

# 基于S&P和Rec-Net的图像隐蔽通信 主动防御方法

马媛媛<sup>1,2)</sup>

赵颖澳<sup>1)</sup>

张 祎<sup>3)</sup>

张倩倩<sup>1,3)</sup>

罗向阳<sup>3)</sup>

<sup>1)</sup>(河南师范大学计算机与信息工程学院 河南 新乡 453007)

<sup>2)</sup>(河南师范大学河南省教育人工智能与个性化学习重点实验室 河南 新乡 453007)

<sup>3)</sup>(信息工程大学河南省态势感知重点实验室 郑州 450001)

**摘要** 近年来,隐蔽通信在社交网络中的广泛应用加剧了网络安全风险,使得可靠防御面临新的挑战。在防御方面,以往的研究主要集中在隐写检测等被动防御。然而,目前被动防御存在一些漏洞,一方面,在低负载率下,隐写检测的虚警率和漏检率过高,使得隐写检测尚未达到百分百正确率,导致无法可靠判别;另一方面,因在社交网络等现实场景中无法获得载密图像的负载率、质量因子和隐写算法种类等先验知识,导致隐写分析人员难以对秘密信息进行定位和提取。因此,隐写检测为主的被动防御无法及时应对实际应用场景。针对上述问题,本文提出一种针对图像隐写的隐蔽通信主动防御方法,使得在通信双方毫无察觉的情况下彻底阻断秘密信息的传输。首先,分析不同噪声模型对载密图像的破坏程度,选取效果最好的椒盐噪声(Salt-and-Pepper noise, S&P)进行秘密信息的破坏,得到噪声图像;其次,通过对中值滤波层和高斯滤波层的原理性分析,发现中值滤波层和高斯滤波层适用于恢复噪声图像质量和破坏秘密信息,基于此,设计一个端到端的图像恢复网络(Recovery Network, Rec-Net),得到高质量的“干净”图像。Rec-Net既维持社交网络通信双方传递图像的视觉效果和秘密信息破坏效果,又不改变图像的存储空间大小;最后,鉴于误码率和清除率准则在未知隐写和完整秘密信息序列等先验知识前提下无法度量主动防御效果,本文提出一种新的基于改变率的隐写主动防御图像评价准则,同时能够快捷准确度量图像中秘密信息破坏的主动防御效果。所提方法不仅对不同隐写的隐蔽通信具有通用性,而且满足社交网络实时性要求。实验结果表明,在不同数据集下,本文方法均具有高秘密信息破坏效果和高图像质量,能够达到100%的防御成功率,阻断社交网络中的隐蔽通信,其中“干净”图像的误码率最高可达到53%。同时,在不同负载率的数据集下,本方法与SC-Net方法和AO-Net方法进行对比,在秘密信息破坏方面各提升5.61%和0.56%;在图像质量方面各提升4.44%和34.8%。

**关键词** 主动防御;鲁棒隐写;卷积神经网络;S&P噪声;隐蔽通信

中图法分类号 TP18 DOI号 10.11897/SP.J.1016.2025.00985

## An Active Defense Method for Image Covert Communication Based on S&P and Rec-Net

MA Yuan-Yuan<sup>1,2)</sup> ZHAO Ying-Ao<sup>1)</sup> ZHANG Yi<sup>3)</sup> ZHANG Qian-Qian<sup>1,3)</sup> LUO Xiang-Yang<sup>3)</sup>

<sup>1)</sup>(Department of Computer and Information Engineering, Henan Normal University, Xinxiang, Henan 453007)

<sup>2)</sup>(Engineering Lab of Intelligence Business & Internet of Things, Henan Normal University, Xinxiang, Henan 453007)

<sup>3)</sup>(Henan Province Key Laboratory of Cyberspace Situation Awareness, Information Engineering University, Zhengzhou 450001)

**Abstract** In recent years, the wide application of covert communication in social networks has exacerbated network security risks, making reliable defense face new challenges. In terms of

收稿日期:2024-05-31;在线发布日期:2025-02-25。本课题得到河南省优秀青年科学基金(222300420058)、国家自然科学基金(620220103, 62202495, 62272163, 62172435, U23A20305)、国家重点研发项目(2022YFB3102900)、中原科技创新领军人才计划(214200510019)、河南省重点研发项目(221111321200)资助。马媛媛(通信作者),博士,副教授,中国计算机学会(CCF)会员,主要研究领域为网络与信息安全、图像隐写分析、粒计算。E-mail: 121100@htu.edu.cn。赵颖澳,硕士研究生,中国计算机学会(CCF)会员,主要研究领域为隐写分析。张 祎,博士,讲师,主要研究领域为图像隐写与分析技术。张倩倩,博士生,讲师,主要研究领域为粗糙集、粒计算。罗向阳,博士,教授,主要研究领域为网络信息安全、图像隐写、隐写分析。

defense, previous research mainly focuses on passive defense such as steganalytic. However, there are some loopholes in passive defense. On the one hand, under low payload, the false alarm rate and miss detection rate of steganalytic are too high, so that steganalytic has not yet reached the 100% accuracy rate, resulting in the inability to reliably discriminate. On the other hand, due to the impossibility of obtaining the a priori knowledge of the payload of the stego image, the quality factor, the type of steganography algorithms in the real scenarios such as the social networks. It is difficult for steganalysts to locate and extract secret messages. Therefore, the passive defense based on steganalytic can not cope with the practical application scenarios in a timely manner. To address the above problems, this paper proposes an active defense method for image steganography, which can completely block the transmission of secret messages without the awareness of both parties. Firstly, we analyze the damage degree of different noise models to the stego image, and select the most effective Salt-and-Pepper noise (S&P) to destroy secret messages and obtain the noise image. Secondly, we analyze the original rationality of the median and gaussian filter layers, and find that the median and gaussian filter layers are suitable for restoring the quality of noise image and destroying secret messages. Based on this, an end-to-end image recovery network (Rec-Net) is designed to obtain high-quality “clean” image. Rec-Net maintains the visual effect of the image and the effect of destroying secret messages transmitted by both parties in the social networks without changing the size of the image storage space. Finally, given that the bit error rate and removal rate criteria are unable to measure the active defense effect under the premise of a priori knowledge of the unknown steganography and the complete sequence of secret messages, this paper proposes a new active defense steganographic image evaluation criterion based on the change rate. The evaluation criterion can quickly and accurately measure the active defense effect of the destruction of secret messages in the image. The proposed method is not only generalized for covert communications with different steganography, but also meets the real-time requirements of social networks. The experimental results show that the proposed method has high secret messages destruction effect and high image quality in different datasets, and can achieve 100% success rate in blocking covert communication in social networks, and the highest BER of the “clean” image can reach 53%. Meanwhile, under the datasets with different payloads, this method is compared with the SC-Net method and the AO-Net method, and each of them improves 5.61% and 0.56% in terms of secret messages destruction; and each of them improves 4.44% and 34.8% in terms of image quality.

**Keywords** active defense; robust steganography; convolutional neural network; salt-and-pepper noise; covert communication

## 1 引 言

隐写<sup>[1]</sup>是一门将秘密信息隐藏到图像、视频、文本等载体中进行隐蔽通信的技术。作为图像隐写的对立面,隐写分析<sup>[2-5]</sup>是对它的防御,它通过对载体进行检测,判断其中是否隐藏秘密信息,以及指出载体中秘密信息的容量和位置,提取秘密信息,甚至破坏秘密信息。根据防御方式不同,本文将其分为被动防御技术和主动防御技术。被动防御技术不修改

载密图像的内容,主要包括隐写检测<sup>[6]</sup>、隐写定位<sup>[7-9]</sup>、隐写提取等。其中,隐写检测判断图像为载体图像还是载密图像;隐写定位是判断载密图像中秘密信息的位置;隐写提取是通过具体了解先验知识的前提下从载密图像中提取秘密信息。主动防御技术是对被动防御技术的补充和提升,通过主动修改载密图像内容来破坏秘密信息,利用神经网络的表征能力实现对秘密信息的过滤,阻断社交网络平台中不法用户隐蔽通信。目前,在社交网络等现实场景中载密图像的负载率参数、隐写算法种类等先

验知识获取难度较大,使得隐写分析人员对载密图像的检测以及秘密信息的定位和破译较为困难。一方面,被动防御技术对载密图像的检测无法阻止隐蔽通信;另一方面,即使技术人员花费大量时间与精力得到载密图像中的秘密信息,仍然难以保证破译秘密信息的时效性,无法有效阻止隐蔽通信滥用事件发生。因此,在社交网络中隐蔽通信主动防御更具有实际应用优势。

针对隐蔽通信主动防御,研究者已经提出了一些方法。本文将其分为两类,第一类是利用卷积神经网络的主动防御方法。朱等人<sup>[10]</sup>受到DnCNN<sup>[11]</sup>网络的启发,提出了一种包含攻击模块和优化模块的深度学习网络(简称AO-Net),AO-Net能够在破坏秘密信息的能力和图像质量两方面达到较好的均衡。张等人<sup>[12]</sup>受到DnCNN网络架构的启发,提出一种针对水印的主动防御方法,通过测试三种不同的鲁棒水印算法Block DCT、BSS-C、ULPM证明该网络可以对水印图像达到实时防御的目的。李等人<sup>[13]</sup>提出一种基于残差学习的卷积神经网络主动防御方法,通过在网络中提取7个特征块的方式提高水印主动防御能力,利用超深分辨率(Very Deep Super Resolution, VDSR)技术<sup>[14]</sup>,提升图像质量。第二类是利用对抗神经网络的主动防御方法。王等人<sup>[15]</sup>提出针对隐写的生成式对抗网络(Generative Adversarial Networks, GAN)主动防御方法,通过GAN训练n种算法的载密图像到载体图像的映射模型,使得在应用阶段载密图像经过n次无替换抽样的映射模型下达到破坏秘密信息的目的。该文<sup>[16]</sup>提出一种DDSP(Deep Digital Steganography Purifier)网络,DDSP网络利用预训练的自动编码器作为生成器破坏秘密信息,并且通过判别器对生成的干净图像与对应的载体图像进行判别,从而在不牺牲图像质量的情况下针对隐写进行主动防御。Zhu等人<sup>[17]</sup>提出两种针对隐写的Scaling-Net和SC-Net主动防御方法,其中,Scaling-Net适用于社交网络中的超大尺寸图像,SC-Net适用于普通尺寸图像。以上两类主动防御方法都是依靠神经网络强大的端到端学习能力和表征能力,一方面在破坏秘密信息的能力上取得较好的效果,另一方面也在图像的不可感知方面有较大提升。

图像隐写主动防御技术目前面临着诸多难点和挑战。其中,难点在于:(1)现有主动防御方法主要关注于固定模式和正态分布扭曲等破坏方式,以达到对秘密信息破坏的目的,而嵌入秘密信息的模式

及分布无迹可寻,导致破坏未知隐写嵌入时秘密信息误码率低,从而降低图像隐秘信息的阻断率;(2)卷积层的局部性和低冗余性使得中心像素值在感受野内映射时发生改变,使得现有方法缺失有效像素值,导致无法重构干净图像,从而不可避免地造成图像退化的问题,阻碍隐写主动防御在实际场景的推广和使用。(3)现有主动防御评价准则所需的隐写算法、嵌入率以及完整秘密信息序列等,在公开信道场景下通常难以获得,导致无法定量评估主动防御效果。挑战在于:(1)随着社交网络的发展,在主动防御平台破坏秘密消息中,往往存在大量冗余网络,造成计算资源的浪费和时空成本的增加,难以在实际应用中满足图像隐写防御方法对时效性的需求;(2)目前隐写方法层出不穷,例如生成式隐写等均给隐写主动防御带来新的挑战。

目前,图像隐蔽通信主动防御的研究引起广泛关注,其方法虽然能够在一定程度上对秘密信息的提取进行干扰,但是仍存在一定的局限性:(1)现有主动防御方法对载密图像处理后,“干净”图像的误码率仍较低。(2)现有主动防御方法生成的“干净”图像质量被削弱。(3)现有主动防御方法使用频率最高的误码率难以操作,无法给出定量评价结果。

考虑到上述问题,本文拟提出一种基于S&P和Rec-Net的图像隐蔽通信主动防御方法。该方法的应用场景是社交网络中公开通讯方式传递的图像。本文通过加入S&P噪声的方式破坏潜在的载密图像中的秘密信息。另外,将破坏后的图像进行一定程度的恢复,使得最终得到的“干净”图像质量不低于“发布”图像质量,能够综合考虑秘密信息破坏与恢复图像质量两个因素。本方法能够使接收方即使获得载密图像但仍然无法成功获取秘密信息,达到在社交网络中图像隐蔽通信主动防御的目的。本文的主要贡献如下:

(1) 对比分析不同噪声模型对社交网络中“发布”图像的破坏程度后,选取破坏能力最好的S&P噪声无差别攻击载体图像和载密图像,即在载体图像和载密图像中叠加椒噪点和盐噪点,使得载密图像中存在的秘密信息被覆盖和过滤,从而得到破坏秘密信息后的噪声图像,进而达到主动防御的目的。

(2) 鉴于噪声图像中存在明显黑白像素点,影响图像的视觉效果,在图像恢复卷积神经网络Rec-Net中,本文增加更适用于优化S&P噪声图像质量兼顾破坏秘密信息的中值滤波层和高斯滤波层,既

维持社交网络通信双方传递图像的视觉效果又不改变图像的存储空间大小。

(3) 误码率无法在未知隐写算法和未知秘密信息序列的情况下得到评价结果,在实际应用中具有较大局限性。针对上述问题,本文提出一种隐写主动防御图像评价准则——改变率,该准则能够在未知隐写算法与完整秘密信息序列等先验条件下更加快速简便地度量载密图像中秘密信息的破坏效果。

本文的其余部分组织如下:第2节介绍本工作的动机;第3节阐述本文提出的主动防御方法;第4节介绍隐写主动防御图像评价准则;在第5节中,验证本文方法的有效性;最后一部分对本文进行总结与展望。

## 2 动 机

本文动机主要包括两个方面,分别是“加噪”主动防御的动机和隐写主动防御图像评价准则的动机。

### 2.1 “加噪”主动防御的动机

近些年来,图像的隐写检测取得长足的发展,检测准确率也已达到较高的水平。目前,一些层出不穷的生成式隐写算法<sup>[18]</sup>给检测带来巨大挑战。鲁棒图像隐写技术是指在非完美传输通道下嵌入隐秘信息,实现隐蔽通信。Luo等人<sup>[2]</sup>提出一种抗JPEG压缩检测的自适应隐写算法DMAS。DMAS利用基于量化表的自适应抖动调制算法、基于侧信息的嵌入代价计算算法以及RS码,能够保证嵌入信息对JPEG压缩的鲁棒性。在DMAS基础上,Chen等人<sup>[4]</sup>提出自适应隐写算法GMAS。该算法通过双层STCs的三元嵌入并合理扩展嵌入域,显著提高GMAS的鲁棒性与安全性。与此同时,社交网络存在的图像数量巨大,且规格类型多样,不乏存在隐写算法与负载率未知的载密图像,这导致载体失配问题。上述问题使得隐写分析从基础科学研究到实际应用的转化过程减慢。对于经过普通隐写算法嵌入秘密信息后的载密图像而言,裁剪、美化、融合、拼贴等轻量级的数字图像处理能够达到破坏载密图像中秘密信息的效果。然而,对于目前较为流行的鲁棒隐写技术<sup>[19-25]</sup>而言,一方面,简单的有损信道无法达到破坏其秘密信息的目的;另一方面,由于在低载荷下隐写检测结果的虚警率和漏检率过高,导致最终对载密图像的分类结果不够准确,即图像隐写检测在防止鲁棒隐写秘密通信方面尚未达到实际应用要

求。此外,安全部门即使确认社交网络传输的是载密图像,可能仍无法使用IP定位找到接收方的位置,依然不能阻止接收方的不法行为。

因此,图像隐写分析需要主动防御技术。该技术对社交网络中所有的图像进行处理,使得无论是否检测到载密图像,接收方都无法从载密图像中提取秘密信息,从而达到图像隐蔽通信主动防御的目的。图1展示在社交网络中图像隐写检测和图像隐蔽通信主动防御的一般过程。

由图1(a)可知,社交网络中图像隐写检测的一般过程分为两大类。第一类是传统的隐写检测方法,第二类是基于深度学习的隐写检测方法。传统的隐写检测方法对上传到社交网络中的图像进行特征提取、特征增强和分类检测。基于深度学习的隐写检测方法主要通过神经网络强大的表征学习能力对图像进行分类检测。无论是传统还是基于深度学习的图像隐写检测方法,接收方无法直接阻止秘密信息的传输。由图1(b)可知,主动防御能够使得接收方无法提取秘密信息,从而成功阻止秘密信息的传输。

本文根据方式不同,将主动防御分为两类,如图1(b)所示。第一类方法是通过去除隐写噪声的主动防御方法,简称“去噪”方法。由于隐写可被认为是在图像中加入一些高频弱噪声信号,因此,使用去噪网络或者图像滤波能够破坏其中的秘密信息。此类方法具有减少人工干预的优势,但在破坏秘密信息效果方面有待提高。一方面,载密图像具有特征隐蔽性的特点,与载体图像高度相似,通过常规图像领域的去噪网络进行训练较为困难;另一方面,使用图像滤波对秘密信息进行破坏则会降低图像的视觉质量。第二类方法是通过叠加噪声的形式破坏秘密信息,简称“加噪”方法。“加噪”方法本质是通过孤立噪声点对载密图像像素点进行无差别覆盖,实现破坏秘密信息的目的。该方法的优势在于无论载密图像中的秘密信息嵌入在哪个区域,“加噪”方法都能够全面地破坏秘密信息。鉴于此,本文考虑使用“加噪”方法对载密图像进行攻击,达到隐蔽通信主动防御的目的。

### 2.2 隐写主动防御图像评价准则的动机

隐写主动防御图像评价准则是在一定的标准下对秘密信息破坏效果进行定量度量,为主动防御方法的有效性提供判定依据。目前较为常用的准则有误码率(Bit Error Rate, BER)<sup>[12]</sup>和清除率(Removal Rate, RR)<sup>[10]</sup>。误码率是秘密信息中错误的比特数

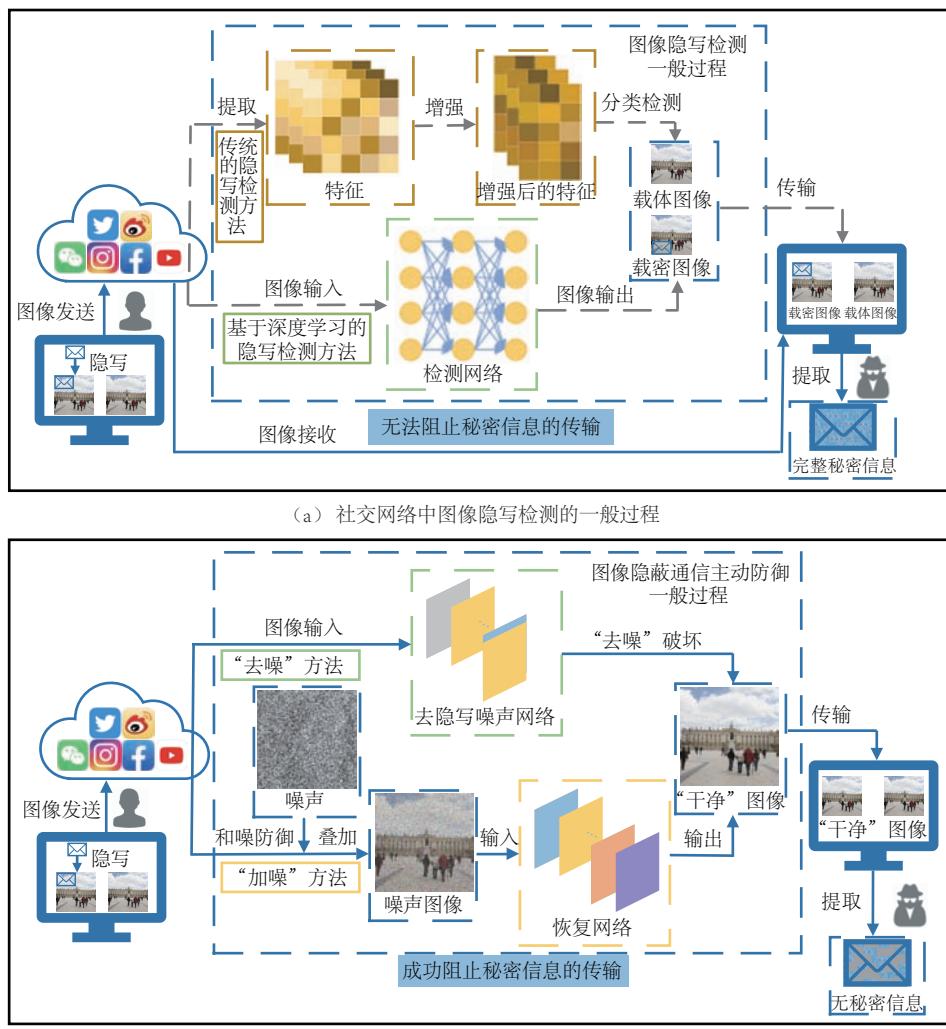


图1 社交网络中图像隐写检测和图像隐蔽通信主动防御的一般过程

占总比特数的比率。计算误码率时需要提前知晓隐写算法、隐写负载率、载密图像质量因子与完整的秘密信息序列，并对载密图像中嵌入的秘密信息进行提取，才能与发送方嵌入的秘密信息序列进行比较，从而得到错误比特数，导致该准则在社交网络等实际场景下难以应用。即使侥幸得到上述各种先验知识，提取秘密信息的过程也较为烦琐，无法快速给出定量评价结果。清除率评价方法的计算过程如下所示。

$$RR = 1 - \|2 \times BER - 1\| \quad (1)$$

可见：清除率评价方法的计算需要使用到误码率，因此在误码率无法计算时，也同样无法得到清除率的结果。因此清除率和误码率类似，都具有相同的局限性。鉴于上述可知，本文提出一种在无任何先验条件下即可对秘密信息破坏效果进行度量的评价准则。

### 3 提出的方法

针对被动防御的局限性，本文提出基于S&P和Rec-Net的图像隐蔽通信主动防御方法（简称SPRN方法），在“发布”图像中加入S&P噪声，破坏社交网络中潜在载密图像的秘密信息，以达到主动防御的目的。本节将从SPRN方法原理、S&P主动防御和Rec-Net图像恢复三个方面对SPRN方法进行详细阐述。

#### 3.1 SPRN方法原理

针对隐写检测在社交网络中的局限性，本文提出SPRN方法，通过修改图像内容的方式破坏秘密信息，使接收方即使获得载密图像仍无法成功获取秘密信息。该方法既维持图像质量，又破坏隐藏在其中的秘密信息，从而实现图像隐蔽通信主动防御的目的。本方法主要包括两部分，如图2所示。

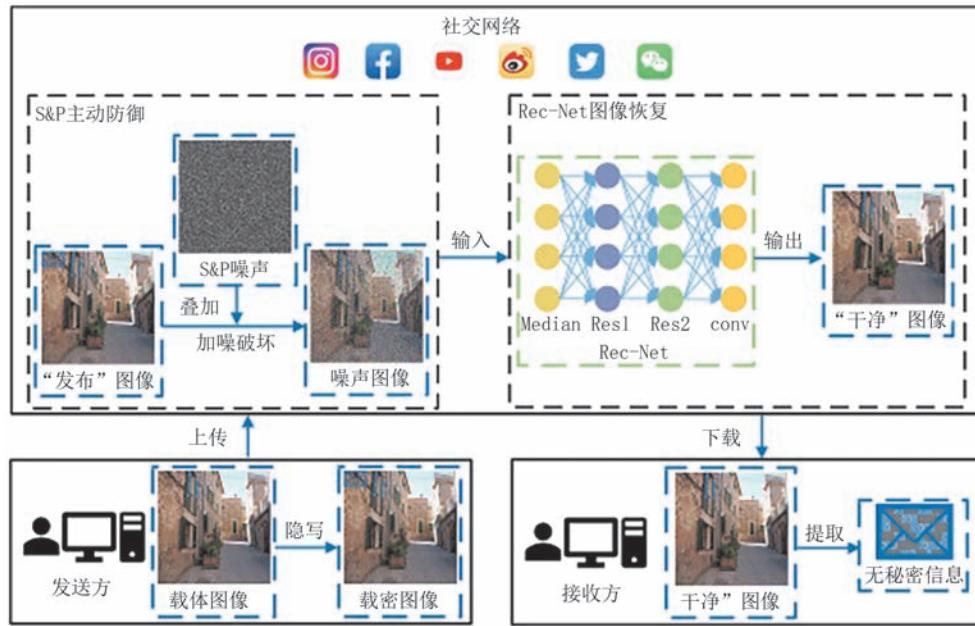


图2 本方法的整体框架

图2中,第一部分为S&P主动防御,在该阶段“发布”图像中加入S&P噪声,得到噪声图像,达到破坏秘密信息的目的。第二部分为Rec-Net图像恢复,通过训练好的带中值滤波层与高斯滤波残差块的Rec-Net对噪声图像进行优化,使得在无人察觉的情况下恢复该图像的质量,并维持第一部分中主动防御效果。最终,接收方将会接收到“干净”图像,且无法提取秘密信息。

### 3.2 S&P主动防御

在计算机视觉领域,图像噪声和图像滤波都对图像产生不同程度的影响。隐蔽通信主动防御可以通过加入噪声和滤波的方式破坏载密图像中的秘密信息。图像噪声通常是存在于图像中的干扰信息或者无关信息,对图像的分辨率会有较大影响。常见的噪声有S&P噪声、高斯白噪声、泊松噪声等。通过分析文献[26]发现,如果恰当使用噪声就能够成功防御隐蔽通信,这是因为在载密图像中加入噪声点便会覆盖掉该图像中嵌入秘密信息的像素点,噪声点越随机,秘密信息被破坏得越多。由于高斯白噪声与泊松噪声的概率密度函数分别符合高斯分布与泊松分布,它们的噪声点不具有严格随机性,破坏秘密信息较少。并且,泊松噪声依赖于信号本身,对于不同强度的信号会导致噪声点出现概率不同,无法达到破坏秘密信息的目的,不能成功防御隐蔽通信。由于S&P噪声是脉冲噪声,它以纯噪声的形式出现在被污染的位置(称为缺失像素),并且随机地改变一些像素值,能够在缺失像素上,擦除载密图像

在该位置的所有秘密信息,故可成功防御隐蔽通信。为了验证上述分析的可信性,本文采用DMAS和GMAS隐写算法,得到质量因子为95的30张载密图像,在其中分别加入均值和方差均为0.01的高斯白噪声、泊松噪声、噪声水平为0.5的S&P噪声以及标准差为0.5的高斯滤波,以此为例,使用误码率准则评价主动防御效果,如表1所示。

表1 主动防御方法效果对比

隐写 算法	主动防御 方法	负载率				
		0.01	0.02	0.03	0.04	0.05
DMAS	高斯滤波	0.4572	0.4556	0.4809	0.4897	0.4856
	泊松噪声	0.4573	0.4541	0.4679	0.4815	0.4875
	S&P噪声	<b>0.5041</b>	<b>0.5087</b>	<b>0.4948</b>	<b>0.4951</b>	<b>0.4999</b>
	高斯白噪声	0.4771	0.483	0.4802	0.4811	0.4826
GMAS	高斯滤波	0.4587	0.4891	0.4656	0.4815	0.4954
	泊松噪声	0.4867	0.4473	0.4804	0.4751	0.4819
	S&P噪声	<b>0.4979</b>	<b>0.4967</b>	<b>0.5074</b>	<b>0.5048</b>	<b>0.4972</b>
	高斯白噪声	0.4607	0.4708	0.4826	0.4867	0.4784

表1可知,在主动防御DMAS隐写时,加入高斯滤波、泊松噪声、S&P噪声与高斯白噪声得到图像误码率最小值分别为0.4556、0.4541、0.4948和0.4771,其中S&P噪声主动防御DMAS隐写得到图像误码率值最高。在主动防御GMAS隐写时,加入高斯滤波、泊松噪声、S&P噪声与高斯白噪声得到图像误码率最小值分别为0.4587、0.4473、0.4967和0.4607,其中S&P噪声主动防御GMAS

隐写得到图像误码率值最高。以上数据均表明S&P噪声能够成功破坏秘密信息,其主动防御效果远高于其他主动防御方法。

本文以主动防御DMAS隐写为例,在载密图像中加入不同噪声水平的S&P噪声(实验环境设置见5.1),主动防御效果如图3所示。

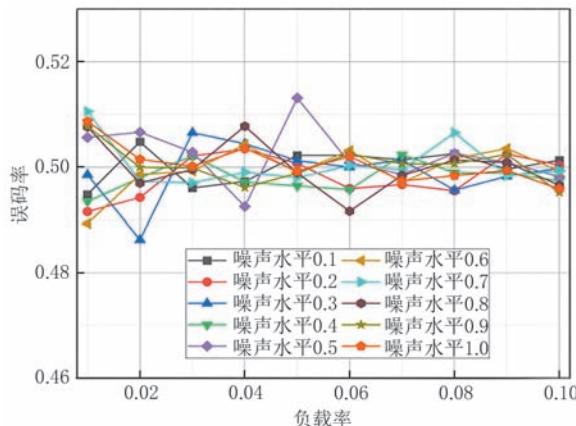


图3 不同S&P噪声水平的主动防御效果对比

由图3可知,横坐标表示不同负载率,纵坐标表示误码率,不同噪声水平使用不同颜色的折线表示,颜色由图例所示。以上折线数据表明S&P噪声主动防御DMAS隐写的误码率结果在0.485与0.515之间,即均在0.5处波动。以噪声水平0.5为例,当负载率为0.05时,载密图像的误码率最大为0.5131;当负载率为0.04时,载密图像的误码率最小为0.4925。由此可知不同负载率下载密图像的误码率均值为0.5021,秘密信息序列不能被恢复。因此,不同噪声水平的S&P噪声均能够对载密图像中秘密信息进行破坏,实现主动防御隐写的目的。

综合上述原因,在本阶段采用叠加S&P噪声进行主动防御,这里S&P主动防御方式定义如下:

$$I(m, n) = \begin{cases} 0 & r_1 < p \& r_2 < 0.5 \\ 255 & r_1 < p \& r_2 \geq 0.5 \\ I(m, n) & r_1 \geq p \end{cases} \quad (2)$$

其中,  $I(m, n)$  表示图像位于坐标  $(m, n)$  处的像素值,  $p \in (0, 1)$  为给定的噪声水平,  $r_1$  和  $r_2$  分别是在每个像素上生成的两个随机值,前者决定像素是否会被污染,后者控制像素是最大值还是最小值。

在此阶段,本文在“发布”图像中加入S&P噪声,实现对图像的无差别攻击,从而达到主动防御的目的。这一过程具体表述如下。

$$\begin{cases} C + M = S \\ C \cup S = \theta \\ \theta + I = \theta_n \end{cases} \quad (3)$$

其中,  $\theta$  和  $\theta_n$  分别表示“发布”图像和噪声图像,  $C$  和  $S$  分别表示载体图像和载密图像,  $M$  和  $I$  分别表示秘密信息和噪声信息。

### 3.3 Rec-Net图像恢复

在SPRN方法的S&P主动防御阶段中得到噪声图像,该图像中的秘密信息虽然已经遭受到攻击,但是由于S&P噪声中随机的椒噪点和盐噪点的叠加会对图像视觉效果层面造成一定的影响,针对上述问题,受到梁等人<sup>[27]</sup>的启发,本文设计一个端到端的图像恢复网络(Recovery Network, Rec-Net),用于维持社交网络中“干净”图像的质量效果。

#### 3.3.1 Rec-Net整体架构

噪声图像中存在的随机黑白点在像素层面会表现在图像像素峰点上的最大值和最小值,使得图像邻域像素点间的值相差过大,这势必会影响社交网络通信双方的正常传递,使得接收方能够轻易察觉到图像的变化。因此,本文设计Rec-Net,目的是使接收方接收到的“干净”图像既无秘密信息又无像素极值点(椒噪点和盐噪点),且更加接近“发布”图像,在不影响视觉观感的同时,存储空间大小的变化也在可接受范围。Rec-Net的整体架构如图4所示。

在Rec-Net中,输入为S&P主动防御阶段生成的噪声图像,输出为秘密信息被破坏掉的“干净”图像。Rec-Net首先从三层中值滤波层(Median Layers)开始,得到与噪声图像大小一致的中值滤波特征图,目的是去除该图像中的极值点;其次是一层高斯滤波层(Gauss Layer),得到由64个特征通道组成的特征图,目的是消除中值滤波层产生的窗口模糊化和块化;然后是16组高斯滤波残差块(Gauss Residual blocks)与中值滤波层(Median Layers, Gauss+BN, Gauss+BN+Prelu 和 Conv+BN+Relu)、16组高斯滤波残差块(Gauss+BN, Gauss+BN+Prelu 和 Conv+BN+Relu),目的是将去极值点与去窗口模糊化相结合,得到邻域像素差值更小的特征图。网络中最后一层为卷积层(Conv Layer),目的是生成三个通道RGB的图像。

Rec-Net能够实现对噪声图像的恢复,使得社交网络接收方能够接收没有秘密信息并且视觉质量较好的图像,本文将该网络反向传播的问题定义为公式(4)。

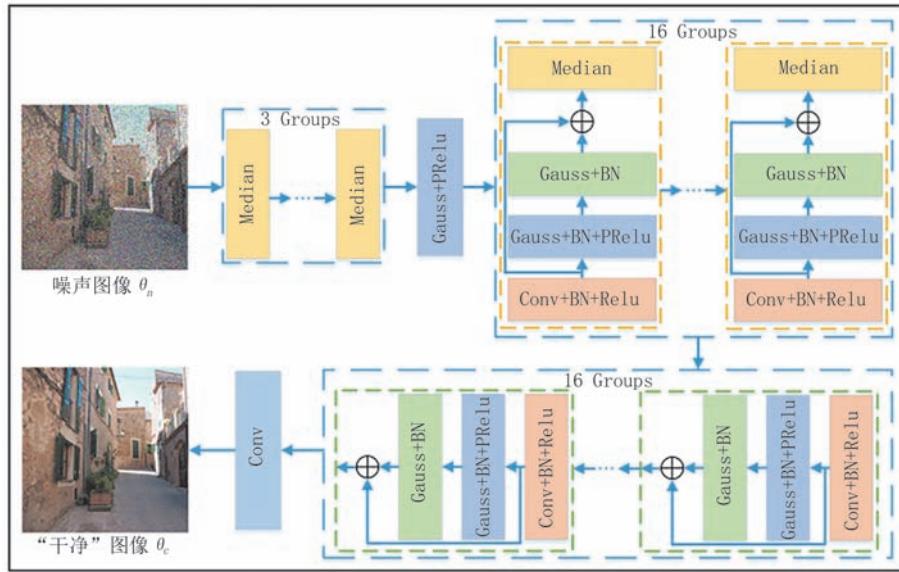


图4 Rec-Net的整体架构

$$Loss = \|\theta - \theta_c\|_2^2 \quad (4)$$

其中,  $\theta$  为“发布”图像,  $\theta_c$  为网络生成的“干净”图像。Rec-Net 在整个网络训练过程中都使用公式(4)作为损失函数, 能够使训练过程更加容易收敛。具体过程是通过公式(4)计算出  $\theta_c$  和  $\theta$  之间的差异值, 在反向传播过程中再通过公式(4)去更新 Rec-Net 中每层的各个参数, 以此用来降低  $\theta_c$  和  $\theta$  与之间的差异, 最后 Rec-Net 生成的  $\theta_c$  更加接近  $\theta$ , 从而达到恢复具有不同程度 S&P 噪声污染的潜在  $\theta_c$  的目的。下面将详细描述 Rec-Net 中的中值滤波层、高斯滤波层以及高斯滤波残差块。

### 3.3.2 基于中值滤波层的极值点优化

对于本文方法的第一阶段破坏秘密信息之后, 噪声图像中的 S&P 噪声点较为明显, S&P 噪声点在图像像素层面体现在极值点上, 椒噪点为最小极值点, 盐噪点为最大极值点。极值点在图像上的表现为黑色像素点和白色像素点, 影响视觉效果。而一般的卷积层主要用于提取图像特征, 不能够清除图像中的所有极值点, 由于存在上述问题, 因此, 需要考虑其他方法。

中值滤波作为一种传统的非线性滤波, 本质上是一种统计排序滤波器, 它用局部窗口的中值替换给定窗口中居中的像素, 能够在消除噪声的同时保持较好的图像细节。一个 2-D 中值滤波的输出可用公式(5)表示。

$$g_{median}(x, y) = \underset{(m, n) \in N(x, y)}{median} [\theta_n(m, n)] \quad (5)$$

其中,  $\theta_n(m, n)$  和  $g_{median}(x, y)$  分别是噪声图像和中

值滤波处理后的图像,  $N(x, y)$  为  $(x, y)$  的  $N$  邻域, 邻域越大图像平滑的效果越好, 但邻域过大, 平滑会使边缘和轮廓信息损失得越大, 从而使输出的图像变得模糊, 因此需合理选择邻域的大小, 通常为  $3 \times 3$  或  $5 \times 5$ 。受文献[27]启发, 在图像上重复应用中值滤波, 直到所有极值点被固定大小的局部窗口中的中值所取代, 使得噪声图像中的极值点被消除。因此, 本文使用中值滤波来代替卷积层, 称为中值滤波层。

Rec-Net 不是直接对图像进行中值滤波操作, 而是作为网络中的中值滤波层对图像中不同特征通道的椒盐噪声点进行过滤。网络使用中值滤波层的滤波核大小为  $3 \times 3$ 。通过这种方式, Rec-Net 基本上去除了不同特征图中的极值点, 然后结合中值滤波特征用于预测更好的“干净”图像。其中, 中值滤波层和传统卷积层相同, 都是以移动窗口的方式应用于特征通道的每个元素, 如图 5 所示。

图 5 中, 由 RGB 通道组成的 Lena 图像对应于 3 个特征通道, 分别是红绿蓝三个特征通道, 中值滤波层在该图像中应用三次。以 R 通道为例说明中值滤波层的应用过程, 在该通道中取  $3 \times 3$  的像素点矩

阵, 如  $\begin{bmatrix} 198 & 215 & 214 \\ 198 & 215 & 214 \\ 199 & 216 & 215 \end{bmatrix}$ , 找出该矩阵像素点中的中

值(214), 用该值替换中心像素点(像素值 215)的

值, 得到  $\begin{bmatrix} 198 & 215 & 214 \\ 198 & 214 & 214 \\ 199 & 216 & 215 \end{bmatrix}$ 。通过改变噪声图像中的

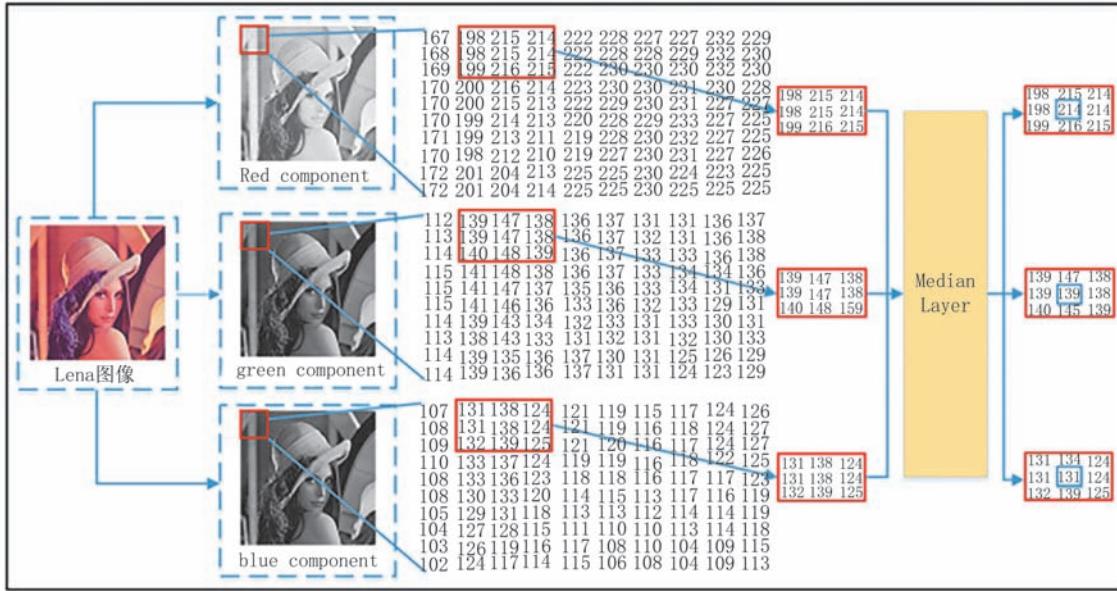


图5 中值滤波层的应用示例

局部窗口像素中心点的值,使得局部窗口中邻域像素点的值差异变小,从而达到消除最大极值点和最小极值点的目的。G和B通道采用相同的处理过程,消除每个通道的极值点,从而对图像实现中值滤波。

### 3.3.3 基于高斯滤波层与高斯滤波残差块的图像恢复

经过Rec-Net中值滤波层的噪声图像虽然能够去除噪声图像中90%以上的椒盐噪声,但是,中值滤波层会使得该图像中部分区域变得窗口模糊化和块化,未能有效保持噪声图像的边界信息与细节的处理。从而造成接收方收到的图像质量下降,并影响秘密信息破坏效果。而与中值滤波不同,高斯滤波是一种线性平滑滤波,是对整幅图像进行加权平均的过程。经过高斯滤波处理后的图像中每一个像素点的值,都由其本身和邻域内的其他像素值经过加权平均后得到,能够在过滤秘密信息的同时降低图像相邻局部窗口中边界像素点值的差别,并且能够使噪声图像中边缘细节的保持效果更好,从而达到减少窗口模糊化和块化的目的。因此,本文通过使用高斯滤波层破坏秘密信息,并改善图像中存在模糊化和块化的区域。高斯滤波层是使用一组可学习的卷积运算代替高斯滤波器(固定参数的平滑滤波器),在该层中使用64个高斯滤波器生成64个特征通道,其中高斯滤波器的滤波核大小固定为3×3用于Rec-Net网络的训练。

传统的线性整流单元(Rectified Linear Unit,

ReLU)激活函数见公式(6),它能增加网络特征提取能力,但会导致神经元坏死问题,使得网络训练过程中拟合能力下降。为了缓解这一困局,对ReLU激活函数进行改进,在该激活函数中加入变量,称为参数整流线性单元(Parametric Rectified Linear Unit, PReLU)激活函数,见公式(7)。PReLU激活函数区别于ReLU激活函数,如图6所示。

$$ReLU(x) = \begin{cases} x, & x > 0 \\ 0, & x \leq 0 \end{cases} \quad (6)$$

$$PReLU(x) = \begin{cases} x, & x > 0 \\ ax, & x \leq 0 \end{cases} \quad (7)$$

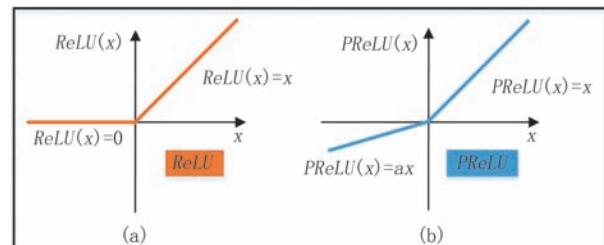


图6 ReLU与PReLU激活函数对比

这里,  $a$ 的值会随着网络训练的反向传播进行更新。PReLU能够在不增加计算量的前提下,改善Rec-Net过拟合问题,使得网络训练过程中收敛更快。进而一定程度上缓解神经元坏死的问题。于是,本文在高斯滤波层中选择PReLU作为激活函数,用来为Rec-Net提供非线性关系,提高该网络的表征和拟合能力。

由于Rec-Net网络的层数较多,单独训练会使

得模型的复杂度上升,也使得训练变得困难,进而导致网络退化。因此,本文在网络中加入残差学习的思想,使得网络中的数据可以实现跨层流动,避免因网络层数增加导致网络梯度消失问题。本文将嵌入高斯滤波层的残差块称为高斯滤波残差块。它由Gauss+BN、Gauss+BN+Prelu和Conv+BN+Relu组成。如图4所示,Rec-Net中带中值层的高斯滤波残差块与不带中值层的高斯滤波残差块分别训练16组,能更好地去除噪声图像中的极值点以及全面恢复该图像中的残缺信号,使得接收方能够得到视觉质量佳且提取不出秘密信息的“干净”图像。

## 4 隐写主动防御图像评价准则

为了在无任何隐写先验条件下更加快捷的描述载密图像中秘密信息的破坏效果,本文提出一种新的隐写主动防御图像评价准则——改变率(Change Rate, CR)。另外,为了验证所提准则的可信性,本文采用相关性分析来对其有效性进行阐明。

### 4.1 改变率

CR是“发布”图像与接收方接收到的“干净”图像的像素改变量占总像素数的比率。它是对图像整体变化程度的一种描述。CR的计算公式如下所示。

$$CR = \frac{1}{HW} \sum_{m,n} D(m,n) \quad (8)$$

$$D(m,n) = \begin{cases} 1, & \theta(m,n) \neq \theta_c(m,n) \\ 0, & \text{otherwise} \end{cases} \quad (9)$$

其中,  $H$ 与  $W$ 分别表示图像的高度与宽度,  $m$ 与  $n$ 分别表示图像像素点的横坐标和纵坐标,  $D(m,n)$ 表示衡量图像像素改变次数,若改变的次数越多,则破坏秘密信息越多。同时,CR越大,隐写主动防御效果越好。

### 4.2 相关性分析

为了描述CR与BER对评价结果的一致性,通常通过两事物之间联系紧密程度关系进行相关性分析和评估,因此,本文采用协方差Cov和皮尔森相关系数 $\rho$ 对两者进行相关性度量,并衡量它们之间的偏差关联程度。Cov和 $\rho$ 的计算公式如下所示。

$$Cov(X, Y) = E(XY) - E(X) \cdot E(Y) \quad (10)$$

$$\rho_{XY} = \frac{Cov(X, Y)}{\sqrt{D(X)} \sqrt{D(Y)}} \quad (11)$$

其中,变量  $X$  和  $Y$  分别表示为  $X = [x_1, x_2, x_3, \dots, x_n]$

和  $Y = [y_1, y_2, y_3, \dots, y_n]$ ,  $E(X)$  为  $X$  的期望,  $D(X)$  为  $X$  的方差,  $\bar{X}$  表示  $X$  的平均值,  $P$  为概率。公式(11)中  $\rho_{XY}$  是描述变量  $X$  和  $Y$  之间的线性相依性和联动性强弱的一种度量。 $\rho \in [-1, 1]$ ,  $\rho$  越接近 1, 表明  $X$  和  $Y$  之间正相关越强, 越接近 -1, 表明  $X$  和  $Y$  之间负相关越强。根据 4.1 节中的实验设置和过程, 本节数据来源随机选择 30 幅载密图像(质量因子为 95), 并在载密图像中加入噪声水平 0.4(即  $\rho=0.4$ ) 的 S&P 噪声得到相应的噪声图像, 通过 Rec-Net 得到“干净”图像, 以 30 幅“干净”图像为例, 得到一组实验数据(包括 30 幅“干净”图像的 CR 和 BER), 如图 7 所示。

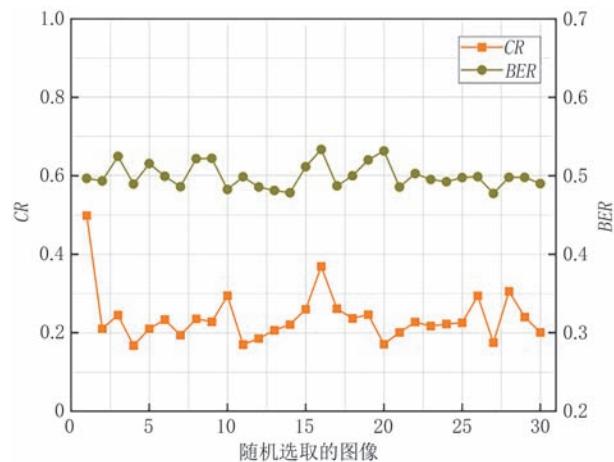


图 7 CR 与 BER 的结果对比

在图 7 中, 横坐标表示随机选择的图像, 左纵轴表示 CR, 右纵轴表示 BER, 橙色正方形点表示 CR 值, 绿色圆形点表示 BER 值, 橙色线条和绿色线条分别表示 CR 值和 BER 值的折线。

根据图 7 中的数据点对 CR 和 BER 的 Cov 进行计算, 得到它们的相关性度量结果, 过程如下:

(1) 计算 CR 与 BER 的方差, 分别为  $D(CR) = 4.368 \times 10^{-3}$ ,  $D(BER) = 2.4983 \times 10^{-4}$ 。

(2) 由公式(10)得到 CR 与 BER 的协方差  $Cov(CR, BER)$ , 结果为  $Cov(CR, BER) = 6.4 \times 10^{-4}$ , 可得该结果为正, 由此可知 CR 与 BER 之间的协同关系是同向变化的, 即为正相关。

(3) 根据  $D(CR)$ 、 $D(BER)$  和  $Cov(CR, BER)$  计算得到  $\rho = 0.6132$ 。

根据上述结果, 能够看出 CR 和 BER 数据点值之间具有较高的相似度。为了更好地对 CR 和 BER 进行相关性分析, 本节从函数角度说明 CR 与 BER 的关联程度, 对图 7 中 30 张图像的 CR 值与 BER 值

使用Allometric1函数进行非线性拟合,其中CR与BER拟合曲线参数值如表2所示。

表2 CR与BER拟合曲线参数值

拟合参数 Equation	Allometric1函数拟合参数值 $y=ax^b$		
	Plot	CR	BER
$a$	0.3151±	0.504 69±	
	0.040 46	0.009 21	
$b$	-0.114 31±	-0.0038±	
	0.051 55	0.006 96	
Reduced Chi-Sqr	0.004	2.560 24E-4	
Residual Sum of Squares	0.1119	0.007 17	
R-Square(COD)	0.116 63	0.010 56	

表2中,Allometric1函数的原型为 $y=ax^b$ , $a$ 和 $b$ 分别是函数的两个系数;Reduced Chi-Sqr为标准误差;Residual Sum of Squares是残差平方和,用于度量二者曲线的差异程度;R-Square(COD)表示回归拟合效果。其中,根据CR与BER的拟合曲线函数能够得到CR趋近BER的0.5倍,故CR的基准值为0.05。根据表2中的参数值,生成CR与BER的拟合曲线图,如图8所示。

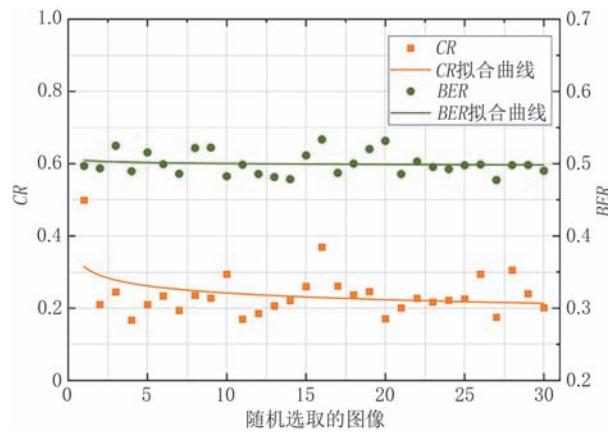


图8 CR与BER的拟合曲线趋势图

如图8所示,横坐标表示随机选取的图像,左纵轴表示CR,右纵轴表示BER,橙色正方形数据点和绿色圆形数据点分别代表CR值和BER值,橙色曲线和绿色曲线分别代表CR拟合曲线和BER拟合曲线。由图8可知,CR拟合曲线和BER拟合曲线走向趋势基本一致,可得二者之间的相关程度较高。结合图7中折线走向也看出,在误差允许范围内,CR与BER的折线走向趋势大致相同,意味着两者结果一致。这也说明本文提出的隐写主动防御图像评价准则CR与BER的评价结果具有一致性。

综合上述可知,CR与BER具有相同的评价能力,CR能够在未知隐写算法、负载率与完整秘密信息序列下发挥优势,且无需提取秘密信息。因此,图像隐蔽通信主动防御中能够使用CR对秘密信息破坏效果进行评价。

#### 4.3 基于CR的秘密信息破坏度量

为了更加详细地描述隐写主动防御图像度量过程,本节给出一个基于CR的秘密信息破坏度量算法,具体步骤如算法1所示。

##### 算法1. 基于CR的秘密信息破坏度量算法

输入:发送方发送的“发布”图像 $\theta$ 与接收方接收到的“干净”图像 $\theta_c$   
输出: $\theta$ 和 $\theta_c$ 的改变率CR值

```

1: For m=1: H Do /*H为图像的高度/
2:   For n=1: W Do /*W为图像的宽度/
3:     If  $\theta(m, n) \neq \theta_c(m, n)$  Then
4:       D(m, n)=1
5:     Else D(m, n)=0
6:   End If
7: End For
8: End For
9: For m=1: H Do
10:  For n=1: W Do
11:    While D(m, n)=1 Do
12:      Sum=Sum+1 /*Sum为累加变量/
13:    End While
14:  End For
15: End For
16: Return Sum/HW

```

利用算法1可以得到CR评价准则的计算结果,为后续在实验过程中充分利用CR进行度量提供了理论依据。此外,本节对计算CR与BER结果所需要的先验知识进行对比,如表3所示。

表3 CR与BER拟合曲线参数值

先验知识	CR	BER
载密图像所对应的隐写算法	不需要	需要
载密图像的负载率	不需要	需要
载密图像的质量因子	不需要	需要
原始嵌入秘密信息长度	不需要	需要
原始嵌入秘密信息完整序列	不需要	需要

由表3可知,CR皆不需要上述的先验知识即可计算结果,而BER需要上述全部的先验知识才能得到结果。因此,在社交网络等实际应用场景中本文提出的CR隐写防御评价准则具有较大优势。

BER同CR类似,都是对秘密信息破坏程度进行

度量,但是二者计算过程中时间开销有所区别。由于BER在实际场景中无法计算,为了便于与CR对比,本文以具体的鲁棒隐写DMAS、固定隐写负载率

为0.1、质量因子95的载密图像、载密图像分辨率 $m \times n$ 以及 $s$ 个秘密信息序列为为例,对CR与BER的计算过程和时间复杂度进行对比分析,如表4所示。

表4 CR与BER计算过程和时间复杂度对比

计算过程	CR		BER	
	计算过程	时间复杂度	计算过程	时间复杂度
获得“发布”图像	√	--	√	--
获得“干净”图像	√	--	√	--
计算“干净”图像的宽度和高度	√	$O(1)$	×	--
比较“发布”图像和“干净”图像像素点值	√	$O(1)$	×	--
提取“发布”图像中秘密信息的比特位值	×	--	√	$O(mn)$
计算“发布”图像中嵌入秘密信息的总长度	×	--	√	$O(1)$
提取“干净”图像中秘密信息的比特位值	×	--	√	$O(mn)$
遍历信息比特位值	×	--	√	$O(s)$
计算“干净”图像中秘密信息错误比特位值的总长度	×	--	√	$O(1)$

表4中“计算过程”表示CR与BER需要的计算过程,“时间复杂度”表示计算CR与BER值的时间复杂度。“√”表示需要该计算过程,“×”表示不需要该计算过程,“--”表示不存在时间复杂度。这里获得主动防御方法的“发布”图像和“干净”图像,无论使用哪种评价准则,都是必须要有的步骤。计算“干净”图像的宽度和高度是图像属性值的计算,时间复杂度为 $O(1)$ 。从表4中能够看出CR需要的时间复杂度仅为 $O(1)$ 。而BER需要的时间复杂度为 $O(mn + s)$ 。显而易见 $O(CR) \ll O(BER)$ ,因此CR需要的计算过程较少,时间复杂度较低。

以图7中选取的30张载密图像为例,计算此30张载密图像CR值与BER值的平均运行时间,如表5所示。

表5 CR与BER平均计算时间对比

评价方法	平均计算时间/s
CR	0.014
BER	4.535

从表5中能看出CR时间开销中需要花费的时间较少,仅在 $10^{-2}$ 数量级;而BER的平均计算时间达到4.545 s。因此,CR评价方法的计算时间远远低于BER的计算时间。

综上可知,在CR与BER对图像隐蔽通信主动防御的评价结果具有一致性的基础上,CR的时间复杂度与平均计算时间远低于BER。并且与BER不同,CR的计算过程不必获取隐写算法、负载率和秘密信息序列等先验知识,因此,在社交网络中使用CR对秘密信息破坏程度度量更具有实用性。

## 5 实验结果与分析

为了分析SPRN方法的有效性,本文从实验设置和评价准则、秘密信息破坏程度、图像质量效果、消融实验与对比实验四个方面进行一系列实验验证及相应说明。

### 5.1 实验设置和评价准则

本节包括两部分:实验设置和评价准则。下面分别对这两部分进行详细的说明。

#### 5.1.1 实验设置

实验过程中所使用的处理器型号为11th Gen Intel(R) Core(TM) i5-11300H 3.11 GHz。为了更好验证实验效果的准确性,本文在BOSSbase 1.01和BOWS2数据集中各取100张图像进行验证,将“发布”图像转换成质量因子为65、75、85和95的频域图像,并选取六种算法进行隐写(Payloads=0.01~0.1,步长0.01),如表6所示。为了便于表达实验结果,本文在图表中将BOSSbase 1.01数据集简称B1,BOWS2数据集简称B2,鲁棒隐写算法DMAS和GMAS分别简写D和G,普通隐写算法J-UNIWARD、nsF5、S-UNIWARD和WOW分别简写成J、F、S和W。

表6 六种隐写算法

鲁棒隐写算法	普通隐写算法	
	空域	频域
DMAS <sup>[2]</sup> (D)	S-UNIWARD <sup>[3]</sup> (S)	J-UNIWARD <sup>[3]</sup> (J)
GMAS <sup>[4]</sup> (G)	WOW <sup>[5]</sup> (W)	nsF5 <sup>[6]</sup> (F)

在网络训练阶段,为了加快网络的训练速度,本文在BOSSbase 1.01中任意选择100张图像进行预处理。首先,对每张图像使用三次样条插值方法缩放成大小为 $256 \times 256$ 的预处理图像;其次,对每张预处理图像按照步长为25划分成9个 $64 \times 64$ 的小块;最后,对每个小块按照0.1~0.9的噪声级别生成9张噪声图像作为网络的训练集。Rec-Net网络的参数设置如表7所示。

表7 Rec-Net参数设置

参数名称	数值
图像数量	8100
图像大小	$64 \times 64$
训练集比例	0.8
验证集比例	0.2
S&P噪声级别	0.1~0.9(步长0.1)
学习率Learning rate	0.001
批量大小Batch_size	32

### 5.1.2 评价准则

评价准则主要包括两个方面,一方面是秘密信息的破坏效果评价准则,本文采用BER与所提CR进行衡量。BER见公式(17)和(18)。第二方面是秘密信息破坏后“干净”图像的质量恢复效果评价准则,本文采用均方误差(Mean Square Error, MSE)、峰值信噪比(Peak Signal to Noise Ratio, PSNR)和结构相似度(Structural Similarity Index Measure, SSIM)来进行度量,公式如下所示。

$$BER = \frac{M_e}{M_t} = \frac{\sum_i M(i)}{M_t} \quad (12)$$

其中, $M_t$ 表示嵌入秘密信息的总长度, $M_e$ 表示秘密信息的序列。BER的结果越大,表示清除秘密信息越多,隐写破坏的效果越好。文献[8]中描述当 $BER > 0.1$ 时,秘密信息序列不能被恢复。因此, $BER > 0.1$ 时,说明图像中的秘密信息被破坏。

$$MSE = \frac{1}{HW} \sum_{m=0}^{H-1} \sum_{n=0}^{W-1} \left[ \theta(m, n) - \theta_c(m, n) \right]^2 \quad (13)$$

$$PSNR = 20 \log_{10} \left( \frac{MAX_\theta}{\sqrt{MSE}} \right) \quad (14)$$

$$SSIM = \frac{(2\mu_\theta\mu_{\theta_c} + k_1^2 L^2)(2\sigma_{\theta\theta_c} + k_2^2 L^2)}{(\mu_\theta^2 + \mu_{\theta_c}^2 + k_1^2 L^2)(\sigma_\theta^2 + \sigma_{\theta_c}^2 + k_2^2 L^2)} \quad (15)$$

其中, $\mu_\theta$ 和 $\mu_{\theta_c}$ 分别是 $\theta$ 和 $\theta_c$ 的平均值, $\sigma_\theta^2$ 和 $\sigma_{\theta_c}^2$ 分别是 $\theta$ 和 $\theta_c$ 的方差, $\sigma_{\theta\theta_c}$ 是 $\theta$ 和 $\theta_c$ 的协方差, $L$ 为图像像

素值的范围, $k_1$ 和 $k_2$ 是常数,分别取0.01和0.03。 $MSE$ 越小,代表该图像相互映射的误差越小,图像间的相似程度越高,也就是该图像的质量越好。相反地,PSNR与SSIM越大,表示该图像的质量越好。

## 5.2 秘密信息破坏实验

本节针对鲁棒隐写和普通隐写进行实验,验证SPRN方法对图像隐蔽通信主动防御的有效性,包括两部分:主动防御鲁棒隐写和主动防御普通隐写。

### 5.2.1 主动防御鲁棒隐写

根据5.1节中的实验设置,本文方法对两种鲁棒隐写算法DMAS和GMAS进行主动防御,破坏其嵌入的秘密信息。在不同负载率下,SPRN方法生成50张“干净”图像的BER和CR取均值作为最终的实验结果,如表8~表9所示。

由表8可见,在B1数据集与不同负载率下,本方法破坏秘密信息后,“干净”图像的BER和CR均值为0.4849和0.3014,分别是基准值(秘密信息序列不能被恢复时 $BER=0.1$ , $CR=0.05$ )的4.849倍和6.028倍,表明本方法成功破坏秘密信息,实现主动防御。在B1数据集与不同质量因子下,随着负载率的减小,“干净”图像的误码率不断增加,即主动防御能力不断增强,也验证本方法在低负载率下的有效性。在B2数据集与不同负载率下,在质量因子95、负载率为0.06 bpnac时,“干净”图像的BER值最大为0.5025。并且“干净”图像的BER和CR均值为0.4814和0.2868,分别是基准值的4.814倍和5.736倍。以上数据表明,在不同数据集与不同质量因子下,本方法都能够达到破坏秘密信息的目的,成功实现主动防御DMAS鲁棒隐写。

由表9可见,在B1数据集和不同负载率下,以质量因子85为例,本方法破坏秘密信息后,“干净”图像的BER和CR最大值分别是0.5050和0.2717。本方法破坏秘密信息后,在不同质量因子(65、75、85和95)下,“干净”图像的BER均值分别是0.4873、0.4900、0.5011和0.4905,CR均值分别为0.2774、0.2671、0.2710和0.2226。以上数据表明在不同质量因子下,本方法均能够成功破坏秘密信息。在B2数据集和不同质量因子下,本方法破坏秘密信息后,“干净”图像的BER均值分别为0.4924、0.4824、0.4818和0.5003,CR均值分别为0.3066、0.3122、0.2993和0.2023。结果表明本方法能够成功防御GMAS鲁棒隐写。在B2数据集与不同质量因子下,随着负载率的减小,“干净”图像的误码率不断增加,即主动防御能力不断增强。以上数据表明,在不

表8 主动防御DMAS得到“干净”图像的BER和CR值

数据集	质量因子	指标	Payloads									
			0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.10
B1	65	BER	0.4957	0.4952	0.4857	0.4828	0.4761	0.4755	0.4770	0.4729	0.4730	0.4726
		CR	0.3143	0.3148	0.3152	0.3156	0.3165	0.3171	0.3177	0.3185	0.3193	0.3203
	75	BER	0.4956	0.4909	0.4837	0.4782	0.4725	0.4698	0.4659	0.4654	0.4618	0.4606
		CR	0.2990	0.2994	0.2998	0.3003	0.3006	0.3011	0.3017	0.3024	0.3030	0.3036
	85	BER	0.5014	0.4990	0.4956	0.4930	0.4913	0.4909	0.4893	0.4901	0.4887	0.4861
		CR	0.3034	0.3036	0.3037	0.3039	0.3040	0.3044	0.3046	0.3049	0.3052	0.3055
	95	BER	0.4959	0.4958	0.4940	0.4942	0.4937	0.4904	0.4891	0.4892	0.4876	0.4874
		CR	0.2825	0.2825	0.2826	0.2827	0.2831	0.2834	0.2835	0.2837	0.2843	0.2846
B2	65	BER	0.4970	0.4970	0.4976	0.4989	0.4976	0.4990	0.4997	0.4998	0.4996	0.5013
		CR	0.2596	0.2598	0.2602	0.2610	0.2613	0.2620	0.2627	0.2633	0.2639	0.2645
	75	BER	0.4814	0.4823	0.472	0.4647	0.4578	0.4558	0.4527	0.4513	0.4449	0.4443
		CR	0.2517	0.2518	0.2522	0.2528	0.2531	0.2536	0.2541	0.2544	0.2551	0.2555
	85	BER	0.4875	0.4771	0.4701	0.4698	0.4629	0.4582	0.4573	0.4599	0.4563	0.4581
		CR	0.2857	0.2859	0.2861	0.2862	0.2866	0.2869	0.2872	0.2876	0.2878	0.2882
	95	BER	0.4962	0.4995	0.4999	0.4976	0.5018	0.5025	0.5017	0.5024	0.5020	0.5018
		CR	0.2770	0.2775	0.2776	0.2779	0.2781	0.2787	0.2790	0.2795	0.2800	0.2803

表9 主动防御GMAS得到“干净”图像的BER和CR值

数据集	质量因子	指标	Payloads									
			0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.10
B1	65	BER	0.5025	0.4885	0.4919	0.4842	0.4858	0.4804	0.4878	0.4878	0.4810	0.4831
		CR	0.2766	0.2768	0.2767	0.2770	0.2772	0.2773	0.2778	0.2776	0.2783	0.2785
	75	BER	0.4937	0.4938	0.4895	0.4905	0.4952	0.4913	0.4896	0.4854	0.4868	0.4843
		CR	0.2664	0.2667	0.2667	0.2669	0.2669	0.2673	0.2672	0.2676	0.2676	0.2681
	85	BER	0.5015	0.5050	0.5035	0.4962	0.5015	0.5009	0.5012	0.5004	0.5011	0.4996
		CR	0.2706	0.2705	0.2705	0.2708	0.2707	0.2710	0.2712	0.2713	0.2714	0.2717
	95	BER	0.4889	0.4917	0.4900	0.4947	0.4936	0.4921	0.4923	0.4883	0.4884	0.4850
		CR	0.2219	0.2222	0.2221	0.2224	0.2226	0.2225	0.2228	0.2228	0.2233	0.2237
B2	65	BER	0.4981	0.4981	0.4963	0.4952	0.4935	0.4918	0.4899	0.4887	0.4867	0.4859
		CR	0.3059	0.3062	0.3060	0.3061	0.3064	0.3065	0.3068	0.3071	0.3074	0.3076
	75	BER	0.4957	0.4978	0.4972	0.4936	0.4897	0.4817	0.4771	0.4716	0.4607	0.4592
		CR	0.3116	0.3117	0.3116	0.3117	0.3120	0.3121	0.3124	0.3126	0.3127	0.3131
	85	BER	0.4942	0.4967	0.4932	0.4905	0.4886	0.4810	0.4756	0.4724	0.4668	0.4588
		CR	0.2988	0.2989	0.2987	0.2991	0.2993	0.2992	0.2996	0.2997	0.2998	0.3000
	95	BER	0.5038	0.5002	0.4989	0.5001	0.5001	0.4999	0.5008	0.4993	0.5005	0.4996
		CR	0.2017	0.2018	0.2018	0.2022	0.2022	0.2021	0.2025	0.2028	0.2031	

同数据集、负载率以及质量因子下,本方法破坏秘密信息后,“干净”图像具有不同的BER和CR值,且该结果远远超过秘密信息的不可恢复值。因此,本方法能够在社交网络中实现防御GMAS鲁棒隐写的目的。

### 5.2.2 主动防御普通隐写

本文方法对四种普通隐写算法J-UNIWARD、nsF5、S-UNIWARD和WOW进行主动防御,破坏其嵌入的秘密信息。本文在使用频域隐写算法时,

对“发布”图像采用4.1节的实验设置中的质量因子进行压缩,而对空域隐写算法的“发布”图像不进行压缩。本文对破坏J-UNIWARD隐写算法采用BER与CR两种评价指标进行评价,而其余普通隐写算法则采用CR评价指标进行评价。实验分别对不同负载率下SPRN方法生成的50张“干净”图像取平均值作为最终的结果,如表10~表11所示。

由表10可知,在B1数据集上,质量因子为95、负载率为0.03bpnac时,本文方法破坏秘密信息后,

表10 主动防御J-UNIWARD得到“干净”图像的BER值

数据集	质量因子	Payloads									
		0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.10
B1	65	0.5040	0.5001	0.5007	0.4981	0.4974	0.4990	0.5023	0.5009	0.4989	0.5011
	75	0.4988	0.4952	0.4990	0.5004	0.4982	0.4960	0.4997	0.5009	0.5019	0.5026
	85	0.5012	0.5019	0.5016	0.4996	0.4988	0.4993	0.5004	0.4994	0.5018	0.5017
	95	0.5008	0.5026	0.5052	0.4996	0.4994	0.4957	0.4995	0.4994	0.4994	0.5008
B2	65	0.5032	0.4961	0.4982	0.4984	0.5000	0.4979	0.5006	0.5017	0.5002	0.4998
	75	0.4994	0.4995	0.4996	0.5046	0.4985	0.4993	0.4981	0.5000	0.5007	0.4983
	85	0.5008	0.4991	0.5002	0.5019	0.5010	0.5000	0.5014	0.5014	0.5008	0.4995
	95	0.5000	0.4991	0.4989	0.5011	0.4994	0.4993	0.5006	0.5001	0.5016	0.4995

表11 主动防御J-UNIWARD、nsF5、S-UNIWARD和WOW得到“干净”图像的CR值

数据集	算法	质量因子	Payloads									
			0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.10
B1	J	65	0.2973	0.2972	0.2973	0.2973	0.2971	0.2974	0.2972	0.2974	0.2972	0.2973
		75	0.2865	0.2864	0.2864	0.2864	0.2864	0.2863	0.2863	0.2863	0.2866	0.2863
		85	0.2908	0.2906	0.2906	0.2907	0.2907	0.2906	0.2905	0.2906	0.2906	0.2906
		95	0.2863	0.2861	0.2860	0.2862	0.2864	0.3036	0.3036	0.3037	0.3037	0.2862
	F	65	0.2972	0.2972	0.2967	0.2969	0.2968	0.2968	0.2969	0.2967	0.2964	0.2965
		75	0.2862	0.2863	0.2862	0.2861	0.2863	0.2862	0.2859	0.2860	0.2859	0.2857
		85	0.2906	0.2906	0.2905	0.2905	0.2906	0.2905	0.2905	0.2905	0.2905	0.2907
		95	0.2863	0.2861	0.2861	0.2861	0.2858	0.2857	0.2859	0.2858	0.2855	0.2856
	S	0	0.2914	0.2912	0.2911	0.2911	0.2911	0.2910	0.2913	0.2911	0.2912	0.2911
		W	0	0.2912	0.2913	0.2911	0.2912	0.2911	0.2909	0.2912	0.2912	0.2912
	J	65	0.3059	0.3059	0.3059	0.3060	0.3058	0.3059	0.3060	0.3060	0.3059	0.3060
		75	0.2936	0.2935	0.2934	0.2935	0.2934	0.2934	0.2935	0.2936	0.2936	0.2935
		85	0.2987	0.2987	0.2986	0.2988	0.2989	0.2988	0.2989	0.2986	0.2986	0.2988
		95	0.2267	0.2265	0.2262	0.2266	0.2264	0.2268	0.2266	0.2263	0.2265	0.2266
	B2	65	0.3062	0.3059	0.3057	0.3058	0.3056	0.3056	0.3055	0.3052	0.3052	0.3052
		75	0.2935	0.2932	0.2932	0.2934	0.2933	0.2932	0.2931	0.2928	0.2932	0.2928
		85	0.2986	0.2988	0.2986	0.2985	0.2986	0.2984	0.2985	0.2986	0.2985	0.2984
		95	0.2247	0.2248	0.2247	0.2248	0.2249	0.2246	0.2249	0.2246	0.2248	0.2246
	S	0	0.3039	0.3040	0.3039	0.3038	0.3040	0.3039	0.3041	0.3038	0.3038	0.3041
		W	0	0.3040	0.3039	0.3039	0.3038	0.3039	0.3040	0.3040	0.3038	0.3039

“干净”图像的BER取得最大值为0.5052;质量因子为75、负载率为0.02bpnac时,“干净”图像的BER取得最小值为0.4952。表明本文方法破坏秘密信息后,“干净”图像的BER值均在0.4952~0.5052之间,由于 $BER > 0.1$ 时,说明图像中的秘密信息被破坏。在B2数据集不同质量因子上,本文方法破坏秘密信息后,“干净”图像的BER最小值分别为0.4961、0.4981、0.4991和0.4989,表明“干净”图像的BER值不低于0.49,说明图像中的秘密信息百分百被破坏。由此可知,SPRN方法能够成功破坏J-UNIWARD嵌入的秘密信息,使得接收方无法完整提取秘密信息序列。故本方法能够在社交网络中达

到主动防御J-UNIWARD隐写的目的。

表11可见,在B1和B2数据集中,SPRN方法破坏J-UNIWARD嵌入的秘密信息后,“干净”图像的CR最大值分别为0.3037和0.3060,最小值分别为0.2860和0.2262,其中最小值分别是基准值的5.72倍和4.524倍;在不同数据集下,本方法破坏隐写算法nsF5嵌入的秘密信息后,“干净”图像的CR最大值分别为0.2972和0.3062,最小值分别为0.2855和0.2246,其中最小值分别是基准值的5.71倍和4.492倍;当负载率为0.07 bpna时,以B2数据集为例,本方法破坏隐写算法S-UNIWARD和WOW嵌入的秘密信息后,“干净”图像的CR值最大为0.3041

和0.3040,分别是基准值的6.082倍和6.08倍。由此可知,在不同负载率下,本文方法能够主动防御多种普通隐写。

综上,对于主动防御鲁棒隐写和普通隐写而言,本文方法破坏秘密信息后,“干净”图像的BER与CR的结果远远高于基准值0.1和0.05(秘密信息序列不能被恢复时BER=0.1,CR=0.05)。并且在不同数据集与负载率下,二者的值均趋于稳定。验证SPRN方法能够有效地破坏社交网络中潜在载密图像中的秘密信息,并无需考虑隐写者使用的具体隐写算法、质量因子以及负载率等因素,实现社交网络中隐蔽通信主动防御的目的。

### 5.3 图像质量恢复实验

在SPRN方法第一阶段使用S&P噪声破坏载密图像中隐藏的秘密信息,将会导致图像的质量发生一定变化,本节针对鲁棒隐写和普通隐写进行一系列实验,验证本文方法对图像质量恢复的有效性,包括三部分:主动防御鲁棒隐写、主动防御普通隐写以及实验结果展示。图像质量恢复实验的过程与秘密信息破坏实验类似,本实验采用PSNR、SSIM以及MSE对“干净”图像进行评价。

#### 5.3.1 主动防御鲁棒隐写

与4.2节中的实验类似,在使用相同数据集下,本文方法主动防御鲁棒隐写DMAS和GMAS,将Rec-Net生成“干净”图像的PSNR、SSIM以及MSE的平均值作为实验结果,如图9~11所示。

图9(a)和(b)中横轴表示DMAS和GMAS的不同负载率,纵轴表示PSNR值,数据用不同颜色的折线显示。这里,具体的算法和质量因子如图例所示。例如黑色折线表示在质量因子为65中,本方法主动防御DMAS后,得到“干净”图像的PSNR。由图9(a)可知,在不同质量因子与不同负载率下,本方法主动防御DMAS后,“干净”图像的PSNR值介于36和39之间。本方法主动防御GMAS后,“干净”图像的PSNR值介于41和43.5之间。因此,在B1数据集中不同质量因子下,实验结果表明本方法对图像质量恢复的有效性。由图9(b)可知,在不同质量因子与不同负载率下,本方法主动防御DMAS后,在质量因子为75时,“干净”图像的PSNR值达到45.5,在质量因子为95时,“干净”图像的PSNR值为39.5。故“干净”图像的PSNR值分布虽然较为分散,但都达到社交网络传输图像的质量要求。本方法主动防御GMAS后,“干净”图像的PSNR值介于41和45之间,因此,由上述可知在B2数据集中

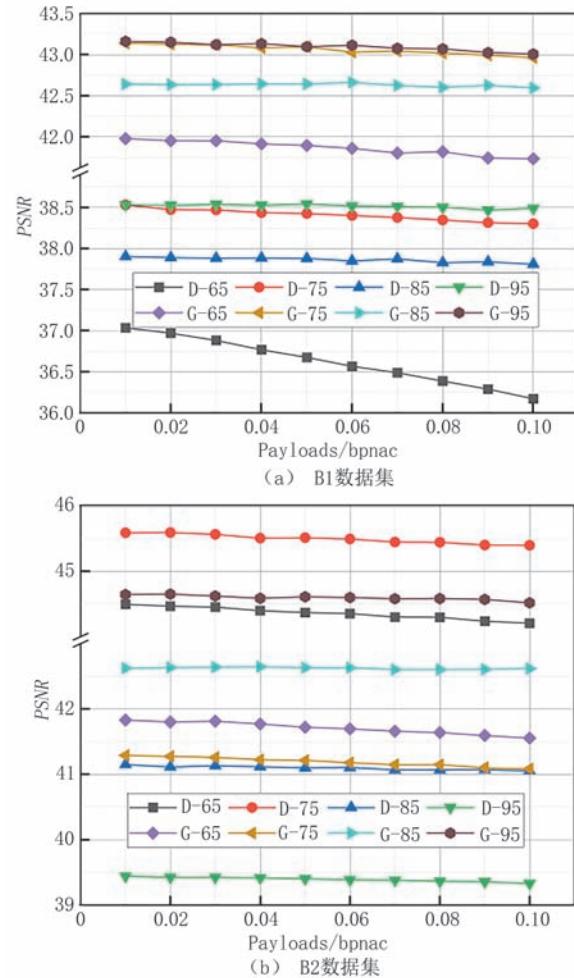


图9 主动防御DMAS和GMAS得到“干净”图像的PSNR

不同质量因子下,本文方法都能够恢复噪声图像的质量。

图10(a)和(b)中横轴代表DMAS和GMAS的不同负载率,纵轴表示SSIM值,不同质量因子的SSIM值用不同颜色的折线显示。这里,具体的算法和质量因子如图例所示。由图10(a)可知,在不同质量因子和不同负载率下,本方法主动防御DMAS后,“干净”图像的SSIM值均介于0.93和0.96之间。本方法主动防御GMAS后,“干净”图像的SSIM值在0.975波动。在B1数据集以及不同质量因子下,以上数据表明本文方法能够恢复噪声图像质量。由图10(b)可知,在不同质量因子和不同负载率下,本方法主动防御DMAS后,在质量因子为75时,“干净”图像的SSIM值达到0.9915,在质量因子为95时,“干净”图像的SSIM值为0.976。“干净”图像的SSIM值都符合社交网络传输图像的质量要求。本方法主动防御GMAS后,“干净”图像的SSIM值0.99处波动。因此,在B2数据集中,以

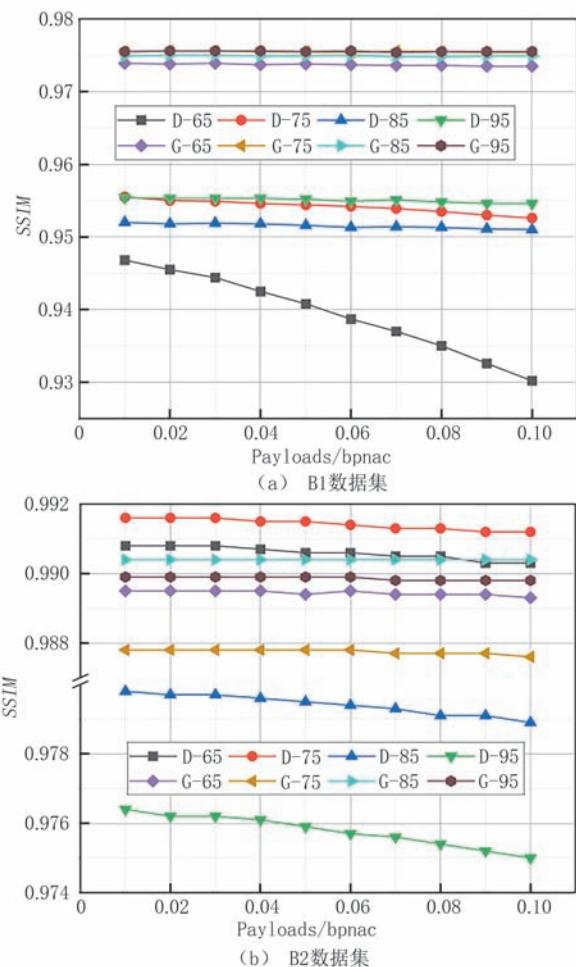


图10 主动防御DMAS和GMAS得到“干净”图像的SSIM

上数据均表明本文方法恢复噪声图像质量的有效性。

图11(a)和(b)中横轴代表DMAS和GMAS的不同负载率,纵轴表示MSE值,不同质量因子的MSE值用不同颜色的折线显示。这里,具体的算法和质量因子如图例所示。由图11(a)可知,在不同质量因子与不同负载率下,本方法主动防御DMAS后,“干净”图像的MSE值均介于14和24.5之间。本方法主动防御GMAS后,“干净”图像的MSE值在6以下。以上表明在B1数据集中不同质量因子下,本文方法都能够恢复噪声图像的质量。由图11(b)可知,在不同负载率下,本方法主动防御DMAS后,质量因子95时,“干净”图像的MSE值在9.8和10.5之间,质量因子为75时,“干净”图像的MSE值在2.1以下。本方法主动防御GMAS后“干净”图像的MSE值在2.8和7之间。以上数据表明在B2数据集中,SPRN能够恢复噪声图像的质量。

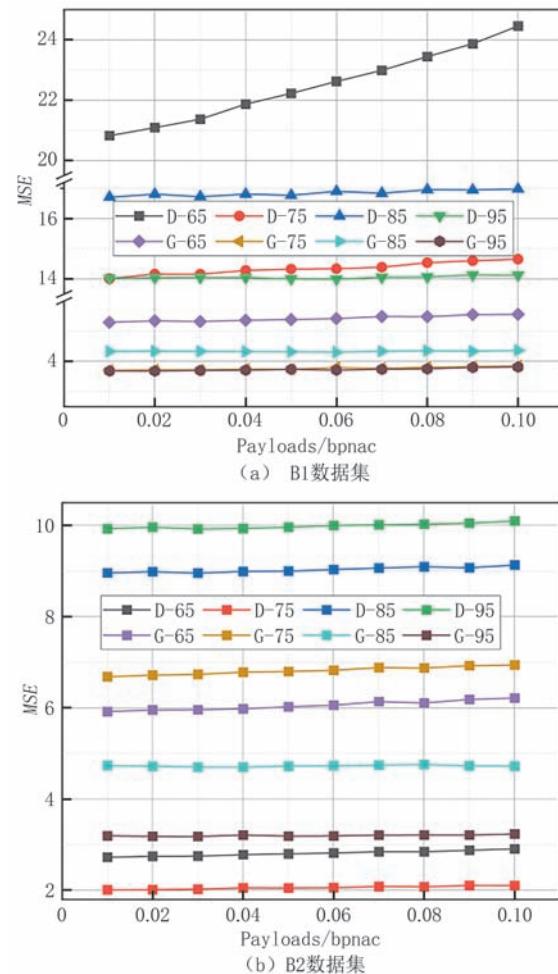


图11 主动防御DMAS和GMAS得到“干净”图像的MSE

### 5.3.2 主动防御普通隐写

与4.2节中的实验类似,在相同数据集下,本文方法主动防御普通隐写,将Rec-Net生成“干净”图像的PSNR、SSIM以及MSE的平均值作为实验结果,如图12~图16所示。

如图12~图16所示,横坐标表示负载率,图12~图16(a)坐标表示不同负载率下PSNR值,(b)纵坐标表示不同负载率下SSIM值,(c)纵坐标表示不同负载率下MSE值。其中,图12~图15表示本方法主动防御J-UNIWARD和nsF5后,“干净”图像的不同图像质量指标,图16表示本方法主动防御S-UNIWARD和WOW后,“干净”图像的不同图像质量指标。由图12~图16可知,本方法主动防御四种普通隐写,各组图像的PSNR值介于40~45之间,SSIM值介于0.95~1之间,MSE值都在6以内,同时在不同负载率下的差距很小,在一定程度上表明本方法不仅对恢复噪声图像有效,也表明本方法对不同隐写算法具有普适性。

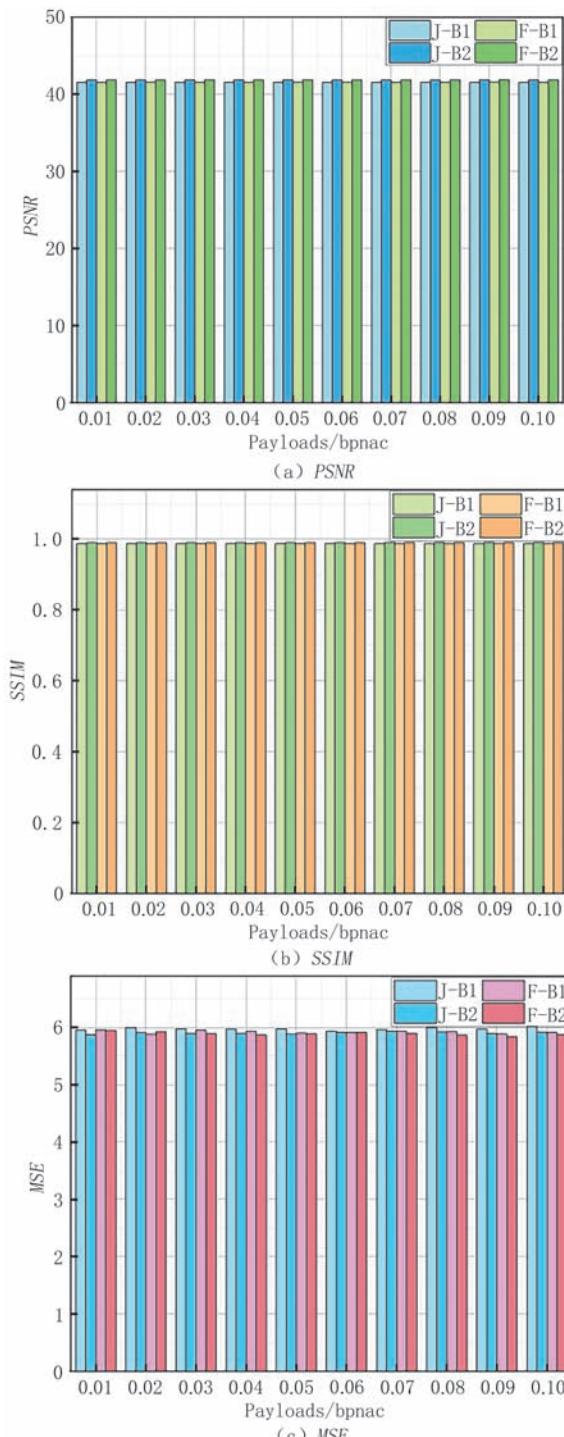


图12 主动防御 J-UNIWARD 和 nsF5(QF65)得到“干净”图像的评价结果

### 5.3.3 实验结果展示

从测试图像中随机选择8幅图像,视觉效果与具体评价标准值在图17与表12分别展示。

图17从上到下依次是载密图像、噪声图像、“干净”图像。通过实验结果表明,破坏秘密信息之后,通过网络恢复的“干净”图像和载密图像没有明显视觉差别。

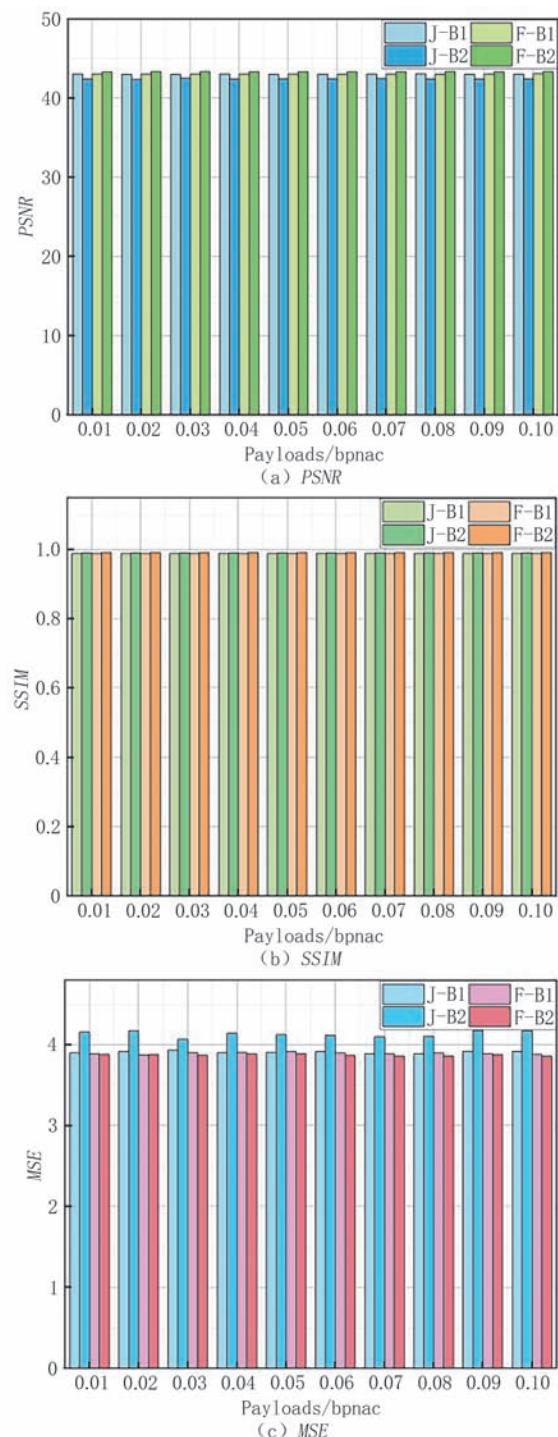


图13 主动防御 J-UNIWARD 和 nsF5(QF75)得到“干净”图像的评价结果

如表12所示,图6.jpg的BER值与CR值分别达到0.5336和0.3691;PSNR值与MSE值分别达到40.4960和5.8007。随机选择的8张图像的BER值都达到0.5以上,同时,图像的PSNR值在40以上。以上数据表明本方法对秘密信息破坏和图像质量恢复是有效的。同时,从另一方面也说明本文的方法在主动防御社交网络中的隐蔽通信方面是有效的。

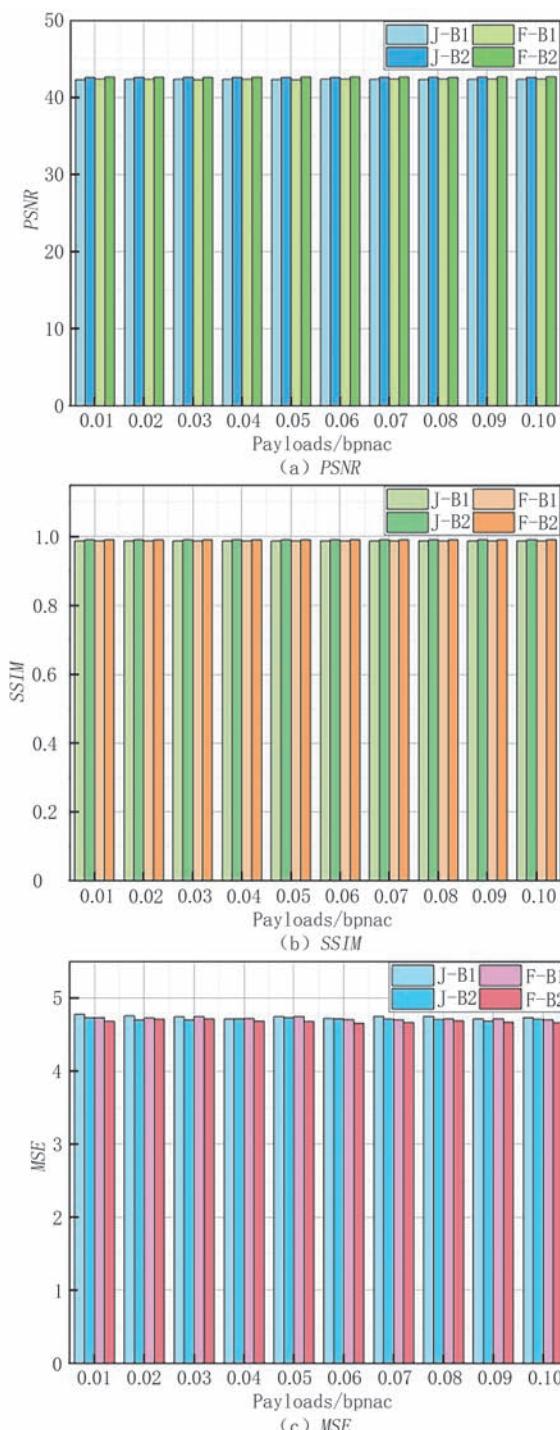


图14 主动防御J-UNIWARD和nsF5(QF85)得到“干净”图像的评价结果

综上可知,对于鲁棒隐写和普通隐写而言,本文方法破坏秘密信息后,“干净”图像的PSNR、SSIM和MSE值均符合社交网络中发送图像的要求,并在不同数据集与负载率下,“干净”图像的PSNR、SSIM和MSE值均趋于稳定。以上数据表明SPRN方法能够有效恢复噪声图像的质量。因此,针对图像隐蔽通信,本文方法能够成功进行防御,并使得通

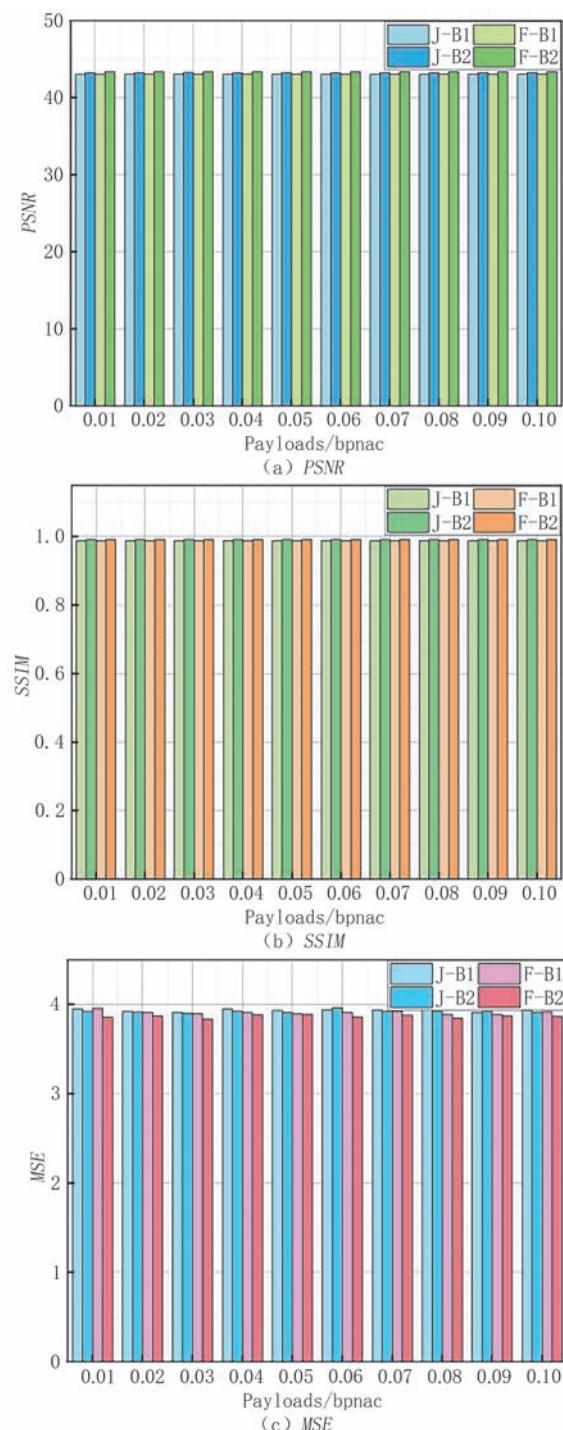


图15 主动防御J-UNIWARD和nsF5(QF95)得到“干净”图像的评价结果

信双方在毫无察觉的情况下彻底阻断秘密信息的传输。

#### 5.3.4 社交网络实验结果

由于实际场景图像往往经过多次压缩,为了验证在社交网络中所提方法的有效性,本文对质量因子为65与75的测试图像,再次采用质量因子65与75进行压缩。在该情况下对DMAS隐写进行防

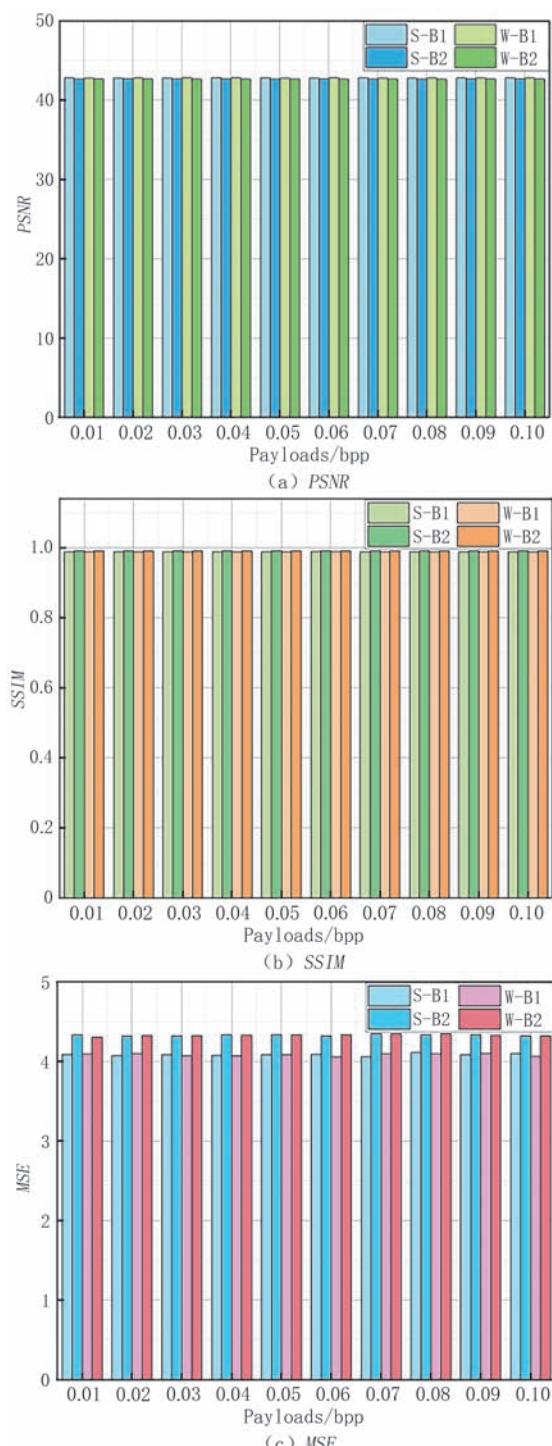


图 16 主动防御 S-UNIWARD 和 WOW 得到“干净”图像的评价结果

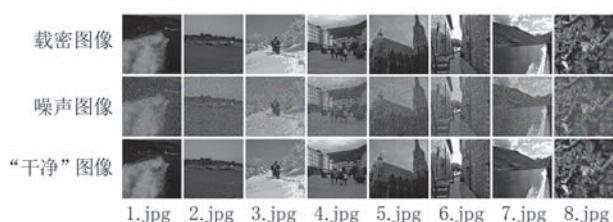


图 17 随机选取 8 幅图像

表 12 “干净”图像评价结果对比

“干净”图像	BER	CR	PSNR	SSIM	MSE
1.jpg	0.5248	0.2451	42.3801	0.9742	3.7589
2.jpg	0.5156	0.2102	44.2492	0.9910	2.4443
3.jpg	0.5218	0.2356	41.9672	0.9909	4.1339
4.jpg	0.5223	0.2278	42.2294	0.9901	3.8917
5.jpg	0.5116	0.2599	41.0880	0.9828	5.0615
6.jpg	0.5336	0.3691	40.4960	0.9911	5.8007
7.jpg	0.5203	0.2463	42.1201	0.9870	3.9909
8.jpg	0.5317	0.1709	46.3700	0.9926	1.4999

御,得到不同负载率下的主动防御评价效果,如表 13 和表 14 所示。

表 13 SPRN 主动防御重压缩载密图像(QF65)得到“干净”图像的评价结果

Payloads	评价指标				
	PSNR	SSIM	MSE	CR	BER
0.01	37.6653	0.9825	15.9402	0.3725	0.4938
0.02	37.6484	0.9825	15.8913	0.3727	0.4964
0.03	37.5881	0.9824	16.0828	0.3733	0.4947
0.04	37.5907	0.9821	16.0582	0.3742	0.4902
0.05	37.5504	0.9820	16.1930	0.3748	0.4863
0.06	37.5277	0.9817	16.2760	0.3759	0.4809
0.07	37.4999	0.9816	16.3345	0.3767	0.4798
0.08	37.4627	0.9813	16.4514	0.3776	0.5050
0.09	37.4501	0.9810	16.5166	0.3785	0.4810
0.1	37.3864	0.9807	16.6878	0.3798	0.4815

表 14 SPRN 主动防御重压缩载密图像(QF75)得到“干净”图像的评价结果

Payloads	评价指标				
	PSNR	SSIM	MSE	CR	BER
0.01	41.5585	0.9891	4.9377	0.3096	0.4953
0.02	41.5105	0.9891	4.9680	0.3100	0.4903
0.03	41.5219	0.9890	4.9576	0.3105	0.4724
0.04	41.4669	0.9888	5.0251	0.3110	0.4724
0.05	41.4189	0.9888	5.0626	0.3116	0.4613
0.06	41.3720	0.9886	5.1152	0.3124	0.4510
0.07	41.3940	0.9885	5.1090	0.3128	0.4470
0.08	41.3361	0.9883	5.1631	0.3134	0.4464
0.09	41.3209	0.9881	5.1903	0.3143	0.4419
0.1	41.1939	0.9879	5.3355	0.3152	0.4340

由表 13 和表 14 可知,在经过多次压缩下,SPRN 方法主动防御 DMAS 隐写得到“干净”图像的 PSNR、SSIM 和 MSE 均符合社交网络传递图像的要求。干净图像的 BER 和 CR 均超过破坏秘密信

息的基准值。因此,SPRN方法能够抵御社交网络中的多次压缩情况。

#### 5.4 消融实验和对比实验

本节分为两部分:Rec-Net消融实验和对比实验。

##### 5.4.1 Rec-Net消融实验

为了验证Rec-Net高斯滤波层、高斯滤波残差块和中值滤波层对秘密信息破坏与图像质量恢复的影响,本小节选取BOWS2数据集中50张“发布”

图像验证,然后将图像压缩成质量因子为85的载体图像。同时使用鲁棒隐写算法DMAS进行不同负载率的隐写,其他的实验设置与4.1节中一致。实验结果如表15所示。其中,本节将去掉高斯滤波层(Without Gauss Layers)网络简写为WGL-Net,将去掉高斯滤波残差块(Without Gauss Residual Blocks)网络简写为WGRB-Net,将去掉中值滤波层(Without Median Layers)网络简写为WML-Net。

表15 Rec-Net消融实验

网络	指标	Payloads									
		0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.10
WGL-Net	PSNR	25.28	25.26	25.27	25.26	25.25	25.27	25.25	25.25	25.24	25.24
	SSIM	0.67	0.70	0.70	0.70	0.70	0.70	0.69	0.69	0.69	0.69
	MSE	231.1	231.6	231.6	231.4	232.1	231.3	231.8	232.1	232.3	232.4
	BER	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
	CR	0.308	0.309	0.309	0.309	0.309	0.309	0.310	0.310	0.310	0.310
WGRB-Net	PSNR	25.39	25.38	25.38	25.39	25.38	25.38	25.37	25.37	25.37	25.37
	SSIM	0.68	0.68	0.68	0.68	0.68	0.68	0.68	0.68	0.68	0.68
	MSE	229.2	229.3	229.2	229.4	229.26	229.5	229.6	229.6	229.6	229.8
	BER	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
	CR	0.303	0.303	0.303	0.304	0.304	0.304	0.304	0.305	0.305	0.305
WML-Net	PSNR	22.68	22.60	22.68	22.68	22.68	22.70	22.65	22.68	22.60	22.66
	SSIM	0.68	0.68	0.68	0.68	0.67	0.67	0.67	0.67	0.67	0.67
	MSE	389.3	390.7	388.7	389.0	389.4	388.3	391.0	389.2	389.6	390.6
	BER	0.49	0.49	0.50	0.50	0.50	0.50	0.50	0.51	0.50	0.51
	CR	0.276	0.276	0.276	0.277	0.277	0.277	0.278	0.278	0.278	0.278
Rec-Net	PSNR	42.63	42.64	42.65	42.63	42.63	42.59	42.59	42.58	42.57	42.58
	SSIM	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
	MSE	4.7	4.7	4.7	4.7	4.8	4.8	4.8	4.8	4.8	4.8
	BER	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
	CR	0.299	0.299	0.299	0.300	0.300	0.300	0.301	0.301	0.301	0.302

图像质量方面,在不同负载率下,Rec-Net所得“干净”图像的PSNR、SSIM和MSE的均值分别是42.209、0.99和4.76。其中WGL-Net所得“干净”图像的PSNR、SSIM和MSE的均值分别为25.257、0.693和231.77。WGRB-Net所得“干净”图像的PSNR、SSIM和MSE的均值分别为25.378、0.68和229.446。WML-Net所得“干净”图像的PSNR、SSIM和MSE的均值分别为22.661、0.674和389.58。以上数据表明WGL-Net、WGRB-Net和WML-Net所得“干净”图像的PSNR、SSIM和MSE值较低,说明它们不能够较好的恢复图像质量。验证Rec-Net生成的图像在经过S&P隐写防御后都能得到较好恢复。

破坏秘密信息方面,在不同负载率下,四种网络所得“干净”图像的BER的均值相等均为0.5。WGL-Net、WGRB-Net、WML-Net和Rec-Net所得“干净”图像的CR均值分别为0.3093、0.304、0.2771和0.3002。以上结果表明图像的BER和CR值均超过破坏秘密信息的要求,也验证本网络的架构设置都不影响第一阶段S&P主动防御的效果。综上,Rec-Net架构的设置不仅保持第一阶段破坏秘密信息的效果,也将噪声图像较好地恢复成“发布”图像的质量。

##### 5.4.2 对比实验

对比实验采用BOWS2数据集中30张载体图像,并将图像压缩成质量因子为75的JPEG图像,在

每个负载率下分别生成 30 张载密图像。实验以鲁棒隐写算法 DMAS 和 GMAS 以及普通频域隐写算法 J-UNIWARD 和 nsF5 为例, 分别对比本文方法与 AO-Net<sup>[6]</sup>方法、SC-Net<sup>[13]</sup>方法、基于高斯白噪声的主动防御方法以及基于图像 JPEG 压缩的主动防御方法。其中高斯白噪声的均值和方差为 0.01, 图像 JPEG 的压缩因子为 65。

### (a) 主动防御 DMAS 和 GMAS 对比实验

在不同负载率下, 五种方法主动防御 DMAS 和 GMAS 的对比结果如表 16~17 所示。由表 16 可知, 在不同负载率下, 与 AO-Net<sup>[6]</sup>方法相比, SPRN 方法破坏秘密信息后, “干净”图像的 PSNR、SSIM、MSE、BER 与 CR 的结果分别提升 2.85%、0.957%、6.958%、5.461% 和 24.258%。与 SC-Net<sup>[13]</sup>方法相比, SPRN 方法破坏秘密信息后, “干净”图像的 PSNR、SSIM、MSE、BER 与 CR 的结果分别提升 34.795%、8.189%、92.577%、0.559% 和 20.43%。与基于高斯白噪声的主动防御方法相比, SPRN 方法

破坏秘密信息后, “干净”图像的 PSNR、SSIM 与 MSE 的结果分别提升 53.017%、66.8% 和 14450%。“干净”图像的 BER 与 CR 基本一致。与基于图像 JPEG 压缩的主动防御方法相比, SPRN 方法破坏秘密信息后, “干净”图像的 PSNR、SSIM、MSE、BER 与 CR 的结果分别提升 7.585%、0.451%、50.13%、12.602% 和 18.913%。由此可知, 与其他方法主动防御 DMAS 相比, SPRN 方法既能高误码率破坏秘密信息, 又能提升图像视觉质量。由表 17 可知, 不同负载率下, 与 AO-Net<sup>[6]</sup>方法相比, SPRN 方法破坏秘密信息后, “干净”图像的 PSNR、SSIM、BER 与 CR 的结果分别提升 1.292%、0.664%、0.923% 和 18.01%。与 SC-Net<sup>[13]</sup>方法相比, SPRN 方法破坏秘密信息后, “干净”图像的 PSNR、SSIM、BER 与 CR 的结果分别提升 33.345%、7.841%、0.071% 和 14.055%。并且, 在负载率为 0.02、0.03、0.05、0.06、0.07 和 0.09 中, “干净”图像的 BER 值远超过其他方法, 而“干净”图像的 MSE 值介于 AO-Net<sup>[6]</sup>方

表 16 五种方法主动防御 DMAS 的结果对比

方法	指标	Payloads									
		0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.1
SPRN方法	PSNR	42.770	42.756	42.716	42.714	42.691	42.659	42.633	42.607	42.579	42.587
	SSIM	0.9907	0.9906	0.9906	0.9905	0.9905	0.9904	0.9903	0.9902	0.9901	0.9900
	MSE	4.3199	4.3561	4.3881	4.3963	4.3991	4.4294	4.4679	4.5066	4.5044	4.5069
	BER	0.4989	0.4997	0.4987	0.5016	0.5002	0.4998	0.4976	0.4984	0.4989	0.4986
	CR	0.3041	0.3046	0.3050	0.3054	0.3060	0.3064	0.3070	0.3080	0.3082	0.3085
AO-Net <sup>[7]</sup> 方法	PSNR	41.689	41.680	41.622	41.579	41.523	41.476	41.419	41.364	41.288	41.232
	SSIM	0.9830	0.9829	0.9824	0.9821	0.9815	0.9809	0.9803	0.9797	0.9789	0.9783
	MSE	4.5567	4.5646	4.6175	4.6610	4.7169	4.7691	4.8258	4.8837	4.9628	5.0275
	BER	0.4984	0.4853	0.4683	0.4570	0.4448	0.4378	0.4263	0.4132	0.4108	0.4044
	CR	0.0633	0.0632	0.0635	0.0634	0.0637	0.0637	0.0639	0.0640	0.0642	0.0645
SC-Net <sup>[8]</sup> 方法	PSNR	31.744	31.726	31.711	31.590	31.679	31.663	31.637	31.624	31.606	31.584
	SSIM	0.9188	0.9184	0.9178	0.9171	0.9163	0.9154	0.9143	0.9132	0.9120	0.9110
	MSE	58.940	59.090	59.222	59.412	59.515	59.709	59.898	60.037	60.213	60.400
	BER	0.4851	0.4966	0.4979	0.4931	0.4934	0.4915	0.4910	0.4947	0.4971	0.4961
	CR	0.1012	0.1013	0.1015	0.1016	0.1018	0.1020	0.1023	0.1025	0.1028	0.1032
基于高斯白噪声的主动防御方法	PSNR	20.047	20.051	20.045	20.050	20.051	20.046	20.050	20.045	20.047	20.048
	SSIM	0.3279	0.3281	0.3284	0.3284	0.3286	0.3288	0.3290	0.3290	0.3293	0.3297
	MSE	644	644	644	644	644	645	644	645	644	644
	BER	0.4974	0.5002	0.5048	0.4987	0.4985	0.5038	0.5023	0.5000	0.4963	0.4991
	CR	0.5595	0.5594	0.5596	0.5596	0.5598	0.5600	0.5597	0.5600	0.5599	0.5596
基于图像 JPEG 压缩的主动防御方法	PSNR	39.495	39.485	39.472	39.459	39.444	39.430	39.416	39.398	39.380	39.364
	SSIM	0.9861	0.9861	0.9861	0.9860	0.9869	0.9858	0.9857	0.9856	0.9855	0.9854
	MSE	8.7776	8.7933	8.8158	8.8353	8.8599	8.8831	8.9108	8.9383	8.9724	8.9974
	BER	0.4575	0.4565	0.4261	0.3903	0.3773	0.3575	0.3346	0.3215	0.3112	0.2997
	CR	0.1170	0.1171	0.1171	0.1171	0.1172	0.1172	0.1172	0.1173	0.1173	0.1174

法和SC-Net<sup>[13]</sup>方法之间。与基于高斯白噪声的主动防御方法相比,SPRN方法破坏秘密信息后,“干净”图像的PSNR、SSIM与MSE的结果分别提升52.5%、66.83%和13352%。“干净”图像的BER与CR基本一致。与基于图像JPEG压缩的主动防御方

法相比,SPRN方法破坏秘密信息后,“干净”图像的PSNR、SSIM、MSE、BER与CR的结果分别提升3.267%、0.225%、24.1%、1.008%和14.366%。由此可知,与其他方法主动防御GMAS相比,SPRN方法具有明显优势。

表17 五种方法主动防御GMAS的结果对比

方法	指标	Payloads									
		0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.10
SPRN方法	PSNR	42.221	42.199	42.222	42.204	42.217	42.194	42.188	42.218	42.196	42.232
	SSIM	0.9897	0.9896	0.9896	0.9896	0.9896	0.9896	0.9896	0.9896	0.9896	0.9896
	MSE	4.7952	4.7833	4.7911	4.7995	4.7779	4.8032	4.8005	4.7848	4.7927	4.7268
	BER	0.4986	0.5017	0.5020	0.4999	0.4989	0.5007	0.5012	0.4995	0.4995	0.5004
	CR	0.2431	0.2436	0.2433	0.2439	0.2437	0.2436	0.2435	0.2436	0.2438	0.2434
AO-Net <sup>[7]</sup> 方法	PSNR	41.724	41.727	41.690	41.701	41.691	41.665	41.656	41.640	41.626	41.584
	SSIM	0.9832	0.9832	0.9830	0.9832	0.9831	0.9831	0.9830	0.9831	0.9830	0.9829
	MSE	4.5267	4.5238	4.5580	4.5464	4.5591	4.5846	4.5922	4.6055	4.6184	4.6611
	BER	0.5002	0.5006	0.4996	0.4957	0.4951	0.4925	0.4885	0.4871	0.4786	0.4722
	CR	0.0632	0.0632	0.0633	0.0633	0.0634	0.0635	0.0635	0.0635	0.0637	0.0639
SC-Net <sup>[8]</sup> 方法	PSNR	31.747	31.735	31.716	31.696	31.676	31.660	31.624	31.591	31.567	31.528
	SSIM	0.9188	0.9187	0.9185	0.9183	0.9180	0.9177	0.9173	0.9169	0.9164	0.9160
	MSE	58.874	58.948	59.057	59.236	59.396	59.518	59.820	60.103	60.335	60.701
	BER	0.4997	0.4972	0.4990	0.5009	0.4976	0.4998	0.4988	0.5015	0.4987	0.5021
	CR	0.1016	0.1018	0.1021	0.1024	0.1027	0.1030	0.1034	0.1039	0.1043	0.1048
基于高斯白噪声的主动防御方法	PSNR	20.0419	20.0487	20.0489	20.0492	20.0542	20.0545	20.0436	20.0507	20.0499	20.0511
	SSIM	0.3276	0.3280	0.3280	0.3282	0.3283	0.3283	0.3280	0.3284	0.3285	0.3288
	MSE	645	644	644	644	643	643	645	644	644	644
	BER	0.5057	0.5020	0.5031	0.4972	0.4927	0.5022	0.4974	0.5066	0.5022	0.4979
	CR	0.5597	0.5598	0.5596	0.5595	0.5594	0.5596	0.5596	0.5595	0.5596	0.5592
基于图像JPEG压缩的主动防御方法	PSNR	40.8558	40.8581	40.8501	40.8440	40.8417	40.8328	40.8254	40.8075	40.8028	40.7825
	SSIM	0.9874	0.9874	0.9874	0.9874	0.9874	0.9874	0.9874	0.9874	0.9874	0.9874
	MSE	5.9109	5.9090	5.9168	5.9221	5.9266	5.9358	5.9438	5.9605	5.9707	5.9925
	BER	0.4974	0.4974	0.4988	0.4919	0.4999	0.4903	0.4853	0.4849	0.4778	0.4779
	CR	0.0999	0.0998	0.0999	0.0998	0.0998	0.0999	0.0999	0.1000	0.1000	0.1000

## (b) 主动防御J-UNIWARD和nsF5对比实验

在不同负载率下,五种方法主动防御J-UNIWARD和nsF5的对比结果如表18~19所示。

由表18可知,SPRN方法破坏秘密信息后,“干净”图像的PSNR、SSIM以及CR值均比其他方法高,“干净”图像的MSE值比其他方法的值低,在负载率为0.01、0.02、0.04、0.05和0.06下,“干净”图像的BER值远高于AO-Net方法和SC-Net方法。与AO-Net<sup>[6]</sup>方法相比,SPRN方法破坏秘密信息后,“干净”图像的PSNR、SSIM、MSE、BER与CR的结果分别提升1.659%、0.662%、8.852%、0.34%和23.711%。与SC-Net<sup>[13]</sup>方法相比,SPRN方法破坏秘密信息后,“干净”图像的PSNR、SSIM、MSE、BER与CR的结果分别提升33.544%、

7.69%、92.976%、0.437%和19.903%。与基于高斯白噪声的主动防御方法相比,SPRN方法破坏秘密信息后,“干净”图像的PSNR、SSIM、MSE与BER的结果分别提升52.723%、67.51%、15.507%和0.532%。“干净”图像的CR的值略低。与基于图像JPEG压缩的主动防御方法相比,SPRN方法破坏秘密信息后,“干净”图像的PSNR、SSIM、MSE、BER与CR的结果分别提升3.668%、0.232%、43.168%、0.551%和20.041%。由此可知,与上述两种方法相比,本文方法对秘密信息破坏的同时仍然保持较高的图像质量。

由表19可知,一方面,在不同负载率下,SPRN方法破坏秘密信息后,“干净”图像的PSNR和SSIM值明显高于AO-Net<sup>[6]</sup>方法和SC-Net<sup>[13]</sup>方法,而“干

表 18 五种方法主动防御 J-UNIWARD 的结果对比

方法	指标	Payloads									
		0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.10
SPRN方法	PSNR	42.393	42.341	42.494	42.391	42.435	42.436	42.458	42.406	42.381	42.372
	SSIM	0.9897	0.9896	0.9897	0.9896	0.9897	0.9897	0.9898	0.9897	0.9896	0.9897
	MSE	4.1520	4.1679	4.0645	4.1391	4.1212	4.1145	4.0944	4.1040	4.1743	4.1735
	BER	0.5118	0.5099	0.5004	0.5092	0.5103	0.5123	0.4922	0.4976	0.5030	0.4966
	CR	0.3003	0.2999	0.3000	0.3008	0.3001	0.3003	0.3001	0.3004	0.3006	0.3009
AO-Net <sup>[7]</sup> 方法	PSNR	41.723	41.717	41.721	41.721	41.725	41.711	41.713	41.711	41.724	41.720
	SSIM	0.9832	0.9832	0.9832	0.9832	0.9832	0.9831	0.9831	0.9831	0.9832	0.9832
	MSE	4.5221	4.5384	4.5316	4.5286	4.5268	4.5371	4.5387	4.5388	4.5261	4.5289
	BER	0.4985	0.5010	0.5007	0.5015	0.4991	0.5003	0.5018	0.5012	0.5031	0.5021
	CR	0.0632	0.0632	0.0632	0.0632	0.0632	0.0633	0.0633	0.0633	0.0632	0.0632
SC-Net <sup>[8]</sup> 方法	PSNR	31.760	31.759	31.759	31.760	31.758	31.759	31.756	31.756	31.757	31.754
	SSIM	0.9190	0.9190	0.9190	0.9190	0.9190	0.9190	0.9190	0.9190	0.9191	0.9190
	MSE	58.793	58.784	58.798	58.797	58.800	58.803	58.829	58.823	58.792	58.834
	BER	0.4908	0.5027	0.4983	0.5033	0.4996	0.5026	0.4992	0.5017	0.5015	0.4999
	CR	0.1013	0.1014	0.1013	0.1013	0.1013	0.1013	0.1013	0.1013	0.1013	0.1013
基于高斯白噪声的主动防御方法	PSNR	20.0498	20.0503	20.0513	20.0438	20.0514	20.0484	20.0498	20.0534	20.0556	20.0440
	SSIM	0.3278	0.3280	0.3285	0.3277	0.3279	0.3283	0.3280	0.3282	0.3285	0.3277
	MSE	645	645	644	645	644	645	645	644	644	645
	BER	0.4860	0.5029	0.5083	0.4978	0.4992	0.4991	0.5001	0.4977	0.4982	0.5008
	CR	0.5597	0.5596	0.5597	0.5596	0.5596	0.5594	0.5595	0.5595	0.5594	0.5600
基于图像 JPEG 压缩的主动防御方法	PSNR	40.8557	40.8557	40.8551	40.8544	40.8554	40.8557	40.8551	40.8537	40.8542	40.8546
	SSIM	0.9874	0.9874	0.9874	0.9874	0.9874	0.9874	0.9874	0.9874	0.9874	0.9874
	MSE	5.9130	5.9129	5.9131	5.9137	5.9132	5.9130	5.9132	5.9141	5.9146	5.9132
	BER	0.5070	0.4877	0.5003	0.4939	0.4958	0.5063	0.5010	0.4977	0.4982	0.5003
	CR	0.0999	0.0999	0.1000	0.0999	0.0999	0.0999	0.1000	0.1000	0.0999	0.0999

表 19 五种方法主动防御 nsF5 的结果对比

方法	指标	Payloads									
		0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.10
SPRN方法	PSNR	42.322	42.340	42.349	42.316	42.313	42.335	42.347	42.354	42.332	42.369
	SSIM	0.9904	0.9904	0.9904	0.9904	0.9904	0.9904	0.9904	0.9904	0.9904	0.9904
	MSE	4.7335	4.7495	4.7291	4.7627	4.7612	4.7236	4.7114	4.7144	4.7349	4.7031
	CR	0.3131	0.3128	0.3127	0.3129	0.3128	0.3127	0.3125	0.3123	0.3125	0.3122
	PSNR	41.734	41.706	41.720	41.718	41.720	41.716	41.716	41.694	41.701	41.708
AO-Net <sup>[7]</sup> 方法	SSIM	0.9832	0.9831	0.9831	0.9832	0.9831	0.9831	0.9831	0.9830	0.9831	0.9830
	MSE	4.5148	4.5429	4.5311	4.5318	4.5265	4.5359	4.5336	4.5551	4.5544	4.5434
	CR	0.0630	0.0632	0.0631	0.0632	0.0632	0.0632	0.0633	0.0634	0.0634	0.0633
	PSNR	31.764	31.767	31.77	31.774	31.776	31.781	31.785	31.788	31.792	31.797
SC-Net <sup>[8]</sup> 方法	SSIM	0.9190	0.9191	0.9191	0.9191	0.9192	0.9192	0.9192	0.9193	0.9193	0.9194
	MSE	58.740	58.693	58.654	58.612	58.567	58.517	58.452	58.407	58.352	58.290
	CR	0.1012	0.1012	0.1011	0.1011	0.1010	0.1010	0.1009	0.1008	0.1008	0.1007
	PSNR	20.0490	20.0444	20.0523	20.0545	20.0420	20.0504	20.0473	20.0503	20.0457	20.0520
基于高斯白噪声的主动防御方法	SSIM	0.3278	0.3275	0.3277	0.3282	0.3276	0.3277	0.3276	0.3277	0.3274	0.3280
	MSE	645	645	644	644	646	645	645	645	644	644
	CR	0.5601	0.5597	0.5592	0.5592	0.5600	0.5594	0.5598	0.5595	0.5596	0.5596
	PSNR	40.8578	40.861	40.8639	40.868	40.8714	40.8758	40.8801	40.8849	40.8893	40.8951
基于图像 JPEG 压缩的主动防御方法	SSIM	0.9874	0.9874	0.9874	0.9874	0.9874	0.9874	0.9875	0.9875	0.9875	0.9875
	MSE	5.9096	5.9052	5.9009	5.8948	5.8896	5.883	5.8775	5.8714	5.8655	5.8575
	CR	0.0999	0.0999	0.0998	0.0998	0.0998	0.0997	0.0997	0.0996	0.0996	0.0995

净”图像的MSE值介于AO-Net<sup>[6]</sup>方法和SC-Net<sup>[13]</sup>方法之间。与AO-Net<sup>[6]</sup>方法相比,SPRN方法破坏秘密信息后,“干净”图像的PSNR、SSIM与CR的结果分别提升1.497%、0.746%和24.942%。与SC-Net<sup>[13]</sup>方法相比,SPRN方法破坏秘密信息后,“干净”图像的PSNR、SSIM与CR的结果分别提升33.223%、7.747%和21.167%。与基于高斯白噪声的主动防御方法相比,SPRN方法破坏秘密信息后,“干净”图像的PSNR、SSIM与MSE的结果分别提升53.047%、66.91%和13617%。“干净”图像的CR的值略低。与基于图像JPEG压缩的主动防御方法相比,SPRN方法破坏秘密信息后,“干净”图像的PSNR、SSIM、MSE与CR的结果分别提升3.455%、0.298%、24.468%和21.29%。以上结果表明本方法生成“干净”图像的质量满足社交网络需求,且质量较好。另一方面,在不同负载率下,SPRN方法破坏秘密信息后,“干净”图像的CR结果均高于另外两种方法,以上数据表明本文方法对秘密信息破坏程度较高。

综上可知,在秘密信息破坏和图像质量恢复方面,SPRN方法均优于其他两种方法。

## 6 总 结

现有的隐蔽通信防御大多数基于检测等被动防御方法。针对检测在低负载率下虚警率和漏检率过高,并且,面对社交网络中未知隐写、负载率等先验知识情况下无法实时有效阻断隐蔽通信的问题,本文提出SPRN主动防御方法。在第三方毫无察觉的情况下,清除秘密信息,主动防御社交网络中的隐蔽通信。所提方法分析不同的噪声对秘密信息的破坏效果,选择叠加S&P噪声的方式破坏秘密信息,达到主动防御的目的。由于S&P噪声中随机的椒噪点和盐噪点的叠加会对图像视觉效果层面造成一定的影响。因此,利用Rec-Net对图像进行视觉质量恢复,实现对图像的优化。最后,得到无法提取出秘密信息的“干净”图像,既保持图像的高视觉质量,又不增加存储空间大小。此外,本文提出一种新的基于CR的隐写主动防御图像评价准则,该准则能够在未知隐写的先验条件下度量秘密信息破坏效果,弥补误码率的不足。实验结果表明,在不同数据集和负载率下,所提方法能够有效地破坏社交网络中潜在载密图像中的秘密信息,实现隐蔽通信主动防御的目的。同时,与先进方法SC-Net和AO-Net在

BOWS2测试集上进行对比,结果充分表明在秘密信息破坏和图像质量恢复方面,SPRN方法具有较大优势。该方法应用于社交网络平台,可通过主动防御的方式实现阻断不法用户隐蔽通信的目的。

在未来的工作中,我们将致力于在秘密信息破坏精准度与时效性两大方面进一步提升主动防御能力,使主动防御具有更强的适用性,并为完善主动防御评价体系提供理论支撑。

## 参 考 文 献

- [1] Xiang Lingyun, Yang Shuanghui, Wang Rong, et al. A constraint-based natural language information hiding method from sequence to steganographic sequence. Chinese Journal of Computers, 2023, 46(8): 1650-1669 (in Chinese)  
(向凌云, 杨双辉, 王蓉等. 基于序列到隐写序列的约束型自然语言信息隐藏方法. 计算机学报. 2023, 46(8): 1650-1669)
- [2] Ma Yuanyuan, Xu Jiucheng, Zhang Yi, et al. Feature selection method for rich model steganalysis detection based on W2ID criterion. Chinese Journal of Computers, 2021, 44(4): 724-740 (in Chinese)  
(马媛媛, 徐久成, 张祎等. 基于W2ID准则的Rich Model隐写检测特征选取方法. 计算机学报, 2021, 44(4): 724-740)
- [3] Yang Yu, Zhang Ziwei, Wen Juan. Multi-task and few-sample text steganalysis based on capsule network. Chinese Journal of Computers, 2022, 45(12):2592-2604 (in Chinese)  
(杨雨, 张梓葳, 文娟. 基于胶囊网络的多任务少样本隐写分析. 计算机学报, 2022, 45(12):2592-2604)
- [4] Tian Hui, Wu Junyan, Yan Yan, et al. Adaptive multi-rate speech stream steganalysis based on fractional pitch delay correlation. Chinese Journal of Computers, 2022, 45(6): 1308-1325 (in Chinese)  
(田晖, 吴俊彦, 严艳等. 基于小数基音延迟相关性的自适应多速率语音流隐写分析. 计算机学报, 2022, 45(6):1308-1325)
- [5] Ma Yuanyuan, Xu Lige, Zhang Yi, et al. Steganalysis feature selection with multidimensional evaluation & dynamic threshold allocation . IEEE Transactions on Circuits and Systems for Video Technology, 2023, 34(3), 1954-1969
- [6] Zhou Zhili, Ding Chun, Li Jin, et al. Research on Generative Steganography. Chinese Journal of Computers, 2023, 46(9): 1855-1887 (in Chinese)  
(周志立, 丁淳, 李进等. 生成式隐写研究. 计算机学报, 2023, 46(9): 1855-1887)
- [7] Li Ruixiang, Xu Rui, Ma Yuanyuan, et al. LandmarkMiner: Street-level Network Landmarks Mining Method for IP Geolocation. ACM Transactions on Internet of Things, 2021, 2(3): 1-22
- [8] Zhou Fan, Wang Tianliang, Zhong Ting, et al. Identifying user geolocation with hierarchical graph neural networks and explainable fusion . Information Fusion, 2022, 81: 1-13
- [9] Zhang Fan, Liu Fenlin, Luo Xiangyang. Correction: Geolocation of covert communication entity on the Internet for post-steganalysis . EURASIP Journal on Image and Video

- Processing, 2023, 2023(1): 2
- [10] Zhu Zhiying, Li Sheng, Qian Zhenxing, et al. Destroying robust steganography in online social networks . Information Sciences, 2021, 581: 605-19
- [11] Zhang Kai, Zuo Wangmeng, Chen Yunjin, et al. Beyond a gaussian denoiser: residual learning of deep CNN for image denoising . IEEE Transactions on Image Processing, 2017, 26(7): 3142-3155
- [12] Geng Linfeng, Zhang Weiming, Chen Haozhe, et al. Real-time attacks on robust watermarking tools in the wild by CNN. Journal of Real-Time Image Processing, 2020, 17(3): 631-641
- [13] Li Qi, Wang Chunpeng, Wang Xiaoyu, et al. A new imperceptible watermark attack method based on residual learning. Journal of Software, 2023, 34(9): 4351-4361 (in Chinese)  
(李琦, 王春鹏, 王晓雨, 等. 基于残差学习的新型不可感知水印攻击方法. 软件学报, 2023, 34(9): 4351-4361)
- [14] Kim J, Lee J K, Lee K M. Accurate image super-resolution using very deep convolutional networks//Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, USA, 2016, 1646-1654
- [15] Wang Huaiqi, Qian Zhenxing, Feng Guorui, et al. Defeating data hiding in social networks using generative adversarial network . EURASIP Journal on Image and Video Processing, 2020, 2020(1): 30
- [16] Corley I A, Lwowski J, Hoffman J. Destruction of image steganography using generative adversarial networks . ArXiv, 2019, abs/1912.10070 arxiv preprint arxiv:1912,10070,2019
- [17] Zhu Zhiying, Wei Ping, Qian Zhenxing, et al. Image sanitization in online social networks: A general framework for breaking robust information hiding . IEEE Transactions on Circuits and Systems for Video Technology, 2023, 33(6): 3017-3029
- [18] Wei Ping, Li Sheng, Zhang Xinpeng, et al. Generative steganography network// Proceedings of the 30th ACM International Conference on Multimedia. Lisboa, Portugal, 2022: 1621-1629
- [19] Zhu Liyan, Luo Xiangyang, Zhang Yi, et al. An asymmetric distortion steganography method based on super-pixel filtering. Chinese Journal of Computers, 2023, 46(7): 1473-1493 (in Chinese)  
(朱利妍, 罗向阳, 张祎, 等. 基于超像素滤波的非对称失真隐写方法. 计算机学报, 2023, 46(7): 1473-1493)
- [20] Lu Wei, Zhang Junhong, Zhao Xianfeng, et al. Secure robust JPEG steganography based on autoEncoder with adaptive BCH encoding . IEEE Transactions on Circuits and Systems for Video Technology, 2021, 31(7): 2909-2922
- [21] Tao Jinyuan, Li Sheng, Zhang Xinpeng, et al. Towards robust image steganography. IEEE Transactions on Circuits and Systems for Video Technology, 2019, 29(2): 594-600
- [22] Yu Xinzhi, Chen Kejiang, Wang Yaofei, et al. Robust adaptive steganography based on generalized dither modulation and expanded embedding domain . Signal Processing, 2020, 168: 107343
- [23] Zhang Yi, Luo Xiangyang, Yang Chunfang, et al. Joint JPEG compression and detection resistant performance enhancement for adaptive steganography using feature regions selection . Multimedia Tools and Applications, 2017, 76(3): 3649-68
- [24] Li Jiyu, Fu Zhangjie, Wang Fan. A canny-gauss general-domain image steganography algorithm. Chinese Journal of Computers, 2024, 47(1): 213-230 (in Chinese)  
(李季瑀, 付章杰, 王帆. Canny-Gauss通用域图像隐写算法. 计算机学报, 2024, 47(1): 213-230)
- [25] Zhang Yi, Zhu Xiaodong, Qin Chuan, et al. Dither modulation based adaptive steganography resisting jpeg compression and statistic detection . Multimedia Tools and Applications, 2018, 77(14): 17913-35
- [26] Langelaar G C, Lagendijk R L, Biemond J. Removing spatial spread spectrum watermarks by non-linear filtering// Proceedings of the 9th European Signal Processing Conference (EUSIPCO 1998). Island of Rhodes, Greece, 1998: 1-4
- [27] Liang Luming, Deng Seng, Gueguen L, et al. Convolutional neural network with median layers for denoising salt-and-pepper contaminations . Neurocomputing, 2021, 442: 26-35



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

**ZHAO Ying-Ao**, M. S. candidate. Her main research interest is image steganalysis.

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

**ZHANG Qian-Qian**, Ph. D. , candidate, lecturer. Her main research interests include rough set and granular computing.

**LUO Xiang-Yang**, Ph. D. , professor. His main research interests include network information security, image steganography, steganalysis.

## Background

The content of this article belongs to the issue of steganography defense in the field of network information security. At present, research on steganography defense mainly focuses on passive steganography defense, and there is relatively little research on active steganography defense. This method can enable the receiver to successfully obtain secret messages even if they obtain the stego image, thus achieving the goal of active defense against image covert communication in social networks. This project has been supported by

the Henan Province Science Foundation for Youths (No. 222300420058), the National Natural Science Foundation of China (No. 6202002103, 62202495, 62272163, 62172435, U23A20305), the National Key Research and Development Program of China (No. 2022YFB3102900), the Central Plains Science and Technology Innovation Leading Talent Program (No. 214200510019), and the Key Research and Development Project of Henan Province (No. 221111321200).