中图法分类号:TP309.2　　**文献标识码**:A　　**文章编号**:1006-8961(2024)02-0355-14

论文引用格式:Ma B, Li K, Xu J, Wang C P, Li J and Zhang L W. 2024. High-security image steganography with the combination of multiple competition and channel attention. Journal of Image and Graphics, 29(02):0355-0368(马宾, 李坤, 徐健, 王春鹏, 李健, 张立伟. 2024. 联合多重对抗与通道注意力的高安全性图像隐写. 中国图象图形学报, 29(02):0355-0368)[DOI:10.11834/jig.230134]

联合多重对抗与通道注意力的高安全性图像隐写

马宾[1,2],李坤[1,2],徐健[3*],王春鹏[1,2],李健[1,2],张立伟[4]

1. 齐鲁工业大学(山东省科学院)网络空间安全学院,济南　250353; 2. 山东省计算机网络重点实验室,济南　250098;
3. 山东财经大学计算机科学与技术学院,济南　250014; 4. 积成电子股份有限公司,济南　250104

摘　要:**目的**　现有基于对抗图像的隐写算法大多只能针对一种隐写分析器设计对抗图像,且无法抵御隐写分析残差网络(steganalysis residual network,SRNet)、Zhu-Net等最新基于卷积神经网络隐写分析器的检测。针对这一现状,提出了一种联合多重对抗与通道注意力的高安全性图像隐写方法。**方法**　采用基于U-Net结构的生成对抗网络生成对抗样本图像,利用对抗网络的自学习特性实现多重对抗隐写网络参数迭代优化,并通过针对多种隐写分析算法的对抗训练,生成更适合内容隐写的载体图像。同时,通过在生成器中添加多个轻量级通道注意力模块,自适应调整对抗噪声在原始图像中的分布,提高生成对抗图像的抗隐写分析能力。其次,设计基于多重判别损失和均方误差损失相结合的动态加权组合方案,进一步增强对抗图像质量,并保障网络快速稳定收敛。**结果**　实验在BOSS Base 1.01数据集上与当前主流的4种方法进行比较,在使用原始隐写图像训练后,相比于基于U-Net结构的生成式多重对抗隐写算法等其他4种方法,使得当前性能优异的5种隐写分析器平均判别准确率降低了1.6%;在使用对抗图像和增强隐写图像再训练后,相比其他4种方法,仍使得当前性能优异的5种隐写分析器平均判别准确率降低了6.8%。同时也对对抗图像质量进行分析,基于测试集生成的2 000幅对抗图像的平均峰值信噪比(peak signal-to-noise ratio,PSNR)可达到39.925 1 dB,实验结果表明本文提出的隐写网络极大提升了隐写算法的安全性。**结论**　本文方法在隐写算法安全性领域取得了较优秀的性能,且生成的对抗图像具有很高的视觉质量。

关键词:隐写;隐写分析;对抗图像;通道注意力;生成对抗网络(GAN)

High-security image steganography with the combination of multiple competition and channel attention

Ma Bin[1,2], Li Kun[1,2], Xu Jian[3*], Wang Chunpeng[1,2], Li Jian[1,2], Zhang Liwei[4]

1. School of Cyber Security, Qilu University of Technology (Shandong Academy of Sciences), Jinan　250353, China;
2. Shandong Provincial Key Laboratory of Computer Networks, Jinan　250098, China;
3. School of Computer Science and Technology, Shandong University of Finance and Economics, Jinan　250014, China;
4. Integrated Electronic Systems Lab Co., Ltd., Jinan　250104, China

收稿日期:2023-03-22;**修回日期**:2023-08-10;**预印本日期**:2023-08-17

*通信作者:徐健　sdfixj@126.com

基金项目:国家自然科学基金项目(62272255);国家重点研发计划资助(2021YFC3340602);山东省自然科学基金创新发展联合基金项目(ZR2022LZH011);山东省科技型中小企业能力提升工程项目(2022TSGC2485);济南市带头人工作室项目(2020GXRC056);济南市引进创新团队项目(202228016);山东省高校青年创新团队项目(2022KJ124);教育部"春晖计划"科研合作项目(HZKY20220482);山东省自然科学基金项目(ZR2020MF054)

Supported by: National Natural Science Foundation of China (62272255); National Key R&D Program of China (2021YFC3340602); Shandong Provincial Natural Science Foundation Innovation and Development Joint Fund (ZR2022LZH011); Youth Innovation Team of Colleges and Universities in Shandong Province (2022KJ124); Natural Science Foundation of Shandong Province, China (ZR2020MF054)

Abstract: Objective The advancement of current steganographic techniques has been facing many challenges. The method of modifying the original image to hide the secret information is traceable, rendering it susceptible to detection by steganalyzers. The coverless steganographic method improves the security of steganography. However, it has limitations, such as small embedding capacity, large image database, and difficulty extracting secret information. The cover image generative steganography method also produces small and unnatural generated images. Introducing adversarial examples provides a new approach to address these limitations by adding subtle perturbations to the original image to form an adversarial image that is not visually distinguishable and causes wrong classification results to be outputted with high confidence. Thus, the security of image steganography is enhanced. However, most existing steganographic algorithms based on adversarial examples can only design adversarial samples for one steganalyzer, making them vulnerable to the latest convolutional neural network-based steganalyzers, such as SRNet and Zhu-Net. In response to this problem, a high-security image steganography method with the combination of multiple competition and channel attention is proposed in this study. **Method** In the proposed method, we generate the adversarial noise V using the generator G, which employs the U-Net architecture with added channel-attention modules. Subsequently, the adversarial noise V is added to the original image $\boldsymbol{X}$ to obtain the adversarial image. The pixel space minimum mean square error loss *MSE_loss* is adopted to train the generator network G. Thus, high-quality and semantically meaningful adversarial images are generated. Then, we generate the stego image from the original image $\boldsymbol{X}$ using the steganography network (SN) and input the original image $\boldsymbol{X}$ and its corresponding stego image into the steganalysis optimization network to optimize its parameters. Moreover, we build multiple steganalysis adversarial networks (SANs) to discriminate the original image $\boldsymbol{X}$ and its adversarial image and assign different scores to the adversarial and original images, providing multiple discriminant losses *SDO_loss1*. Furthermore, we embed secret messages into the adversarial image through the SN to generate the enhanced stego image. The adversarial image and the enhanced stego image are reinput into the optimized multiple steganalyzers to improve the antisteganalysis performance of the adversarial image. The SAN evaluates the data-hiding capability of the adversarial image and provides multiple discriminant losses *SDO_loss2*. Additionally, the weighted superposition of the MSE_loss, namely, the multiple steganalysis discrimination losses *SDO_loss1* and *SDO_loss2*, is employed as the cumulative loss function of generator G to improve the image quality of the adversarial image and its antisteganalysis ability. Finally, the proposed method enables fast and stable network convergence and high stego image visual quality and antisteganalysis ability. **Result** First, we select four high-performance deep-learning steganalyzers, namely, Xu-Net, Ye-Net, SRNet, and Zhu-Net, for simultaneous adversarial training to improve the antisteganalysis ability of adversarial images. However, simultaneously conducting experiments with four steganalysis networks may sharply increase the number of model parameters, resulting in slow training speed and long training period. Furthermore, each iteration of adversarial noise is generated according to the gradient feedback of the four steganalysis networks during the adversarial image generation process. A consequence of this approach is that the original image is subjected to excessive, unnecessary adversarial noise, leading to low-quality adversarial images. In response to this issue, we execute ablation experiments on different steganalysis networks employed in training. These experiments aim to decrease model parameters, reduce training time, and ultimately enhance the quality of adversarial images for their antisteganalysis capability improvement. The role of the generator is to produce adversarial noise, which is subsequently incorporated into the original image to generate adversarial images. Different positions of adversarial noise in the original image can cause distinct perturbations to the steganalysis network, and the quality of the generated adversarial images can be influenced differently. This study introduces ablation experiments by altering the addition of the channel attention module at various positions of the generator to examine the effectiveness of the channel attention module. The parameters of the generator loss function are fine-tuned by conducting the ablation experiment. Subsequently, we generate 2 000 adversarial images using the proposed model and evaluate the quality of these images. The results reveal that the average peak signal-to-noise ratio (PSNR) value of the 2 000 generated adversarial images is 39.925 1 dB. Furthermore, more than 99.55% of these images have a PSNR value greater than 39 dB, and more than 75% of the generated adversarial images have a PSNR value greater than 40 dB. Additionally, the average structural similarity index measure (SSIM) value of the generated adversarial images is 0.962 5. Among these images, more than 69.85% have an SSIM value greater than 0.955, and more than 55.6% of the adversarial samples have an SSIM value greater than 0.960. These results indicate that compared with

the original images, the generated adversarial images exhibit high visual similarity. Finally, we conduct a comparative study of the proposed method with the current state-of-the-art methods on the BOSS Base 1.01 dataset. The experiments are conducted on the BOSS Base 1.01 dataset, and the results are compared with those of the current state-of-the-art methods. Compared with the four methods, the five steganalysis methods show decreased average accuracy by 1.6% after training on the original steganographic images. Compared with other four methods, the five steganalysis methods show decreased average accuracy by 6.8% after further training with adversarial images and enhanced steganographic images. The experimental results indicate that the proposed steganographic method significantly improves the security of the steganographic algorithm. **Conclusion** In this study, we propose a steganographic architecture based on the U-Net framework with lightweight channel attention modules to generate adversarial images, which can resist multiple steganalysis networks. The experiment results demonstrate that the security and generalization of the algorithm we propose exceed those of the compared steganographic methods.

Key words: steganography; steganalysis; adversarial images; channel attention; generative adversarial network(GAN)

0 引　言

图像隐写术作为信息隐藏(Petitcolas 等,1999)的一个分支,一直受到广泛关注,它旨在以一种不可察觉的方式将秘密信息嵌入到载体中。根据不同的隐写机制,现有的图像隐写术可以分为基于原始图像嵌入的隐写术、无载体隐写术和生成对抗图像隐写术。

基于原始图像嵌入的隐写术可以根据嵌入域的不同分为空间域隐写术和变换域隐写术。空间域隐写术通过修改载体图像的像素值来隐藏秘密信息。一些常用的空间域隐写术包括最低有效位(least significant bit,LSB)隐写术(Mielikainen,2006)、高难度检测(highly-undetectable stegosystems,HUGO)隐写术(Pevný 等,2010)和小波求权(wavelet obtained weights,WOW)隐写术(Holub 和 Fridrich,2013)。在载体图像的不同位置进行嵌入可能会产生不同的影响,因此选择适当的嵌入位置非常重要。为了解决这个问题,Filler等人(2011)提出了最小化加性失真的校验子格编码(syndrome trellis codes,STC),可以与任何加性失真代价函数结合使用来开发隐写方法。此后,研究人员专注于提升失真函数的安全性。另一方面,变换域隐写术通过修改载体图像的频域系数来隐藏秘密信息。一些典型的变换域隐写术方法包括通用小波相对畸变(JPEG universal wavelet relative distortion,J-UNIWARD)隐写术(Holub 等,2014)、均匀嵌入失真度量(uniform embedding distortion metric,UED)隐写术(Guo 等,2012)以及尹晓琳等人(2022)的方法。基于原始图像嵌入的隐写术方法因为通过改变像素值或变换域中的系数将秘密数据嵌入到载体图像中,所以不可避免地会给载体图像带来失真,这使得它们难以抵御基于统计分析的隐写分析器的检测。

为了解决基于原始图像嵌入的隐写术的限制,研究者提出了无载体图像隐写术。传统的无载体隐写方法不需要修改载体,而是通过设计图像特征和秘密数据之间的映射规则(Zhang 等,2018a;Luo 等,2021)或使用特定算法将秘密数据合成到图像纹理中来隐藏秘密信息(Wu 和 Wang,2015;Liu 等,2022)。尽管无载体图像隐写术具有高安全性的优点,但仍然存在一些需要克服的挑战,例如嵌入容量低、需要大量图像数据库和图像质量不理想等问题。最近,基于深度学习的无载体方法(Chen 等,2022;Peng 等,2022)已被开发用于提高嵌入容量,但它们仍然存在图像失真和秘密信息提取准确率低等问题。

随着深度学习的发展,基于卷积神经网络(convolutional neural network,CNN)的图像隐写分析已经取得了很高的性能,例如 Xu-Net(Xu 等,2016)、Ye-Net(Ye 等,2017)、SRNet(steganalysis residual network)(Boroumand 等,2019)、Zhu-Net(Zhang 等,2020),检测性能已经超越了空域富模型(spatial rich model,SRM)(Fridrich 和 Kodovsky,2012)这种优异的传统隐写分析器。传统的图像隐写方法已经无法抵抗这些隐写分析器的检测,因此迫切需要增强图像隐写术的安全性。受深度学习算法的启发,Goodfellow 等人(2014)提出了生成对抗网络(generative adversarial network,GAN),为图像隐写方法带来了新的灵感。一系列基于 GAN 的图像隐写方法

(郑钢 等,2021;Tan 等,2022)开始涌现并表现良好。对抗样本的提出也为图像隐写提供了新的思路:通过对载体图像添加细微的对抗噪声扰动来构造一个新的对抗图像,这种扰动在视觉上是无法察觉到的。将基于对抗图像生成的隐写图像输入到隐写分析器中,隐写分析器输出具有很高置信度的错误结果。然而,向隐写图像添加对抗噪声可能会破坏原始分布,并影响由编码和解码算法的性质(如纠错码)导致的秘密信息提取效果。为了解决这个问题,现有的大多数基于对抗图像的隐写方法在信息隐藏之前向载体图像添加对抗噪声。

为提高现有隐写方法的安全性,本文提出一种联合多重对抗与通道注意力的高安全性图像隐写方法。原始图像可经过本文方法生成适合隐写的对抗图像,使用该对抗图像生成的隐写图像可抵御基于深度学习和特征统计的隐写分析器的检测。本文模型由生成器、隐写器、多重隐写分析器网络和通道注意力模块构成,且受生成对抗网络的启发,本文模型设计了隐写分析优化网络、隐写分析对抗网络,通过网络间的对抗学习提高隐写图像抗隐写分析检测的能力,生成更适合隐写的对抗图像。本文选用U-Net(Ronneberger 等,2015)作为生成器网络框架来生成对抗图像,同时为了更好地调整对抗噪声在原始图像中的分布,引入多个压缩激励网络(squeeze-and-excitation networks,SENet)(Hu 等,2018)通道注意模块来显式地建模通道依赖关系,这使网络能够将对抗噪声集中在更关键和有效的通道特征中,提高抗隐写分析能力和对抗图像的视觉质量。

本文工作的主要贡献如下:1)设计了一种同时对抗多重隐写分析器的训练网络,首先对多重隐写分析器网络进行性能优化,采用优化后的隐写分析器构建隐写分析对抗网络,通过生成网络与多重隐写分析对抗网络间的多重对抗迭代训练,生成对隐写分析网络扰动最大的对抗图像。2)发现通道注意力机制对提升生成对抗图像视觉质量和抗隐写分析能力具有促进作用。通道注意力模块可以在图像的网络激活中学习通道相关性并自适应调整通道特征,它使对抗噪声专注于嵌入对神经网络扰动最大、对原始图像改动最小的位置。3)不同于相关方法使用已训练好的隐写分析器来获得对抗图像,本文采用隐写分析器与生成器同时训练的策略,对抗图像随着隐写分析器参数的更新也不断迭代优化。同时引入均方误差损失,在保证抗隐写分析能力的前提下,提升对抗图像视觉质量。

1 相关工作

根据隐写图像生成阶段的不同,对抗样本在图像隐写中的应用可分为两种。第1种如图1所示,是基于隐写图像的对抗样本,它是指在隐写图像中添加精心构造且人眼难以察觉的细微干扰而形成的样本,这些样本会导致基于神经网络构造的隐写分析器以高置信度输出错误的分类结果;第2种如图2所示,是基于原始图像的对抗样本,它是指在原始载体图像中添加精心构造且人眼难以察觉的细微干扰而形成的样本,使用这些样本作为新的载体图像嵌入秘密信息生成的隐写图像会使基于神经网络构造的隐写分析器以高置信度输出错误的分类结果。第1种对抗样本因为在隐写图像中添加扰动,可能破坏了秘密信息在隐写图像中的原有分布,导致提取秘密信息准确率下降,所以一般采用第2种对抗样本进行隐写研究。

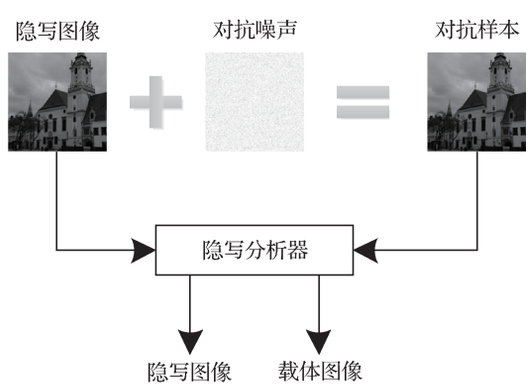

图1 基于隐写图像的对抗样本

Fig. 1 The adversarial example based on stego image

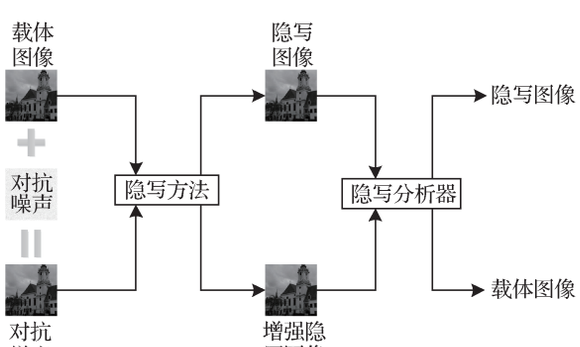

图2 基于原始图像的对抗样本

Fig. 2 The adversarial example based on original image

Zhang 等人(2018b)提出了一种基于对抗样本的图像隐写算法,利用快速梯度下降模型(fast gradient sign model,FGSM)来将输入的噪声迭代生成对抗载体图像,并采用经典的自适应隐写算法嵌入秘密信息,并可针对多个隐写分析生成对抗图像。

Zhou 等人(2020)采用全卷积神经网络(fully convolutional neural network,FCN)作为载体图像生成器,构建了一种可以快速生成对抗载体图像的网络模型,其设计的新的损失函数,使得对抗载体图像和隐写图像能够欺骗隐写网络的分析。该隐写模型一定程度上提高了载体图像的生成速度和质量,增强了隐写图像的安全性。

Liu 等人(2021)提出了一种增强现有隐写方法安全性的对抗嵌入隐写方法,首先结合多个载体图像梯度和生成的隐写图像来确定成本修改的方向。然后该方法并没有调整全部或随机部分的嵌入成本,而是根据载体图像梯度的幅度及其成本仔细选择候选成本。通过调整一小部分嵌入成本,可以显著提高在重新训练的基于 CNN 和传统隐写分析器上评估的现代隐写方法的安全性。此外,在不同图像数据库上的安全性能评估表明,该方法的泛化性良好。

马宾等人(2023)提出了一种基于 U-Net 结构的生成式多重对抗隐写算法。该算法解决了现有基于生成式对抗网络的隐写算法存在的生成图像质量尺寸小、内容不可控的问题。该方法利用了基于 U-Net 的生成网络模型,将参考图像中的详细信息传输到生成的载体图像中。该方法可控地生成较高质量的目标载体图像,从而增强载体图像的信息隐藏能力。

然而,这些工作虽然都取得了优秀的性能表现,但仍存在可改进之处。Zhang 等人(2018b)的隐写模型没有训练生成器,需要对每幅图像进行迭代操作来生成对抗图像,当面对大量需要嵌入秘密信息的图像时,耗费时间巨大,因而该模型只适用于少量载体图像的情况。Zhou 等人(2020)的模型只能针对一种隐写分析器生成对抗图像,且其抗隐写分析能力鲁棒性不够理想。Liu 等人(2021)的方法因为在生成对抗图像的过程中缺少迭代过程,所以在安全性方面弱于其他方法。马宾等人(2023)的方法直接使用生成器生成一幅对抗图像,没有添加对抗噪声的过程,所以生成图像质量较差,平均峰值信噪比(peak signal-to-noise ratio,PSNR)仅有 36 dB 左右。针对这些不足之处,本文通过引入通道注意力机制和多重对抗思想,训练了一个可以快速有效地生成大量鲁棒对抗图像的生成器,有效提升了图像隐写术的安全性。

2 提出的方法

受生成对抗网络判别器与生成器的对抗训练和对抗样本对神经网络的高干扰性的启发,提出了一种基于生成对抗图像的提升图像隐写术安全性方法。首先,概述了模型体系结构和基本思想,然后详细描述了模型各组成部分的组成。最后,说明了各网络的损失函数。

2.1 模型的总体概述

如图 3 所示,本文模型由 3 部分组成:1)生成器 G,输入原始图像 $\boldsymbol{X}$,输出对抗扰动噪声 $\boldsymbol{V}$,将对抗扰动噪声 $\boldsymbol{V}$ 添加到原始图像中生成对抗图像 $\boldsymbol{X}_V$;2)隐写器 SN,输入原始图像 $\boldsymbol{X}$ 或对抗图像 $\boldsymbol{X}_V$,输出隐写图像 $\boldsymbol{X}_S$ 或增强隐写图像 $\boldsymbol{X}_{VS}$;3)多重隐写分析网络 SD(steganalysis discriminator),它试图通过给隐写图像和载体图像分配不同的分数来区分二者,其包含两个子网络:隐写分析优化网络 SON(steganalysis optimization network)和隐写分析对抗网络 SAN(steganalysis adversarial network)。

2.2 生成器网络结构

当前存在很多性能优异的生成模型。其中,VAE 代表变分自编码器(variational autoencoder),可以从输入数据中学习潜在变量,并生成新的样本。它主要用于数据的增广分布,例如生成新的图像、视频、声音和文本等。U-Net 与 VAE 类似,都是编码解码(encoder-decoder)结构,该网络由相互对称的用于获得前后文信息的收缩路径和用于精确定位的扩展路径组成。U-Net 通过同层跳接的网络结构,可以将原始图像的细节信息传递到生成模型中,从而实现图像高品质重建。

在本文方法中,生成器生成的不是图像,而是对抗噪声,对抗噪声添加到原始图像中生成对抗图像,对抗噪声在原始图像中分布位置的不同也会大大影响生成对抗图像的图像质量和抗隐写分析能力,需要精准地定位对抗噪声的分布位置。因此,本文采用 U-Net 网络结构,在加入对抗噪声的同时,减

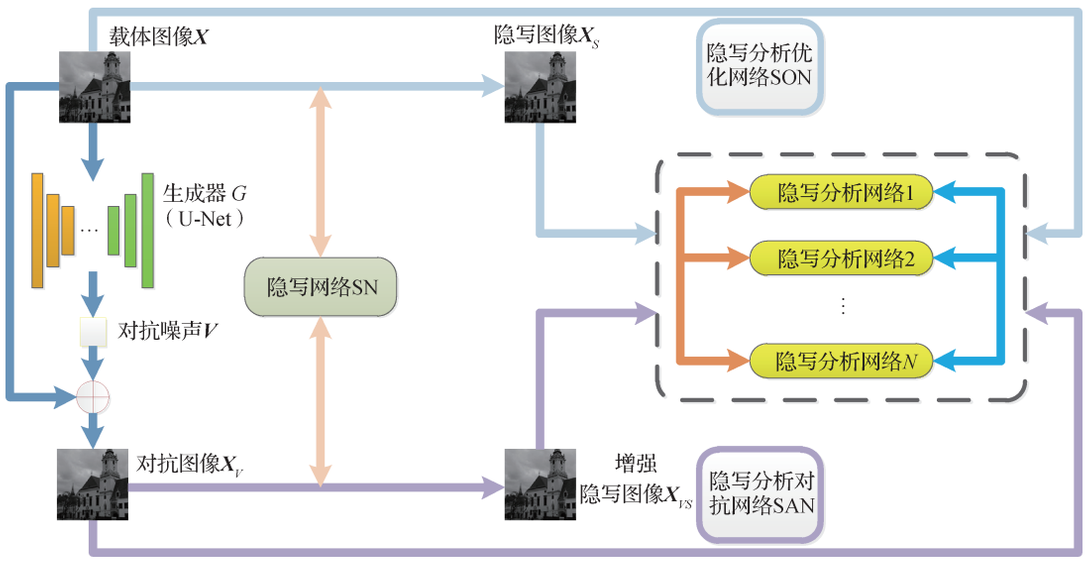

图3　模型总体框架

Fig. 3　The framework of the model

少对原始图像的扰动，从而生成更适合信息隐写的载体图像。如图4所示，本文使用U-Net构建生成器网络结构。通过调整对抗噪声的大小，可以将其强度控制在一定范围内。最后，本文将对抗噪声添加到原始图像，并将像素值约束在特定范围内，生成对抗图像。

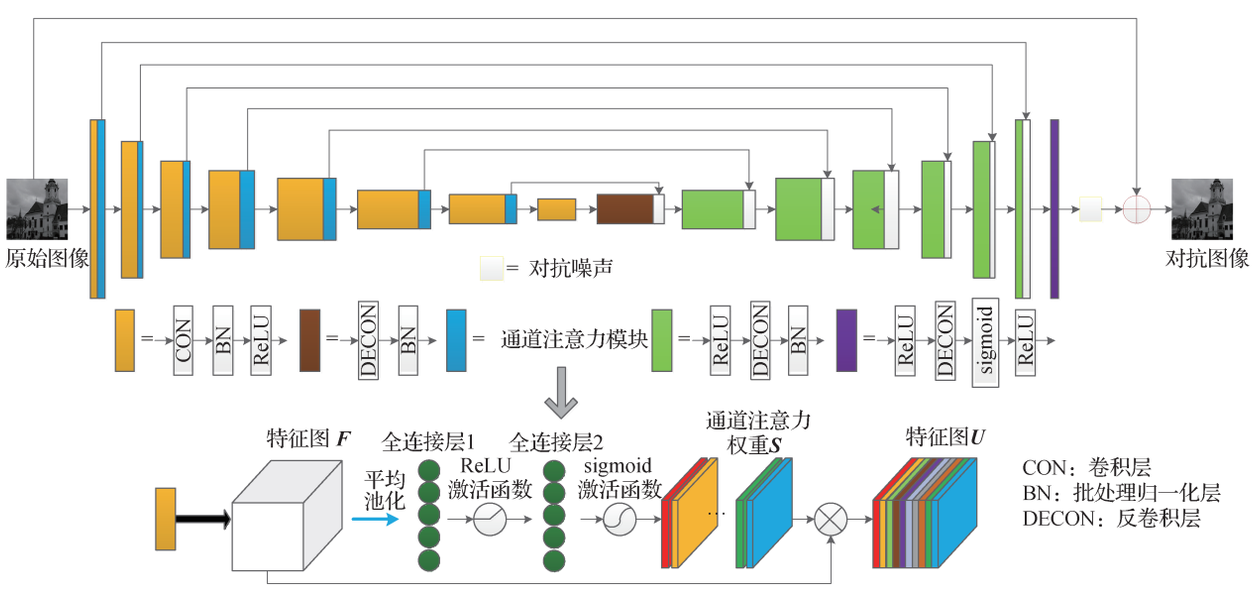

图4　生成器网络结构

Fig. 4　The network structure of the generator

2.3　隐写方法

近年来，随着STC隐写码的问世，自适应隐写（adaptive steganography）框架在主流隐写算法中得到广泛应用。这一框架的核心是结合“STC编码+代价函数”，通过代价函数计算每个载体元素的修改代价，并在完成消息嵌入的同时运用STC编码算法，以最小化总修改代价。实验中分别采用两种性能优异的主流隐写方法来生成嵌入代价。其中使用生成对抗网络的自动隐写失真学习框架（automatic steganographic distortion learning framework with GAN，ASDL-GAN）（Tang等，2017）模拟了加性失真的隐写术和基于深度学习的隐写分析之间的对抗竞争。在

ASDL-GAN 框架下，学习到的失真函数与对抗隐写分析器的不可检测性直接相关。UT-GAN（Yang 等，2019）跟踪了 ASDL-GAN 的研究，将生成器替换为高效的 U-Net 结构。此外，通过引入无需预训练的 Tanh-simulator 函数，在不损失安全性能的前提下，使得训练时间大大减少。

2.4 通道注意力模块

深度学习中的注意力机制使网络学习关注重要特征而忽略不相关的特征。在生成对抗图像的过程中，对抗噪声以与原始图像特征融合的方式嵌入。这些特征最终对隐写分析网络的扰动重要性不同，因此注意力机制可能有助于提高对抗图像抗隐写分析能力，同时在不同通道合理位置添加对抗噪声对生成对抗图像质量的影响也不同，通道注意力机制可能有助于提升生成对抗图像质量。硬注意力机制选择输入数据元素的一个子集，完全丢弃其余元素。由于其不可微性，它通常与强化学习相关联。在生成对抗图像的过程中，载体图像的信息需要尽可能保留，而不是丢弃。此外，强化学习的训练往往效率低下。自我注意力机制捕获计算机视觉领域中图像块之间的内在关联。由于对抗图像的生成考虑了整体图像信息，因此对其的直接贡献很小。因此本文选择软注意力机制，它为特征分配一个介于 0 和 1 之间的权重，以指示需要注意的程度。

软注意力机制主要包括空间注意力和通道注意力。前者允许网络找到合适的图像区域，而后者有助于在特征图中聚焦于有利的通道。空间注意力使用注意力模型生成一个掩膜，表明原始图像的注意敏感性，这是一种空间注意。掩膜中的值越大，意味着相应像素的变化将导致视觉检测的风险越高。然而，空间注意力显示的注意力不太敏感的区域并不是复杂的纹理或边缘区域，它们被认为对自适应隐写术是安全的。由于卷积层本身具有边缘检测和纹理提取的效果，本文认为附加空间注意力模型的功能是有限的，这促使本文研究通道注意力对图像隐写术的影响。在本文方法中，卷积层是基本的构建块。它将图像转换为多通道特征图，以便在这些通道特征中加入对抗噪声。在处理输入特征图时，经典卷积运算无法捕获通道内的整体信息和通道之间的依赖关系，导致输出特征图中出现一些无意义的通道。在输出的对抗图像中，无意义的通道可能进一步转化为不必要的噪声，这对对抗图像的图像质量不利。因此，应强调重要的通道，抑制无意义的通道。为此，本文引入了 SENet（squeeze-and-excitation networks）通道注意力模块，根据通道的重要性调整通道。将输入特征图 $\boldsymbol{F}$ 表示为 $\boldsymbol{F} \in \mathbf{R}^{M \times H \times W}$，它首先利用通道间的相互依赖性来推导权重向量。每个权重反映了每个通道的重要性。然后将权重乘以相应的通道以缩放特征，输出重新校准的特征图 $\boldsymbol{U}$。通道注意力模块的结构如图 4 的下半部分所示。SENet 首先使用平均池化在特征图 $\boldsymbol{F}$ 的每个通道中聚合空间信息以获得 $\boldsymbol{f}_{\text{avg}} \in \mathbf{R}^{M \times 1}$，具体为

$$\boldsymbol{f}_{\text{avg}}^{m} = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} \boldsymbol{F}_{i,j}^{m} \tag{1}$$

式中，m 表示 $\boldsymbol{f}_{\text{avg}}$ 中的第 m 个元素，(i, j) 表示坐标。池化操作将每个通道的全局信息压缩为一个标量，作为空间特征统计。为了从这些统计信息中得出表示每个通道重要性的权重，SENet 对它们执行线性和非线性操作。具体来说，SENet 使用由两个全连接层组成的共享网络来传播 $\boldsymbol{f}_{\text{avg}}$。接下来通过 sigmoid 函数将特征向量转换为通道权重向量。即权重向量 $\boldsymbol{s} \in \mathbf{R}^{M \times 1}$ 计算为

$$\boldsymbol{s} = \sigma(\boldsymbol{W}_2(\delta(\boldsymbol{W}_1 \boldsymbol{f}_{\text{avg}}))) \tag{2}$$

式中，δ 和 σ 分别代表 ReLU 激活函数和 sigmoid 激活函数，$\boldsymbol{W}_1$ 和 $\boldsymbol{W}_2$ 指每层的权重。需要注意的是，隐藏层对输入进行降维，以平衡模型的性能和计算复杂性。本文参考相关通道注意力机制的研究，将折断系数 τ 设置为 16，这是最佳配置。最后，$\boldsymbol{s}$ 的每个元素（作为标量）乘以 $\boldsymbol{F}$ 的每个通道以计算 $\boldsymbol{U} = [\boldsymbol{U}^1, \boldsymbol{U}^2, \cdots, \boldsymbol{U}^m]$。$\boldsymbol{U}$ 的第 m 个通道的计算可表示为

$$\boldsymbol{U}^m = \boldsymbol{s}^m \boldsymbol{F}^m \tag{3}$$

这样，通过与较低权重相乘来抑制无用通道，反之亦然。修改通道后，特征图 $\boldsymbol{U}$ 具有更强的与对抗噪声融合的能力。

2.5 损失函数

在本文网络中，生成器 G 输入原始图像，输出对抗噪声，对抗噪声再添加到原始图像中生成对抗图像，并使用隐写分析网络 SD 判别生成的对抗图像是否为原始图像，输出概率越大，生成对抗图像越倾向于被认定为原始图像。多重隐写分析网络的训练目标是：当输入为对抗图像时，期望输出概率接近于 0；当输入原始图像时，期望输出概率接近于 1。它的损失函数 L_{SD_1} 可以表示为

$$L_{\mathrm{SD}_1}=-\sum_{i=1}^{2}x_i'\log(x_i) \tag{4}$$

式中，x_1和x_2分别是原始图像和生成对抗图像的概率，x_1'和x_2'对应的则是原始图像和生成对抗图像的标签。

此外，多重隐写分析网络SD判别生成的对抗图像与基于对抗图像生成的增强隐写图像的概率，输出概率越大，增强隐写图像越倾向于被认定为对抗样本（隐写分析器认为其为原始图像）。多重隐写分析网络的训练目标为：当输入为增强隐写图像时，期望输出概率接近于0；当输入对抗图像时，期望输出概率接近于1。它的损失函数L_{SD_2}可以表示为

$$L_{\mathrm{SD}_2}=-\sum_{i=1}^{2}y_i'\log(y_i) \tag{5}$$

式中，y_1和y_2是隐写分析器SD最后经过softmax层的输出，分别是对抗样本与增强隐写图像的概率，y_1'和y_2'分别是输入的对抗样本和增强隐写图像对应的标签。

在本文方法中，需要保证生成的对抗样本$\boldsymbol{X}_V$和原始图像$\boldsymbol{X}$的视觉不可区分性。为了实现这一目标，本文用均方差损失（*MSE_loss*）来表示图像失真损失，具体为

$$L_m=MSE\_loss(\boldsymbol{X},\boldsymbol{X}_V)=\frac{1}{C\times H\times W}\left\|\boldsymbol{X}-\boldsymbol{X}_V\right\|_2^2 \tag{6}$$

本文方法的生成器通过生成对抗图像来干扰多重隐写分析器的判别，输出错误的结果，所以两个判别损失L_{SD_1}和L_{SD_2}取相反数相加作为生成器总损失的一部分。同时，为了提升生成对抗图像的视觉质量，像素空间均方差损失*MSE_loss*添加到总损失中，促使生成的对抗图像取得更理想的PSNR值。最终生成器G的总损失为

$$\begin{aligned}L_G=&k_1\left(-\alpha\times L_{\mathrm{SD}_1}-\beta\times L_{\mathrm{SD}_2}\right)+\cdots+\\&k_n\left(-\alpha\times L_{\mathrm{SD}_1}-\beta\times L_{\mathrm{SD}_2}\right)+\lambda\times L_m\end{aligned} \tag{7}$$

式中，n代表本文训练过程中同时对抗的隐写分析器数量，k，α，β，λ均为权重系数。本文研究初始开展时同时对抗四重隐写分析网络，训练速度缓慢，生成对抗样本图像质量较差，经过消融实验和损失函数权重调整将隐写分析器数量降为2。

本文方法在初始构思时只是将生成器生成对抗图像，隐写分析器判别输入图像是原始图像还是对抗图像做了对抗训练。但在后续实验中发现，使用这种方式生成的对抗图像在嵌入秘密信息后抗隐写分析能力大大下降，分析认为嵌入的秘密信息破坏了原有对抗图像中的对抗噪声的分布。于是，在本文方法中，将对抗图像嵌入秘密信息生成隐写图像的过程和隐写分析器判别输入图像是载体图像还是隐写图像做了对抗训练。通过本文提出的损失函数也可以看出，原始图像和对抗图像，以及对抗图像和增强隐写图像都会输入到隐写分析器中，通过对抗训练让网络学习，将秘密信息尽可能嵌入到不会影响对抗图像抗隐写能力的位置。

3 实验结果与分析

为了通过实验验证本文方法，选择当前流行的用于隐写基准测试的BOSS Base数据集。它由10 000幅未压缩的大小为512×512像素的灰度图像组成，本文方法将其降维为256×256像素的图像集，并随机选取10 000幅开展研究，训练集和验证集分别为8 000幅和2 000幅。秘密信息嵌入步骤使用了两种典型的隐写方法，即ASDL-GAN和UT-GAN。本文使用了4种基于CNN的隐写分析器，即Xu-Net、Ye-Net、SRNet和Zhu-Net以及一种传统的基于特征统计的隐写分析器SRM。基于各自论文中描述的实验结果和本文的实验验证，这些隐写分析器的检测能力按以下顺序排列：Zhu-Net > SRNet > Ye-Net > SRM > Xu-Net。使用Adam作为生成器的优化器，学习率 = 0.000 1，beta_1 = 0.5，beta_2 = 0.99。隐写方法和隐写分析网络的优化器及其使用的参数在原始论文中给出。

3.1 隐写分析网络消融实验

为提升对抗图像抗隐写分析能力，本文初始选择同时对抗Xu-Net、Ye-Net、SRNet、Zhu-Net 4种高性能深度学习隐写分析器进行训练。考虑到同时对抗4重隐写分析网络会使模型参数量急剧增加，导致训练速度缓慢、训练周期长，且生成对抗图像的过程中对抗噪声每次迭代要按照4重隐写分析网络的梯度反馈来叠加生成，这可能会导致有大量冗余的对抗噪声被添加到原始图像中，最终使得生成的对抗图像质量较差。

为了在保证对抗图像抗隐写分析能力良好的前提下，减少模型参数量，缩短训练时间，提高生成对抗图像的视觉质量，本文在实验中对训练时同时对抗的隐写分析网络数量进行了消融实验。如表1所

示，本文分别采用 Xu-Net + Ye-Net + SRNet + Zhu-Net、Ye-Net + SRNet + Zhu-Net、Ye-Net + SRNet、Ye-Net + Zhu-Net、SRNet + Zhu-Net、SRNet、Zhu-Net 的隐写分析网络结构来进行训练。实验中采用 UT-GAN 作为隐写方法，嵌入容量为 0.4 bit/像素。

表 1 显示了消融实验结果，其中“原始隐写”表示隐写分析器识别原始隐写图像的准确率，“增强隐写”表示隐写分析器识别基于对抗图像生成的隐写图像的准确率。最优结果应为 0.5，这表明隐写分析器无法区分输入图像是载体图像还是隐写图像。表 1 数据表明，选择 SRNet + Zhu-Net 作为隐写分析网络的组合进行训练可以实现隐写安全和图像质量之间的最佳平衡。因此，本文将隐写分析网络数量 n 设置为 2，并选择此组合进行后续实验。由实验结果可知，针对 SRNet 和 Zhu-Net 生成的对抗图像，在面对 Xu-Net、Ye-Net 的检测时也可以取得很好的抗隐写分析效果，这表示我们的模型具有很好的泛化性，即使面对未知的隐写分析器的检测时也可以取得很好的效果。

表 1　隐写分析网络消融实验结果

Table 1　The results of steganalytic networks ablation experiments

不同隐写分析网络组合	平均 PSNR/dB	平均 SSIM	Xu-Net/%		Ye-Net/%		SRNet/%		Zhu-Net/%		训练时间/ min
			增强隐写	原始隐写	增强隐写	原始隐写	增强隐写	原始隐写	增强隐写	原始隐写	
Xu-Net + Ye-Net + SRNet + Zhu-Net	34.893 7	0.873 6	50.10	84.90	50.20	86.60	49.60	89.10	50.50	89.40	1 956
Ye-Net + SRNet + Zhu-Net	37.198 6	0.910 1	50.10	84.90	49.80	86.60	50.40	89.10	50.50	89.40	1 745
Ye-Net + SRNet	38.201 9	0.937 9	49.60	84.90	50.40	86.60	50.60	89.10	52.30	89.40	1 328
Ye-Net + Zhu-Net	39.219 4	0.955 9	50.50	84.90	49.70	86.60	52.10	89.10	50.80	89.40	1 289
SRNet + Zhu-Net	39.925 1	0.960 1	50.20	84.90	50.40	86.60	49.60	89.10	50.50	89.40	1 469
SRNet	43.391 1	0.979 2	49.60	84.90	49.10	86.60	50.80	89.10	52.80	89.40	996
Zhu-Net	43.045 8	0.978 2	51.20	84.90	51.10	86.60	52.50	89.10	50.90	89.40	903

3.2　通道注意力模块消融实验

在深度学习中使用通道注意力机制，可以让网络学会强调关键的特征而忽略不相关的特征。在生成对抗图像时，对抗噪声与原始图像特征相结合，这些特征对对抗图像具有不同的意义。因此，使用通道注意力机制可以增强对抗图像的抗隐写分析能力。此外，在不同的通道中加入对抗噪声对图像视觉质量的影响是不同的，因此，使用通道注意力机制可以潜在地提高对抗图像的质量。

本文进行消融实验来评估通道注意力添加位置对本文提出的模型的影响，本文测试了 4 种变体：1）在编码阶段添加通道注意力模块；2）在解码阶段添加通道注意力模块；3）在编码和解码阶段都添加通道注意力模块；4）不添加通道注意力模块。信息嵌入方法为 UT-GAN，嵌入容量为 0.4 bit/像素。实验结果如表 2 所示，当选择在编码阶段加入通道注意力模块时，对抗图像质量和抗隐写分析能力都达到了最佳效果。

表 2　通道注意力模块消融实验结果

Table 2　The results of channel attention modules ablation experiments

添加位置	PSNR /dB	Xu-Net /%	Ye-Net /%	SRNet /%	Zhu-Net /%
不添加	39.052 3	50.4	50.6	49.4	50.6
编码阶段	**39.925 1**	**50.2**	**50.4**	**49.6**	**50.5**
解码阶段	39.429 4	49.5	50.5	50.6	50.7
编码阶段 + 解码阶段	39.530 8	50.3	49.4	50.5	**49.5**

注：加粗字体表示各列最优结果。

3.3　生成器损失函数权重实验

3.3.1　MSE 损失 L_m 权重 λ

采用均方差（mean squared error，MSE）作为生成网络的损失参数时，损失的权值也会对生成对抗样本的图像质量和抗隐写分析能力产生影响。表 3 为不同 λ 取值下生成对抗图像的 PSNR 和结构相似性（structural similarity，SSIM）评价结果以及相应 4 种

隐写分析网络的准确率。由表3可以看出，当λ = 0.2时，生成图像平均PSNR为39.355 6 dB，SSIM为0.960 0，使Xu-Net、Ye-Net、SRNet和Zhu-Net的准确率也分别达到了49.8%、50.3%、49.5%和50.4%，此时生成对抗图像质量最优，且抗隐写分析能力也较强，本文认为达到了最优平衡，是最理想的权重。因而，实验中将生成器损失函数中*MSE_Loss*的权重λ设置为0.2。

表3　权重λ实验结果

Table 3　The results of weight λ

λ	PSNR /dB	SSIM	Xu-Net /%	Ye-Net /%	SRNet /%	Zhu-Net /%
0.1	36.451 0	0.908 3	50.30	50.50	49.50	50.70
0.2	**39.355 6**	**0.960 0**	**49.80**	**50.30**	**49.50**	**50.40**
0.3	37.798 9	0.911 5	50.40	49.40	50.60	49.30
0.4	38.605 4	0.937 9	49.30	50.70	49.40	50.80
0.5	36.237 7	0.905 4	50.40	49.30	50.70	50.60
0.6～1.0	失去安全性					

注：加粗字体表示各列最优结果。

3.3.2　判别损失权重α和β

表4展示了不同α和β设置下，达到的最优PSNR与SSIM值以及4种隐写分析网络的准确率。由表4可以看出，在α = 0.1，β = 0.9时，对抗图像的PSNR达到了39.431 2 dB，SSIM达到了0.962 0，4种隐写分析网络的准确率分别达到50.3%、49.6%、50.4%和50.3%，此时生成对抗图像质量最优，且抗隐写分析能力也较强，本文认为达到了最优平衡，是最理想的判别损失权重。因而，实验中将判别损失权重设置为α = 0.1，β = 0.9。

表4　权重α和β实验结果

Table 4　The results of weight α and β

判别损失权重	PSNR /dB	SSIM	Xu-Net /%	Ye-Net /%	SRNet /%	Zhu-Net /%
α=0.5 β=0.5	39.351 6	0.961 6	50.4	**50.4**	**49.6**	50.6
α=0.4 β=0.6	38.994 3	0.957 2	49.4	50.8	49.3	50.5
α=0.3 β=0.7	39.287 5	0.961 4	50.5	49.4	50.5	49.3
α=0.2 β=0.8	39.137 5	0.959 4	49.3	50.7	49.4	50.8
α=0.1 β=0.9	**39.431 2**	**0.962 0**	**50.3**	**49.6**	**50.4**	**50.3**

注：加粗字体表示各列最优结果。

3.3.3　SRNet和Zhu-Net权重分配k_1和k_2

本文在前面对多重隐写分析网络结构做了消融实验，最终只留下SRNet和Zhu-Net，在生成器损失函数中，SRNet和Zhu-Net也存在权重分配问题，SRNet损失对应权重为k_1，Zhu-Net损失对应权重为k_2。通过对两种隐写分析网络分配不同的权重，生成对抗图像对两种隐写分析网络的扰动程度也不相同。表5展示了SRNet、Zhu-Net不同权重分配下生成对抗图像的图像质量和对不同隐写分析网络的扰动效果。本文期望通过调整权重分配来使生成的对抗图像对每个隐写分析网络都达到很好的扰动效果。由表5数据可以看出，当k_1 = 0.5，k_2 = 0.5时，生成对抗样本对4种隐写分析网络的扰动效果总体上

表5　权重k_1和k_2实验结果

Table 5　The results of weight k_1 and k_2

k_1和k_2	PSNR/dB	SSIM	Xu-Net/%	Ye-Net/%	SRNet/%	Zhu-Net/%
k_1 = 0.5 k_2 = 0.5	40.355 6	**0.962 5**	**50.20**	**50.30**	**50.50**	49.50
k_1 = 0.6 k_2 = 0.4	40.326 2	0.959 0	50.30	**50.30**	49.30	50.80
k_1 = 0.4 k_2 = 0.6	**40.566 5**	0.960 2	49.70	49.60	50.60	49.60
k_1 = 0.7 k_2 = 0.3	39.584 7	0.949 8	50.50	50.40	49.20	50.70
k_1 = 0.3 k_2 = 0.7	39.360 8	0.946 6	49.40	49.50	50.70	50.90
k_1 = 0.8 k_2 = 0.2	37.482 1	0.922 1	49.50	50.60	49.40	**49.70**
k_1 = 0.2 k_2 = 0.8	37.352 8	0.921 6	50.40	50.60	50.70	50.90

注：加粗字体表示各列最优结果。

最强，同时具有良好的图像质量。

3.4 对抗图像质量分析

本节中，使用之前实验确定的隐写分析网络数量和参数来训练模型，然后基于测试集生成2 000幅对抗图像。图5显示了原始图像和生成对抗图像之间的差异。为了更加细致地观察这些差异，本文还展示了它们相应的直方图。如图5所示，原始图像和对抗图像非常相似，人类视觉系统无法区分它们。

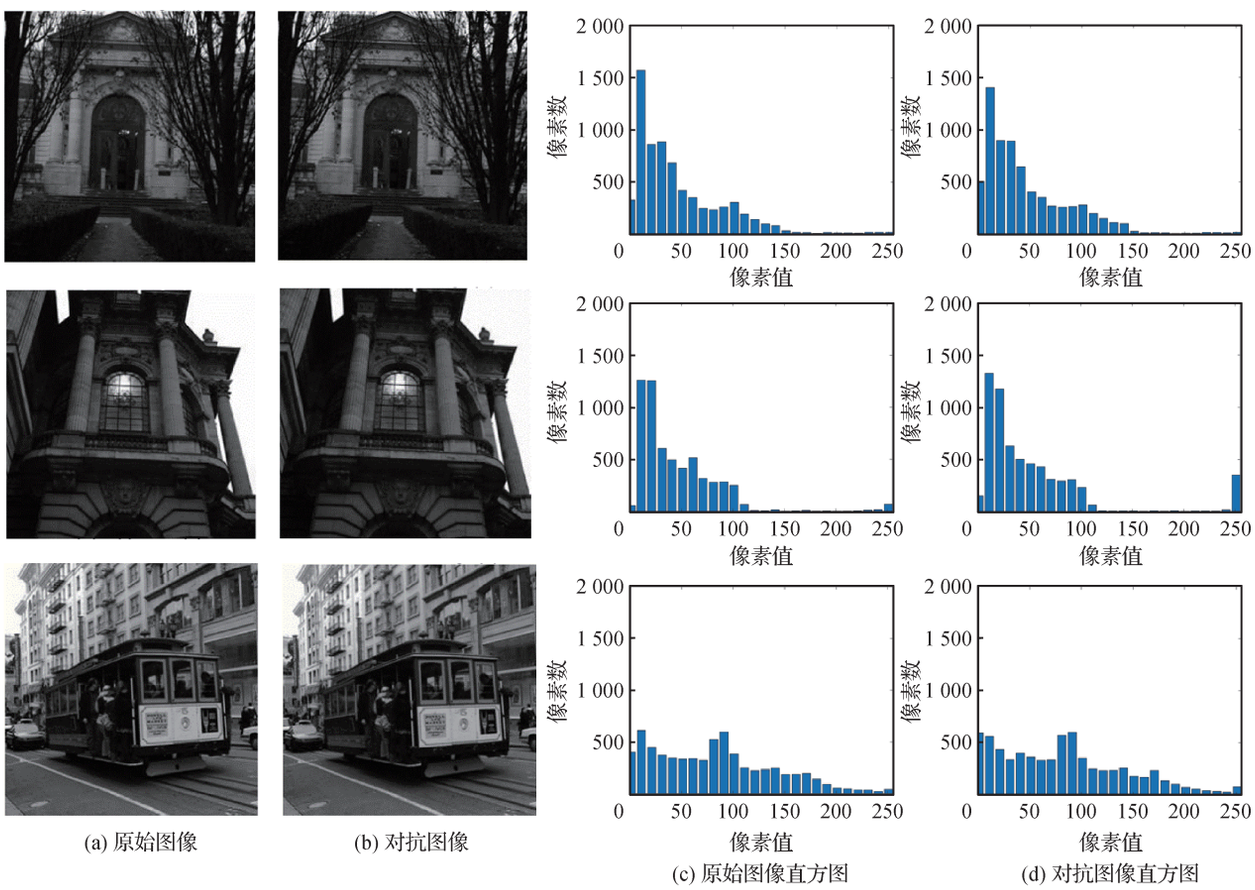

图5 原始图像与对抗图像及其直方图

Fig. 5 The original images, adversarial images and their corresponding histograms

((a)original images; (b) adversarial images; (c) histograms of original images; (d) histograms of adversarial images)

图6展示了随机挑选的800幅对抗图像的PSNR和SSIM值的散点图。根据统计，平均PSNR值为39.925 1 dB，平均SSIM值为0.960 1。实验结果表明，本文设计的基于U-Net添加通道注意力模块的生成器可以输出高视觉质量的对抗图像。

3.5 抗隐写分析能力比较

抗隐写分析能力是评估对抗图像性能的核心指标。考虑到Zhang等人(2018b)、Zhou等人(2020)、Liu等人(2022)和马宾等人(2023)的方法都通过生成对抗图像增强图像隐写术安全性，因此本文将所提方法与这4种当前性能较优异方法的实验结果进行比较。实验中使用的秘密信息嵌入的隐写方法分别是ASDL-GAN和UT-GAN，嵌入容量为0.4 bit/像素。实验中，为了保证可比性，本文比较的其他4种方法也分别根据SRNet和Zhu-Net的梯度反馈生成2 000幅对抗图像，且与本文方法在噪声强度方面相似。实验结果如表6所示。

表6展示了已经针对原始隐写图像进行过训练的隐写分析器对对抗图像的检测能力。用于检测的预训练隐写分析器与被攻击的隐写分析器具有相同的框架。从表6中可以看出，当用于检测的隐写分析器是Xu-Net或Ye-Net或SRM时，与未经修改的原始图像相比，使用5种方法生成的对抗图像，经过秘密信息嵌入生成的增强隐写图像，都可以将隐写分析网络的准确性降低至0.5左右。但是，当面对SRNet或Zhu-Net的检测时，基于其他4种方法的增

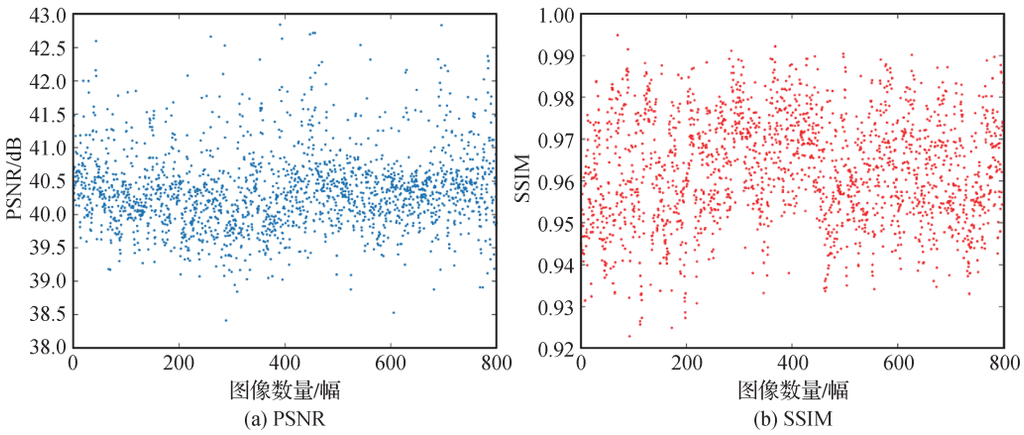

图6　800幅对抗图像的PSNR和SSIM值散点图

Fig. 6　The scatter plots of the PSNR and SSIM of 800 adversarial images((a)PSNR; (b)SSIM)

表6　使用原始隐写图像训练过的隐写分析器的检测准确率

Table 6　Accuracy of the steganalyzers trained on original images

/%

方法	ASDL-GAN					UT-GAN				
	Xu-Net	Ye-Net	SRNet	Zhu-Net	SRM	Xu-Net	Ye-Net	SRNet	Zhu-Net	SRM
	86.80*	88.40*	92.10*	92.60*	87.40*	84.90*	86.60*	89.10*	89.40*	85.80*
Zhang等人(2018b)	50.30	50.30	45.90	46.20	50.40	**50.20**	49.70	53.10	53.80	**50.20**
Zhou等人(2020)	**50.20**	49.20	46.00	46.40	50.40	**49.80**	**50.20**	53.40	46.70	49.70
Liu等人(2022)	50.30	49.30	46.80	47.10	50.30	49.80	50.20	46.40	46.90	49.70
马宾等人(2023)	49.70	49.50	47.10	47.70	50.40	50.30	49.70	47.10	53.60	49.70
本文	49.70	**50.20**	**50.50**	**49.40**	**49.80**	**50.20**	50.40	**49.60**	**50.50**	**49.80**

注:加粗字体表示各列最优结果,带*数值表示原始隐写数据。

强隐写图像都无法欺骗这两种当前最为先进的隐写分析器。而本文方法在面对这两种隐写分析器的检测时表现良好,有效地提高了图像隐写的安全性。

此外,本文将生成的对抗图像和增强的隐写图像作为训练集,重新训练了5个隐写分析器,并使用重新训练后的隐写分析器来检测5种方法生成的对抗图像的抗隐写分析能力。实验结果如表7所示。由表7数据可得,5种方法的抗隐写分析能力都有不同程度的下降。因为这5五种方法都是通过添加对抗噪声来修改图像的敏感区域,使隐写分析器错误分类增强隐写图像,从而增强抗隐写分析能力。然而,添加噪声会影响图像的高频区域,不可避免地添加冗余的“特征”信息,使重新训练的隐写分析器更容易检测到这些信息。但是本文方法仍然比其他4种方法表现更好,因为其他4种方法在训练生成器时固定了隐写分析器的参数,而本文方法在隐写分析器和生成器之间进行对抗训练,这意味着生成器不断根据优化后的隐写分析器的反馈迭代地生成对抗图像,使它们具备更好的抗隐写分析能力,从而更显著地提高图像隐写术安全性。

4　结　论

针对现有隐写方法很难抵御基于深度学习的隐写分析器的检测的问题,本文提出一种基于生成对抗图像的提升图像隐写术安全性的新方法。使用基于U-Net框架的添加通道注意力模块的生成器,经过与多重隐写分析网络的对抗训练,可以生成高质量的对抗图像。相较于使用原始图像嵌入秘密信息生成的隐写图像,使用本文方法生成的对抗图像作为载体图像生成的增强隐写图像可以抵御当前先进的隐写分析器的检测。实验结果显示,本文方法生

表 7　使用对抗图像再训练过的隐写分析器的检测准确率

Table 7　Accuracy of the steganalyzers retrained on adversarial images

/%

方法	ASDL-GAN					UT-GAN				
	Xu-Net	Ye-Net	SRNet	Zhu-Net	SRM	Xu-Net	Ye-Net	SRNet	Zhu-Net	SRM
	89.40*	91.70*	94.50*	94.90*	90.60*	89.00*	89.80*	93.60*	94.10*	90.20*
Zhang等人(2018b)	61.40	63.80	70.30	71.20	61.40	60.20	60.90	69.60	71.70	60.20
Zhou等人(2020)	58.40	59.60	67.60	67.80	58.40	57.80	58.40	66.30	66.70	57.80
Liu等人(2022)	58.60	59.60	65.30	66.90	59.20	57.40	58.20	64.80	65.30	58.00
马宾等人(2023)	57.90	57.20	64.50	64.90	57.40	56.80	58.10	63.50	64.80	57.40
本文	**53.50**	**53.80**	**59.40**	**60.70**	**53.50**	**52.10**	**52.70**	**58.60**	**59.70**	**52.10**

注:加粗字体表示各列最优结果,带*数值表示原始隐写数据。

成的对抗图像的平均PSNR值可以达到39.925 1 dB,图像质量极高。同时,在BOSS Base数据集上与其他4种基于生成对抗图像的方法的比较实验显示,本文方法性能更为优异,对图像隐写术安全性的提升更大。

虽然本文方法可以有效提升图像隐写术的安全性,但仍存在需要改进之处。本文方法的生成器使用的是传统U-Net结构,网络层数较深,导致模型参数量较大,训练速度有待提升。未来工作中,将围绕设计更加轻巧的生成器结构开展研究,提升模型训练速度。

参考文献(References)

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

Chen X Y, Zhang Z T, Qiu A Q, Xia Z H and Xiong N N. 2022. Novel coverless steganography method based on image selection and starGAN. IEEE Transactions on Network Science and Engineering, 9(1): 219-230 [DOI: 10.1109/TNSE.2020.3041529]

Filler T, Judas J and Fridrich J. 2011. Minimizing additive distortion in steganography using syndrome-trellis codes. IEEE Transactions on Information Forensics and Security, 6(3): 920-935 [DOI: 10.1109/TIFS.2011.2134094]

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

Goodfellow I J, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A and Bengio Y. 2014. Generative adversarial networks//Proceedings of the 27th International Conference on Neural Information Processing Systems. Montreal, Canada: MIT Press: 2672-2680

Guo L J, Ni J Q and Shi Y Q. 2012. An efficient JPEG steganographic scheme using uniform embedding//Proceedings of 2012 IEEE International Workshop on Information Forensics and Security. Costa Adeje, Spain: IEEE: 169-174 [DOI: 10.1109/WIFS.2012.6412644]

Holub V and Fridrich J. 2013. Designing steganographic distortion using directional filters//Proceedings of 2012 IEEE International Workshop on Information Forensics and Security. Costa Adeje, Spain: IEEE: 234-239 [DOI: 10.1109/WIFS.2012.6412655]

Holub V, Fridrich J and Denemark T. 2014. Universal distortion function for steganography in an arbitrary domain. EURASIP Journal on Information Security, 2014(1): #1 [DOI: 10.1186/1687-417X-2014-1]

Hu J, Shen L and Sun G, 2018. Squeeze-and-excitation networks//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 7132-7141 [DOI: 10.1109/CVPR.2018.00745]

Liu M L, Luo W Q, Zheng P J and Huang J W. 2021. A new adversarial embedding method for enhancing image steganography. IEEE Transactions on Information Forensics and Security, 16: 4621-4634 [DOI: 10.1109/TIFS.2021.3111748]

Liu Q, Xiang X Y, Qin J H, Tan Y and Zhang Q. 2022. A robust coverless steganography scheme using camouflage image. IEEE Transactions on Circuits and Systems for Video Technology, 32(6): 4038-4051 [DOI: 10.1109/TCSVT.2021.3108772]

Luo Y J, Qin J H, Xiang X Y and Tan Y. 2021. Coverless image steganography based on multi-object recognition. IEEE Transactions on Circuits and Systems for Video Technology, 31(7): 2779-2791 [DOI: 10.1109/TCSVT.2020.3033945]

Ma B, Han Z W, Xu J, Wang C P, Li J and Wang Y L. 2023. Generative multiple adversarial steganography algorithm based on U-Net

structure. Journal of Software, 34(7): 3385-3407 (马宾, 韩作伟, 徐健, 王春鹏, 李健, 王玉立. 2023. 基于U-Net结构的生成式多重对抗隐写算法. 软件学报, 34(7): 3385-3407) [DOI: 10.13328/j.cnki.jos.006537]

Mielikainen J. 2006. LSB matching revisited. IEEE Signal Processing Letters, 13(5): 285-287 [DOI: 10.1109/LSP.2006.870357]

Peng F, Chen G F and Long M. 2022. A robust coverless steganography based on generative adversarial networks and gradient descent approximation. IEEE Transactions on Circuits and Systems for Video Technology, 32(9): 5817-5829 [DOI: 10.1109/TCSVT.2022.3161419]

Petitcolas F A P, Anderson R J and Kuhn M G. 1999. Information hiding-a survey. Proceedings of the IEEE, 87(7): 1062-1078 [DOI: 10.1109/5.771065]

Pevný T, Filler T and Bas P. 2010. Using high-dimensional image models to perform highly undetectable steganography//Proceedings of the 12th International Conference on Information Hiding. Calgary, Canada: Springer: 161-177 [DOI: 10.1007/978-3-642-16435-4_13]

Ronneberger O, Fischer P and Brox T. 2015. U-Net: convolutional networks for biomedical image segmentation//Proceedings of the 18th International Conference on Medical Image Computing and Computer-Assisted Intervention. Munich, Germany: Springer: 234-241 [DOI: 10.1007/978-3-319-24574-4_28]

Tan J X, Liao X, Liu J T, Cao Y and Jiang H B. 2022. Channel attention image steganography with generative adversarial networks. IEEE Transactions on Network Science and Engineering, 9(2): 888-903 [DOI: 10.1109/TNSE.2021.3139671]

Tang W X, Tan S Q, Li B and Huang J W. 2017. Automatic steganographic distortion learning using a generative adversarial network. IEEE Signal Processing Letters, 24(10): 1547-1551 [DOI: 1547-155110.1109/LSP.2017.2745572]

Wu K C and Wang C M. 2015. Steganography using reversible texture synthesis. IEEE Transactions on Image Processing, 24(1): 130-139 [DOI: 10.1109/TIP.2014.2371246]

Xu G S, Wu H Z and Shi Y Q. 2016. Structural design of convolutional neural networks for steganalysis. IEEE Signal Processing Letters, 23(5): 708-712 [DOI: 10.1109/LSP.2016.2548421]

Yang J H, Ruan D Y, Huang J W, Kang X G and Shi Y Q. 2019. An embedding cost learning framework using GAN. IEEE Transactions on Information Forensics and Security, 15: 839 - 851 [DOI: 10.1109/TIFS.2019.2922229]

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

Yin X L, Lu W, Zhang J H and Luo X Y. 2022. Robust JPEG steganography based on lossless carrier and robust cost. Journal of Image and Graphics, 27(1): 238-251 (尹晓琳, 卢伟, 张俊鸿, 罗向阳. 2022. 无损载体和鲁棒代价结合的JPEG图像鲁棒隐写. 中国图象图形学报, 27(1): 238-251) [DOI: 10.11834/jig.210406]

Zhang R, Zhu F, Liu J Y and Liu G S. 2020. Depth-wise separable convolutions and multi-level pooling for an efficient spatial CNN-based steganalysis. IEEE Transactions on Information Forensics and Security, 15: 1138-1150 [DOI: 10.1109/TIFS.2019.2936913]

Zhang X, Peng F and Long M. 2018a. Robust coverless image steganography based on DCT and LDA topic classification. IEEE Transactions on Multimedia, 20(12): 3223-3238 [DOI: 10.1109/TMM.2018.2838334]

Zhang Y W, Zhang W M, Chen K J, Liu J Y, Liu Y J and Yu N H. 2018b. Adversarial examples against deep neural network based steganalysis//Proceedings of the 6th ACM Workshop on Information Hiding and Multimedia Security. Innsbruck, Austria: ACM: 67-72 [DOI: 10.1145/3206004.3206012]

Zheng G, Hu D H, Ge H and Zheng S L. 2021. End-to-end image steganography and watermarking driven by generative adversarial networks. Journal of Image and Graphics, 26(10): 2485-2502 (郑钢, 胡东辉, 戈辉, 郑淑丽. 2021. 生成对抗网络驱动的图像隐写与水印模型. 中国图象图形学报, 26(10): 2485-2502) [DOI: 10.11834/jig.200404]

Zhou L C, Feng G R, Shen L Q and Zhang X P. 2020. On security enhancement of steganography via generative adversarial image. IEEE Signal Processing Letters, 27: 166-170 [DOI: 10.1109/LSP.2019.2963180]

作者简介

马宾,男,教授,博士生导师,主要研究方向为信息隐藏与多媒体安全、数字图像处理。E-mail: sddxmb@126.com

徐健,通信作者,女,副教授,主要研究方向为信息隐藏与多媒体安全、数字图像处理。E-mail:sdfixj@126.com

李坤,男,硕士研究生,主要研究方向为隐写和隐写分析。E-mail:990630753@qq.com

王春鹏,男,副教授,主要研究方向为数字图像水印和数字图像处理。E-mail:mpeng1122@163.com

李健,男,副教授,主要研究方向为多媒体数字安全和数字图像处理。E-mail:ljian_20@163.com

张立伟,男,工程师,主要研究方向为光伏并网接口装置与高标准环网柜开发。E-mail:kunxxx2021@163.com