中图分类号：　　单位代号：10280
密　　级：　　学　　号：20721449

# 上海大学

# 专业学位硕士学位论文

# SHANGHAI UNIVERSITY
# MASTER'S DISSERTATION

| 题目 | 基于图神经网络的<br>图像隐写分析技术研究 |
|---|---|

作　者　柳琦云
学科专业　电子信息
导　师　吴汉舟
完成日期　2023 年 6 月

姓　　名： 柳琦云　　学号：20721449

论文题目： 基于图神经网络的图像隐写分析技术研究

# 上海大学

本论文经答辩委员会全体委员审查,确认符合上海大学硕士学位论文质量要求。

答辩委员会签名：

主任：

委员：

导　师：

答辩日期：

姓　　名： 柳琦云　　学号：20721449

论文题目： 基于图神经网络的图像隐写分析技术研究

# 原 创 性 声 明

本人声明：所呈交的论文是本人在导师指导下进行的研究工作。除了文中特别加以标注和致谢的地方外，论文中不包含其他人已发表或撰写过的研究成果。参与同一工作的其他同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

签　名：_____日 期：_____

# 本论文使用授权说明

本人完全了解上海大学有关保留、使用学位论文的规定，即：学校有权保留论文及送交论文复印件，允许论文被查阅和借阅；学校可以公布论文的全部或部分内容。

（**保密的论文在解密后应遵守此规定**）

签 名：_____导师签名：_____日期：_____

上海大学工程硕士学位论文

# 基于图神经网络的
# 图像隐写分析技术研究

姓　　名：柳琦云
导　　师：吴汉舟
学科专业：电子信息

上海大学通信与信息工程学院

二〇二三年六月

A Dissertation Submitted to Shanghai University for the Degree of Master in Engineering

# Image Steganalysis Based on Graph Neural Networks

Candidate：Qiyun Liu

Supervisor：Hanzhou Wu

Major：Electronic Information

**School of Communication and Information Engineering**

**Shanghai University**

**June, 2023**

# 摘 要

图像隐写是一种以图像为载体的信息隐藏技术。图像隐写分析作为对抗图像隐写的技术，主要目的在于提取数字图像中可能存在的隐写特征，以判断其中是否包含秘密信息。当前基于深度学习的图像隐写分析算法都是结合深度卷积神经网络设计的，大量工作也证明了其适用性。近年来，图神经网络因其适用于非结构化数据而蓬勃发展，且具有能够同时捕捉局部特征和全局特征的优点。为了挖掘图神经网络提取隐写特征的潜能，本文提出将其应用于图像隐写分析，并在空域和 JPEG 域两个修改域对其进行了研究，取得的成果如下：

(1) 针对图神经网络能够更好地挖掘图像块之间关联信息的优势，本文提出了一种基于图表示学习的空域图像隐写分析算法，该算法能够利用图神经网络分析图像的全局特征以达到更好的检测性能。在具体框架中，设计了将图像转化为图的构建方法，以节点映射图像块，边表示图像块之间的局部关联。每个节点都与由浅层卷积神经网络从相应图像块确定的特征向量相关联。利用图神经网络学习图像隐写信息的全局相关性，进而实现含密图像的高效检测。实验表明，与基线卷积神经网络模型相比，所提出的方法展现了更加优异的性能，这展示了图表示学习在图像隐写分析任务中的潜能。

(2) 针对卷积层的大量堆叠可能会造成隐写特征减弱的问题，本文提出了一种结合图神经网络优势和卷积神经网络优势的 JPEG 域图像隐写分析算法，该算法主要由一个图注意力学习模块和一个特征增强模块组成。其中，图注意学习模块的设计是为了避免卷积神经网络依靠深度堆叠来扩展感知域的局部特征学习所造成的全局特征损失；特征增强模块的设计是为了防止卷积层的堆叠削弱隐写信息，两者相结合以保证同时学习隐写信息的局部特征和全局特征。此外，作为一种有效的网络参数初始化方式，预训练也被引入以提高网络提取鉴别性特征的能力。实验结果表明，所提出的算法在检测精度上优于以往的工作，验证了所提出的方法的优越性和适用性。

**关键词：** 图像隐写分析、图神经网络、图注意力机制、信息隐藏

# ABSTRACT

Image steganography is an information hiding technique using images as carriers. The main purpose of image steganalysis as a technique to combat image steganography is to extract the possible steganographic features in digital images to determine whether they contain secret information. Current deep learning-based image steganalysis algorithms are designed in combination with deep convolutional neural networks, and a large amount of works have demonstrated their applicability. In recent years, graph neural networks have flourished due to their applicability to unstructured data and have the advantage of being able to capture both local features and global features. In order to exploit the potential of graph neural networks for extracting steganographic features, this thesis proposes to apply them to image steganalysis, and investigates them in two modified domains, the spatial domain and the JPEG domain, with the following results:

（1）For the advantage that graph neural networks can better tap the association information between image blocks, this thesis proposes an algorithm for steganalysis of spatial domain images based on graph representation learning, which can use graph neural networks to analyze the global features of images to achieve better detection performance. In the specific framework, a graph construction method is designed to transform images into graphs. Each node is associated with a feature vector determined by a shallow convolutional neural network from the corresponding image block. And each edge represents local associations between image blocks. The graph neural network learns the global correlation of image steganographic information, enabling efficient detection of secret images. Experimental results demonstrate that the proposed method performs competitively compared to the baseline convolutional neural network model, showcasing the potential of graph representation learning in steganography analysis.

（2）To address the problem that the large number of stacked convolutional layers may cause the steganographic features to be weakened, this thesis proposes a JPEG domain image steganalysis algorithm that combines the advantages of graph neural network and convolutional neural network, which mainly consists of a graph attention learning module and a feature enhancement module. Among them, the graph attention learning module is designed to prevent the loss of global features caused by the convolutional neural network relying solely on depth stacking to extend local feature learning in the perceptual domain; the feature enhancement module is designed to prevent the stacking of convolutional layers from weakening the steganographic information. In addition, as a way to initialize the network weights with large-scale datasets, pre-training is also introduced for use to improve the network's ability to extract discriminative features. Experimental results show that the proposed algorithm outperforms previous work in terms of detection accuracy, which also validates the superiority and applicability of the proposed method.

**Keywords:** Image Steganalysis, Graph Neural Networks, Graph Attention Mechanism, Information Hiding

# 目　录

# 第一章 绪论

## 1.1 研究的背景和意义

互联网和多媒体技术的蓬勃发展给信息交流带来了极大的便利，同时也带来了不少安全方面的风险与隐患。由于机密信息在通信过程中被恶意窃取或篡改、不法分子利用隐蔽通信技术实现非法目的等事件的不断发生，信息安全与隐私保护问题愈发严峻。因此，如何在互联网中安全地进行通信以及如何对抗不法分子利用互联网进行隐蔽通信等问题都亟待解决，这对维护社会安全稳定和保障国家信息安全具有重要的意义。为了保障通信安全，信息隐藏技术应运而生，这是一种将秘密数据隐藏在正常媒体中实现隐蔽通信的技术[1]。隐写作为信息隐藏的重要分支，能够将秘密信息嵌入到诸如数字图像和音视频等多媒体信号中而不会造成显著失真[2]。其与密码学的区别在于，密码学利用加密技术将消息进行加密而难以被破解[3]，而隐写利用“正常”的数字媒体实现秘密信息的伪装来掩盖隐蔽通信的存在[4]。

尽管隐写能够用于隐蔽通信，但也易被不法分子利用，用于非法目的。作为对抗隐写的安全技术，隐写分析主要通过从数字载体中提取并分析敏感的统计特征，以判断载体是否含有机密信息。隐写分析技术的发展有助于保证隐写技术的合规使用，具备预防国家和个人机密泄露、防止非法信息肆意传播的重要功能，对保护经济社会安全稳定和我国信息安全具有重要实用价值。此外，隐写分析有助于挖掘当前隐写方案的缺陷，可以用于评估隐写方案的安全性，协助研究人员开发更具安全性能的隐写算法。同样，隐写算法可以作为隐写分析算法的评价指标，推动隐写分析技术的进步。因此，隐写和隐写分析技术的发展是相辅相成、密不可分的。

通常情况下，任何类型的数字信号都可以作为隐写算法的载体。然而，从应用的角度来看，数字图像是最常用的隐写载体，因为它易于编辑且具有大量容纳秘密信息的冗余空间。因此，大多数现有的隐写方法都是针对数字图像设计的。图像隐写通常将秘密信息隐藏在人类视觉系统难以察觉的像素变化或图

像频域变化中。深度学习[5]在计算机视觉及模式识别领域取得了巨大的成功，如何利用深度学习技术实现高效的隐写分析是当下的研究热点。本文旨在研究面向数字图像的隐写分析技术，通过深度学习技术实现隐写图像的高效检测。

## 1.2 图像隐写和图像隐写分析概述

### 1.2.1 图像隐写

(1) 隐写的基本概念

隐写(steganography)一词来自于希腊词源，“stegos”意味着“覆盖”，“grafia”意味着“书写”，它被定义为“隐蔽写作”[6]。隐写是一种实现安全防护的信息安全技术，秘密数据被嵌入到诸如图像、文本或视频的多媒体载体信号中，并在公共信道中进行传输以实现隐蔽通信[7]。“数据隐藏”或“隐写术”的内涵可以通过 Simmons 于 1984 年提出的“囚犯困境”模型来理解[8]。如图 1.1 所示，囚犯 Alice 和 Bob 被关押在不同的牢房中，他们正在商讨如何越狱，但是他们的通信内容受到狱警 Eve 的监视，一旦他们的交流内容引起怀疑，越狱计划就会失败。为了应对这一风险，他们事先共享了一个密钥，Alice 利用共享的密钥将秘密信息嵌入到通信载体中得到含密载体，再通过公共信道(被监控的信道)将含密载体传输给 Bob。当 Bob 成功接收到含密载体，他将通过共享密钥从含密载体中重构秘密信息，从而实现了隐写通信。隐写通信最显著的特点是它隐藏了秘密信息的存在性。

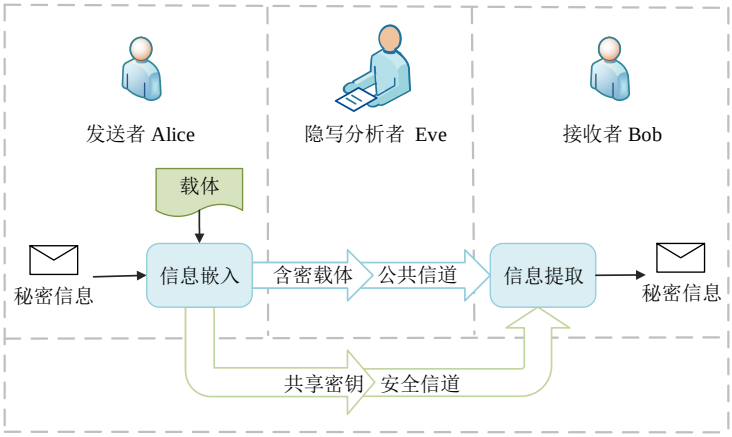



图1.1 “囚犯困境”模型

在隐写过程中，秘密信息是需要被隐藏的数据，载体是可随通信媒介的改变而相应变化的多媒体信号，常见的载体有文本、图像和视频等。文本作为人类语言的符号化编码，是人们使用最频繁的交流工具，因此利用文本进行隐写因其使用场景的普遍性而具有广阔的应用前景，但由于文本的高度编码性和低信息冗余性导致很难承载较大的信息量，且轻微的修改很容易造成语义异常，因此文本隐写具有很大的技术挑战。

数字图像作为信息交流中不可或缺的数字媒介常用于隐写，与文本相比，图像具有较大的信息冗余，且编辑后更不易被察觉。近年来，网络带宽的增加和各种高效视频压缩算法的涌现促使视频成为移动终端和社交网络中流行的信息载体。相对于图像，视频可视为多幅图像的集合。从隐写的角度看，利用视频进行隐蔽通信能够传输更多的数据量。然而，视频有损编解码过程异常复杂，导致秘密信息的嵌入复杂度高，且秘密信息重构出现错误的可能性相对较高。综上所述，同文本和视频相比，利用图像进行隐写能够在数据量、计算复杂度和隐蔽性等方面取得更好的平衡。

隐写的常见评价指标包括不可感知性、嵌入量、安全性、计算复杂度和普适性等，具体描述如下：

a) 不可感知性：隐写的一个基本要求是经过隐写的载体不易引起异常，即不可感知性好。不可感知性可通过构建关于自然载体和含密载体的统计失真来定量分析。以图像为例，均方误差（Mean Square Error, MSE）和峰值信噪比（Peak Signal-to-Noise Ratio, PSNR）是常见的图像失真度量方法，可用于评估隐写图像的视觉失真[9]。一般而言，均方差越小或者峰值信噪比越大表明隐写后的图像的视觉质量越好，即不可感知性相对更好。其他度量方法还包括均方根误差（Root Mean Square Error, RMSE）、加权 PSNR（Weighted PSNR, WPSNR）[10]、结构相似性（Structural SIMilarity, SSIM）指数[11]和欧氏距离（Euclidean Distance）等。

b）嵌入量：嵌入量是指经过隐写后载体承载的秘密信息量，其值越大越好。不失一般性，以图像为例，若隐写算法直接修改空域图像的像素值，则通常采用单位像素所承载的比特数（bits per pixel, bpp）度量嵌入量，若隐写算法修改的是图像的频域系数，则可用单位频域系数所承载的比特数（bits per frequency

coefficient，bpfc）来度量嵌入量。通常情况下，频域图像隐写算法多采用DCT（Discrete Cosine Transform）非零交流系数进行信息嵌入，此时可采用单位非零交流 DCT 系数（bits per non-zero alternating-current coefficient, bpnzac）度量嵌入量[12]。总体而言，在保证不可感知性的前提条件下，我们总是希望嵌入量越大越好，即载体可承载的秘密信息量值越大越好。

c）安全性：隐写需要保证隐写统计特征的不可检测性，如果该隐写算法能够抵抗各种检测攻击，则说明它是安全的。一般情况下，安全性使用隐写检测的准确率（Accuracy）或最小平均错误率（Minimum average decision error ratio）来度量，其值越小表示该隐写系统的安全性越高。

d) 计算复杂度：计算复杂度指隐写算法的嵌入复杂度和提取复杂度，分别为秘密信息嵌入载体和从载体中提取秘密信息所需要的计算工作量，可用时间复杂度（运行时间）和空间复杂度（储存空间）来度量，其值都应越小越好。

d) 普适性：普适性刻画了隐写算法的可扩展性。一般而言，我们希望所设计的隐写算法具有一般性，即算法可用于不同的载体类型或通信环境。

（2）图像隐写的发展现状

隐写早在许多古西方文字记载中就出现过，最早可以追溯到公元前四百多年，古希腊士兵将密信刻在奴隶的头皮上，等到其头发长出后再送出去以实现隐蔽通信[13]。近代也涌现了许多有关隐写的应用，如两次世界大战中就频繁地使用隐形墨水和伪装物品等方法进行隐蔽通信。自 20 世纪 90 年代以来，随着互联网和多媒体技术的迅猛发展，古老的隐写术逐渐演变出全新的方式，特别是在数字多媒体技术的基础上，隐写获得了系统性的研究，并成为信息安全领域备受关注的热点之一。

图像隐写作为隐写最主要的研究内容，根据秘密信息嵌入的作用域进行划分，可分为空域图像隐写和变换域图像隐写。空域图像隐写直接在图像空域像素上进行隐写嵌入修改，而变换域方法通过使用离散余弦变换（Discrete Cosine Transform, DCT）、离散小波变换（Discrete wavelet transform, DWT)、哈达玛变换（Hadamard Transformation）等任意一种变换将图像从空域转换为频域再进行秘密信息嵌入，这一类最常见的是 JPEG（Joint Photographic Experts Group）域隐写。

早期的图像隐写技术主要通过减少载体图像的修改数据量来保证安全性。最早提出的图像隐写是 LSB（Least Significant Bit）替换隐写[14]，信息嵌入可以通过将图像中随机选择的像素的最低比特位替换为二进制秘密信息来实现。为了扩大嵌入量同时提高安全性，后续研究者设计了许多隐写编码方法，如矩阵编码[15]、EMD（Exploiting Modification Direction）编码[16]、双层编码[17]和 MME（Modified Matrix Encoding）编码[18]等。也有研究者通过保持数字图像的统计分布特性不变来保证安全性，这类算法的主要思想是设计一个描述图像统计特性的模型，并使得嵌入信息后的模型保持不变，例如 HPDM（Histogram-Preserving Data Mappings）[19]、MB（Model Based）[20]、F5[21]和 nsF5[22]等方法。

为了选择隐蔽性更好的位置嵌入秘密信息，后续研究者使用最小化失真嵌入框架来增强安全性，该框架由失真代价函数和隐写编码两个部分组成，失真代价函数用于度量整幅载体图像被修改后的失真程度，隐写编码（如 Syndrome-Trellis Codes, STCs[23]）用于最小化失真函数，并生成相应的含密图像。基于这一框架设计的隐写算法通常被称为内容自适应隐写算法，因为它们可以根据事先定义好的失真代价函数灵活地选择纹理较为复杂的区域进行隐写嵌入，从而更好地对抗基于统计方法的隐写检测技术。

早期建立在最小化失真嵌入框架下的一种经典空域图像隐写技术是 HUGO（Highly Undetectable steGO）[24]，它通过减小载体图像和含密图像在 SPAM（Subtractive Pixel Adjacency Matrix）[25]特征空间上的差异性有效降低了隐写图像的被检测概率。2014 年，Holub 等人提出了 WOW（Wavelet Obtained Weights）[26]算法和 S-UNIWARD（Spatial UNIversal Wavelet Relative Distortion）[27]算法，都利用了多个方向的滤波器来提取载体图像上的差异信息，再根据不同方向的残差构建失真代价函数，能够适应各种复杂图像的特点。同年，Li 等人[28]提出的 HILL（HIgh-pass, Low-pass, and Low-pass）算法则引入失真扩散原则，通过平滑滤波来调节失真代价值的变化，进一步提升了隐蔽性。这些算法都使用量化嵌入变化对邻域像素的影响计算像素的嵌入成本。

与前文中基于启发式经验设计的图像隐写算法不同，Fridrich 等人[29]提出的 MVG（MultiVariate Gaussian model）算法采用特定的高斯图像模型作为失真代价

函数，能够度量统计检测性，这在一定程度上提高了隐蔽性。随后在 2016 年他们又提出了 MiPOD（Minimizing the Power of Optimal Detector）算法[30]，通过最小化嵌入对载体模型的影响来确定嵌入概率。

JPEG 域图像隐写也可基于最小化失真框架进行设计，如 Guo 等人[31]提出的均匀嵌入失真隐写 UED（Uniform Embedding Distortion）利用失真代价函数将隐写修改均匀地分布在所有可嵌入的 DCT 系数上。Holub 等人[27]将 UNIWARD 算法从空域拓展到 JPEG 域提出了 J-UNIWARD，在保证嵌入质量前提下提高了隐蔽性。2020 年，Cogranne 等人[32,33]基于 MiPOD 算法，在空域的基础上提出了 JPEG 域隐写算法 J-MiPOD，旨在提高 JPEG 隐写图像的隐蔽性和鲁棒性。

近年来，机器学习和深度学习在计算机视觉领域取得了巨大的成功。这一进展为隐写的研究带来了新的思路和挑战，推动着图像隐写形成了一系列全新的理念和方法。基于深度学习的图像隐写算法通常由隐写网络和提取网络共同构成，现有的方法主要分为基于卷积神经网络（Convolutional Neural Network, CNN）和基于生成对抗网络（Generative Adversarial Networks）[34]两种。基于卷积神经网络的图像隐写方法能够通过将图像隐藏到图像中的方法提高信息隐藏的容量，但此类方法的不可察觉性较差[35,36]。基于生成对抗网络的隐写通过不同对象之间的对抗训练来学习含密图像的生成或秘密信息的嵌入方式，主要包含直接生成载体图像的方法[37]、直接生成含密图像的方法[38,39]和学习载体图像元素修改代价的方法[40]。

总的来说，基于最小化失真嵌入框架和基于深度学习的隐写算法已成为图像隐写领域的重要研究内容，这些算法的研究不仅有助于提高隐写的隐蔽性和鲁棒性，也为发展更加安全、稳健的信息隐藏技术提供了思路。

### 1.2.2 图像隐写分析

（1）隐写分析的基本概念

作为隐写的一种对抗技术，隐写分析随着隐写技术的发展也得到深入研究，其目的是揭示载体中隐藏信息的存在。大多数现有的隐写分析方法将载体中隐藏信息的检测建模为一个二分类问题，即确定一个给定的载体中是否包含秘密

信息。如图 1.2 所示，隐写分析者通常可以收集一些自然载体和含密载体，利用这些有标签的载体提取特征训练所设计的隐写分析分类器，训练好的隐写分析器即可以分辨一些未知的载体是否含有秘密信息。

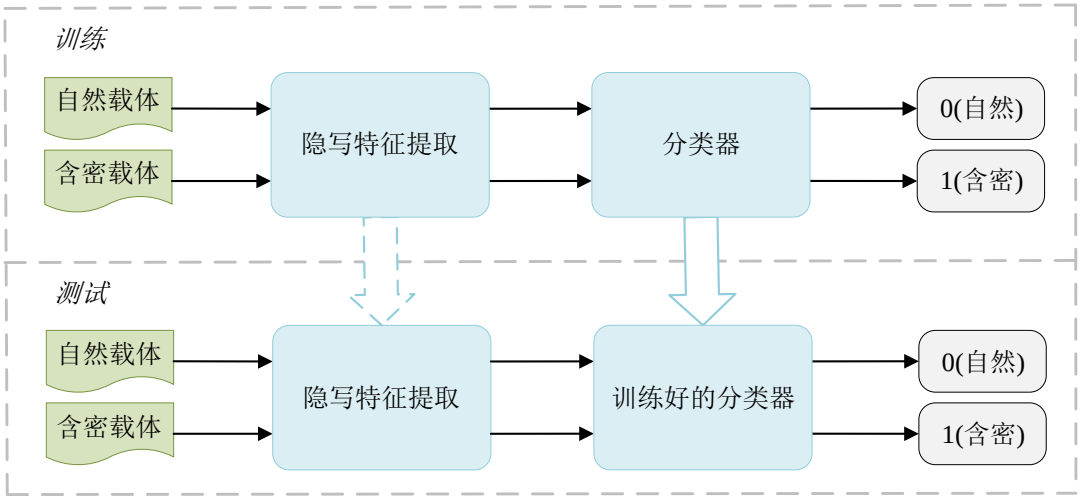



图 1.2 隐写分析的通用框架

隐写分析作为一种二分类任务，其评价指标与常见二分类问题一致，假设自然载体(Cover)为阴性类(Negative)，含密载体(Stego)为阳性类(Positive)，通常使用以下几个指标进行性能评估：

a）准确率(Accuracy, ACC)：指隐写分析分类结果正确的样本数与总样本数的比值，即：

$$ACC = \frac{N_{TP} + N_{TN}}{N_{TP} + N_{FP} + N_{TN} + N_{FN}} \tag{1.1}$$

其中，$N_{TP}$ 表示正确分类的含密载体数，$N_{FP}$ 表示错误分类的含密载体数，$N_{TN}$ 表示正确分类的自然载体数，$N_{FN}$ 表示错误分类的自然载体数。准确率越高表示该隐写分析器的性能越好。

b）对(Minimum average decision error ratio)：

$$P_E = \min_{P_{FA}} \frac{1}{2}\left(P_{FA} + P_{MD}\right) \tag{1.2}$$

其中，$P_{FA}$ 表示误报率(False AlarmRate, FAR)，也称假阳率(False positive rate)，表示将自然载体误认为含密载体的比例；$P_{MD}$ 表示漏检率(Missed Detection Ratio, MDR)，表示将含密载体误判为自然载体的比例。

（2）图像隐写分析的发展现状

隐写分析和隐写是相辅相成、共同发展的，为了对抗快速发展的图像隐写，研究者们也对图像隐写分析进行了大量研究。现有的数字图像隐写分析按照其发展进程可分为传统图像隐写分析和基于深度学习的图像隐写分析。

传统图像隐写分析方法一般分为两步，第一步是构建隐写分析特征，第二步是采用分类器进行分类。在分类器的选择方面，早期的方法多采用支持向量机（Support Vector Machines, SVM）[41]对隐写分析特征进行分类。然而，随着高维隐写分析特征的不断增加，SVM 分类的计算复杂度也随之增加。因此，Fridrich 团队提出了一种基于 Fisher 线性判别分析（Fisher Linear Discriminant, FLD）[42]的集成分类器，其优势在于能够以较低的计算时间复杂度获得与 SVM 相媲美的隐写分析性能。

目前传统的图像隐写分析研究主要关注如何构建隐写分析特征。在空域图像隐写分析领域，应用比较广泛的方法是 Fridrich 团队提出的空域“富”模型特征 SRM（Spatial Rich Model）[43]和 PSRM（Projection Spatial Rich Model）[44]。SRM 利用多种类型的滤波器提取图像中的残差图像，再对残差图像进行量化和截断，从而构建高维“富”模型特征。针对已提出的 SRM 特征，一些改进版的自适应隐写分析特征也被陆续提出，例如 tSRM[45]、maxSRM[46]、AdaptiveSRM[47]以及 $\sigma$SRM[48]等。由于隐写分析者也可以获取选择信道（Selection Channel），因此近年来，一些研究者提出了结合选择信道的自适应隐写分析方法，在特征提取阶段引入选择信道，并赋予纹理复杂区域更大的权重，以降低平滑区域的干扰。

对于 JPEG 域图像隐写分析，研究者们也提出了许多专用的隐写分析特征，如 CC-PEV（Cartesian-calibrated PEV）、CC-JRM（Cartesian-calibrated JPEG Rich Model）[49]等。后续 Fridrich 等人[50]提出了基于 JPEG 相位特征的残差特征方法，该方法利用 JPEG 块在解压缩过程中的相位信息来构建特征，进一步提高了隐写分析性能。其他类似特征还有 DCTR（Discrete Cosine Transform Residual）[51]、PHARM（Phase AwaRe projection Model）[52]、GFR（Gabor Filter Residual）[53]和 MD-CFR（Maximum Diversity-Cascade Filter Residual）[54]等。Denemark 等人[55]还提出了基于选择信道感知（Selection Channel Aware）的 JPEG 相位特征。

近年来，深度学习在计算机视觉领域展现出令人瞩目的成就，并在图像隐写分析中得到应用。2014 年，Tan 等人[57]提出了一种使用深度学习技术的隐写分析模型。随后 2015 年，Qian 等人[58]提出了一种使用监督学习方法中卷积神经网络的模型，该方法较成功地将卷积神经网络(CNN)应用于隐写分析任务，并取得了与传统 SRM 手工提取特征相当的结果。这些研究表明，深度学习在图像隐写分析方面具有巨大的潜力。

随后，Xu 等人[59]提出了第一个性能超过 SRM 的基于 CNN 的隐写分析器 XWS-CNN，并进一步使用集成学习[60]提高了检测性能。Ye 等人[61]提出了空域隐写分析网络 YeNet，其性能显著优于 SRM。Boroumand 等人[63]提出了适用于空域和 JPEG 域的完全端到端的隐写分析网络 SRNet(Steganalysis Residual Network)。2019 年，Deng 等人[63]提出一种快速高效的 CNN，该结构引入了全局协方差池化来提高检测性能。Tan 等人[64]提出了一种基于通道剪枝的深度残差网络结构搜索方法 CALPA-NET (ChAnneL-Pruning-Assisted deep residual NETwork)。2020 年，Zhang 等人[65]提出了 ZhuNet，该网络通过深度可分离卷积和空间金字塔池化(Spatial Pyramid Pooling, SPP)来提升检测性能。2021 年，You 等人[66]提出了基于孪生网络的隐写分析方法 SiaStegNet(Siamese Steganalysis Network)。2022 年，Ren 等人[67]提出了基于对比学习的隐写分析框架。

针对 JPEG 域隐写分析，Jang 等人[68]设计了基于特征聚合的 JPEG 域图像隐写分析网络 FANet(Feature Aggregation Network)。Yousfi 等人[69]研究了经预训练的网络在 ALASKA#2 隐写分析竞赛[70]中的表现情况，其中包括 EfficientNet[71]、MixNet[72]和 ResNet[73]等网络模型，实验证明在 ImageNet[74]上进行预训练的模型在 ALASKA#2 数据集上能实现比 SRNet 更好的 JPEG 域图像隐写分析性能。

综上所述，传统图像隐写方法主要在构建隐写分析特征方面进行了大量的研究，此类方法隐写特征的构建和分类通常是分两步进行的。基于深度学习的图像隐写分析相较于传统算法因其端到端的网络特性，将构建隐写分析特征和特征分类融为一体，节约了人工成本，为图像隐写分析研究开辟了新的道路。基于深度学习的图像隐写分析算法已经成为主流的研究内容，能够在一定程度上满足更复杂的隐写检测需求。

## 1.3 论文的主要研究内容与安排

### 1.3.1 论文主要研究内容

本论文主要研究基于图神经网络的图像隐写分析。由于深度学习技术在计算机视觉领域的快速发展，将其应用于图像隐写分析领域逐渐成为研究热点之一。近年来，图神经网络不断发展，其处理全局的图信息相较于深度卷积神经网络设具有更大的优势，因此本文结合了卷积神经网络和图神经网络各自的优点，提出了基于图神经网络的空域图像隐写分析和 JPEG 域图像隐写分析方法。本论文具体研究内容如下：

(1) 基于图神经网络的空域图像隐写分析

针对卷积神经网络需要通过增加卷积层来扩大感受野，可能会导致全局信息丢失的问题，本文提出了基于图表示学习的空域图像隐写分析算法，该算法能够利用图神经网络分析图像的全局特征以达到更好的检测性能。在详细的网络结构中，首先设计了将每个图像转换为图数据的方法，其中用节点表示图像块的隐写特征，节点之间的边表示图像块之间的局部关系。随后利用图神经网络学习图像隐写的局部信息和全局相关性，从而实现含密图像的高效检测。实验表明，与基线 CNN 模型相比，所提出的网络实现了有竞争力的性能，验证了图表示学习的优异性，也展示了图表示学习在图像隐写分析中的巨大潜力。

(2) 基于图神经网络的 JPEG 域图像隐写分析

结合了 CNN 的局部特征提取的特点和图注意神经网络全局化的优势，设计了一种更适于图像隐写特征传输和增强的网络模型。与其他方法相比，该模型能够较好地解决由于卷积层的堆叠而导致的隐写特征弱化的问题，同时还能利用图注意力网络更好地学习隐写信号的全局特征，提高了微弱的图像隐写特征在网络中的传输。此外，该方法先在规模更大的数据集上预训练以学习先验知识，再通过微调预训练模型的方式将模型的先验知识迁移到对样本的隐写检测分类器中，从而提升其检测性能。与现有方法相比，所提出的方法在检测性能与收敛速度上均得到了提升。

### 1.3.2 论文结构安排

本论文各章内容结构安排如下：

第一章阐述了图像隐写分析的研究背景和研究意义，介绍了隐写和隐写分析的基本概念，并着重介绍了图像隐写和图像隐写分析的发展现状。

第二章首先介绍了图像隐写的代表性成果，随后回顾了图像隐写分析中的传统方法和基于深度学习的方法，最后介绍了图神经网络相关技术。

第三章将图神经网络和卷积神经网络相结合，设计了将图像转换为图的方法，并提出了一种基于图表示学习的空域图像隐写分析方法。

第四章将基于图神经网络的图像隐写分析的方法从空域拓展到 JPEG 域，提出了一种基于隐写特征增强和图注意力学习的 JPEG 域图像隐写分析模型。

第五章总结了本文所提出的基于图神经网络的图像隐写分析研究内容，并对图像隐写分析进行了展望。

# 第二章 相关技术介绍

## 2.1 图像隐写技术

### 2.1.1 基于失真代价函数的图像隐写

早期的图像隐写算法通常随机选择一些像素进行修改，如 LSB 替换算法、LSBM（LSB Matching）算法[75]和 EMD 算法[76]等，这些算法对图像中所有像素的选择概率是一样的，可以看作是非自适应算法。然而研究表明，图像中纹理区域也就是噪声较高的区域相比于平滑区域具有更好的隐藏特性和抗检测性[27]，因此研究者们提出了基于内容自适应的隐写算法，将秘密信息嵌入到视觉系统难以察觉的图像纹理区域。

现在最流行的内容自适应隐写算法基本都是基于最小化失真框架设计的[23]，如图 2.1 所示是最小化失真隐写框架示意图。具体而言，每个载体元素通过计算嵌入成本的方式赋予一个被修改的代价函数，通过在信息嵌入的过程中最小化全局代价生成对应的含密图像。基于最小化失真框架的隐写方法与非自适应方法的不同之处在于，后者期望最小程度地修改原载体图像，而前者的目标是在理论失真和实际编码方案的限制下，选择失真代价函数最小化的最佳嵌入方案。

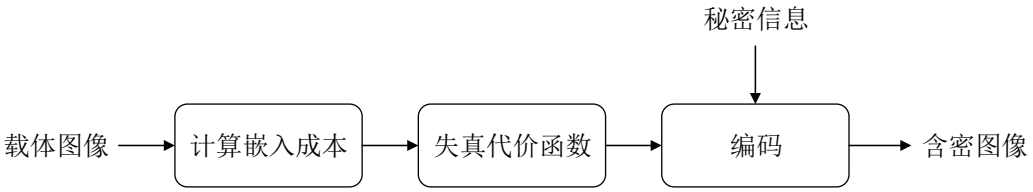



图2.1 最小化失真框架

为了更好地选择引起图像失真最小的像素进行修改，基于失真代价函数的图像隐写的主要遵循三项原则：第一项是复杂性原则，即对图像内容复杂区域的元素分配低成本，而对图像光滑区域的元素分配高成本。第二项是扩散原则，即具有高嵌入优先级的元素周围的元素也具有较高的优先级，同理具有低嵌入优先级的元素周围的元素优先级则较低。最后一项是集群原则，该原则指出聚类嵌入的不可察觉性优于分散嵌入[78]。

基于最小化失真框架的图像隐写的设计要点在于其失真代价函数的构建。失真代价函数又称为损失函数或成本函数，是对图像进行隐写时所造成的失真程度的衡量标准。具体来说，它衡量的是将秘密信息嵌入到载体图像中所引入的扭曲和改变，因此它越小则表示隐藏的信息越难以察觉。它的实现方式一般来说有两种：一种是使用一些先验的方式来度量嵌入成本，例如 UNIWARD[27]、HILL[28]和 UED[31]等隐写算法；另一种则是直接度量图像的统计可检测性，如 MiPOD[32]和 J-MiPOD[33]等隐写算法。以下将对这些算法中所使用的失真代价函数做详细介绍。

（1）S-UNIWARD 和 J-UNIWARD

S-UNIWARD[27]是使用加性失真函数设计的空域隐写算法，它首先使用方向性小波 DB-8 来构造三个不同方向的一级分解小波滤波器组$\{\boldsymbol{K}^{(k)}, k=1,2,3\}$。然后假设含密图像$\boldsymbol{Y}$由载体图像 $\boldsymbol{X}$ 修改一个像素值后得到，则 S-UNIWARD 的失真代价函数定义为载体图像经过定向小波滤波器组滤波后系数的相对变化，具体计算如式(2.1)：

$$D(\boldsymbol{X},\boldsymbol{Y})=\sum_{k=1}^{3}\sum_{u,v}\frac{\left|W_{u,v}^{(k)}(\boldsymbol{X})-W_{u,v}^{(k)}(\boldsymbol{Y})\right|}{\left|W_{u,v}^{(k)}(\boldsymbol{X})\right|+\varepsilon} \tag{2.1}$$

其中，$W_{u,v}^{(k)}(\boldsymbol{X})$和$W_{u,v}^{(k)}(\boldsymbol{Y})$分别表示使用小波滤波器$\boldsymbol{K}^{(k)}$对$\boldsymbol{X}$和$\boldsymbol{Y}$滤波后的第$(u,v)$个小波系数，$\varepsilon$表示一个接近于 0 的正实数。

J-UNIWARD[27]是 S-UNIWARD 隐写算法在 JPEG 域实现的版本，它与空域的 S-UNIWARD 类似，首先将 JPEG 载体图像解压缩到空域，然后使用小波分解系数的相对变化计算失真代价函数。其失真代价函数可以表示为式(2.2)：

$$D(\mathbf{X},\boldsymbol{Y})=\sum_{k=1}^{3}\sum_{u,v}\frac{\left|W_{u,v}^{k}\left(J^{-1}(\boldsymbol{X})\right)-W_{u,v}^{k}\left(J^{-1}(\boldsymbol{Y})\right)\right|}{\left|W_{u,v}^{k}\left(J^{-1}(\boldsymbol{X})\right)\right|+\varepsilon} \tag{2.2}$$

其中，$J^{-1}$表示 JPEG 解压缩，$W_{u,v}^{k}\left(J^{-1}(\boldsymbol{X})\right)$和$W_{u,v}^{k}\left(J^{-1}(\boldsymbol{Y})\right)$分别表示使用小波滤波器$\boldsymbol{K}^{(k)}$对解压缩的$\boldsymbol{X}$和$\boldsymbol{Y}$滤波后的第$(u,v)$个小波系数。

（2）HILL

WOW 和 S-UNIWARD 倾向于利用纹理区域的像素来隐藏信息，然而纹理区域中可能适合嵌入信息的少数像素修改代价很高，将它们修改可能会造成生成的含密图像失真较明显。为此，Li 等人[28]提出了 HILL 算法，其新颖之处在于设计了一个失真函数来降低修改代价，该函数共有三个滤波器，包括一个高通滤波器和两个低通滤波器。在嵌入数据时，HILL 算法不会过度依赖图像的特定区域或纹理结构，而是利用整个图像的统计特性进行嵌入，这使得嵌入的内容更为隐蔽且难以被检测。载体图像首先经过一个高通滤波器 $\boldsymbol{K}_{\mathrm{HP}}$ 滤波从而得到纹理残差，然后使用一个低通滤波器 $\boldsymbol{K}_{\mathrm{LP1}}$ 对纹理残差的绝对值进行滤波得到嵌入适应度，将嵌入适应度取倒数可以得到图像的失真代价，再通过另一个低通滤波器 $\boldsymbol{K}_{\mathrm{LP2}}$ 即可得到邻域内像素的失真代价的均匀分布，表达式如公式(2.3)：

$$D(\boldsymbol{X},\boldsymbol{Y})=\frac{1}{\left|\boldsymbol{X}*\boldsymbol{K}_{\mathrm{HP}}\right|*\boldsymbol{K}_{\mathrm{LP1}}}*\boldsymbol{K}_{\mathrm{LP2}} \tag{2.3}$$

其中，低通滤波器 $\boldsymbol{K}_{\mathrm{LP1}}$ 和 $\boldsymbol{K}_{\mathrm{LP2}}$ 分别为 $3\times3$ 和 $15\times15$ 的均值滤波器，$\boldsymbol{K}_{\mathrm{HP}}$ 是一个 $3\times3$ 的高通滤波器，如公式(2.4)：

$$\boldsymbol{K}_{\mathrm{HP}}=\begin{bmatrix}-1 & 2 & -1\\ 2 & 4 & 2\\ -1 & 2 & -1\end{bmatrix} \tag{2.4}$$

（3）UED-SC、UED-JC 和 UED-SI

均匀嵌入失真隐写算法 UED(Uniform Embedding Distortion)[77]指出 DCT 系数在嵌入修改时，若任意选择嵌入位置，小值 JPEG 系数的数量较多将导致被选择的可能性更大。但是小值系数分布上的剧烈变化更容易造成系数分布的形变，从而引起较大的抗检测隐患。因此，UED 优先选择大值系数进行修改。GUO 等人[77]基于 DCT 系数本身及其邻域和块间相邻系数的取值构造嵌入失真函数，并使用 STC 编码嵌入，提出了减少失真并提高嵌入量的基于单系数的均匀嵌入失真(Uniform Embedding Distortion based on Single Coefficient, UED-SC)、基于联合系数的均匀嵌入失真(UED based on Joint Coefficients, UED-JC)和基于旁路信息的均匀嵌入失真(Side-Informed UED, SI-UED)。

UED-SC 被设计用来有效地缓解由于随机嵌入引起的统计特征的突然变化，特别是一阶的统计特征的突然变化。因此 UED-SC 主要考虑 DCT 系数本身幅度的影响，其失真代价函数如公式(2.5)：

$$\rho_{ij}^{\mathrm{SC}} = \left|c_{ij}\right|^{-1} \tag{2.5}$$

其中，$c_{ij}$ 代表位置 $(i, j)$ 上的 JPEG 量化系数。

除了 DCT 系数本身的幅度外，UED-JC 指出它的内邻域和块间邻域的系数与该系数也具有很强的相关性，因此应该共同考虑其影响。假设定义 DCT 系数 $c_{ij}$ 的内邻域为 $N_{\mathrm{ia}} = \{c_{i+1,j}, c_{i+1,j}, c_{i,j+1}, c_{i,j+1}\}$，块间邻域为 $N_{\mathrm{ir}} = \{c_{i+8,j}, c_{i+8,j}, c_{i,j+8}, c_{i,j+8}\}$，则基于联合系数的均匀嵌入失真代价函数可被定义为公式(2.6)：

$$\begin{aligned}\rho_{ij}^{\mathrm{JC}} = &\sum_{d_{\mathrm{ia}} \in N_{\mathrm{ia}}} \left(\left|c_{ij}\right| + \left|d_{\mathrm{ia}}\right| + \alpha_{\mathrm{ia}}\right)^{-1} \\ &+ \sum_{d_{\mathrm{ir}} \in N_{\mathrm{ir}}} \left(\left|c_{ij}\right| + \left|d_{\mathrm{ir}}\right| + \alpha_{\mathrm{ir}}\right)^{-1}\end{aligned} \tag{2.6}$$

其中，$\alpha_{\mathrm{ia}}$ 与 $\alpha_{\mathrm{ir}}$ 为待定常数，有实验结果分别确定为 1.3 与 1；$c_{ij}$ 为系数值，以上失真函数限制了小值区的取值。

通常情况下，以原始空域图像作为先验信息可以帮助设计高效的 JPEG 隐写方案[79]。借助原始的空域图像，可以很容易地获得由于 JPEG 压缩引起的舍入误差 $R$ 和由数据嵌入引起的嵌入误差 $R'$。因此，可以得到加性舍入误差 $E = \left|R' - R\right|$（嵌入误差和舍入误差之间的绝对值），该误差为基于旁路信息的均匀嵌入失真(SI-UED)的重要组成部分，并且需要尽可能小。最终 SI-UED 的失真函数被设计为同时满足均匀嵌入和最小绝对失真，即加性舍入误差也要最小化，其失真代价函数被定义为每个像素加性舍入误差和 UED-JC 失真代价函数的乘积，具体表达式如公式(2.7)：

$$\begin{aligned}\rho_{ij}^{\mathrm{SI}} = e_{ij} \cdot \Bigg[&\sum_{d_{\mathrm{ia}} \in N_{\mathrm{ia}}} \left(\left|c_{ij}\right| + \left|d_{\mathrm{ia}}\right| + \alpha_{\mathrm{ia}}\right)^{-1} \\ &+ \sum_{d_{\mathrm{i}} \in N_{\mathrm{i}}} \left(\left|c_{ij}\right| + \left|d_{\mathrm{ir}}\right| + \alpha_{\mathrm{ir}}\right)^{-1}\Bigg]\end{aligned} \tag{2.7}$$

其中，$e_{ij}$ 代表 $c_{ij}$ 的加性舍入误差，$\alpha_{\mathrm{ia}}$ 与 $\alpha_{\mathrm{ir}}$ 为待定常数，有实验结果确定为 0.2，其他参数与 UED-JC 相同。

（4）MiPOD/J-MiPOD

MiPOD[30]算法设计了一种基于统计模型的失真代价函数，该算法建立在一种统计模型下，该模型假设载体图像$\boldsymbol{X}$中的每个像素是独立分布的，且分布遵循高斯分布，即$x_{i,j} \sim Gauss\left(\mu_{i,j},\sigma_{i,j}^2\right)$，其中$\mu_{i,j}$和$\sigma_{i,j}^2$分别为像素$x_{i,j}$的均值和方差。基于这种像素统计模型，MiPOD 本质上在于找到每个像素的嵌入概率$\beta_{i,j}$，同时最小化最大似然比测试(Likelihood Ratio Test)的反射系数$\varrho^2$，可以表示为公式(2.8)：

$$\varrho^2 = \sum_{i,j} \frac{\beta_{i,j}^2}{\sigma_{i,j}^4} \tag{2.8}$$

假设隐写者可以估计像素方差$\sigma_{i,j}^2$，则最小化最大似然比测试的反射系数可以转化为在嵌入负载$\alpha$的约束下，求解使反射系数$\varrho^2$最小的最优嵌入概率$\beta_{i,j}$，该问题可以表示为公式(2.9)：

$$\begin{aligned} &\min_{\boldsymbol{\beta}} \varrho^2 = \min_{\boldsymbol{\beta}} \sum_{i,j} \frac{\beta_{i,j}^2}{\sigma_{i,j}^4} \\ &\text{s.t.} \sum_{i,j} H\left(\beta_{i,j}\right) = \alpha \end{aligned} \tag{2.9}$$

其中，$\boldsymbol{\beta} = (\beta_{i,j})$，$\sum_{i,j} H\left(\beta_{i,j}\right)$表示嵌入的信息量，$\alpha$为给定的嵌入负载。

当求解得到嵌入概率$\beta_{i,j}$之后，为了使用 STCs 编码方式，隐写者需要将嵌入修改概率$\beta_{i,j}$转化为嵌入成本$\rho_{i,j}$，转化关系为公式(2.10)：

$$\rho_{i,j} = \frac{1}{\lambda} \ln\left(\frac{1}{\beta_{i,j}} - 2\right) \tag{2.10}$$

其中$\lambda$是为了满足嵌入负载限制的拉格朗日乘数。

J-MiPOD[33]是将空域隐写算法 MiPOD 拓展到 JPEG 域的 JPEG 隐写算法，该算法首先将 JPEG 载体图像解压缩至空域，然后求解使最大似然比测试的反射系数最小的嵌入修改概率。MiPOD 和 J-MiPOD 的创新之处在于其嵌入修改概率(或者嵌入成本)是通过最小化嵌入对载体统计模型的影响来确定的。相比之下，一些其他内容自适应隐写算法(WOW，UNIWARD，HILL 等)都是通过量化嵌入变化对邻域元素的影响来计算图像元素的嵌入成本。

### 2.1.2 基于深度学习的图像隐写

基于失真代价函数的图像隐写算法主要通过轻微修改载体图像，通过最小化统计特性的变化来提高安全性，但是此类方法比较依赖启发式设计。相比之下，深度学习技术在图像特征提取和表示方面具有端到端的优势，并且能够通过大量数据迭代学习以减少人工成本。因此，研究者们借鉴卷积神经网络的思想构建了许多用于图像隐写的模型。同时，深度学习技术也被广泛应用于图像隐写分析领域，并取得了良好的成果。为了增强隐写算法的安全性，一些方法还引入了对抗训练[34]。本节将重点介绍基于卷积神经网络和基于生成对抗网络这两种图像隐写算法的实现原理与应用。

(1) 基于卷积神经网络的图像隐写

2015 年， Baluja[35]提出了一种将秘密彩色图像隐藏到另外一幅彩色图像中的隐写算法，该算法由预处理网络、隐写网络和提取网络三个网络共同组成，且这三个网络同时训练。实验证明了该隐写网络不像 LSB 算法一样直接修改像素值以嵌入信息，而是将秘密信息隐藏到载体图像的每一个图像通道中，但是该算法的缺点在于得到的含密图像不可察觉性较差。2018 年，Rehman 等人[36]提出了将灰度图像隐藏到彩色图像中的隐写算法，该隐写网络通过主副两个分支分别处理秘密灰度图像和载体彩色图像，主分支输入载体图像的同时还输入副分支的输出，但得到的含密图像有一定泛黄的问题。

为了提高安全性，Sharma 等人[80]于 2019 年提出了与 Baluja 类似但预处理网络作用不同的隐写算法，其预处理网络只对秘密图像进行加密但不参与训练。但是该网络得到的含密图像能看出图像加密的痕迹，安全性的提升效果并不明显。2021 年，Yang 等[81]也提出了针对秘密图像的隐写算法，该算法将载体图像中提取的特征在发送端作为密钥对秘密图像加密，在接收端通过提取含密图像的特征来进行解密，与 Sharma 的算法相比，安全性得到了显著提高。

为了降低含密图像的不可察觉性，Duan 等人[82]提出了一种基于改进 Xception 算法的隐写网络，该网络使用改进的 Xception 结构提高网络适应性的同时获得不同尺度的特征，并且使用密集连接以加速网络的收敛，与先前的方法相比，得到的含密图像具有较高的视觉质量。受 U-Net 网络和 Inception 模块

的启发，Kich 等人[83]提出了将 U-Net 网络中卷积层替换为 Inception 结构的隐写网络，利用 Inception 结构中不同尺寸的卷积核以获得不同维度的隐写特征，有助于将秘密图像的细节隐藏到载体图像中。虽然该算法可以适用于不同尺寸的图像，且含密图像的视觉质量较高，但是重建的秘密图像存在颜色失真的现象。Subramanian 等人[84]提出了一种基于自编码器的图像隐写算法，该算法使用预处理网络分别处理载体图像和秘密图像以提高隐写网络的处理效率，后续的隐写网络和提取网络都使用了卷积神经网络。虽然模型较为简单，但同样可以得到较高质量的含密图像，不过重建的秘密图像存在失真较大的问题。

除了降低含密图像的不可察觉性，提高隐写容量也是研究重点。Baluja 等人[85]在之前的研究基础上提出了一种一图藏多图的隐写模型，该模型能够将两幅彩色图像隐藏到另外一幅彩色图像中，而且利用隐写图像的残差图像降低了被检测的概率。Lu 等人[86]提出了两图藏一图的隐写算法 ISN，该算法将图像的隐写和提取看作为一个逆问题，使用 ISN 的共享参数以实现高效的图像隐写和重建，具有较好的不可察觉性。虽然 Baluja 和 Lu 的算法实现了多图隐写，但较大的隐写容量也使得图像的抗隐写分析能力下降。

(2) 基于生成对抗网络的图像隐写

生成对抗网络(Generative adversarial networks, GAN)[34]的原理是通过对抗训练来生成所需要的图像，主要分为生成网络和鉴别网络两部分。Volkhonskiy 等人[87]于 2016 年提出了 SGAN(Steganalysis GAN)算法，该算法通过生成网络生成载体图像，然后使用$\pm 1$的方式嵌入秘密信息，再使用鉴别网络用来衡量所生成含密图像的不可察觉性，两个网络共同训练实现了高隐蔽性的秘密信息嵌入。但是该网络的缺点在于生成的载体图像视觉质量不高。与 SGAN 类似，Shi 等人[37]提出了具有收敛更快的损失函数的 SSGAN，该网络具有更好的稳定性，且含密图像视觉质量更高。这两个隐写算法都是通过生成对抗网络生成适于隐写的载体图像，然后使用传统隐写算法进行秘密信息的嵌入。图像隐写得到含密图像也可以等同于一个生成图像的过程，隐写网络可以类比于生成网络。因此，在后续的研究中，研究者主要直接采用生成对抗网络来生成含密图像，这类隐写算法通常由三部分组成：隐写网络、提取网络和隐写分析网络。

Zhu 等人[39]受到神经网络对于图像微小扰动敏感性的启发提出了 HiDDeN 算法，该算法采用生成对抗网络生成含密图像，并通过对抗训练提高隐写图像的安全性。此外，为了增强其鲁棒性，在 HiDDeN 中还加入了噪声层，使网络在多种攻击下仍能安全地进行隐写。Wang 等人[88]提出了 HidingGAN，通过使用多个 Inception-ResNet 模块来融合不同尺度的特征，在隐写容量和视觉质量上相较 HiDDeN 有了明显提升，损失函数也引入了感知损失，以提高含密图像的不可察觉性。Zhang 等人[89]提出了 SteganoGAN 模型，可以将任意二进制数字隐藏在载体图像中，嵌入率达到 4.4 比特每像素。其实验结果表明，密集连接方式可以有效地提高含密图像的质量。Wang 等人[90]将 UNet++结构应用于自编码器隐写网络，通过叠加和整合 DCT 非零交流系数不同层的特征来提高含密图像的视觉质量。Zheng 等人[91]提出了 IIH-GAN 模型，该模型使用 SE-ResNet（Squeeze and Excitation ResNet）作为隐写和提取网络，最大限度地保留了载体图像的信息，从而确保了含密图像的视觉质量和准确性。

以上隐写算法相较传统算法，在安全性和隐写容量方面都有一定提升。然而，由于隐写容量仍然存在限制，因此许多研究者提出了嵌入率更高的隐写算法，将秘密图像隐藏在尺寸相同的载体图像中。Zhang 等人[38]提出了 ISGAN 算法，可以将灰度图像隐藏在彩色图像中，并引入了 Y 通道方案以解决隐写图像泛黄的问题。此外，为了提高抗隐写分析能力，他们将修改后的 XWS-CNN 作为隐写分析网络，并提出组合式的损失函数来提高含密图像的视觉质量。Chen[92]提出的隐写算法将灰度图像隐藏到载体图像的 B 通道，以解决颜色失真问题，并同样使用修改后的 XWS-CNN[59]作为隐写分析网络。Fu 等人[93]提出了 HIGAN 算法，可以将彩色图像隐藏到载体图像中。该算法使用多个残差块组成隐写网络，并将修改过的 XWS-CNN 作为隐写分析网络，通过加深网络的深度，提高了含密图像的视觉质量。与 ISGAN 相比，这些算法都采用了不同的方法来解决颜色失真的问题，并且使用类似的隐写分析网络来提高抗隐写分析能力。

综上所述，基于深度学习的图像隐写算法通常由隐写网络和提取网络两部分组成，各个方法之间的主要区别在于采用了不同的网络结构，其相较于传统算法在不可察觉性和容量方面表现更优。虽然增加网络的宽度和使用残差结构

来加深网络深度等手段可以提高隐写算法的性能，但仍然存在一些问题，比如含密图像颜色失真以及重建的秘密图像视觉质量一般等。此外，这两种方式还会增加计算复杂度和内存消耗。在大容量图像隐写的情况下，仍然需要进一步提高含密图像的抗隐写分析能力。因此，基于深度学习的图像隐写算法仍需要继续研究和优化，以满足更安全的隐蔽通信需求。

## 2.2 图像隐写分析技术

当前，图像隐写分析方法可分为基于人工特征的方法和基于深度学习的方法。基于人工特征的隐写分析主要包括构造隐写分析特征和分类器的训练两个步骤，其中用于隐写分析特征构建的滤波器通常需要经过人工设计调整以达到最优效果。此类方法虽然具有良好的性能表现，但人工成本较大以及泛化能力有限一定程度限制了其发展。而基于深度学习的方法能够弥补这个问题，成为研究者广泛采用的新兴技术。基于深度学习的方法使用端到端的技术，即通过神经网络来自动提取隐写分析特征来完成图像二分类任务。通过深层表示学习，这种方法能够有效地利用图像的多层次特征信息，避免了传统方法中需要人工挑选特征的难题。同时，在模型训练过程中，神经网络基于大量数据进行自主优化，可以进一步提高分类准确性。下面将分别介绍传统图像隐写分析和基于深度学习的图像隐写分析。

### 2.2.1 基于人工特征的图像隐写分析

基于人工特征的隐写分析方法分为空域图像隐写分析和 JPEG 域图像隐写分析。空域隐写分析方法通常利用空域像素值的统计特征来分析图像的隐写信息。其中，空域“富”模型 SRM[43]和 PSRM[44]是非常典型的代表。SRM 采用平滑卷积操作，通过像素相邻差值的熵来描述图像的统计特征。PSRM 在 SRM 的基础上，加入了投影距离的权重，进一步提升了分类准确率。JPEG 域特征隐写分析方法则是基于图像变换域的特征进行分析。其中，JPEG 相位特征[50]是比较经典的一种方法，其他类似方法还包括 DCTR[51]、PHARM[52]和 GFR[53]等。这些方法主要关注 DCT 系数的变化，并采用不同的统计特征描述图像的隐写信息。例

如，DCTR 采用二阶统计量来刻画 DCT 系数的变化，PHARM 在 DCTR 基础上加入了相邻有效系数的位置信息，GFR 则使用了全局像素均值和方差构建典型结构。下面将对经典的 SRM 和 JPEG 相位特征作简要介绍。

（1）SRM

SRM 提取隐写特征的流程如图 2.2 所示，待检测图像首先通过多个高通滤波器计算得到残差特征图，然后对每个残差图进行量化和截断，最后构造每个特征子模型的四阶共生矩阵并使用特征融合降维。SRM 特征的具体提取方法可以分为以下几步。

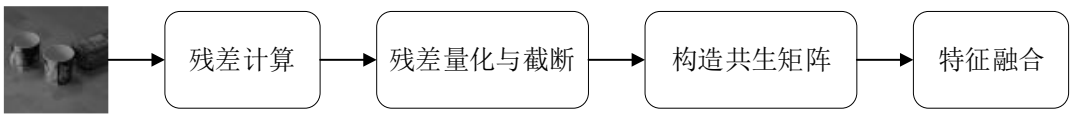



图2.2 SRM特征提取

a）残差计算

SRM 设计了多组高通滤波器来计算噪声残差特征图，设待检测图像为$\boldsymbol{X}$，定义$x_{i,j}$为图像第$(i,j)$个像素点，则待检测图像的线性残差可由公式(2.11)计算：

$$r_{i,j} = pred(ne_{i,j}) - RCx_{i,j} \tag{2.11}$$

其中，$RC$为残差阶数，$(i,j)$为像素的横纵坐标，$ne_{i,j}$为像素点$x_{i,j}$的邻接像素，$pred(ne_{i,j})$为像素点$x_{i,j}$的预测值，$r_{i,j}$为图像的噪声残差。噪声残差共包括一阶、二阶、SQUARE、EDGE$3\times3$和EDGE$5\times5$六种类型，每种残差类型下还分为线性滤波残差和非线性滤波残差。

b)量化与截断

第一步噪声残差计算的结果通常数值跨度较大，因此隐写分析中一般会对噪声残差再进行一步量化和截断，具体计算如公式(2.12)：

$$r_{i,j} \leftarrow trunc_T\left(round\left(\frac{r_{i,j}}{q}\right)\right) \tag{2.12}$$

其中，$q$为量化系数，$round(\cdot)$表示四舍五入取整函数，$trunc_T(\cdot)$表示阶段函数，$T$为截断阈值，截断函数能够将量化残差的数值范围限定在$-T$和$T$之间。

c)构造共生特征矩阵

SRM 在水平和垂直两个方向上给每个特征子模型构造了四阶共生矩阵，如表达式(2.13)：

$$\begin{aligned} f_d^{(\mathrm{h})} &= \frac{1}{Nor}\sum_{i,j=1}^{M,N-3}\left[r_{i,j+k-1}=d_k,\forall k=1,\cdots,4\right] \\ f_d^{(\mathrm{v})} &= \frac{1}{Nor}\sum_{i,j=1}^{M-3,N}\left[r_{i+k-1,j}=d_k,\forall k=1,\cdots,4\right] \end{aligned} \tag{2.13}$$

其中，$d_k\in\{-T,-T+1,\cdots,T\}$，$\boldsymbol{d}=(d_1,d_2,d_3,d_4)$，$M\times N$ 为图像的尺寸大小。$[Z]$表示艾佛森括号，当 $Z$ 成立时，$[Z]=1$，否则为零。$f_d^{(\mathrm{h})}$ 和 $f_d^{(\mathrm{v})}$ 分别为水平方向和垂直方向的四阶共生特征矩阵，$Nor$ 为归一化参数。

d)特征融合

SRM 算法采用特征合并方法，成功将 SRM 的特征维度减小至 34671，为后续的分类降低了计算复杂度。

(2) JPEG 相位特征

JPEG 相位是进行 JPEG 压缩的过程中所采用的一种编码方式，基于 JPEG 相位的隐写分析特征是通过收集 JPEG 相位残差的直方图特征来实现的。隐写信息对解压缩 JPEG 图像中像素的影响取决于 JPEG 相位，因此 DCTR 和 GFR 两个特征都是基于相位划分的方式构建的。其中，DCTR 使用 64 个 DCT 基函数作为滤波器对解压缩图像进行滤波，然后对每个滤波后的图像基于 JPEG 相位进行划分，并将划分好的 JPEG 相位残差构造成直方图特征。GFR 使用二维 Gabor 滤波器来替换 DCTR 中的 DCT 基函数滤波器。尽管这两种特征构造方法略有不同，但是它们都采用了基于 JPEG 相位的方式来进行隐写分析，因此可以采用统一的方式来介绍方法步骤。

a) 计算残差

首先，需要选择合适的大小为 $k_1\times k_2$ 的滤波器 $\boldsymbol{K}\in\mathbb{R}^{k_1\times k_2}$ 对大小为 $M\times N$ 的解压缩图像 $\boldsymbol{X}$ 进行滤波处理，如式(2.14)：

$$r(\boldsymbol{X},\boldsymbol{K})=\boldsymbol{X}\star\boldsymbol{K} \tag{2.14}$$

其中，$r(\boldsymbol{X},\boldsymbol{K})$表示解压缩图像的噪声残差，$\star$ 为卷积操作。

b）量化和截断

对滤波后的图像基于 JPEG 相位进行量化划分，如公式(2.15)：

$$r(\boldsymbol{X},\boldsymbol{K},Q)=Q_Q\left(\frac{r(\boldsymbol{X},\boldsymbol{K})}{q}\right) \tag{2.15}$$

其中，$Q_Q(\cdot)$为量化器，$Q=\{0,1,\cdots,T\}$，$q$ 为固定量化步长。

c）构造 JPEG 相位直方图$h_g^{(a,b)}(\boldsymbol{X},\boldsymbol{K},Q)$，如公式(2.16)：

$$h_g^{(a,b)}(\boldsymbol{X},\boldsymbol{K},Q)=\sum_{m=1}^{[M/8][N/8]}\sum_{n=1}\left[r_{a,b}^{(m,n)}(\boldsymbol{X},\boldsymbol{K},Q)=g\right] \tag{2.16}$$

其中，$0\le a,b\le 7$，$0\le g\le T$，$T$ 为截断阈值，[·]为向下取整函数。最后将直方图进行合并和融合得到 JPEG 相位隐写分析特征。

### 2.2.2 基于深度学习的图像隐写分析

近年来，深度学习因其处理图像特征的优异表现而被广泛应用于计算机视觉任务。由于深度学习技术具有很强的特征迁移能力，学者们也开始将其引入到图像隐写分析中。基于深度学习的图像隐写分析一般由三个部分组成，分别是噪声残差计算、特征学习和二分类，其通用框架如图 2.3 所示。其中，特征提取的目的是提取隐写特征图，一般使用高通滤波器或者神经网络进行残差计算得到特征图。特征学习的目的是提取特征分类阶段的判别特征，一般由神经网络层组成，最后进行二分类判断是否隐写。

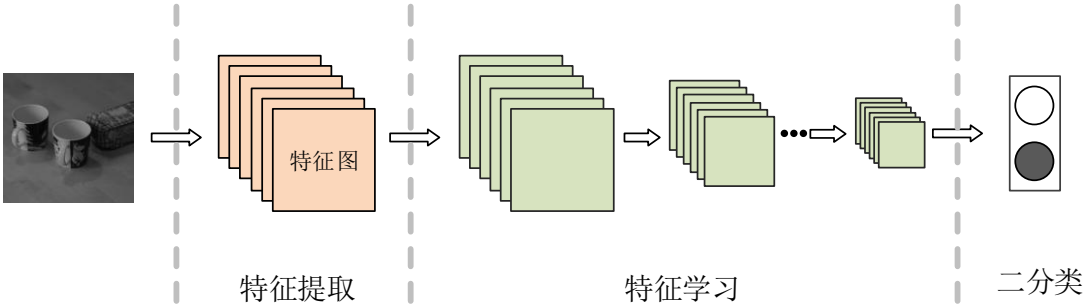



图2.3 基于深度学习的图像隐写分析通用框架

基于深度学习的图像隐写分析主要分为空域隐写分析和 JPEG 域隐写分析。空域隐写分析模型通常基于数字图像的像素值进行隐写分析，模型通过训练神经网络学习提取与隐写信息相关的特征，更好地识别出隐写图像，从而达到良好的检测性能。而 JPEG 域隐写分析模型可通过 JPEG 图像的 DCT 系数的相位信息来构建隐写特征，利用 JPEG 图像压缩的特性基于深度学习技术进行分析，判断是否存在隐写信息。也可以通过先将 JPEG 图像解压缩到空域，再通过卷积神经网络进行隐写分析。以下将对两种类型中的经典方法做相应介绍。

(1) 基于深度学习的空域图像隐写分析

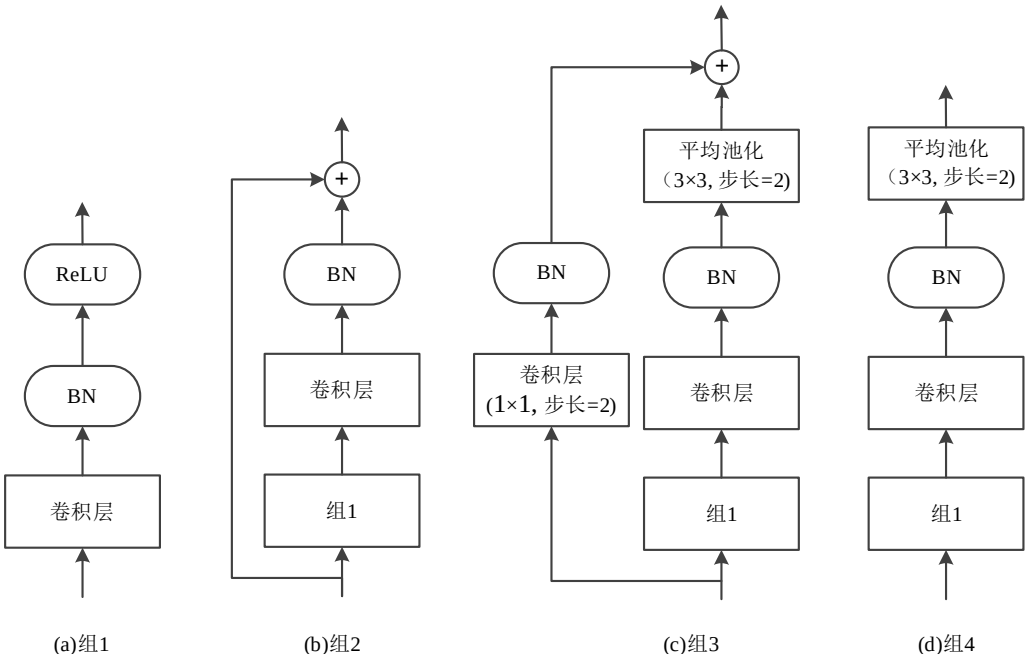
2014 年，Tan 等人[57]尝试将深度学习方法应用于隐写分析，并试图弥补传统方法在低信噪比环境下的不足。他们提出的模型的核心思想是直接利用卷积神经网络对图像进行特征提取和处理，从而检测出其中是否隐藏秘密信息。然而该模型缺少预处理环节，导致隐写嵌入的一些特征没有被完全挖掘。为了进一步提高检测性能，Qian 等人[58] 于 2015 年设计了一个具备预处理环节的卷积神经网络模型。该模型的预处理层采用高通滤波器来提取信号中的高频噪声以提高模型训练的收敛速度，最终取得了较好的性能。

2016 年，Xu 等人[59] 基于上述工作进一步优化，并提出了将深度学习应用于隐写分析任务且性能超越 SRM 的模型 XWS-CNN。该模型在第一组卷积池化组中使用绝对值激活函数（ABSolute activation, ABS）生成特征映射中元素的绝对值，以促进后续层的统计建模。为了防止过度拟合，在网络的早期阶段使用双曲正切（TanH）的饱和区域约束数据值的范围，在更深的层中使用整流线性单元（Rectified Linear Unit, ReLU）[94]激活函数，且使用 1×1 卷积降低建模的强度。同时，模型还引入了批量归一化（Batch Normalization, BN）模块以解决卷积神经网络中的梯度下降的问题。此外，还使用了全局平均池化（Global Average Pooling, GAP）降低维度来加速计算。

2017 年，Ye 等人[61]提出了空域隐写分析网络 YeNet，其检测性能显著优于 SRM。该网络不使用传统的高通滤波器获取残差信号，而是利用 30 个 SRM 滤波器对图像进行预处理，以获得更加多样化的残差信号，并且滤波器参数会根据训练进行更新。此外，在预处理层中，YeNet 引入了一种新的激活函数——

截断线性单元(Truncated Linear Unit, TLU)，它可以帮助模型更好地适应隐写信号分布，并强制模型关注小幅度的隐写信号，而不是大幅度的图像内容信号。

Boroumand 等人[62]于 2018 年提出了一个完全端到端的隐写分析模型 SRNet，该网络结构不同于先前方法，放弃了使用手工设计的滤波器来初始化第一层卷积核，而是采用随机初始化方法来初始化所有的卷积层，SRNet 中的卷积池化组如图 2.4 所示。为了避免抑制隐写信息，SRNet 的前七层卷积层并未使用类低通滤波器的平均池化层。同时引入残差连接后，SRNet 成功缓解了梯度消失的问题，并且使得模型更容易学习到图像的噪声残差特征。




(a)组1　(b)组2　(c)组3　(d)组4

图2.4 SRNet中的卷积池化组

2020 年，Zhang 等人[65]提出了 ZhuNet，采用了小尺寸的 3×3 卷积核和大尺寸的 5×5 卷积核共同预处理图像。这种预处理策略利用小尺寸卷积核更好地关注局部区域的细节信息，同时利用大尺寸卷积核更好地捕捉全局结构特征。此外，ZhuNet 还引入了深度可分离卷积层以增强特征之间的通道相关性，并应用空间金字塔池化[95]聚合局部特征以提高其表示能力。

2021 年，You 等人[66]提出了一种基于孪生网络的隐写分析方法 SiaStegNet。该网络使用两个相同结构的子网络对输入图像进行处理，同时设计了两个损失函数来帮助网络收敛。其中一个损失函数用于测量孪生网络的相似度，而另一个损失函数则用来激励网络在提取特征时关注到隐写信息。Ren 等人[67]基于对

比学习提出了一种隐写分析框架，通过最大化不同类别样本特征之间的距离和最小化同一类别样本特征之间的距离来改善隐写分析的特征表示。

(2) 基于深度学习的 JPEG 域图像隐写分析

2017年，Zeng等人[96]受到XWS-CNN[59]的启发提出了基于深度学习的JPEG域图像隐写分析框架，该网络与 XWS-CNN 类似对图像进行了预处理操作，使其检测性能优于传统方法 DCTR。同年，Xu[97]在 XWS-CNN[59]基础上设计了适用于JPEG隐写分析深度残差网络J-XuNet，其深度达到了20层。并通过实验表明，使用卷积层替换池化操作可以改善检测精度。

受到JPEG相位特征[52]影响，Chen等人[98]提出了一种基于JPEG相位感知的隐写分析网络。该网络修改了空域隐写分析网络 XWS-CNN，并将特征图划分为 64 个并行通道，以便于隐写分析网络感知 JPEG 相位。作者还引入了两种将相位感知整合到网络架构中的方法，分别为P-Net和V-Net。

适用于空域隐写分析的SRNet[62]同样适用于JPEG域隐写分析，其输入为未取整的解压缩像素。相比 J-XuNet 和 Chen 等方法，SRNet 具有更高的 JPEG 域隐写分析检测性能。Jang 等人[68]提出了基于特征聚合的 JPEG 域隐写分析网络FANet。FANet 扩展了输入图像的卷积块的通道数，并将各种层级和不同分辨率的特征图进行聚合，以利用丰富的信息来提高隐写分析性能。

2020年，Yousfi等人[69]调研了预训练过的网络结构在ALASKA#2隐写分析竞赛[70]中的表现，这些网络结构包括 EfficientNet[71]、MixNet[72]和 ResNet[73]等。这些在 ImageNet[74]上预训练的模型可以快速收敛于 JPEG 隐写分析数据集上，并获得比只在原隐写分析数据集上训练的隐写分析 CNN(如 SRNet[62])更好的性能。研究者发现移除第一层卷积层的池化操作可以更好地保留微弱的隐写信号，从而获得更高的隐写分析准确率。

综上所述，基于深度学习的图像隐写分析算法从一开始需要单独借助高通滤波器对图像进行预处理，然后再使用卷积神经网络进行分类，发展到完全使用神经网络对图像进行端到端的处理。与传统图像隐写分析算法相比，基于深度学习的图像隐写分析算法利用了损失函数帮助检测结果收敛，通过多轮模型训练实现了同步优化，节约人工资源的同时提升了隐写分析器的性能。

## 2.3 图神经网络

### 2.3.1 图论基础

图作为一种可以广泛描述各类关系型数据的数据结构，许多实际情况可以方便地用一个图来描述，每个图由一组点和连接这些点的线组成。例如，点可以是通信中心，用线代表通信联系。在数学中，图由节点和连接节点的边构成，节点表示研究对象，边表示两个对象间的联系。图可以表示为节点和边的集合，记作 $G=(V,E)$，其中 $V=\{v_1,v_2,v_3,\ldots,v_n\}$ 是数量为 $n$ 的节点的集合，$E$ 为边的集合。即 $v_i,v_j\in V$ 为图的两个节点，$e_{ij}(v_i,v_j)\in E$ 则表示连接节点 $v_i$ 和 $v_j$ 的边。

当根据边是否有指向来划分时，图可以分为有向图和无向图。有向图中所有边为有向边，$e_{ij}$ 和 $e_{ji}$ 表示两条指向相反的边；无向图中 $e_{ij}$ 和 $e_{ji}$ 则有相同的含义。当根据图中的每条边是否含有权重来划分，图可分为加权图和非加权图。加权图的每条边都与一个实数相对应，该实数成为对应边上的权重。一般情况下，将权重抽象为两个节点之间的关联强度。与之相反的是非加权图，非加权图中一般使用 0 代表边不存在，1 代表边存在，存在的边可以认为权重相同。

为了将图数据能够方便的储存在计算机中，通常可以使用邻接矩阵来储存节点之间的关系。设图 $G=(V,E)$，$E=\{e_1,e_2,e_3,e_4,e_5,e_6\}$，如图 2.5 所示。

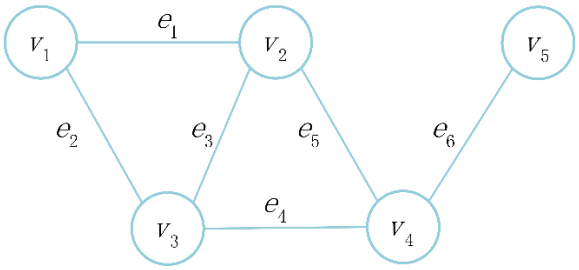



图2.5 图$G$示例

可用邻接矩阵 $A$ 描述图中节点之间的关联，$A\in\mathbb{R}^{5\times5}$ 可由式(2.17)表示：

$$A_{ij}=\begin{cases}1,(v_i,v_j)\in E\\0,(v_i,v_j)\notin E\end{cases}\tag{2.17}$$

用邻接矩阵存储图的时候，需要使用一个一维数组表示节点集合，一个二维数组来表示邻接矩阵。需要特别说明的是，在实际的图数据中，邻接矩阵由

于图数据的特殊性往往会出现大量的 0 值，因此可以用稀疏矩阵的格式来存储邻接矩阵。图2.6给出了图 $G$ 的邻接矩阵存储表示，通过该图可以看出，无向图的邻接矩阵是沿主对角线对称的，即 $A_{ij} = A_{ji}$。

| | $v_1$ | $v_2$ | $v_3$ | $v_4$ | $v_5$ |
|---|---|---|---|---|---|
| $v_1$ | 0 | 1 | 1 | 0 | 0 |
| $v_2$ | 1 | 0 | 1 | 1 | 0 |
| $v_3$ | 1 | 1 | 0 | 1 | 0 |
| $v_4$ | 0 | 1 | 1 | 0 | 1 |
| $v_5$ | 0 | 0 | 0 | 1 | 0 |

图 2.6 图 $G$ 的邻接矩阵

### 2.3.2 图数据深度学习

图数据相关的任务根据其结构的特殊可以划分为不同层面的问题，主要分为节点层面、边层面和图层面三种类型。下面将简单介绍三种类型的任务：

(1) 节点层面(Node Level)的任务

对于节点层面的任务，通常需要预测每个节点的性质，包括分类任务和回归任务。这类任务常常涉及到图数据中的单个节点特征，但是建模往往建立在整个图数据上，因此节点之间的关系也需要考虑。例如，论文引用网络中，可以使用节点分类来判断每个论文的类型；在线社交网络中，可以通过用户标签的分类和恶意账户检测来保证社交网络中的安全性。

(2) 边层面(Link Level)的任务

边层面的任务主要包括边的分类和预测任务。边的分类是指对边的某种性质进行预测；边的预测是指评估给定的两个节点之间是否会构成边。常见的应用场景如在社交网络中，将用户作为节点，用户之间的关注关系建模为边，通过边预测实现社交用户的推荐。目前，边层面的任务主要集中在推荐业务中。

(3) 图层面(Graph Level)的任务

图层面的任务不针对某个节点或者某条边单独的属性，而是通过整体结构来实现分类、表示和生成等任务。这类任务通常应用在自然科学研究领域，例如对药物分子和酶的分类等，从而提高药物设计和生物学研究的效率。

随着图数据的重要性日益凸显，许多图学习理论也专注于解决与图相关的任务。其中，谱图理论(Spectral Graph Theory)将图论与线性代数相结合，以解决图的分类或节点的聚类问题，而统计关系学习则是一种机器学习理论，将关系表示和似然表示相结合，并打破了传统机器学习算法中对数据独立同分布假设的限制。为了更好地挖掘异构图中丰富的信息，研究者们提出了异构信息网络(Heterogeneous Information Network)，用于分析异构图的结构信息和语义信息。近年来，随着深度学习在实际应用领域取得的巨大成就，表示学习和端到端学习获得越来越多的重视。因此，网络表示学习(Network Embedding)方法被广泛研究，以从复杂的图数据中学习到包含充分信息的向量化表示。然而，网络表示学习面临的一个难题是如何设计将表示学习与任务学习相结合的端到端系统。因此，图数据的端到端学习系统仍然是一个重要的研究课题。

由于图数据的结构复杂性，直接定义出可导的计算框架并不容易。相比之下，像图像、语音和文本这些规则化数据结构是定义在欧式空间的，因此它们适用于基于张量计算体系的处理方式。而对于文本数据来说，它是一种序列数据，所以循环神经网络的工作机制更为适合处理这种序列结构。与这些常见的数据类型相比，图数据可呈现出规则的 2D 栅格结构，这种栅格结构与卷积神经网络的作用机制非常匹配。

近年来，受图信号处理领域对于图信号卷积滤波的定义的启发，图神经网络(Graph Neural Network, GNN)在很大程度上是基于图卷积操作而不断衍生而来[99]。2005 年，Marco Gori 等人[100]首次提出了图神经网络的概念。在这之前，处理图数据的方法是将图转换为一组向量表示，并且通常会导致结构信息的丢失和结果高度依赖于预处理过程。因此，GNN 的发展旨在能够将学习直接建立在图数据之上。随后 2008 年，Gori 等人[101]进一步阐述了图神经网络并提出了监督学习的方法用于训练 GNN。但是，在早期的研究中，采用迭代的方式通过循环神经网络传递邻居信息，直到达到固定状态以学习节点的表征。然而这种方式比较耗费内存，相关研究开始关注如何改进方法以减少计算量。

2012 年，卷积神经网络在视觉领域取得重要进展，因此人们开始研究如何将卷积运算应用于图神经网络中。2013 年，Bruna 等人[102]首次在基于频域卷积

的概念下提出了图卷积网络模型，将可学习的卷积操作用于图数据之上。此后，越来越多的方法被提出并逐渐流行，此时图卷积神经网络已经可分为基于空域和基于谱域两大类。基于空域的图卷积方法便是针对处理大规模图数据学习任务时频域卷积方法存在的时间复杂度问题而提出的改进，它从节点的空间关系出发，利用消息传播机制处理图数据。基于谱域的方法核心思想是通过应用图信号处理中的滤波器，将图卷积操作视为一种去噪处理的过程。实际上，图卷积操作本质上就是对节点的邻接矩阵进行傅里叶变换，并引入一个“谱”的概念，然后进行滤波操作以过滤掉无关的噪声等，这种方法能够有效地提取出数据集中的特征，从而为更深层的网络学习提供支持。

在 2017，J Gilmer 等人[103]提出了消息传播机制神经网络（Message Passing Neural Network, MPNN），这是一种基于空域操作的图卷积神经网络的通用框架，它几乎奠定了现有空域图卷积网络的基础。它将图卷积视为一个信息传递的过程，其中每个节点都可以通过边向其相邻节点传递信息，经过 $k$ 轮信息传递后，该节点的信息可传递给距离更远的节点。这种信息传播机制也被称为空域图卷积，可由式(2.18)表示：

$$\boldsymbol{h}_{v_i}^{k}=U_k\left(\boldsymbol{h}_{v_i}^{k-1},\sum_{u\in N(v_i)}M_k\left(\boldsymbol{h}_{v_i}^{k-1},\boldsymbol{h}_{u}^{k-1},\boldsymbol{e}_{uv_i}\right)\right) \tag{2.18}$$

其中，$\boldsymbol{h}_{v_i}^{k}$ 表达经过 $k$ 论传递后 $v_i$ 的节点向量，$\boldsymbol{h}_{v_i}^{0}$ 表示 $v_i$ 的初始节点向量，$N(v_i)$ 为节点 $v_i$ 所有邻接节点的集合。$M_k(\bullet)$ 定义了节点收集邻域节点传递消息的方式，$U_k(\bullet)$ 定义了各节点通过收集的消息更新自身节点信息的方式。

迭代后的节点向量可用于节点层面的任务如节点分类，如果继续使用一个读出函数聚合所有节点向量则可以得到图层面的特征表示，如式(2.19)，可以用于图层面的分类任务。

$$\boldsymbol{h}_G=R\left(\boldsymbol{h}_v^k\boldsymbol{v}\in\boldsymbol{V}\right) \tag{2.19}$$

现有的空域图卷积网络大都属于以上描述的这个框架，只是对函数 $M_k(\bullet)$，$U_k(\bullet)$，$R(\bullet)$ 的定义方式有所区别。

随后，越来越多的基于空域图卷积的神经网络模型被设计出来，如图卷积神经网络(Graph Convolution Network, GCN)[99]、图注意力网络图注意网络

(Graph Attention Network, GAT)[104]等，这些模型极大地增强了深度学习系统对各类图数据的适应性，为诸多图数据的应用场景下的任务提供了一个极具竞争力的学习方案。因此，图数据与深度学习之间的结合也迎来了第一次真正意义上的成功，并为诸多领域的自动化问题解决提供了强有力的支持。

## 2.4 本章小结

本章介绍了内容自适应图像隐写算法的最小化失真框架和一些代表性成果，并介绍了将深度学习应用于图像隐写的发展现状，接着介绍了基于手工提取特征的传统隐写分析方法，其中详细介绍了空域“富”模型和 JPEG 相位特征。并且分别介绍了将深度学习应用于空域和 JPEG 域的图像隐写分析技术。最后介绍了图论的基础和图神经网络的发展概况。

# 第三章 基于图神经网络的空域图像隐写分析

## 3.1 引言

隐写痕迹可以通过载体图像的固有噪声成分很好地隐藏，这些成分通常位于高频区域。它启发研究者们使用自适应机制[105]或最小失真框架[23]优先将秘密数据嵌入到这些难以察觉到的区域，以更好地抵抗隐写分析。许多工作都是沿着这条思路设计的，例如 HUGO、HILL 和 UNIWARD。与隐写相反，图像隐写分析的任务是确定给定图像是否隐藏秘密信息。从系统设计的角度来看，早期的隐写分析从媒体对象中手工提取的特征，然后使用传统的统计分析工具，如支持向量机和线性判别分析进行分类。例如，基于马尔可夫的特征已广泛用于早期的图像隐写分析[18,31,97,106]。虽然集成和降维可用于增强检测性能[107]，但这些算法严重依赖于复杂的人工特征设计且泛化能力较差。隐写算法通常会修改图像区域中难以检测的像素，为了克服这一困难，近年来，学者们开始研究如何将计算机视觉中的深度卷积神经网络(deep CNNs)[108]所取得的成功转移到图像隐写分析任务中。这些工作可以简要地概括为三个阶段，即残差提取、特征学习和二元分类。具体来说，它们首先滤波输入图像以生成残差图像，通过放大类噪声隐写信号和正常像素之间的比值，从而促进了特征学习的过程。然后通过向深度 CNN 提供残差图像，可以学习判别特征并将其用于二元分类。这整个过程可以通过端到端的方式实现以简化判别过程。

最近，人们越来越有兴趣将深度学习范式扩展到图数据，推动图神经网络(Graph neural network, GNN)成为一个热门话题[109]。GNN 本质上是图表示学习模型，可以很好地应用于节点层面和图层面的任务。通过将数字图像建模为图数据结构，GNN 就可以有效地解决许多计算机视觉领域的问题。已经有研究者使用 GNN 进行图像分类，并取得了一定的成果[110]。受这一点的启发，尽管 CNN 在图像隐写分析中具有优势，但本章朝着基于 GNN 的空域图像隐写分析迈出了一步。实验结果表明，所提出的基于 GNN 的网络结构具有一定竞争力，显示了图表示学习在图像隐写分析中的潜能，而且能启发后续的研究。

## 3.2 基于图神经网络的空域图像隐写分析

### 3.2.1 总体框架

所提出的基于图神经网络的空域图像隐写分析模型结构包括三个阶段，如图 3.1 所示，分别为图像到图的转换、图表示学习和二值分类。图像到图转换的目的是将图像转换为具有节点特征向量的图数据，包含图像分割和图构建两步。然后，图数据可以被输入到图注意网络(Graph Attention Network, GAT)[104]进行表示学习，其输出的特征向量用于最终的二值分类。

具体来说，图像到图的转换由两个部分组成，即图像分割和图构建。图像分割将图像以特定的策略分割为各个图像块，这些图像块将对应图数据的图节点。图构建使用浅层 CNN 将图像块映射为节点特征，再为其添加边特征以构成图数据。图数据被输入两层的图注意力网络进行图表示学习，再经过图读出函数得到图层面的特征向量。最后二值分类由全连接层和 *softmax* 函数构成，输出输入图像的预测结果。各个模块的细节将在下面小节详细介绍。

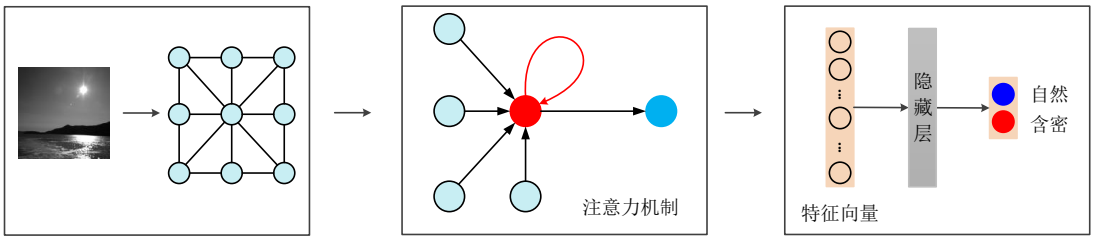



图3.1 基于GNN的空间图像隐写分析框架图

### 3.2.2 图像到图的转换

为了将图像转化为可供图神经网络学习的图数据，本章设计了一种图像到图的转换方法，如图 3.2 所示，具体步骤分为图像分割和图构建两步。图像首先分割为大小相同的图像块，每个图像块由浅层 CNN 映射为特征向量，图构建将特征向量定义为节点特征，再构建边信息共同组成图数据。本节将对图像到图的转换中图像分割和图构建两步进行详细介绍。

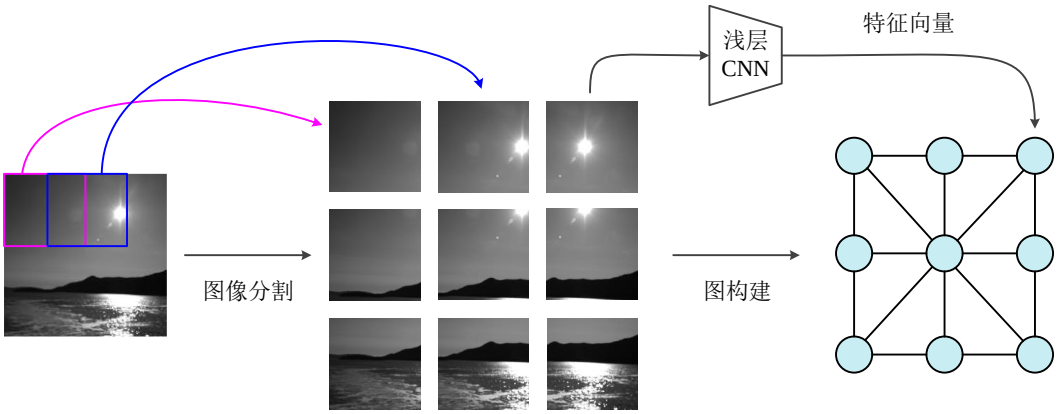



图3.2 图像到图的转换

（1）图像分割

对于输入的灰度图像 $\boldsymbol{I}=\left\{x_{i,j}|1\le i\le h,1\le j\le w\right\}$ 且 $x_{i,j}\epsilon\left\{0,1,...,255\right\}$，首先将 $\boldsymbol{I}$ 划分为 $n\times m$ 个图像块，其中 $n\le h,m\le w$。将一个图像块定义为图像 $\boldsymbol{I}$ 的一个大小为 $h_p\times w_p$ 的子图，同时应满足 $h_p\le h,w_p\le w$。用 $\{\boldsymbol{I}_{u,v}|1\le u\le n,1\le v\le m\}$ 表示所有通过光栅扫描所得的图像块，其中 $\boldsymbol{I}_{u,v}$ 意为该图像块位于位置 $(u,v)$。图像转换到图的第一步就是计算出所有的图像块 $\boldsymbol{I}_{u,v}$，将其定义为：

$$\boldsymbol{I}_{u,v}=\left\{x_{i,j}\middle|i\in\left[f_{u,v},f_{u,v}+h_p\right),j\in\left[g_{u,v},g_{u,v}+w_p\right)\right\} \tag{3.1}$$

其中，$(f_{u,v},g_{u,v})$ 表示 $\boldsymbol{I}_{u,v}$ 左上角像素在 $\boldsymbol{I}$ 中的位置。初始值 $f_{1,1}=g_{1,1}=1$，且

$$f_{u,v}=f_{u,v-1},g_{u,v}=g_{u-1,v},\ \forall u\in[2,n],v\in[2,m] \tag{3.2}$$

对于 $v\in[2,m]$，$g_{u,v}$ 可由下公式求得：

$$g_{u,v}=g_{u,v-1}+(1-\alpha)\cdot w_p \tag{3.3}$$

其中，$\alpha\in[0,1)$ 是控制 $\boldsymbol{I}_{u,v}$ 和 $\boldsymbol{I}_{u,v-1}$ 相交区域的参数。例如：若 $\alpha=0.3$ 意为 $\mathrm{I}_{u,v}$ 中30%的像素也属于 $\boldsymbol{I}_{u,v-1}$。同理，对于 $2\le u\le n$，$f_{u,v}$ 可表示为：

$$f_{u,v}=f_{u-1,v}+(1-\beta)\cdot h_p \tag{3.4}$$

其中，$\beta$ 是控制 $\boldsymbol{I}_{u,v}$ 和 $\boldsymbol{I}_{u-1,v}$ 交集区域的参数。默认情况下，令 $\alpha=\beta$。

例如，当假设$h=w=2h_p=2w_p=512$，且$\alpha=\beta=0$，则$n=m=2$，此时将可以得到四个无交集的图像块，其中$(f_{1,1},g_{1,1})=(1,1)$，$(f_{1,2},g_{1,2})=(1,257)$，$(f_{2,1},g_{2,1})=(257,1)$，$(f_{2,2},g_{2,2})=(257,257)$。若令$\alpha=\beta=0.5$，则可得到 9 个两两之间交集为 50%的子图，相应的左上角的像素点分别为$(1,1)$，$(1,129)$，$(1,257)$，$(129,1)$，$(129,129)$，$(129,257)$，$(257,1)$，$(257,129)$和$(257,257)$。

(2) 图构建：构造一个图$G=(V,E)$分为构建节点和构建边两部分，节点的构建即将每个图像块映射为一个图节点，构建边即给节点对之间建立边的连接。

a) 节点的构建

图节点应与有利于隐写分析的特征向量相关联。为此，本章使用浅层 CNN 将每个高维子图特征映射为一个低维特征向量，并将其分配给相应的节点。本章选用 XWS-CNN[59]进行特征提取。尽管 XWS-CNN 本身在图像隐写分析方面表现出了卓越的性能，但后续实验表明，通过减少卷积层的数量，XWS-CNN 的隐写分析性能将显著下降。当减少层数的浅层 CNN 提取的隐写特征向量输入到图表示学习时，图像隐写分析性能超过原来的浅层 CNN，说明图学习在隐写分析中起着重要作用。

XWS-CNN 共有一个高通滤波层、五个卷积池层和一个线性分类层，各层参数如表 3.1 所示。对于特征提取，本章只使用高通滤波层和卷积池层构成浅层 CNN，将每个二维图像块映射为一个一维特征向量作为节点特征。每组实验中所有图像块将使用相同的浅层 CNN 进行处理，因此只需训练一个浅层 CNN，这样计算成本更低，也减少了多个 CNN 和多个子图之间的错配影响。

表 3.1 所用浅层 CNN 各层参数

| 层名 | 输出大小<br>（通道数×大小） | 卷积核<br>（个数×核大小） | 激活函数 | 池化 |
|---|---|---|---|---|
| 高通滤波层 | 1×(512×512) | - | - | - |
| 卷积池化层1 | 8×(256×256) | 8×(5×5) | ABS+BN+TanH | 平均池化<br>(5×5)<br>步长=2 |
| 卷积池化层2 | 16×(128×128) | 16×(5×5) | BN+TanH | |
| 卷积池化层3 | 32×(64×64) | 32×(1×1) | BN+ReLU | |
| 卷积池化层4 | 64×(32×32) | 64×(1×1) | BN+ReLU | |
| 卷积池化层5 | 128×(1×1) | 128×(1×1) | BN+ReLU | 全局池化 |

浅层 CNN 使用了 XWS-CNN[59]的高通滤波(High-Pass Filtering，HPF)层和卷积池化层(Conv-Pooling Layer，CPL)来构建。XWS-CNN 中有 5 个 CPL，通过从下到上有序组合 CPL，可以构建 5 个不同的浅层 CNN。例如，浅层 CNN 可能仅由 XWS-CNN 的 HPF 层和组 1(即 CPL 1)组成。为了保证由 CNN 输出的特征向量可以输入到后续层，将最后一层的池化操作设置为全局池化。

如图 3.3 展示了由 HPF 和 XWS-CNN 提供的 CPL 组成的四个浅层 CNN。通过应用 HPF 层和所有的 CPL，可以构建第五个浅层 CNN，即 SCNN-V。{SCNN-I， SCNN-II， ...， SCNN-V} 可单独用于隐写分析，方法是添加 XWS-CNN 的分类层。相应的模型本章将其称为 SCNN-I+BC，SCNN-II+BC，SCNN-III+BC，SCNN-IV+BC，SCNN-V+BC。其中，BC 表示二元分类(Binary Classification)，SCNN-V+BC 等同于 XWS-CNN。同时，通过应用本章所提出的框架，基于{SCNN-I，SCNN-II，...，SCNN-V}构建了五个图表示学习模型，并将相应的五个图表示学习模型称为 SCNN-I+GNN+BC，SCNN-II+GNN+BC，...，SCNN-V+GNN+BC。对于所提出的方法，相应浅层 CNN 的输入大小被调整为 $h_p \times w_p$。

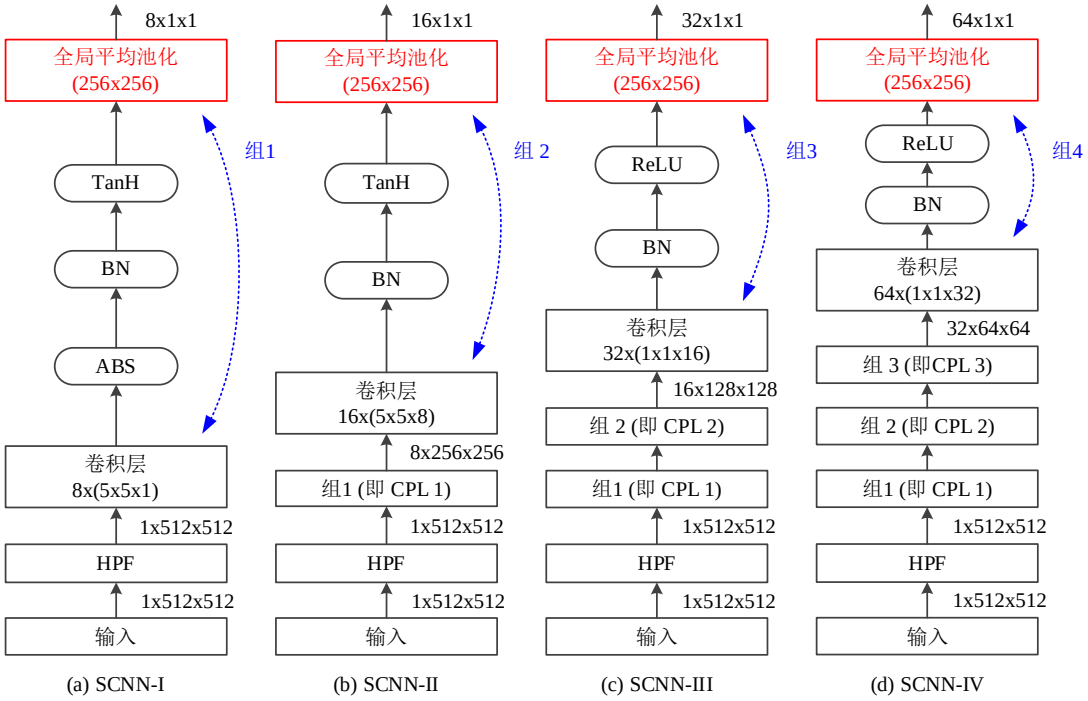



图3.3 基于XWS-CNN 的四种浅层 CNN示意图
(其中 SCNN-X 表示浅层 CNN 中的 CPL 数量为“X”)

(2) 边的构建

对于如何构建最适合的边的方法需要进一步的实验求证。本章定义了以下两种构建边的方式：

第一种是完全图，定义为在任意两个不同的节点之间添加一条边，如图3.4(a)。第二种是格图，该定义利用节点之间的空间关系来构造图，对于两个子图$\mathbf{I}_{a,b}$和$\mathbf{I}_{c,d}$，当$\max\left(\left|a-b\right|,\left|c-d\right|\right)=1$时在对应两节点之间增加一条边，如图 3.4(b)。在具体实验中，为了让每个节点能够进行自学习，还额外给每个节点添加了一条自连接的边。

(a) (b)

图3.4 边的构建：(a)完全图，(b)格图

### 3.2.3 图表示学习

图像到图的转换使我们能够构造一个包含$n\times m$个节点的图，它可以表示为两个矩阵$A\in\{0,1\}^{nm\times nm}$和$W\in\mathbb{R}^{nm\times F}$。其中，$A$表示邻接矩阵，$W$表示矩阵形式的节点特征，$F$为节点的特征个数。图表示学习的目的是用一个图神经网络为上述图中的每个节点生成一个表示(嵌入向量)，然后通过所有节点表示组成的图表示来判断对应图像是否为隐写图像。本章使用图注意网络(Graph Attention Network, GAT)[104]来进行图表示学习。图表示学习需要一个图神经网络来帮助分析不同图像块的特征之间的异同，而图注意网络的优势在于，它为输入图数据的每条边添加了一个注意力系数，并通过训练学习相连两节点之间的关联程度，所以 GAT 能够很好地拟合任务。而且 GAT 遵循邻域聚合范式，以一个图(包括其拓扑结构和描述特征)作为输入，并为每个图节点生成一种表示，每个节点可以表示为一个向量。

图表示学习共使用了两层图注意力层，每层的输入是一组节点特征 $\boldsymbol{h}=\{\vec{h}_1,\vec{h}_2,\ldots,\vec{h}_N\},\vec{h}_i\in\mathbb{R}^F$，其中 $N$ 为节点个数，$F$ 为每个节点的特征个数。该层产生一组新的节点特征(具有不同的特征数 $F'$)，$\boldsymbol{h}'=\{\vec{h}'_1,\vec{h}'_2,\ldots,\vec{h}'_N\},\vec{h}'_i\in\mathbb{R}^{F'}$。

为了获得足够的表达能力，将输入特征转化为更高层次的特征，至少需要一次可学习的线性变换。为此，作为初始步骤，使用一个权重矩阵 $\mathbf{W}\in\mathbb{R}^{F'\times F}$ 作用于每个节点。然后在每个节点上执行自关注，即使用一个共享注意机制 $\alpha:\mathbb{R}^{F'}\times\mathbb{R}^{F'}\to\mathbb{R}$ 计算注意力系数，该系数表示节点 $j$ 特征向量对节点 $i$ 的重要程度，如式(3.5)：

$$e_{ij}=\alpha\left(\mathbf{W}\vec{h}_i,\mathbf{W}\vec{h}_j\right) \tag{3.5}$$

然后通过计算节点 $i$ 与所有邻居节点之间的注意力系数来执行注意力机制，即计算节点 $j\in\mathcal{N}_i$ 的 $e_{ij}$，其中 $\mathcal{N}_i$ 是图中节点 $i$ 的邻居节点合集，在本章所有的实验中，表示节点 $i$ 的一阶邻居合集(包括节点 $i$)。为了使系数在不同的节点之间容易比较，使用 *softmax* 函数对所有的 $e_{ij}$ 进行归一化，如式(3.6)：

$$\alpha_{ij}=\mathrm{softmax}_j\left(e_{ij}\right)=\frac{\exp\left(e_{ij}\right)}{\sum_{k\in\mathcal{N}_i}\exp\left(e_{ik}\right)} \tag{3.6}$$

在具体实验中，注意力机制 $\alpha$ 是一个单层前馈神经网络，由一个权重向量 $\vec{\mathbf{a}}\in\mathbb{R}^{2F'}$ 进行参数化，并应用 LeakyReLU 非线性激活函数。完全展开后，注意力机制计算出的系数可以表示为式(3.7)：

$$\alpha_{ij}=\frac{\exp\left(\mathrm{LeakyReLU}\left(\vec{\mathbf{a}}^T\left[\mathbf{W}\vec{h}_i\,\|\,\mathbf{W}\vec{h}_j\right]\right)\right)}{\sum_{k\in\mathcal{N}_i}\exp\left(\mathrm{LeakyReLU}\left(\vec{\mathbf{a}}^T\left[\mathbf{W}\vec{h}_i\,\|\,\mathbf{W}\vec{h}_k\right]\right)\right)} \tag{3.7}$$

其中，LeakyReLU 非线性激活函数的表达式如公式(3.8)，负输入斜率 $\alpha=0.2$：

$$\mathrm{LeakyReLU}\left(x\right)=\begin{cases}\max(\alpha x,x), x\le 0\\ x, x>0\end{cases} \tag{3.8}$$

得到注意力系数后，可用归一化的注意力系数来计算与之对应的特征的线性组合，再将经过非线性激活函数 $\sigma$ 后的向量作为每个节点的最终输出特征：

$$\vec{h}'_i=\sigma\left(\sum_{j\in\mathcal{N}_i}\alpha_{ij}\mathbf{W}\vec{h}_j\right) \tag{3.9}$$

经过两层图神经网络的消息传递后，节点特征 $\mathbf{h}''=\left\{\overrightarrow{h''}_1,\overrightarrow{h''}_2,\ldots,\overrightarrow{h''}_N\right\}$ 为图注意力网络最后的输出。

### 3.2.4 二元分类

为了实现图层面的分类，需要使用一个函数来将上一节中图注意力网络最终输出的图数据映射生成整个图的表示(图特征向量)，此类函数通常被称为读出函数[109]。读出函数能够将图神经网络最终输出的矩阵形式的节点表示映射到一个实向量，如式(3.10)：

$$\boldsymbol{h}_G = R\left(\boldsymbol{h}_v^k \boldsymbol{v} \in \boldsymbol{V}\right) \tag{3.10}$$

常用的读出函数有求和读出，平均读出和最大读出，也可以使用简单的神经网络进行读出。其中，求和读出可以学习精确的结构信息，平均读出偏向学习分布信息，最大读出偏向学习具有代表性的元素信息[110]。本章选用了平均读出得到图的分布信息以便于图层面的分类，映射后的特征维度等于图注意力网络最后一层输出节点的特征数。图特征向量被输入含有激活函数 ReLU[111]的 64 维全连接层，然后由一个带有 *softmax* 函数的全连接隐藏层处理以输出预测概率。

## 3.3 实验与分析

在本小节，首先介绍了所使用的数据集和搭建模型的环境与超参数，然后测试了模型设计中各个参数对于所提出方法性能的影响。最后展示了本章所提算法与各基线算法的检测性能对比。

### 3.3.1 实验设置

在实验中检测的隐写算法包括 S-UNIWARD[27]和 HILL[28]。数据集选用了 BOSSBase v1.01 [112]，其中包含 10000 幅大小为 512 × 512 的自然图像，并使用 Matlab 生成对应的隐写图像。对于每组实验，在 10000 对自然/隐写图像中，随机选取 4000 对用于模型训练，1000 对用于模型验证，其余 5000 对用于模型测试，且这三个子集没有交集。

模型搭建使用了 PyTorch 框架，并使用单个 TITAN RTX 24 GB GPU 进行训练。小批量大小为 32，迭代轮次为 300 次，因此训练模型的迭代次数总共 75000 次。学习率设置为 0.001，并使用 Adam 优化器[113]（具有两个超参数 $\beta_1 = 0.5$，$\beta_2 = 0.999$）更新模型参数。在实验中，图注意力层的数量为 2，并使用平均读出函数来池化节点表示。3.2.1 节中所提出的完全图和格图都将用于实验评估，并且令参数 $h_p = w_p = 256$，$n = m = 3$。

### 3.3.2 参数最优化

在图构建的图像分割中使用了参数 $\alpha$ 和 $\beta$ 代表相邻两子图横向和纵向的交叠率，具体见 3.2.2 节。子图交叠率会决定分割后的图像块的个数和其映射的节点特征之间的相关性，因此需要对其具体值进行最优化。节点构建所用的浅层 CNN 选用了 3.3.2 节中的 SCNN-V，边构建方法使用了完全图，并令 $\alpha = \beta$。为了能够将图像恰好分为各个交叠相同的图像块，实验中 $\alpha$ 和 $\beta$ 的值选用了 0、0.5、0.66 和 0.75。

实验对 S-UNIWARD 隐写算法 0.4bpp 和 0.1bpp 两个嵌入率进行检测，表 3.2 中展示了使用不同的 $\alpha$ 和 $\beta$ 对于隐写分析性能的影响。当同时平衡计算成本和高检测性能时，$\alpha = \beta = 0.5$ 是图像交叠率最佳的选择。从其对图像块的个数的影响的角度分析，当交叠率为 0 时子图无重叠，因此图像块个数为 $2\times2$，当值为 0.5 时图像将会分为 $3\times3$ 个图像块，为 0.66 和 0.75 时图像块个数分别为 16 个和 25 个。值为 0 时准确率最低可能是因为图像块数过少即图节点过少从而导致图神经网络学习的邻域特征太少进而检测性能较低。而 0.75 时准确率相对较低可能是由于图像块重叠过多使节点相关性太强，从而导致图表示学习对节点之间差异化敏感度降低从而错误分类。参数 $\alpha$ 和 $\beta$ 为 0.5 和 0.66 下的准确率较为接近，可能是同时平衡了节点个数和节点差异化从而能够较好的更新节点特征。

表 3.2 不同参数 $\alpha$ 和 $\beta$ 下模型检测 S-UNIWARD0.4bpp 和 0.1bpp 的准确率

| | $\alpha=\beta=0$ | $\alpha=\beta=0.5$ | $\alpha=\beta=0.66$ | $\alpha=\beta=0.75$ |
|---|---|---|---|---|
| 0.4bpp | 0.7145 | 0.7862 | 0.7836 | 0.7386 |
| 0.1bpp | 0.5212 | 0.5633 | 0.5629 | 0.5378 |

### 3.3.3 图构建对检测性能的影响

在图构建过程中，边的构建方式对于消息的传递、收集和更新等各个环节都会产生不同的影响，因此需要找到最优的构建边的方式来确保图表示学习的有效性和准确性。在 3.2.2 节中提出了两种边的构建方式，分别为完全图和格图。本节选用了具有代表性的嵌入率 0.4bpp 和 0.1bpp 的隐写数据集，以观察完全图和格图两种构建方式对于隐写分析检测模型性能的影响。通过图 3.5 展示的实验结果，可以看到在使用不同模型检测 S-UNIWARD 和 HILL 的 0.4bpp 和 0.1bpp 嵌入率时，完全图和格图两种构建方式所得到的结果也存在明显的区别，因此需要根据具体情况来选择最适合的边的构建方式。

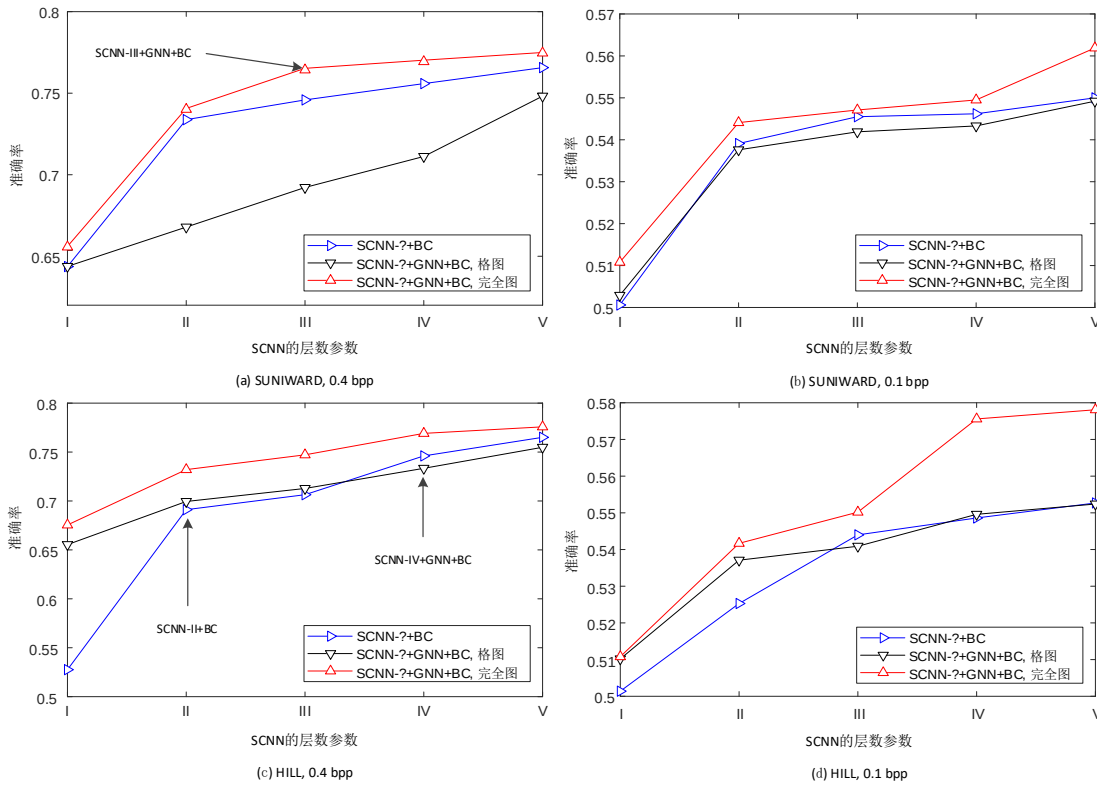



图 3.5 S-UNIWAED 和 HILL 嵌入率为 0.4 bpp 和 0.1bpp 的检测准确率

从图 3.5 中，我们可以得出结论：首先，完全图的性能优于格图，这可能是由于完全图能够使任意两个节点相互连接，以便有效地聚合全局特征并转换为判别特征进行隐写分析。它还验证了不同方式构建的图确实将导致模型性能不同。其次，在完全图的情况下，所提方法在检测精度方面明显优于基准 CNN 模型，显示了图学习对隐写分析的优越性。

### 3.3.4 检测效果评估

为评估所提出模型的检测效果，本节使用完全图模型和基线 CNN 对两种空域隐写算法的 0.1 bpp、0.2 bpp、0.3 bpp、0.4 bpp 和 0.5 bpp 五个嵌入率进行了检测，检测准确率如表 3.3 所示。

表 3.3 所提出模型和基线 CNN 对 S-UNIWARD 和 HILL 的五个嵌入率的检测准确率

| | | 嵌入率(bpp) | | | | |
|---|---|---|---|---|---|---|
| 隐写算法 | 隐写分析模型 | 0.5 | 0.4 | 0.3 | 0.2 | 0.1 |
| S-UNIWARD | SCNN-I-BC | 0.6632 | 0.6498 | 0.5945 | 0.5584 | 0.5001 |
| | SCNN-I-GNN-BC | **0.6719** | **0.6602** | **0.6349** | **0.6008** | **0.5112** |
| | SCNN-II-BC | 0.6703 | 0.6698 | 0.6111 | 0.5776 | 0.5367 |
| | SCNN-II-GNN-BC | **0.7599** | **0.7456** | **0.6944** | **0.6348** | **0.5428** |
| | SCNN-III-BC | 0.7011 | 0.6785 | 0.6338 | 0.5979 | 0.5413 |
| | SCNN-III-GNN-BC | **0.7813** | **0.7662** | **0.7046** | **0.6643** | **0.5452** |
| | SCNN-IV-BC | 0.7229 | 0.6937 | 0.6559 | 0.6112 | 0.5397 |
| | SCNN-IV-GNN-BC | **0.7918** | **0.7847** | **0.7194** | **0.6634** | **0.5478** |
| | SCNN-V-BC | 0.7756 | 0.7321 | 0.6972 | 0.6540 | 0.5452 |
| | SCNN-V-GNN-BC | **0.8149** | **0.7856** | **0.7334** | **0.6697** | **0.5634** |
| HILL | SCNN-I-BC | 0.6704 | 0.6549 | 0.6113 | 0.5578 | 0.5014 |
| | SCNN-I-GNN-BC | **0.6903** | **0.6755** | **0.6532** | **0.5972** | **0.5106** |
| | SCNN-II-BC | 0.6865 | 0.6728 | 0.6364 | 0.5832 | 0.5235 |
| | SCNN-II-GNN-BC | **0.7445** | **0.7321** | **0.6843** | **0.6322** | **0.5437** |
| | SCNN-III-BC | 0.7023 | 0.6813 | 0.6445 | 0.6084 | 0.5376 |
| | SCNN-III-GNN-BC | **0.7662** | **0.7498** | **0.6558** | **0.6127** | **0.5479** |
| | SCNN-IV-BC | 0.7458 | 0.7216 | 0.6994 | 0.6143 | 0.5434 |
| | SCNN-IV-GNN-BC | **0.7902** | **0.7525** | **0.7072** | **0.6328** | **0.5769** |
| | SCNN-V-BC | 0.7539 | 0.7399 | 0.7111 | 0.6430 | 0.5464 |
| | SCNN-V-GNN-BC | **0.8244** | **0.7885** | **0.7492** | **0.6843** | **0.5793** |

从表中可以看出所提出模型在检测精度方面明显优于基线 CNN 模型，且经过图表示学习模型的性能下降率远低于基线 CNN 模型。换句话说，当减少 CPL 的数量时，所提出的方法仍然可以达到相对更高的准确率，例如，嵌入率为 0.4 bpp 的 HILL 算法隐写检测，当 CPL 数量为 2(对应 SCNN-II+GNN+BC)时，检测准确率为 0.7321，当 CPL 数量仅为 1(对应 SCNN-I+GNN+BC)时，检测准确

率为 0.6755，明显高于基准 CNN 模型。这意味着图表示学习有能力更好地利用统计特征和全局信息进行隐写分析。

## 3.4 本章小结

本章提出了一种用于空域图像隐写分析的图表示学习网络，充分利用了卷积神经网络和图神经网络各自的优势。在详细的结构中，首先将每个图像转换为一个图，其中节点表示图像块的隐写特征，节点之间的边表示图像块之间的局部关系。每个节点都与浅层卷积神经网络从相应图像块确定的特征向量相对应。通过图注意力网络，可以学习判别特征以进行有效的隐写分析。实验表明，与基线 CNN 模型相比，所提出的方法实现了更加优异的性能，验证了图表示学习的优越性，也表明了图表示学习在图像隐写分析中的潜能。

# 第四章 基于图神经网络的 JPEG 域图像隐写分析

## 4.1 引言

数字图像有各种格式，从应用的角度来看，JPEG 是社交网络上最流行的图像格式，因为它提供了高视觉质量，同时保持了图像的低存储空间。当应用 JPEG 图像进行隐写时，隐写通信的存在很容易被大量的日常网络社交行为所掩盖，这使得 JPEG 隐写成为一个研究热点。JPEG 隐写的快速发展促使 JPEG 隐写分析成为迫切需要解决的问题。早期的 JPEG 隐写分析方法遵循传统的机器学习框架[53]，需要人工制作统计特征并应用集成分类器进行分类。随着深度卷积神经网络的普及，近年来越来越多的工作[47,58,89,114,115]已经将 CNN 在计算机视觉领域中的成功迁移到了图像隐写分析中。

在探索深度学习对隐写分析的适用性时，有必要考虑到隐写原理。空域图像隐写在数据嵌入过程中，通常直接改变了原始空域中的图像像素。卷积神经网络有可能记住特定的嵌入模式，这将导致训练过的模型的泛化能力较差。与之不同的是，JPEG 隐写通常倾向于通过修改量化的 DCT 系数以在 DCT 域中嵌入秘密信息。当图像转换回空域时，DCT 系数的修改延伸到相应的 $8\times8$ 块的所有像素空间中去。如果神经网络能够学习空域中块之间的统计不一致，JPEG 域图像隐写将比空域图像隐写更容易被检测到。

受图神经网络(GNNs)具有强大的能力来模拟对象之间的统计依赖性这一事实的启发[109]，本章进一步探索图表示学习对于 JPEG 图像隐写分析的适用性。最近 GNN 因为它在处理非欧几里得数据方面的优势而蓬勃发展，越来越多的研究将它们与视觉任务结合起来[116]。由于其能够将图像建模为一个完整的图进行全局分析，因此图表示学习能够更好地把握整体的统计特征和相关性。上一章的工作已经证明了图表示学习可以应用于空域图像隐写分析，所以本章将进一步探索其对 JPEG 域图像隐写分析的适用性。

当前基于 CNN 的工作对适用于 JPEG 隐写分析的网络结构进行了一系列的探索。虽然这些研究结合了现有的深度学习的设计概念，但它们缺乏对图像隐

写分析中涉及的关键因素的关注，其网络深化过程中的特征消失问题被忽视了。在图像隐写分析中，需要识别的隐写特征是非常微弱的，而卷积和池化也会抑制弱信号。微弱的隐写特征在网络中被连续的卷积和池化逐层削弱，有些甚至可能会消失。从另一个角度来看，CNN 通过其卷积层的深度来扩大其局部感知域。在这个过程中，一些全局信息不可避免地会丢失，这也可以理解为会丢失整个图像所有信息之间的相互关系。总而言之，上述两个因素可能是现有隐写分析模型性能有限的主要原因。

基于上述分析，为了设计能够更好地利用隐写特征的网络结构，需要考虑两个关键点。第一是如何防止本就微弱隐写信号被网络本身削弱，第二是如何利用隐写信号的全局特征。为此，本章为 JPEG 域图像隐写分析设计了一个新的深度学习框架，它结合了 CNN 和 GNN 的优点。对于第一点，设计了一个特征增强块作为基本单元，它去除了池化层，并将卷积的步长降低到 1，以更好地增强弱隐写信号。为了利用全局特征，将隐写特征输入一个图注意网络，学习特征之间的全局关系。这样的设计可以进一步从全局角度有效提取残差特征向量，并从全局角度有效地学习隐写特征。此外，预训练作为一种使用大规模数据集初始化网络权重的方式，可以加速模型收敛，提高网络提取鉴别性特征的能力。本章使用 ALASKA#2 对用 BOSSBase v1.01 和 BOWS2 训练的模型进行预训练。实验表明，所提出的方法在不同有效载荷下的检测准确率优于以前的工作。

## 4.2 基于图神经网络的 JPEG 图像隐写分析

### 4.2.1 总体框架

因为所提出的网络模型利用图表示学习进行 JPEG 图像隐写分析，所以将它命名为 JPEG-GraphNet。如图 4.1 给出了 JPEG-GraphNet 的总体框架，从中可以看到整个隐写分析的过程由三个阶段组成，分别是特征提取(左)、特征学习(右下)和特征分类(右上)。特征提取的目的是提取原始隐写特征并进一步增强这些特征，以便在特征学习阶段简化深度表示学习。特征学习的目的是提取判别特征，以供特征分类阶段判断输入是否为隐写。具体来说，特征提取由三个模块

组成，即预处理、隐写特征增强(Steganographic Feature Enhancement, SFE)和图注意力学习(Graph Attention Learning, GAL)。预处理部分对输入 JPEG 图像进行解压缩后提取隐写残差。然后将残差图输入 SFE 模块进行局部特征增强，同时输入 GAL 模块中进行全局特征增强，这两个模块的输出将被合并输入到后续的特征学习模块中。特征学习模块用于将接收到的隐写特征表示为高维特征，该模块由四个具有相似结构的 CNN 块组成，括号中的数字代表卷积核的数量。特征分类阶段将高维特征输入配备 *softmax* 函数的全连接层，最终输出表示图像分类结果的独热(one-hot)向量。下面的小节将详细介绍所设计的 JPEG-GraphNet。

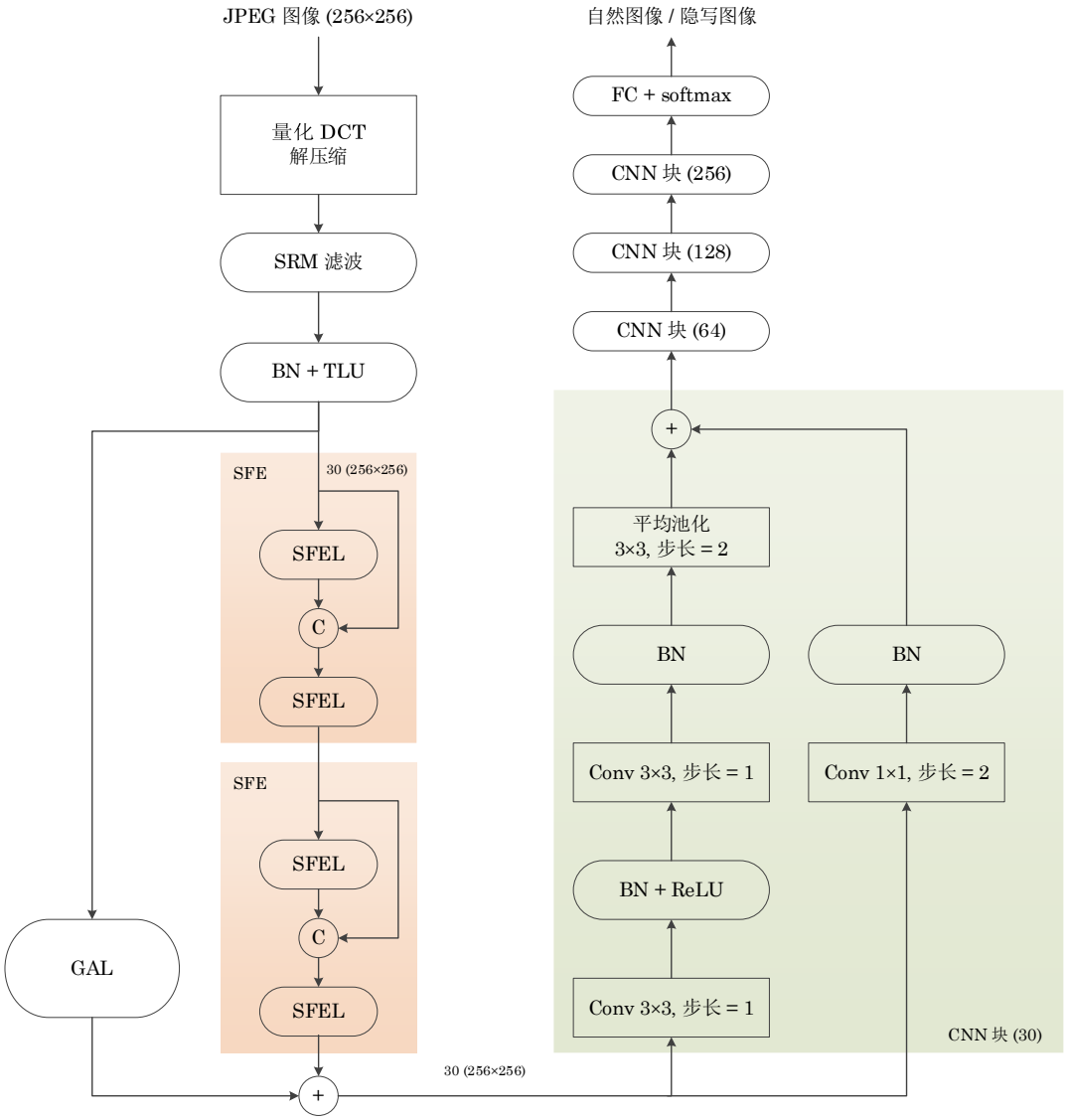



图 4.1 JPEG-GraphNet 的框架示意图

### 4.2.2 预处理

JPEG-GraphNet 的预处理部分涉及量化的 DCT 解压缩、SRM 高通滤波和专用的激活函数。对于一个给定大小为 $h\times w$ 的 JPEG 图像，其中 $h$ 和 $w$ 都是 8 的倍数，在本章的实验中它们都是 256。设 $\boldsymbol{C}=\{c_{i,j}^{n,m}\}$ 是量化 DCT 系数的矩阵，其中 $c_{i,j}^{n,m}$ 是第 $(n,m)$ 个 $8\times8$ 块的第 $(i,j)$ 个元素，其中 $0\le i,j<8$，$0\le n<h/8$，且 $0\le m<w/8$。设 $\boldsymbol{Q}=\{q_{i,j}^{n,m}\}$ 是相应的量化矩阵。量化前的 DCT 系数矩阵 $\boldsymbol{D}$ 可以近似为 $\boldsymbol{C}$ 和 $\boldsymbol{Q}$ 的元素乘积，即 $\boldsymbol{D}=\{d_{i,j}^{n,m}\}$，其中 $d_{i,j}^{n,m}=c_{i,j}^{n,m}q_{i,j}^{n,m}$。需要注意的是，通常情况下，对于任意 $n\neq n'$ 或 $m\neq m'$，都有 $q_{i,j}^{n,m}=q_{i,j}^{n',m'}$。为了确定空间像素(即解压缩的亮度通道)，执行逆 DCT 变换，即式(4.1)：

$$f_{x,y}^{n,m}=\frac{1}{4}\sum_{u=0}^{7}\sum_{v=0}^{7}g_u g_v d_{u,v}^{n,m}\cos\frac{(2x+1)u\pi}{16}\cos\frac{(2y+1)v\pi}{16} \tag{4.1}$$

其中，当 $z=0$ 时，$g_z=1/\sqrt{2}$，否则 $g_z=1$。换言之，$\boldsymbol{C}$ 的空间亮度矩阵可以用 $\boldsymbol{F}=f_{x,y}^{n,m}$ 表示。量化 DCT 解压缩的目标是确定 $\boldsymbol{F}$。要从 $\boldsymbol{F}$ 中确定 $\boldsymbol{D}$，可以应用 DCT 变换，即式(4.2)：

$$d_{u,v}^{n,m}=\frac{1}{4}g_u g_v\sum_{u=0}^{7}\sum_{v=0}^{7}f_{x,y}^{n,m}\cos\frac{(2x+1)u\pi}{16}\cos\frac{(2y+1)v\pi}{16} \tag{4.2}$$

因此，$\boldsymbol{C}$ 可以通过应用 $c_{i,j}^{n,m}=d_{i,j}^{n,m}/q_{i,j}^{n,m}$ 来确定。大量方法已经证明，高通滤波在图像隐写分析中至关重要，它能够帮助模型快速捕获隐写伪影。作为残差提取器，高通滤波能够抑制图像内容并扩大隐写信号的噪声信号比。而高通滤波可以使用一组可训练的卷积核来实现。尽管卷积神经网络的可训练卷积核通常是随机初始化的，但如果随机初始化复杂度较低的 CNN 在相对较小的数据集上训练时往往无法学习出良好的残差特征提取器。这是因为相对较小的数据集可能不足以充分涵盖网络需要学习的复杂模式和特征，从而导致 CNN 模型可能出现欠拟合问题。因此，本章通过手工制作的方式初始化高通滤波卷积核。使用的是在 SRM[43]中描述的共 30 个大小为 $5\times5$ 的基本线性滤波器[43](即“垃圾邮件”滤波器及其对称版本)，滤波器层以 $\boldsymbol{F}$ 作为输入并输出共计 30 个残差特征图, 每个特征图的大小为 $h\times w(\times1)$。

神经网络中的激活函数可以为网络提供非线性建模能力，它可以为神经网络中的每个神经元引入非线性映射以增加特征表示的能力。激活函数有多种选择，如双曲切线、sigmoid 和整流线性单元(ReLU)。在传统的图像隐写分析中，手工制作的特征提取器通常具有某种对称性，例如，残差直方图处于关于零点对称的区间。尽管这些特征将被进一步合并以实现高性能的隐写分析，但它启发将特征对称性引入到初始特征提取中，以便更好地学习特征。基于这种考虑，预处理在早期阶段使用的激活函数应该能够将输入映射到一个对称的区间。此外，为了促进特征学习，将激活函数设计成硬处理函数而不是软处理函数可能会更好。因此，在 JPEG-GraphNet 预处理阶段的 SRM 滤波器滤波后使用截断线性单元(Truncated Linear Unit, TLU)[61]激活函数。TLU 可以表示为式(4.3)：

$$\mathrm{TLU}(x)=\begin{cases}-T, x\in(-\infty,-T)\\ x, x\in[-T,T]\\ T, x\in(T,+\infty)\end{cases} \tag{4.3}$$

其中， T 是预先确定的约束输出的阈值。此外，为了加速模型收敛，默认使用批量归一化(Batch Normalization，BN)[117]。

### 4.2.3 隐写特征增强

为了进一步增强预处理步骤生成的残差图中的残差信息，提出了一个隐写特征增强(SFE)模块，如图 4.1 所示，它被使用了两次。SFE 模块的详细结构如图 4.2 所示，共引入了两个 SFE 层(SFE Layer，SFEL)。其中，“C”表示“拼接”，即将浅层特征与深层特征在深度上相结合，以此来保留更多的隐写信息。SFEL 的设计是受基于 JPEG 隐写分析改进的 EfficientNet[118]的启发，其中提出使用步长为 2 的卷积层和池化将导致浅层 CNN 中特征图的分辨率降低，这会对隐写分析器的检测精度产生负面影响，因为这种结构会增强图像内容同时抑制类噪声的隐写信号。因此，SFEL 的设计遵循两个关键点：一方面，去除池化层以减少特征的损失；另一方面，将所有卷积层的步长设置为 1，并采用配备1×1卷积核的捷径来处理特征，从而尽可能地保留浅层的隐写特征信息。通过这种方式，SFEL 模块可以更加有效地增强隐写残差，提高隐写分析的准确性。

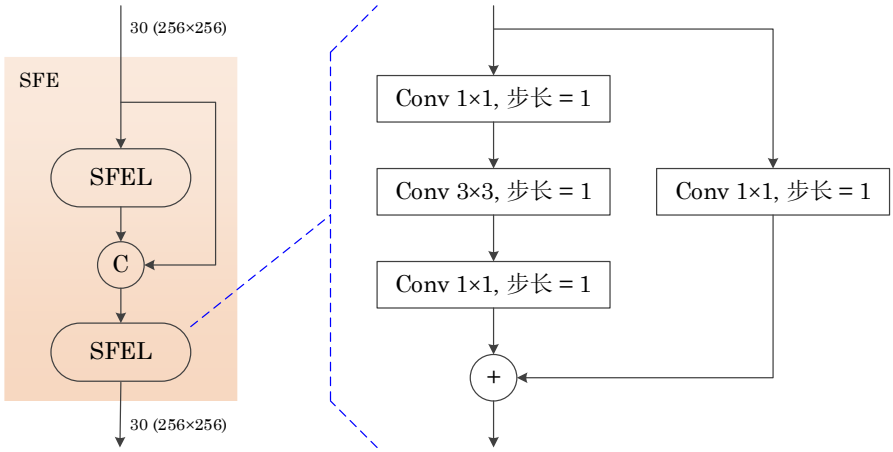



图 4.2　隐写特征增强模块(SFE)和隐写特征增强层(SFEL)结构

当深度学习网络进行深层特征学习时，常常会遇到浅层特征逐渐消失的问题。为了解决这个问题，还在 SFE 中增加了的快捷连接。通过引入这些快捷连接，使得浅层特征能够直接跳过中间层并传递到深层。这样，即使在深层特征学习的过程中，浅层特征仍然能够为网络提供重要的信息，避免其逐渐丢失。且 SFE 中的快捷连接使用了“拼接”，从而更大程度的保留了浅层信息。

为了挖掘全局隐写信息以进行有效的隐写分析，还引入了一个图注意力学习(GAL)模块，其细节在图 4.3 中展示。详细来说，对于给定的 JPEG 图像，首先对其进行预处理，生成一组残差图。然后，每个残差图被分为不相交的 8×8 块，并叠加成一个维度为 $8\times8\times(hw/64)$ 的特征图，例如，如果残差图的宽度和高度都是 256 ，那么特征图的大小就是 $8\times8\times1024$，如图 4.3 所示，就可以构建一个由 64 个节点组成的完全图，每个节点对应一个空间点，完全图中的每个节点又与一个长度为 $(hw/64)$ 的特征向量相对应。换句话说，上述所确定的特征图将被分解成 $(hw/64)$ 维的特征向量，每个特征向量被分配到与该特征向量相同的空间点所对应的节点上。此时可以建立一个与节点特征相对应的完全图，其中节点数为 64 个，边数为 2016 条，每个特征向量的特征个数为 $(hw/64)$ 。

随后将每个构建好的图输入到图注意网络[104]进一步学习全局特征，并输出一组新的节点特征作为结果。图注意学习的层数被设置为2，每个新特征向量的维度与输入特征向量的维度一致。最后，每组新节点特征向量经过还原，重新构成一个大小为 $256\times256$ 的新特征图。通过堆叠所有的新特征图，输出通道数将扩大到 30，并与 SFE 生成的特征图相叠加。

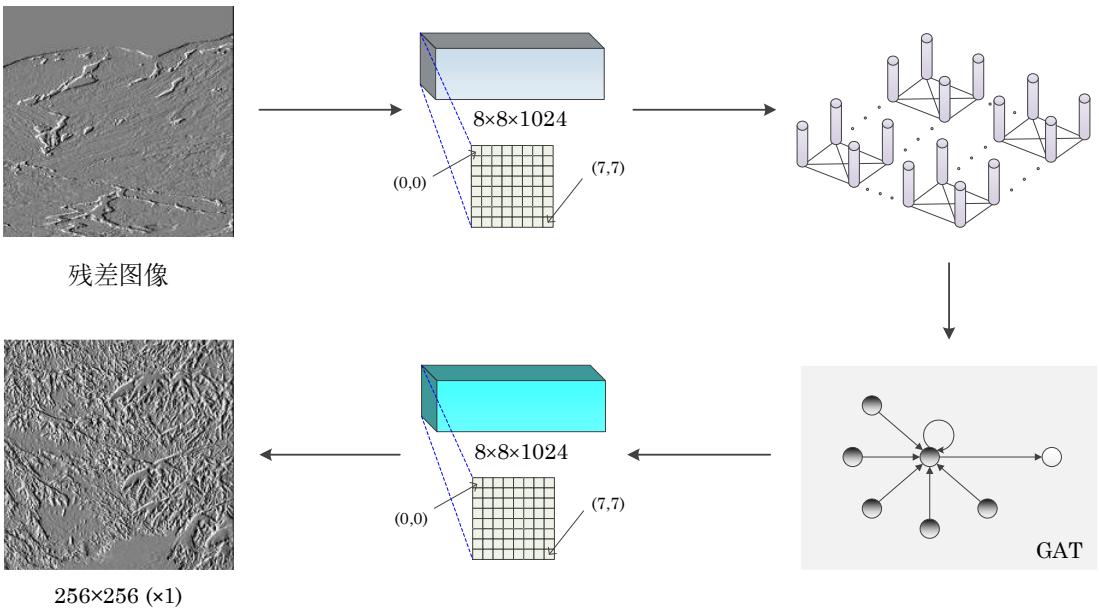



图 4.3　图注意力学习(GAL)模块结构

### 4.2.4　特征学习与分类

经过隐写特征增强处理之后，输出的特征图要经过特征学习模块以提取用于特征分类的判别特征。为了降低特征图的维度，隐写特征学习层采用了步长为 2 的池化和残差捷径。如图 4.1 右侧的 CNN 块所示，在两个步长为 1、卷积核大小为3×3的卷积层之后，应用了步长为 2 的3×3平均池化层，同时，使用步长为 2、卷积核大小为 1 的卷积层作为残差捷径，以将每个 CNN 块输出特征图的宽度和高度同时减少到输入的一半。对于本章所用的 CNN 块的设计，参考了 SRNet 的特征学习部分[62]，其中设计了两种类型的模块用于降维，即 2.2.2 节中介绍的“组 3”和“组 4”。在本章中借鉴了组 3 来设计所用的 CNN 块。

在JPEG-GraphNet中，本章采用四个CNN块用于逐步降维。如图4.1所示，前三个 CNN 块的结构完全相同，每个 CNN 块都包含两个卷积层以及批量标准化和 ReLU 激活函数。其中每个卷积层的卷积核数量在图中用括号内的数字进行表示。最后一个 CNN 块的结构稍有不同，它去掉了捷径并将最后的平均池化改为全局池化，将特征降维到一维。接下来，将该一维特征输入到一个标准的256维全连接层，并通过*softmax*线性分类器输出最终的预测概率。因此，JPEG-GraphNet 可以高效地提取出判别特征，并且快速而准确地对其进行分类。

## 4.3 实验与分析

### 4.3.1 实验设置

实验中使用的数据集来自两个图像源，即 BOSSBase v1.01[112]和 BOWS-2[119]，两个数据集都包含了 10,000 幅灰度图像，分辨率为512×512。由于计算资源有限，所有涉及的图像都被中心裁剪成256×256的分辨率，并用 Matlab 生成对应的隐写图像。训练集包含 14,000 对自然/隐写图像(其中的自然图像 4,000 幅从 BOSSBase v1.01 中随机选择，其余 10,000 幅全部来自 BOWS-2)。包含 1,000 对 BOSSBase v1.01 自然/隐写图像的验证集被用来选择超参数，最终的测试结果是对测试集中 5,000 对 BOSSBase v1.01 自然/隐写图像进行分类后得出的。实验检测的隐写算法包括 J-UNIWARD[27]、UED-JC[31] 和 J-MiPOD[32]。在质量因子(Quality Factor，QF)分别为 75 和 95、有效载荷范围为 0.1 到 0.5 bpnzac 的情况下，评估了所提出的隐写分析模型的性能。

模型使用 PyTorch 框架进行搭建，并用 TITAN RTX 24 GB GPU 进行训练。批量尺寸设置为16对自然/隐写图像，随机梯度下降优化器选用 Adamax[113]。批量归一化参数是通过衰减率为0.9的指数移动平均法学习的。在预训练中，使用 He 初始化[120]和$2\times10^{-4}$的 $L2$ 正则化来初始化卷积网络权重。滤波器的偏置被设为0.2，且没有使用正则化。全连接分类层使用标准偏差为0.01的零均值高斯分布且不添加偏置来初始化权重。训练的网络参数初始化使用预训练的网络权重。

### 4.3.2 预训练策略

预训练可以看作是一种初始化网络权重的方法。参数的正确初始化可以在训练中发挥重要的作用，特别是在复杂的任务中，例如较小有效载荷的隐写分析任务。在深度学习网络处理自然图像的分类问题上，对模型进行预训练可以使模型有能力提取构成自然图像基本元素的过滤器(如边缘、纹理、周期性图案等)。因此，在规模更大的隐写分析数据集上进行预训练可以提高模型提取隐写特征的能力。而且，有效的预训练还能够节约算力，同时提高训练过程中网络的收敛速度。基于上述分析，本节提出了如图 4.4 所示的预训练策略。

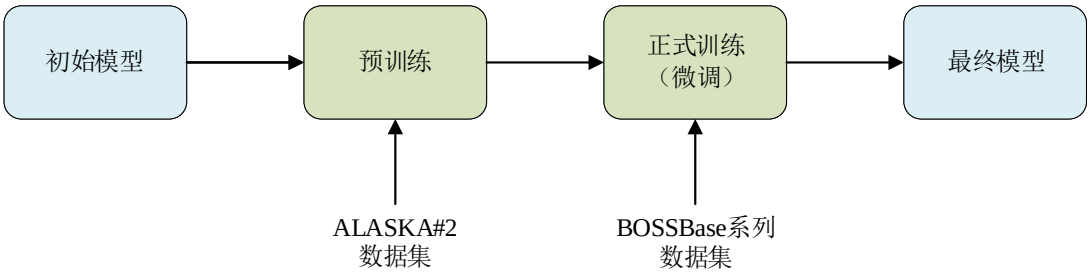



图4.4 预训练框架

预训练数据集选用了最近 Kaggle 比赛中使用的 ALASKA#2 数据集[70]。这个数据集包括用 75、90 和 95 三种质量因子(QF)压缩的彩色图像。对于每种 QF，有 25,000 幅分辨率为 256×256 的图像。 实验使用不同隐写算法生成了有效载荷为 0.1 到 0.5 bpnzac、质量因子分别为 75 和 95 的隐写图像。每种隐写算法下，完整数据集中 25,000 对图像都用于模型预训练。

具体实验步骤如下：首先在预训练数据集上对不同的隐写算法进行了 100 轮次的训练。然后，以 $r_1 = 0.001$ 的初始学习率执行 350,000 次迭代(400 轮次)的训练，之后在基本训练数据集上再进行 87,500 次迭代(100 轮次)，学习率降至 $r_2 = 0.0001$。在最后 100 000 次迭代中在验证集上达到最佳检测准确率的模型参数被作为训练的结果。

为了验证如何预训练对提高隐写分析检测器的检测精度最为有效，设置了 J-UNIWARD 隐写算法的 0.2、0.4 和 0.5 bpnzac 三种不同嵌入率的 ALASKA#2 预训练数据集，图像质量均选用了 QF75。将这三个数据集的预训练中获得的模型参数用于后续的训练，正式训练中模型配置相同，只是数据集替换成图像质量为 QF75 的 BOSSBase v1.01+BOWS-2(后文统称其为 BOSSBase 系列)。模型的检测性能使用测试集上的最小平均错误率 $P_E = \min_{P_{FA}} 1/2(P_{FA} + P_{MD})$ 来衡量，也可称为检测误差。其中，$P_{FA}$和$P_{MD}$分别表示误报率和漏检率。

实验结果如表 4.1 所示。可以看出，在训练集的嵌入率相同的情况下，预训练集的嵌入率越高，最终的模型效果就越好。这可能是由于预训练集的高嵌入率会“教给”模型更多的先验知识，以帮助正式训练中对全新图像的识别。因此，本章在随后的所有实验中均使用 0.5 bpnzac 的 ALASKA#2 作为预训练集，而且检测同一种隐写算法的 JPEG-GraphNet 只需要预训练一次。

表 4.1 JPEG-GraphNet 在 ALASKA#2 预训练并在 BOSSBase 系列训练的检测误差$P_E$

| | ALASKA#2 0.5bpp | ALASKA#2 0.4bpp | ALASKA#2 0.2bpp | 无预训练 |
|---|---|---|---|---|
| BOSSBase 系列 0.5bpp | **.0492** | .0512 | .0564 | .0672 |
| BOSSBase 系列 0.4bpp | **.0931** | .1230 | .1298 | .1332 |
| BOSSBase 系列 0.2bpp | **.1532** | .1679 | .1823 | .1911 |

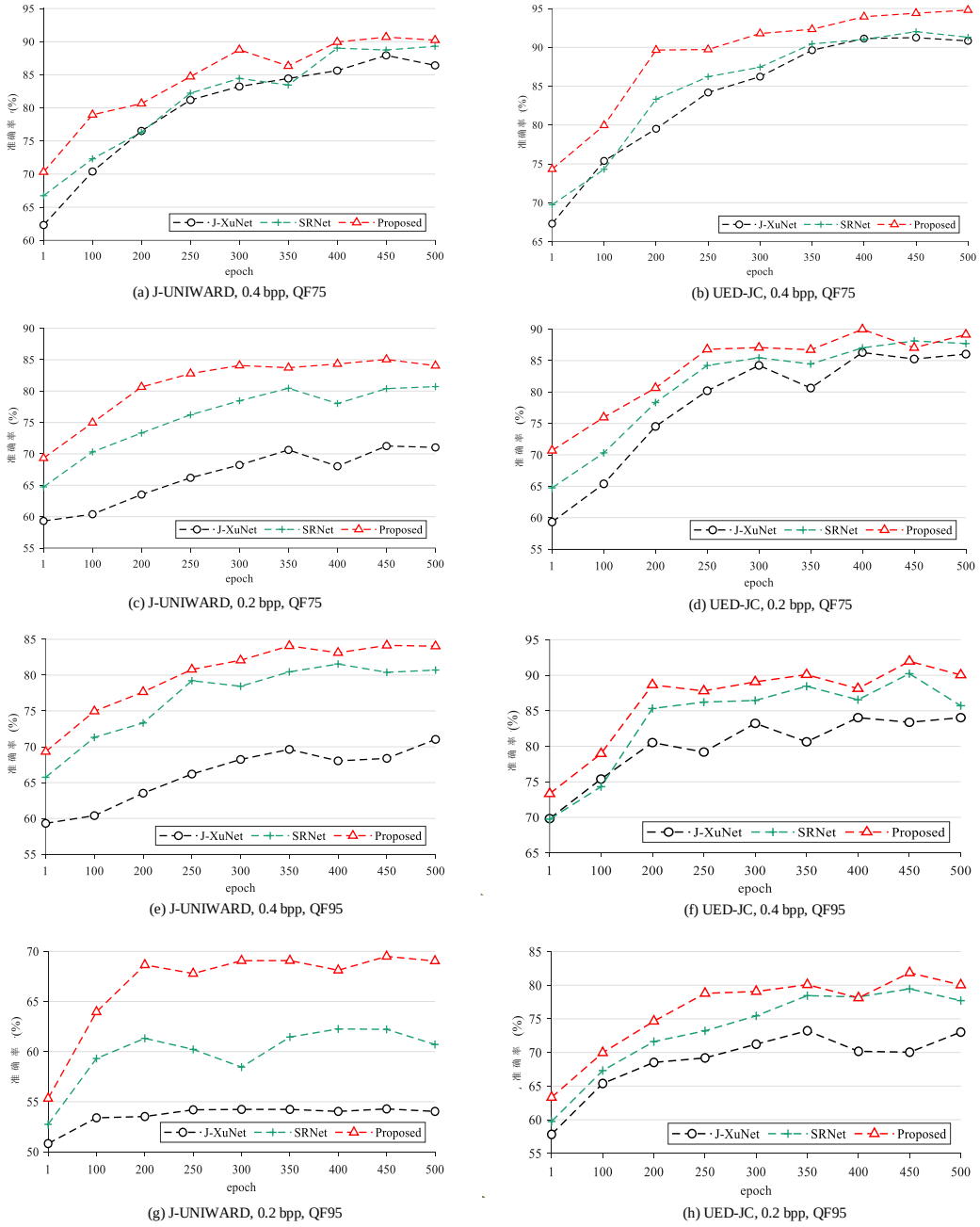
### 4.3.3 检测效果评估

为了评估在 BOSSBase 系列上所提出的 JPEG-GraphNet 的检测效果，将提出的方法与 SRNet[62]和 Xu 提出的 J-XuNet 网络[96]进行了比较。这里比较的所有隐写分析器都是按照相应论文中的描述进行配置的，并在与本文完全相同的数据集上进行训练。网络使用 PyTorch 框架搭建，并在单个 GPU 上运行。实验结果如表 4.2 所示，可以看出在多种嵌入率和嵌入算法下，所提出的方法均具有最低的检测误差，说明同相关算法相对比，JPEG-GraphNet 具有最优的检测性能。

表 4.2 不同隐写算法质量因子为 75 和 95 的五个有效载荷下的检测误差$P_E$

| | | | 嵌入率(bpp) | | | | |
|---|---|---|---|---|---|---|---|
| 隐写算法 | 隐写分析器 | QF | 0.5 | 0.4 | 0.3 | 0.2 | 0.1 |
| J-UNIWARD | J-XuNet | 75 | .0776 | .1437 | .1895 | .2892 | .4310 |
| | | 95 | .2543 | .3162 | .4246 | .4512 | .4812 |
| | SRNet | 75 | .0683 | .1358 | .1543 | .2047 | .3909 |
| | | 95 | .1639 | .2074 | .3367 | .3521 | .4540 |
| | JPEG-GraphNet | 75 | **.0492** | **.0931** | **.1129** | **.1532** | **.3214** |
| | | 95 | **.1112** | **.1639** | **.2407** | **.3110** | **.4173** |
| UED-JC | J-XuNet | 75 | .0473 | .0887 | .1108 | .1372 | .2944 |
| | | 95 | .0983 | .1592 | .2091 | .2983 | .3958 |
| | SRNet | 75 | .0113 | .0789 | .0985 | .1188 | .1923 |
| | | 95 | .0700 | .0978 | .1362 | .2054 | .3369 |
| | JPEG-GraphNet | 75 | **.0105** | **.0532** | **.0798** | **.1003** | **.1799** |
| | | 95 | **.0693** | **.0824** | **.1134** | **.1835** | **.2990** |
| J-MiPOD | J-XuNet | 75 | .0851 | .1367 | .2175 | .3031 | .4509 |
| | | 95 | .2834 | .3576 | .4672 | .4824 | .4861 |
| | SRNet | 75 | .0737 | .1044 | .1750 | .1935 | .4009 |
| | | 95 | .1440 | .2547 | .3352 | .3618 | .4624 |
| | JPEG-GraphNet | 75 | **.0597** | **.1014** | **.1424** | **.1635** | **.3406** |
| | | 95 | **.1235** | **.1754** | **.2632** | **.3214** | **.4219** |

图 4.5 展示了模型训练过程中的准确率曲线，可以看出，在同一训练轮次(epoch)内，所提出的网络具有更高的准确率和更快的收敛速度。而且在训练初期 JPEG-GraphNet 就比其他模型具有更高的准确率，这说明预训练确实带给了模型先验知识，帮助模型更快地收敛。而且在后续的训练中，JPEG-GraphNet 的准确率始终高于其他模型，这也验证了本章提出的 SFE 和 GAL 模块能够使网络能够更好地传递关键的弱隐写信号，从而使提出的模型保持较高的准确率。




(a) J-UNIWARD, 0.4 bpp, QF75

(b) UED-JC, 0.4 bpp, QF75

(c) J-UNIWARD, 0.2 bpp, QF75

(d) UED-JC, 0.2 bpp, QF75

(e) J-UNIWARD, 0.4 bpp, QF95

(f) UED-JC, 0.4 bpp, QF95

(g) J-UNIWARD, 0.2 bpp, QF95

(h) UED-JC, 0.2 bpp, QF95

图4.5 模型训练过程中的准确率曲线

### 4.3.4 消融实验

为了验证所提出的 SFE 模块和 GAL 模块的有效性，需要进一步进行消融实验。本节在比较有代表性的嵌入率 0.4bpnzac 上进行了消融实验，表 4.3 展示了与标准模型相比，无 SFE 模块、无 GAL 模块和无 SFE 和 GAL 的模型的检测误差。从表中的数据可以看出，去除 SFE 模块或去除 GAL 模块都会导致模型的检测误差增加，但无 SFE 的模型的检测误差比无 GAL 的模型更大，可以看出 SFE 在隐写信息增强中具有更强的重要性。GAL 可能通过其提取全局信息信号的能力从另一个维度帮助提高检测性能。同时去除 SFE 和 GAL 后，模型的检测误差增加得更多，说明它们在隐写信息提取部分都非常重要，这也验证了所提出的两个模块的有效性。

表 4.3 与标准模型相比，无 SFE 模块、无 GAL 模块和无 SFE 和 GAL 的检测误差$P_E$

| | J-UNIWARD QF75 | J-UNIWARD QF95 | UED-JC QF75 | UED-JC QF9 |
|---|---|---|---|---|
| JPEG-GraphNet | **.0901** | **.1595** | **.0521** | **.0802** |
| 无SFE | .0232 | .3076 | .0203 | .2897 |
| 无GAL | .0133 | .1830 | .0109 | .1433 |
| 无SFE和GAL | .3987 | .4483 | .3024 | .4075 |

## 4.4 本章小结

在本章中，所提出的JPEG-GraphNet结合了卷积神经网络的特点和图注意神经网络全局化的优势，设计了一种更适于图像隐写信号的特征传输和增强的网络模型。总的来说，本章所提出的方法能够较好地解决由卷积层的堆叠而导致的隐写特征弱化的问题。此外，它还能够更好地利用隐写信号的全局特征。与现有的先进方法相比，经过本章所提出的预训练策略训练出的模型也能够达到更高的检测准确率，证明了图神经网络对于处理JPEG隐写分析任务的可行性。

# 第五章 结论与展望

## 5.1 结论

本篇论文主要聚焦于图像隐写分析领域，讨论了深度学习技术在该领域中的应用和发展，并着重研究了图神经网络在图像隐写分析中的应用。目前，在图像隐写分析中，大多数深度学习模型都是基于卷积神经网络设计的，大量工作也证明了其适用性。针对卷积神经网络在图像隐写分析中存在的一些局限性，本文提出了将卷积神经网络与图神经网络相结合的新方法，并在空域和 JPEG 域两个方面对其进行了研究。

对于空域图像隐写分析，本文在第三章提出了一种基于图表示学习的网络，将每个图像转换为一个图，用节点表示图像块的隐写特征，边表示图像块之间的关系。将图像转换为图的方式可以更好地反映出图像块之间的局部关系，并有效提取出图像块之间的隐写特征差异。通过卷积神经网络来学习图像局部的隐写信息并用图神经网络学习图像信息的全局相关性，以实现含密图像的高效检测。实验结果表明，本文提出的新型隐写分析网络能够更加准确、快速地检测含密图像，同时也证明了图表示学习技术在隐写分析领域的重要性和优越性。为深度学习图像隐写分析提供了一种新的思路。

在第四章，本文提出了一种融合 CNN 和图注意力机制的 JPEG 域图像隐写分析框架。为了能够增强微弱的隐写信号，提出了隐写特征增强模块和图注意力模块，从局部和全局共同增强提取的隐写信号。同时，该框架还通过预训练和微调进行模型迁移，从而提高模型的性能和收敛速度。与现有方法相比，所提出的方法在检测性能和收敛速度方面都表现出明显的优势，并通过消融实验证明了该框架充分利用了图注意力机制，更好地利用隐写信号的全局特征来进一步提升了模型的准确性

综合来看，本文的研究为深度学习技术在图像隐写分析领域的应用提供了一种新的思路，并提出了有效的方法。这项研究具有重要的理论价值和实际应用意义，为后续相关研究提供了参考。

## 5.2 展望

未来，随着数字技术的不断发展和应用，基于图像隐写的信息隐藏和保护方法越来越先进，图像隐写分析领域将变得更加复杂和具有挑战性。因此，对于深度学习技术在该领域中的进一步研究和探索是必要的。

(1) 未来的研究方向可以考虑进一步探究基于图神经网络提取图像隐写特征的不同方法。相对于传统的卷积神经网络，图神经网络能够更好地处理非欧几里得空间数据，并从全局角度理解数据之间的关系，为隐写分析提供更加准确和高效的模型。

(2)面对日益复杂的隐写算法，将强化学习技术引入到图像隐写分析中也是一个很好的想法。通过与环境交互，依赖奖励信号优化模型从而提高模型的鲁棒性和泛化能力。

(3) 随着模式识别、计算机视觉等领域的迅速发展，这些领域的先进方法也可以加以应用到图像隐写分析领域中。例如，多模态信息融合、图像重建和生成、复杂场景下的目标检测等技术都可能为图像隐写分析提供新思路和方法，也将进一步推动该领域的发展。

# 参考文献

[1] BENDER W, GRUHL D, MORIMOTO N, et al. Techniques for data hiding[J]. IBM Systems Journal, 1996, 35(3-4): 313-336.

[2] ARTZ D. Digital steganography: Hiding data within data[J]. IEEE Internet Computing, 2001, 5(3): 75-80.

[3] KATZ J, LINDELL Y. Introduction to modern cryptography[M]. CRC Press, 2020:1-498.

[4] CACHIN C. Digital steganography[J]. Encyclopedia of Cryptography and Security, 2005, 2(2005): 129-168.

[5] LECUN Y, BENGIO Y, HINTON G. Deep learning[J]. Nature, 2015, 521(7553): 436-444.

[6] SARA K. A new steganography method based on HIOP (Higher Intensity Of Pixel) algorithm and Strassen's matrix multiplication[J]. Journal of Global Research in Computer Science, 2011, 2(1): 6-12.

[7] MARVEL L M. Information hiding: Steganography and watermarking[J]. Optical and Digital Techniques for Information Security, 2005, 24(2005): 113-133.

[8] SIMMONS G J. The prisoners' problem and the subliminal channel[C]//Proceedings of Advances in Cryptology, 1984: 51-67.

[9] SWAIN G, LENKA S K. Classification of spatial domain image steganography techniques: A study[J]. International Journal of Computer Science and Engineering Technology, 2014, 5(3): 219-232.

[10] SAU K, BASAK R K, CHANDA A. Image compression based on block truncation coding using clifford algebra[J]. Procedia Technology, 2013, 10: 699-706.

[11] Z. WANG, H.R. SHEIKH AND E.P. SIMONCELLI. Image quality assessment: from error visibility to structural similarity, IEEE Transactions on Image Processing, 2014, 13(4):600-612.

[12] PRADHAN A, SAHU A K, SWAIN G, et al. Performance evaluation parameters of image steganography techniques[C]//Proceedings of the International Conference on Research Advances in Integrated Navigation Systems, 2016: 1-8.

[13] JAMIL T. Steganography: The art of hiding information in plain sight[J]. IEEE Potentials, 1999, 18(1): 10-12.

[14] CELIK M U, SHARMA G, TEKALP A M, et al. Lossless generalized-LSB data embedding[J]. IEEE Transactions on Image Processing, 2005, 14(2): 253-266.

[15] FRIDRICH J, SOUKAL D. Matrix embedding for large payloads[J]. IEEE Transactions on Information Forensics and Security, 2006, 1(3): 390-395.

[16] X. ZHANG, S. WANG. Efficient Steganographic embedding by exploiting modification direction[J]. IEEE Communications Letters, 2006, 10(11): 781-783.

[17] X. ZHANG, W. ZHANG, S. WANG. Efficient double-layered steganographic embedding[J]. Electronics Letters,2007, 43(8): 482-483

[18] Y. KIM, Z. DURIC, D. RICHARDS. Modified matrix encoding technique for minimal distortion steganography[C]//Proceedings of the 8th International Workshop on Information Hiding, 2006: 412-327.

[19] J. EGGERS, R. BAEUML, B. GIROD. Communications approach to image steganography[C] //Proceedings of the International Symposium on Electronic Imaging on ecurityand Watermarking of Multimedia Contents, 2002: 26-37.

[20] S. PHIL. Model-Based steganography[C]//Proceedings of the Second International Workshop on Digital-forensics and Watermarking, 2003: 154-167.

[21] WESTFELD A. F5-a steganographic algorithm[C]//Proceedings of the 4th International Workshop on Information Hiding, 2001: 289-302.

[22] FRIDRICH J, PEVNY T, KODOVSKY J. Statistically undetectable JPEG steganography: Dead ends, challenges, and opportunities[C]//Proceedings of the 9th Workshop on Multimedia and Security, 2007: 3-14.

[23] FILLER T, JUDAS J, FRIDRICH J. Minimizing additive distortion in steganography using syndrome-trellis codes[J]. IEEE Transactions on Information Forensics and Security, 2011, 6(3): 920–935.

[24] PEVNY T, FILLER T, BAS P. Using high-dimensional image models to perform highly undetectable steganography[C]//Proceedings of the 12th International Workshop on

Information Hiding, 2010: 161-177.

[25] PEVNY T, BAS P, FRIDRICH J. Steganalysis by subtractive pixel adjacency matrix[J]. IEEE Transactions on Information Forensics and Security, 2010, 5(2): 215-224.

[26] HOLUB V, FRIDRICH J. Designing steganographic distortion using directional filters[C]//Proceedings of the IEEE International Workshop on Information Forensics and Security, 2012: 234-239.

[27] HOLUB V, FRIDRICH J, DENEMARK T. Universal distortion function for steganography in an arbitrary domain[J]. EURASIP Journal on Information Security, 2014, 2014(1): 1-13.

[28] LI B, WANG M, HUANG J W, et al. A new cost function for spatial image steganography[C]//Proceedings of the IEEE International Conference on Image Processing, 2014: 4206-4210.

[29] FRIDRICH J, KODOVSKY J. Multivariate gaussian model for designing additive distortion for steganography[C]//Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, 2013: 2949-2953.

[30] SEDIGHI V, COGRANNE R, FRIDRICH J. Content-adaptive steganography by minimizing statistical detectability[J]. IEEE Transactions on Information Forensics and Security, 2016, 11(2): 221-234.

[31] GUO L, NI J, SU W, et al. Using statistical image model for JPEG steganography: Uniform embedding revisited[J]. IEEE Transactions on Information Forensics and Security, 2015, 10(12): 2669-2680.

[32] COGRANNE R, GIBOULOT Q, BAS P. Steganography by minimizing statistical detectability: The cases of JPEG and color images[C]//Proceedings of the 8th ACM Workshop on Information Hiding and Multimedia Security, 2020: 161-167.

[33] COGRANNE R, GIBOULOT Q, BAS P. Efficient steganography in JPEG images by minimizing performance of optimal detector[J]. IEEE Transactions on Information Forensics and Security, 2021, 17(2021): 1328-1343.

[34] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial networks[J]. Communications of the ACM, 2020, 63(11): 139-144.

[35] BALUJA S. Hiding images in plain sight: Deep steganography[J]. Advances in Neural Information Processing Systems, 2017: 2066-2076.

[36] UR REHMAN A, RAHIM R, NADEEM S, et al. End-to-end trained CNN encoder-decoder networks for image steganography[C]//Proceedings of the European Conference on Computer Vision, 2019: 723-729.

[37] SHI H, DONG J, WANG W, et al. SSGAN: Secure steganography based on generative adversarial networks [C]//Proceedings of the18th Pacific-Rim Conference on Multimedia Information Processing, 2018: 534-544.

[38] ZHANG R, DONG S, LIU J. Invisible steganography via generative adversarial networks[J]. Multimedia Tools and Applications, 2019, 78(7): 8559-8575.

[39] ZHU J, KAPLAN R, JOHNSON J, et al. HiDDeN: Hiding data with deep networks[C]//Proceedings of the 15th European Conference on Computer Vision, 2018: 682-697.

[40] YANG J, RUAN D, HUANG J. An embedding cost learning framework using GAN[J]. IEEE Transactions on Information Forensics and Security, 2019, 15(1): 839-851.

[41] CHANG C-C, LIN C-J. LIBSVM: A library for support vector machines[J]. ACM Transactions on Intelligent Systems and Technology, 2011, 2(3): 1-27.

[42] KODOVSKY J, FRIDRICH J, HOLUB V. Ensemble classifiers for steganalysis of digital media[J]. IEEE Transactions on Information Forensics and Security, 2011, 7(2): 432-444.

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

[44] KIM Y, DURIC Z, RICHARDS D. Modified matrix encoding technique for minimal distortion steganography[C]//Proceedings of the 8th International Workshop on Information Hiding, 2007: 314-327.

[45] HOLUB V, FRIDRICH J. Random projections of residuals for digital image steganalysis[J]. IEEE Transactions on Information Forensics and Security, 2013, 8(12): 1996-2006.

[46] TANG W, LI H, LUO W, et al. Adaptive steganalysis against WOW embedding algorithm[C]//Proceedings of the 2nd ACM Workshop on Information Hiding and Multimedia

Security, 2014: 91-96.

[47] DENEMARK T, SEDIGHI V, HOLUB V, et al. Selection-channel-aware rich model for steganalysis of digital images[C]//Proceedings of the IEEE International Workshop on Information Forensics and Security, 2014: 48-53.

[48] TANG W, LI H, LUO W, et al. Adaptive steganalysis based on embedding probabilities of pixels[J]. IEEE Transactions on Information Forensics and Security, 2016, 11(4): 734-745.

[49] DENEMARK T, FRIDRICH J, COMESANA-ALFARO P. Improving selection-channel aware steganalysis features[C]//Proceedings of the International Symposium on Electronic Imaging, 2016: 1-8.

[50] PEVNY T, FRIDRICH J. Merging markov and DCT features for multi-class JPEG steganalysis[C]//Proceedings of the Security, Steganography, and Watermarking of Multimedia Contents, 2007: 28-40.

[51] KODOVSKY J, FRIDRICH J. Steganalysis of JPEG images using rich models[C]// Proceedings of the Media Watermarking, Security, and Forensics, 2012, 8083: 81-93.

[52] HOLUB V, FRIDRICH J. Low-complexity features for JPEG steganalysis using undecimated DCT[J]. IEEE Transactions on Information Forensics and Security, 2015, 10(2): 219-228.

[53] HOLUB V, FRIDRICH J. Phase-aware projection model for steganalysis of JPEG images[C]//Proceedings of the Media Watermarking, Security, and Forensics, 2015: 259-269.

[54] SONG X, LIU F, YANG C, et al. Steganalysis of adaptive JPEG steganography using 2D Gabor filters[C]//Proceedings of the 3rd ACM Workshop on Information Hiding and Multimedia Security, 2015: 15-23.

[55] FENG G, ZHANG X, REN Y, et al. Diversity-based cascade filters for JPEG steganalysis[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2020, 30(2): 376-386.

[56] DENEMARK T, BOROUMAND M, FRIDRICH J. Steganalysis features for content-adaptive JPEG steganography[J]. IEEE Transactions on Information Forensics and Security, 2016, 11(8): 1736-1746.

[57] TAN S, LI B. Stacked convolutional auto-encoders for steganalysis of digital images[C]//Proceedings of the Asia-Pacific Signal and Information Processing Association

Annual Summit and Conference, 2014: 1-4.

[58] QIAN Y, DONG J, WANG W, et al. Deep learning for steganalysis via convolutional neural networks[C]//Proceedings of the Media Watermarking, Security, and Forensics, 2015: 171-180.

[59] XU G, WU H-Z, SHI Y-Q. Structural design of convolutional neural networks for steganalysis[J]. IEEE Signal Processing Letters, 2016, 23(5): 708-712.

[60] XU G, WU H-Z, SHI Y Q. Ensemble of CNNs for steganalysis: An empirical study[C]//Proceedings of the 4th ACM Workshop on Information Hiding and Multimedia Security, 2016: 103-107.

[61] YE J, NI J, YI Y. Deep learning hierarchical representations for image steganalysis[J]. IEEE Transactions on Information Forensics and Security, 2017, 12(11): 2545-2557.

[62] BOROUMAND M, CHEN M, FRIDRICH J. Deep residual network for steganalysis of digital images[J]. IEEE Transactions on Information Forensics and Security, 2019, 14(5): 1181-1193.

[63] DENG X, CHEN B, LUO W, et al. Fast and effective global covariance pooling network for image steganalysis[C]//Proceedings of the 7th ACM Workshop on Information Hiding and Multimedia Security, 2019: 230-234.

[64] TAN S, WU W, SHAO Z, et al. CALPA-net: Channel-pruning-assisted deep residual network for steganalysis of digital images[J]. IEEE Transactions on Information Forensics and Security, 2019, 16(2019): 131-146.

[65] ZHANG R, ZHU F, LIU J, 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(2020): 1138-1150.

[66] YOU W, ZHANG H, ZHAO X. A siamese CNN for image steganalysis[J]. IEEE Transactions on Information Forensics and Security, 2021, 16(2021): 291-306.

[67] REN Y, LIU Y, WANG L. Using contrastive learning to improve the performance of steganalysis schemes[C]//Proceedings of the 20th International Workshop on Digital Forensics and Watermarking, 2022: 212-226.

[68] JANG H, OH T-W, KIM K. Feature aggregation networks for image steganalysis[C]//Proceedings of the 8th ACM Workshop on Information Hiding and

Multimedia Security, 2020: 33-38.

[69] YOUSFI Y, BUTORA J, KHVEDCHENYA E, et al. ImageNet pre-trained CNNs for JPEG steganalysis[C]//Proceedings of the IEEE International Workshop on Information Forensics and Security, 2020: 1-6.

[70] COGRANNE R, GIBOULOT Q, BAS P. ALASKA#2: Challenging academic research on steganalysis with realistic images[C]//Proceedings of the IEEE International Workshop on Information Forensics and Security, 2020: 1-5.

[71] TAN M, LE Q. EfficientNet: Rethinking model scaling for convolutional neural networks[C]//Proceedings of the 36th International Conference on Machine Learning, 2019: 6105-6114.

[72] TAN M, LE Q V. MixConv: Mixed depthwise convolutional kernels[J]. arXiv preprint arXiv:1907.09595, 2019.

[73] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 770-778.

[74] DENG J, DONG W, SOCHER R, et al. ImageNet: A large-scale hierarchical image database[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2009: 248-255.

[75] MIELIKAINEN J. LSB matching revisited[J]. IEEE Signal Processing Letters, 2006, 13(5): 285−287.

[76] ZHANG X, WANG S. Efficient steganographic embedding by exploiting modification direction[J]. IEEE Communications Letters, 2006, 10(11): 781-783.

[77] GUO L, NI J, SHI Y Q. Uniform embedding for efficient JPEG steganography[J]. IEEE Transactions on Information Forensics and Security, 2014, 9(5): 814-825.

[78] KIM Y, DURIC Z, RICHARDS D. Modified matrix encoding technique for minimal distortion steganography[C]//Proceedings of the 8th International Workshop on Information Hiding, 2007: 314-327.

[79] SACHNEV V, KIM H J, ZHANG R. Less detectable JPEG steganography method based on heuristic optimization and BCH syndrome coding[C]//Proceedings of the 11th ACM

Workshop on Multimedia and Security, 2009: 131-140.

[80] SHARMA K, AGGARWAL A, SINGHANIA T, et al. Hiding data in images using cryptography and deep neural network[J]. arXiv preprint arXiv:1912.10413, 2019.

[81] 杨晓元, 毕新亮, 刘佳等. 结合图像加密与深度学习的高容量图像隐写算法[J]. 通信学报, 2021, 42(9): 96-105.

[82] DUAN X, GOU M, LIU N, et al. High-capacity image steganography based on improved Xception[J]. Sensors, 2020, 20(24): 7253.

[83] KICH I, TAOUIL Y , BENHFID A. Image steganography scheme using dilated convolutional network[C]//Proceedings of the 12th International Conference on Information and Communication Systems, 2021: 305-309.

[84] SUBRAMANIAN N, CHEHEB I, ELHARROUSS O, et al. End-to-end image steganography using deep convolutional autoencoders[J]. IEEE Access, 2021, 9: 135585-135593.

[85] BALUJA S. Hiding images within images[J]. IEEE Transactions on Pattern Analysis Machine Intelligence, 2020, 42(07): 1685-1697.

[86] LU S P, WANG R, ZHONG T, et al. Large-capacity image steganography based on invertible neural networks[C]//Proceedings of the Conference on Computer Vision and Pattern Recognition, 2021: 10816-10825.

[87] VOLKHONSKIY D, NAZAROV I, BURNAEV E. Steganographic generative adversarial networks[C]//Proceedings of the International Society for Optical Engineering, 2020, 11433: 991-1005.

[88] WANG Z, GAO N, WANG X, et al. HidingGAN: High capacity information hiding with generative adversarial network[J]. Computer Graphics Forum, 2019, 38(7): 393-401.

[89] ZHANG K A, CUESTA-INFANTE A, XU L, et al. SteganoGAN: High capacity image steganography with GANs[J]. arXiv preprint arXiv:1901.03892, 2019.

[90] WANG Y , FU Z, SUN X. High visual quality image steganography based on encoder-decoder model[J]. Journal of Cybersecurity, 2020, 2(3): 115-121.

[91] 郑钢, 胡东辉, 戈辉,等. 生成对抗网络驱动的图像隐写与水印模型[J]. 中国图象图形学报, 2021, 26(10): 2485-2502.

[92] CHEN B, WANG J, CHEN Y , et al. High-capacity robust image steganography via adversarial network[J]. KSII Transactions on Internet Information Systems, 2020, 14(1): 366-381.

[93] FU Z, WANG F, CHENG X. The secure steganography for hiding images via GAN[J]. EURASIP Journal on Image Video Processing, 2020, 2020(1): 1-18.

[94] HE K, ZHANG X, REN S, et al. Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification[C]//Proceedings of the IEEE International Conference on Computer Vision, 2015: 1026-1034.

[95] HE K, ZHANG X, REN S, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(9): 1904-1916.

[96] ZENG J, TAN S, LI B, et al. Large-scale JPEG steganalysis using hybrid deep-learning framework[J]. IEEE Transactions on Information Forensics and Security, 2018, 13(5): 1200-1214.

[97] XU G. Deep convolutional neural network to detect J-UNIWARD[C]//Proceedings of the 5th ACM Workshop on Information Hiding and Multimedia Security, 2017: 67-73.

[98] CHEN M, SEDIGHI V, BOROUMAND M, et al. JPEG-phase-aware convolutional neural network for steganalysis of JPEG images[C]//Proceedings of the 5th ACM Workshop on Information Hiding and Multimedia Security, 2017: 75-84.

[99] ZHANG S, TONG H, XU J, et al. Graph convolutional networks: A comprehensive review[J]. Computational Social Networks, 2019, 6(1): 1-23.

[100] GORI M, MONFARDINI G, SCARSELLI F. A new model for learning in graph domains[C]//Proceedings of the IEEE International Joint Conference on Neural Networks, 2005, 2: 729-734.

[101] SCARSELLI F, GORI M, TSOI A C, et al. The graph neural network model[J]. IEEE Transactions on Neural Networks, 2008, 20(1): 61-80.

[102] BRUNA J, ZAREMBA W, SZLAM A, et al. Spectral networks and locally connected networks on graphs[J]. arXiv preprint arXiv:1312.6203, 2013.

[103] GILMER J, SCHOENHOLZ S S, RILEY P F, et al. Neural message passing for quantum

chemistry[C]//Proceedings of the 34th International Conference on Machine Learning, 2017: 1263-1272.

[104] VELICKOVIC P, CUCURULL G, CASANOVA A, et al. Graph attention networks[J]. arXiv preprint arXiv:1710.10903, 2017.

[105] LUO W, HUANG F, HUANG J. Edge adaptive image steganography based on LSB matching revisited[J]. IEEE Transactions on Information Forensics and Security, 2010, 5(2): 201-214.

[106] SHI Y Q, CHEN C, CHEN W. A Markov process based approach to effective attacking JPEG steganography[C]//Proceedings of the 8th International Workshop on Information Hiding, 2007: 249-264.

[107] WU H. Unsupervised steganographer identification via clustering and outlier detection[M]. Digital Media Steganography, 2020: 295-319.

[108] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. ImageNet classification with deep convolutional neural networks[J]. Communications of the ACM, 2017, 60(6): 84-90.

[109] WU Z, PAN S, CHEN F, et al. A comprehensive survey on graph neural networks[J]. IEEE Transactions on Neural Networks and Learning Systems, 2020, 32(1): 4-24.

[110] AVELAR P H C, TAVARES A R, DA SILVEIRA T L T, et al. Superpixel image classification with graph attention networks[C]//Proceedings of the 33rd Conference on Graphics, Patterns and Images, 2020: 203-209.

[111] NAIR V, HINTON G E. Rectified linear units improve restricted boltzmann machines[C]//Proceedings of the 27th International Conference on Machine Learning, 2010: 807-814.

[112] BAS P, FILLER T, PEVNÝ T. “Break our steganographic system”: The ins and outs of organizing BOSS[C]//Proceedings of the 13th International Conference on Information Hiding, 2011: 59-70.

[113] KINGMA D P, BA J. Adam: A method for stochastic optimization[J]. arXiv preprint arXiv:1412.6980, 2014.

[114] WU H, WANG H, ZHAO H, et al. Multi-layer assignment steganography using graph-theoretic approach[J]. Multimedia Tools and Applications, 2015, 74(2015): 8171-8196.

[115] SU A, HE X, ZHAO X. JPEG steganalysis based on ResNeXt with gauss partial derivative filters[J]. Multimedia Tools and Applications, 2021, 80(2021): 3349-3366.

[116] HAN K, WANG Y, GUO J, et al. Vision gnn: An image is worth graph of nodes[J]. arXiv preprint arXiv:2206.00272, 2022.

[117] IOFFE S, SZEGEDY C. Batch normalization: Accelerating deep network training by reducing internal covariate shift[C]//Proceedings of the 32nd International Conference on International Conference on Machine Learning, 2015: 448-456.

[118] YOUSFI Y, BUTORA J, FRIDRICH J, et al. Improving EfficientNet for JPEG steganalysis[C]//Proceedings of the 2021 ACM Workshop on Information Hiding and Multimedia Security, 2021: 149-157.

[119] BAS P, FURON T. Image database of BOWS-2[DB]. Available at:{http://bows2. gipsa-lab. inpg. fr}, 2007.

[120] HE K, ZHANG X, REN S, et al. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification[C]//Proceedings of the IEEE International Conference on Computer Vision, 2015: 1026-1034.

# 作者在攻读硕士学位期间公开发表的论文

[1] LIU Q, ZHOU L, WU H. Graph representation learning for spatial image steganalysis [C]//Proceedings of the 24th International Workshop on Multimedia Signal Processing, 2022: 1-5. **(EI: 20225013233804)**

[2] LIU Q, YANG Z, WU H. JPEG steganalysis based on steganographic feature enhancement and graph attention learning[J]. Journal of Electronic Imaging, 2023, 23(3): 033032. **（SCI 检索期刊，已发表）**

# 作者在攻读硕士学位期间所参与的项目

[1] 国家自然科学基金青年项目“社交网络多用户协同的行为隐写”(项目编号：61902235)

# 致　谢

在这篇论文完成之际，我要向所有在我学术生涯中帮助和支持我的人们致以最衷心的感谢。

首先，我需要感谢我的导师吴汉舟。他的智慧、耐心和关爱不仅帮助我完成了这篇论文，还指导了我整个研究生阶段。他的批评和鼓励一直是我努力科研的动力。同时，感谢他将那么多珍贵的时间花费在我的研究和毕业论文上。在此，谨向吴老师致以深深的敬意和由衷的感谢。

感谢网络空间安全实验室的同学们在我的研究和写作过程中给予的支持和帮助，我们一起经历了许多有趣的、困难的、挑战的时刻。特别是易标师兄给了我很多启示性的建议，郑晓燕、杨天予和唐雄同学在科研路上给予了我很大鼓励。感谢每一个为我提供帮助的人。

同时，我还要感谢我的家人。他们一直是我生活中的支柱，支持我追求自己的梦想，在我疲倦或失落的时候，给我力量重新振作。我将永远珍惜他们对我的爱和关心。

最后，我要感谢所有为这篇论文提供数据、样本和参考文献的团体和个人。您们的贡献是这篇论文完成的基础，帮助我得到了更丰富全面的研究结论。在此向您们表示深深的感激之情。

路漫漫其修远兮，未来我将不断努力，探索前行。而今日，我要感谢那些一路相伴的人们，在我前进的路上点亮明灯，在我背负重担时伸出援手，感谢你们！