 最后, 采用轻量级的 Transformer 块结合 CNN 架构对噪声信号进行局部-全局的特征提取, 通过在 BOSSbase1.01 和 ALASKA#2 数据集上使用不同隐写算法以及不同嵌入率进行实验对比, 证明了本研究方法的有效性. 下一步的研究方向是提高低嵌入率下的准确率以及减少模型参数, 在轻量化的同时提高检测准确率.

基金项目: 河北省社会科学基金资助项目(HB17GL071); 华北理工大学科研重点资助项目(ZD-YG-202316); 河北省高等教育教学改革研究与实践资助项目(2023GJJG226)

作者简介: 王嘉昊(774634774@qq.com), 华北理工大学硕士研究生. 研究方向: 图像隐写分析、图像处理.

引 文: 王嘉昊, 阎红灿, 谷建涛. 基于 CNN-Transformer 的图像隐写分析[J]. 深圳大学学报理工版, 2025, 42(2): 233-241.

参考文献 / References:

- [1] 黄思远, 张敏情, 柯彦, 等. 基于显著性检测的图像隐写分析方法[J]. 计算机应用, 2021, 41(2): 441-448.
HUANG Siyuan, ZHANG Mingqing, KE Yan, et al. Image steganalysis method based on saliency detection [J]. Journal of Computer Applications, 2021, 41(2): 441-448. (in Chinese)
- [2] KER A D. Steganalysis of LSB matching in grayscale images [J]. IEEE Signal Processing Letters, 2005, 12(6): 441-444.
- [3] LIU Qingzhong, SUNG A H, CHEN Zhongxue, et al. Feature mining and pattern classification for steganalysis of LSB matching steganography in grayscale images [J]. Pattern Recognition, 2008, 41(1): 56-66.
- [4] FRIDRICH J, KODOVSKY J. Rich models for steganalysis of digital images [J]. IEEE Transactions on Information Forensics and Security, 2012, 7(3): 868-882.
- [5] 彭伟, 胡宁, 胡璟璟. 图像隐写分析算法研究概述[J]. 计算机科学, 2020, 47(增刊1): 325-331.
PENG Wei, HU Ning, HU Jingjing. Overview of research on image steganalysis algorithms [J]. Computer Science, 2020, 47(Suppl. 1): 325-331. (in Chinese)
- [6] 陈君夫, 付章杰, 张卫明, 等. 基于深度学习的图像

- 隐写分析综述[J]. 软件学报, 2021, 32(2): 551-578.
- CHEN Junfu, FU Zhangjie, ZHANG Weiming, et al. Review of image steganalysis based on deep learning [J]. Journal of Software, 2021, 32(2): 551-578. (in Chinese)
- [7] TANG Weixuan, LI Haodong, LUO Weiqi, et al. Adaptive steganalysis against WOW embedding algorithm [C]// Proceedings of the 2nd ACM Workshop on Information Hiding and Multimedia Security. New York, USA: ACM, 2014: 91-96.
- [8] QIAN Yinlong, DONG Jing, WANG Wei, et al. Deep learning for steganalysis via convolutional neural networks [C]// Media Watermarking, Security, and Forensics 2015. San Francisco, USA: SPIE, 2015: 94090J.
- [9] XU Guanshuo, WU Hanzhou, SHI Yunqing. Structural design of convolutional neural networks for steganalysis [J]. IEEE Signal Processing Letters, 2016, 23(5): 708-712.
- [10] YE Jian, NI Jiangqun, YI Yang. Deep learning hierarchical representations for image steganalysis [J]. IEEE Transactions on Information Forensics and Security, 2017, 12(11): 2545-2557.
- [11] BOROUHAND M, CHEN Mo, FRIDRICH J. Deep residual network for steganalysis of digital images [J]. IEEE Transactions on Information Forensics and Security, 2019, 14(5): 1181-1193.
- [12] ZHANG Ru, ZHU Feng, LIU Jianyi, et al. Depth-wise separable convolutions and multi-level pooling for an efficient spatial CNN-based steganalysis [J]. IEEE Transactions on Information Forensics and Security, 2020, 15: 1138-1150.
- [13] YOU Weike, ZHANG Hong, ZHAO Xianfeng. A Siamese CNN for image steganalysis [J]. IEEE Transactions on Information Forensics and Security, 2021, 16: 291-306.
- [14] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16×16 words: transformers for image recognition at scale [C]// ICLR 2021. New York, USA: International Conference on Learning Representations, 2021: 3458.
- [15] QIU Xipeng, SUN Tianxiang, XU Yige, et al. Pre-trained models for natural language processing: a survey [J]. Science China Technological Sciences, 2020, 63(10): 1872-1897.
- [16] WEI Kangkang, LUO Weiqi, TAN Shunquan, et al. CTNet: a convolutional transformer network for color image steganalysis [J]. Journal of Computer Science and Technology, 2025, 40(2): 1-15.
- [17] 范文同, 李震宇, 张涛, 等. 基于隐写噪声深度提取的JPEG图像隐写分析[J]. 西安电子科技大学学报, 2023, 50(4): 157-169.
- FAN Wentong, LI Zhenyu, ZHANG Tao, et al. JPEG image steganalysis based on deep extraction of stego noise [J]. Journal of Xidian University, 2023, 50(4): 157-169. (in Chinese)
- [18] SANDLER M, HOWARD A, ZHU Menglong, et al. MobileNetV2: inverted residuals and linear bottlenecks [C]// IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, USA: IEEE, 2018: 4510-4520.
- [19] WANG Qilong, WU Banggu, ZHU Pengfei, et al. ECA-net: efficient channel attention for deep convolutional neural networks [C]// IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, USA: IEEE, 2020: 11531-11539.
- [20] LIU Songtao, HUANG Di, WANG Yunhong. Receptive field block net for accurate and fast object detection [C]// Computer Vision - ECCV 2018. Cham Switzerland: Springer International Publishing, 2018: 404-419.
- [21] LI Jiashi, XIA Xin, LI Wei, et al. Next-ViT: next generation vision transformer for efficient deployment in realistic industrial scenarios [EB/OL]. (2022-08-16) [2024-03-21]. <https://arxiv.org/abs/2207.05501>.
- [22] BAS P, FILLER T, PEVNÝ T. "Break our steganographic system": the ins and outs of organizing BOSS [C]// Information Hiding. Berlin: Springer Berlin Heidelberg, 2011: 59-70.
- [23] COGRANNE R, GIBOULOT Q, BAS P. The ALASKA steganalysis challenge: A first step towards steganalysis [C]// Proceedings of the ACM Workshop on Information Hiding and Multimedia Security. New York, USA: ACM, 2019: 125-137.
- [24] LUO Ge, WEI Ping, ZHU Shuwen, et al. Image steganalysis with convolutional vision transformer [C]// IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway, USA: IEEE, 2022: 3089-3093.
- [25] BRAVO-ORTIZ M A, MERCADO-RUIZ E, VILLAPULGARIN J P, et al. CVTStego-net: a convolutional vision transformer architecture for spatial image steganalysis [J]. Journal of Information Security and Applications, 2024, 81: 103695.

【中文责编: 英子; 英文责编: 木柯】