基于深度学习的图像隐写方法研究

付章杰[1)2)] 王 帆[1)] 孙星明[1)] 王 彦[1)]

[1)](南京信息工程大学计算机与软件学院 南京 210044)

[2)](鹏城实验室 广东 深圳 518000)

摘 要 信息隐藏是保证网络通信数据安全的重要手段之一，不仅能够保证秘密信息本身的安全，还能保证秘密信息的安全传输. 隐写术作为信息隐藏领域的主要技术，受到了国内外学者的广泛关注和深入研究. 现有的空域自适应隐写方法对于待改变像素位置选择大多依赖人为经验设计，需要耗费大量时间与精力. 近年来，随着深度学习的快速发展，深度学习网络对于复杂数据的强表征能力使其被应用到隐写分析领域中，它在快速提取高维特征的同时还能实现分类器的同步优化. 隐写分析模型性能的快速提高，导致隐写术的安全性降低，从而对隐写术的发展带来了极大挑战. 2014 年，生成对抗网络（Generative Adversarial Networks，GAN）的提出，为深度学习与信息隐藏的结合提供了契机. 直至 2016 年，基于深度学习的隐写模型——SGAN 被首次提出，自此以后，利用各类深度学习网络进行信息隐藏的隐写模型大量涌现，使得隐写算法在隐写容量、抗检测性、以及含密图像质量等多方面取得较大提升. 本文首先论述了四类基于深度学习的隐写模型：1）基于生成载体式深度学习隐写方法；2）基于嵌入载体式深度学习隐写方法；3）基于合成载体式深度学习隐写方法；4）基于映射关系式深度学习隐写方法；其次，分别对各类隐写模型进行分析和讨论，并总结模型之间的异同点；然后，探讨了各类隐写模型存在的问题；接着，针对基于深度学习的大容量隐写模型存在的安全问题，提出了基于对抗样本的改进方法；最后，总结了目前基于深度学习网络的隐写模型存在的优缺点并对其未来发展方向进行展望.

关键词 信息隐藏；深度学习；隐写术；生成对抗网络；隐写分析

中图法分类号 TP18 **DOI 号** 10.11897/SP.J.1016.2020.01656

Research on Steganography of Digital Images based on Deep Learning

FU Zhang-Jie[1)2)] WANG Fan[1)] SUN Xing-Ming[1)] WANG Yan[1)]

[1)](*Department of Computer and Software, Nanjing University of Information Science and Technology, Nanjing* 210044)

[2)](*Peng Cheng Laboratory, Shenzhen, Guangdong* 518000)

Abstract Information hiding is one of the important ways to ensure data security during network communication. It not only ensures the security of the secret information itself, but also ensures the security of data transmission. As the main technology in the field of information hiding, steganography has received extensive attention and in-depth research by domestic and foreign scholars. The current spatial adaptive steganography methods mostly rely on human experience for the selection of pixels to be changed, which requires a lot of time and effort. In recent years, with the rapid development of deep learning, the strong representation learning ability for complex data makes it used in the field of steganalysis, so that the steganalysis models based on deep learning can quickly extract high-dimensional features and simultaneously optimize the classifier. The rapid improvement of the performance of the steganalysis model

收稿日期：2019-08-28；在线发布日期：2020-02-06. 本课题得到国家自然科学基金（No. U1836110, No. U1836208）、国家重点研发计划（No. 2018YFB1003205）资助. 付章杰（通信作者），博士，教授，博士生导师，主要研究领域为信息隐藏、数字取证和网络与信息安全. E-mail: wwwfzj@126.com. 王 帆，硕士，主要研究领域为信息隐藏和深度学习. 孙星明，博士，教授，博士生导师，主要研究领域为网络与信息安全. 王 彦，硕士，主要研究领域为信息隐藏和深度学习.

weakens of the security of steganography, which makes a great challenge to the development of steganography. In 2014, the generative adversarial networks (GAN) was proposed, which provided a valuable opportunity for the combination of deep learning and information hiding. Until 2016, a steganographic model based on deep learning--SGAN was proposed firstly. Since then, a large number of steganographic models using various deep learning networks have emerged, which lead to a great improvement in steganography in terms of steganographic capacity, anti-detection, and stego-image's quality. First of all, this paper explains the importance of information hiding for data security and summarizes the historical development of image-based steganography briefly; The second one, this paper discourses the fourth types of steganographic models based on deep learning: 1) the steganography based on deep learning of generating cover images, which generates cover images that are more secure and suitable for hiding by using deep learning networks; 2) the steganography based on deep learning of embedding information in cover images, which replaces the steganography algorithms designed by human experience with deep learning networks to embed and extract secret messages automatically; 3) the steganography based on deep learning of synthesizing cover images, which performs secondary modification and synthesis on the original cover images, and the sender can more securely hide secret messages in the changed area; 4) the steganography based on deep learning of mapping relations, which builds a mapping relationship between a secret message and a cover image(or a noise vector used to generate a cover image), and uses deep learning networks and mapping relationships to extract secret messages almost completely. Then, this paper analyzes and summarizes various type of steganographic models in detail from the aspect of steganographic capacity, anti-detection, and stego-image's quality, and discuss the similarities and differences of various steganographic models. Next, this paper explores the different problems of various steganographic models. Furthermore, this paper proposes an improved method based on adversarial samples for the security problems of large-capacity steganographic models based on deep learning, and briefly describes the method. Last but not least, this paper summarizes the advantages and disadvantages of the current steganographic models based on deep learning and looks forward to its future development directions.

Keywords information hiding; deep learning; steganography; generative adversarial networks; steganalysis

1 引 言

当今时代，大数据为人们的生活带来了诸多便利，企业通过对大量消费者数据的智能分析，可以为他们提供丰富的个性化服务. 然而，大数据也引起了诸多安全问题，比如个人隐私数据泄露、军事机密数据遭受恶意攻击和篡改、商业信息的非法兜售等[1]. 因此,保障数据安全是当今大数据时代数据有效利用的前提和基础. 目前，保证数据安全的主要方法有两种：加密和信息隐藏. 加密是以某种特定的算法将原始的明文信息变成无法识别的密文信息，再利用密钥恢复成明文信息的技术，但是加密之后的数据很容易遭到第三方的怀疑和拦截. 因此，信息隐藏技术应运而生. 信息隐藏是将秘密信息以隐蔽的方式嵌入到载体中的技术，它不仅保证了秘密信息本身的安全，还提升了信息传输过程的安全性. 隐写术作为信息隐藏的重要方法之一，可以保证网络通信过程中数据的安全性，因此成为近年来信息安全领域的热门研究方向之一[2].

信息隐藏将秘密信息通过密钥和特定算法嵌入到载体（比如文本、图像、音频、视频等）中，再由信息接收方通过密钥和提取算法提取出秘密信息，具体原理如图 1 所示. 由于图像的易获取性和多样性使其成为使用最广泛的隐藏载体，空域图像隐写术因其简单有效、可操作性强成为信息隐藏的主流方向之一. 最原始的空域隐写方法包括最低有

效位（Least Significant Bit，简称 LSB）算法，±1嵌入算法等. 其中，LSB 算法是通过修改图像比特层中不太重要的比特信息位来嵌入秘密信息，而 ±1嵌入算法则是通过对图像像素值进行 ±1 操作来嵌入秘密信息.

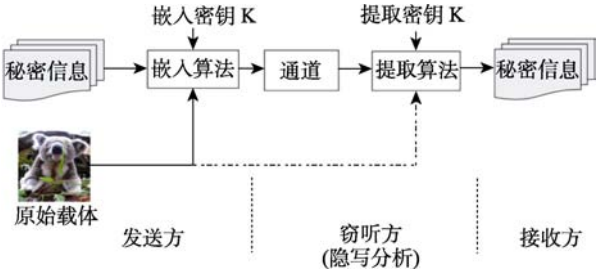

图 1 信息隐藏原理图

然而，上述空域隐写算法得到的含密图像都存在质量下降和统计特性发生变化等问题和缺陷[3]. 这些问题对隐写算法的安全性造成了一定影响. 为了解决上述问题，研究者们对传统隐写方法进行了改进，提出自适应隐写算法（比如 S-UNIWARD[4]、HILL[5]、WOW[6]、HUGO[7]等）来自动寻找图像中适合隐写的区域（比如纹理复杂区域等）. 自适应隐写算法虽然能够减少秘密信息对载体图像统计特性造成的失真，但仍旧未能消除人为设计隐写算法在秘密信息嵌入过程中对载体图像造成的改动痕迹[4].

同时，隐写分析对隐写术也带来了巨大影响. 隐写术与隐写分析二者之间相互矛盾、相互促进，在隐写算法不可知的情况下，隐写分析可以根据观测到的数据信息预测图像中是否含有秘密信息. 隐写分析的具体过程分为两步：特征提取和分类器训练，通常先利用高通滤波器获取残差图像，再利用各种统计模型提取隐写分析特征. 有效特征是决定隐写分析模型检测准确率的关键性因素之一，所以特征提取是隐写分析模型主要的改进方向. 现有的传统隐写分析方法（比如 Spatial Rich Model，SRM[8]、maxSRM[9]等）利用高通滤波器可以提取出上万维复杂特征. 而深度学习（Deep Learning）作为近几年来机器学习领域的热门研究方向[10]，自 2006 年，Hinton 等人[11]采用反向传播算法解决了深层神经网络内部存在的梯度弥散和梯度消失问题，深度学习便在计算机视觉、图像处理、目标跟踪等领域取得了令人瞩目的优异表现. 深度学习能够自动学习复杂数据内部的抽象特征，然后将学习到的高维复杂特征转换到低维特征空间，并保留相同的语义信息，所以，研究者们尝试将深度学习应用到隐写分析领域进行特征提取，在提升隐写分析模型的检测准确率的同时缩减模型的训练时间. 目前，基于深度学习的隐写分析模型的（比如 Ye'Net[12]、SRNet[13]等）快速发展使得隐写术的发展遭遇前所未有的瓶颈.

2014 年，Goodfellow 等人提出的生成对抗网络（GAN）[14]为信息隐藏与深度学习网络的结合提供了契机. Volkhonskiy 等人在 2016 年提出利用深度卷积生成对抗网络（Deep Convolution Generative Adve- rsarial Networks，DCGAN）生成更加适合隐写的载体图像的隐写模型—SGAN[15]. 自此以后，各类基于深度学习的隐写模型不断涌现，有些隐写模型利用生成对抗网络对于复杂数据的建模能力，建模出图像不同像素之间的复杂依赖关系，从而生成更适合隐写且逼真的载体图像；有些隐写模型利用深层卷积神经网络自动学习最小嵌入失真代价，减少嵌入操作对载体图像造成的失真；还有些隐写模型利用对抗训练思想，将隐写术与隐写分析网络作为“对立方”，二者双方进行对抗训练以此提升隐写术的抗隐写分析检测能力. 这些新型的隐写方法不仅能够扩大隐写容量还可以有效的提升隐写安全性，为信息隐藏领域的发展增添了一股新的活力.

本文在深入研究基于深度学习的隐写模型基础上，首先将现有的基于深度学习的隐写模型按照不同的嵌入方式分为以下四类：1）生成载体式深度学习隐写方法；2）嵌入载体式深度学习隐写方法；3）合成载体式深度学习隐写方法；4）映射关系式深度学习隐写方法，具体分类框架如表 1 所示；接着，本文依据实验对比结果，分析各类模型目前存在的问题；然后，本文针对大容量隐写模型的安全问题，提出一种新的解决办法；最后，本文总结目前基于深度学习隐写模型相较于传统隐写算法存在的优势并展望未来的发展方向.

本文结构如下：第 2 节将分析四类隐写模型的网络结构和异同点；第 3 节将实验对比各类隐写方法的性能表现；第 4 节将总结各类隐写模型存在的问题；第 5 节将针对大容量隐写模型的安全问题，提出改进方法；第 6 节总结基于深度学习的图像隐写模型存在的优缺点以及展望其未来发展方向.

表 1　基于深度学习的隐写方法分类

方法	具体实现	优缺点	隐写模型
生成载体式深度学习隐写模型	利用深度学习生成更加适合隐写的载体图像，再利用传统隐写算法实现信息的隐藏和提取	深度学习与隐写术相结合； 载体图像不真实，安全性不高	SGAN[15]、SSGAN[16]等
嵌入载体式深度学习隐写模型	利用深度学习在自然载体图像上完成或辅助完成秘密信息的嵌入和提取	扩大隐写容量、提升隐写安全性； 载体图像中的改动嵌入痕迹较大；	Deep Steganography[17]、HiDDeN[18]、ASDL-GAN[19]、SteganoGAN[20]、ADV-EMB[21]等
合成载体式深度学习以下内模型	利用深度学习对已有载体图像上进行二次改动生成新的图像并完成秘密信息的嵌入和提取	秘密信息的完整提取； 提取时需要依靠额外的 mask	Liu's model[22]、SGSRGAN[23]
映射关系式深度学习隐写模型	秘密信息与随机噪声或目标对象之间建立映射关系，再利用深度学习完成秘密信息的隐藏和提取	载体图像没有嵌入痕迹； 隐写容量小、秘密信息难以恢复	Hu's model[24]、Zhang's model[25]、Meng's model[26]

2　深度学习隐写模型分类

近几年来，基于深度学习的隐写模型不断涌现. 一些隐写模型利用生成对抗网络生成适合隐写的载体图像. 另一些隐写模型利用深度学习网络的反向传播自动学习嵌入失真代价，实现载体图像的最小嵌入失真. 由于隐写术与隐写分析之间的对抗关系与生成对抗网络中的博弈论思想不谋而合，因此还有一些隐写模型通过隐写网络与隐写分析网络的对抗训练来提升隐写的安全性. 除此以外，目前还有很多其他类型的基于深度学习的隐写方法，本文将根据不同的隐写方式对其进行分类.

2.1　基于生成载体式深度学习隐写模型

基于生成载体式深度学习隐写方法，顾名思义，是利用深度学习网络生成适合隐写的载体图像. 该想法在 2017 年 Volkhonskiy 等人[15]提出的 SGAN 模型中被首次提出，模型首先将随机噪声作为输入，通过 DCGAN[27]生成尽可能真实的载体图像，再使用传统的 ±1 嵌入算法实现秘密信息的隐藏，最后将含密图像作为隐写分析网络的输入，在对抗训练过程中提升其抗检测能力. 该模型可用数学公式表示为

$$\begin{aligned}\min_{G}\max_{D}\min_{S} L = {} & \alpha(\mathrm{E}_{x\sim P_{data(x)}}[\log D(x)]+\\ & \mathrm{E}_{x\sim P_{noise(z)}}[\log(1-D(G(z)))]+\\ & (1-\alpha)\mathrm{E}_{z\sim noise(z)}[\log S(stego(G(z)))+\\ & \log(1-S(G(z))]\end{aligned} \tag{1}$$

然而，DCGAN 在训练过程中并不稳定且生成的载体图像质量较差. 于是，2018 年 Shi 等人[16]在 SGAN 的基础上提出了 SSGAN，具体网络框架如图 2 所示. 该模型与 SGAN 模型相似，但是将载体图像的生成网络替换成 Wasserstein GAN(简称为 WGAN), WGAN 的损失函数使用对距离分布更加敏感的 EM 距离(又称为 Wasserstein distance)来提供更加有意义的梯度，从而生成更加符合真实分布的载体图像[28].

上述两种方法的优缺点对比如表 2 所示，它们都是通过与隐写分析网络(GNCNN)进行对抗训练，从而生成更加适合隐写的载体图像. 但是该类隐写方法生成的载体图像在语义上不够自然真实，并且与传统自适应隐写算法相比，隐写的安全性并未获得很大提升.

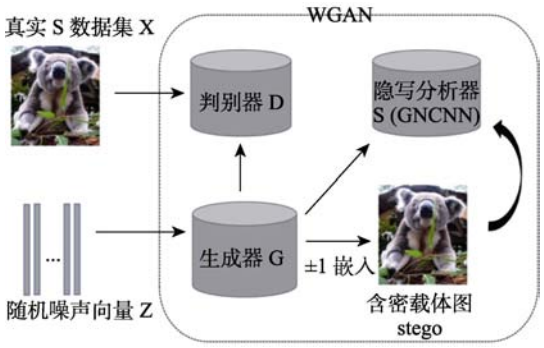

图 2　SSGAN 隐写模型结构

表 2　基于生成载体式深度学习隐写方法分类

方法	实现细节	优点	缺点
SGAN[15]	利用 DCGAN 生成适合隐写的载体图像再与隐写分析对抗，提升隐写安全性	生成的载体图像更加适合隐写	生成载体图像质量较差
SSGAN[16]	在 SGAN 基础上，将 DCGAN 替换成 WGAN	相较于 SGAN，生成载体图像质量更好，抵抗 GNCNN 的检测能力增强	未对隐写算法本身做出改进 生成图像质量一般

2.2 基于嵌入载体式深度学习隐写模型

考虑到基于生成载体式深度学习隐写方法容易受训练过程不稳定或者损失函数等参数的影响，导致生成的载体图像出现质量差、语义混乱等问题. 研究者们提出了基于载体嵌入式的深度学习隐写模型，即在自然载体图像上利用深度学习网络完成或辅助完成秘密信息的嵌入和提取的一种隐写方式. 该类隐写方法目前已经成为利用深度学习进行隐写的主要方式.

2.2.1 自动学习嵌入改变概率以及嵌入代价

目前，现有的自适应隐写算法都是基于最小嵌入失真框架设计的. S-UNIWARD 使用的是一种与嵌入域无关的通用失真函数[4]；HILL 是在像素内嵌入一些信息的效果，为像素分配成本，定义失真函数域，使用加权范数将像素压缩到特征空间[5]；WOW 则是根据复杂区域将秘密信息嵌入到载体图像中，如果图像的区域在结构上比另一区域更加复杂，则该区域像素值被更改[6]. 这些自定义的嵌入失真函数决定了每个待改变像素的嵌入代价，所有像素的嵌入代价之和就是该图像的嵌入最小失真代价.

Li 等人在 2017 年提出利用生成对抗网络自动学习嵌入失真代价，从而寻找最小失真嵌入位置的隐写模型 ASDL-GAN[19]. 该网络模型框架如图 3 所示，由三部分组成：生成器，嵌入模拟器，判别器. 生成器将真实图像（又称为载体图像）作为输入，生成“非 0”的嵌入改动概率图. 接着，把概率图放入嵌入模拟器（TES）“模拟”秘密数据的嵌入，生成与载体图像同尺寸大小的改动位置映射图，该图每个像素点的取值范围都在{–1，0，1}之间. 然后，将载体图像与改动位置映射图进行点对点相加，生成“模拟”的含密载体，判别器检测载体图像是否含有秘密消息.

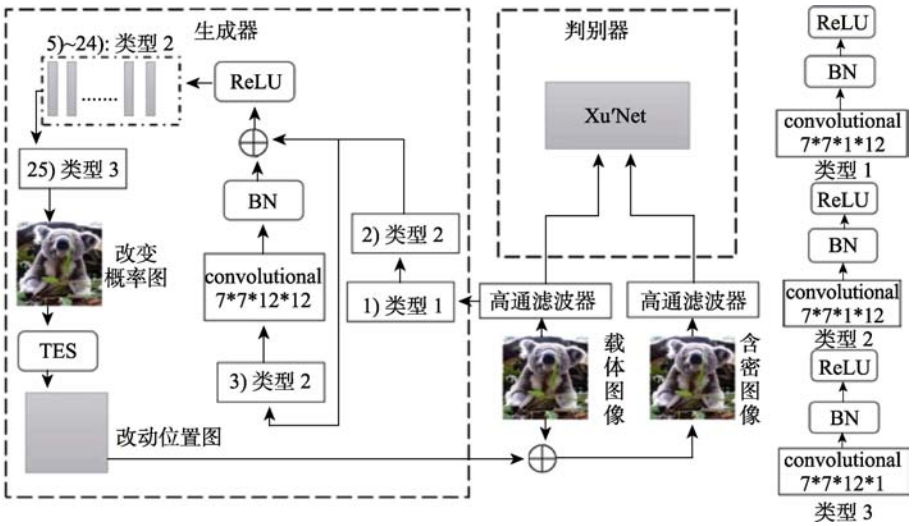

图 3 ASDL-GAN 隐写模型结构

然而，嵌入模拟器 TES 是不可微网络，导致 ASDL-GAN 在训练过程中收敛速度很慢. 所以，Yang 等人[29]对嵌入模拟器进行改进并在 ASDL-GAN 的基础上提出新的隐写模型 UT-SCA-GAN. UT-SCA-GAN 模型使用可微分的 Tanh-simulator 激活函数替代不可微的嵌入模拟网络（TES）来生成改动位置映射图，其中，TES 子网结构与 Tanh-simulator 激活函数如图 4 所示. Tanh-simulator 函数不会影响网络层之间的反向传播，从而使得隐写模型加速收敛并节省了训练时间.

2.2.2 编码-解码网络

编码-解码网络中的编码和解码操作与隐写术中的隐写和提取操作十分相似，都是先将多个数据（比如:文字、数据、图像等）先融合然后再进行抽离或提取. 相较于寻找最小嵌入失真代价或者先生成载体图像再嵌入秘密信息等隐写方法，基于编码-解码网络的隐写方法不需要隐写者具备很多隐写方面的先验知识，它不仅可以自动学习秘密信息的隐写和提取，还能成功隐藏大容量的秘密信息.

Hayes 等人提出的 SteGAN 隐写模型将编码-解码网络运用到隐写领域[30]. 该隐写方法定义了一个 Alice、Bob 和 Eve 的三方对抗游戏，三方分别用来进行信息隐藏、信息提取与隐写分析，它们各自的损失函数可写为

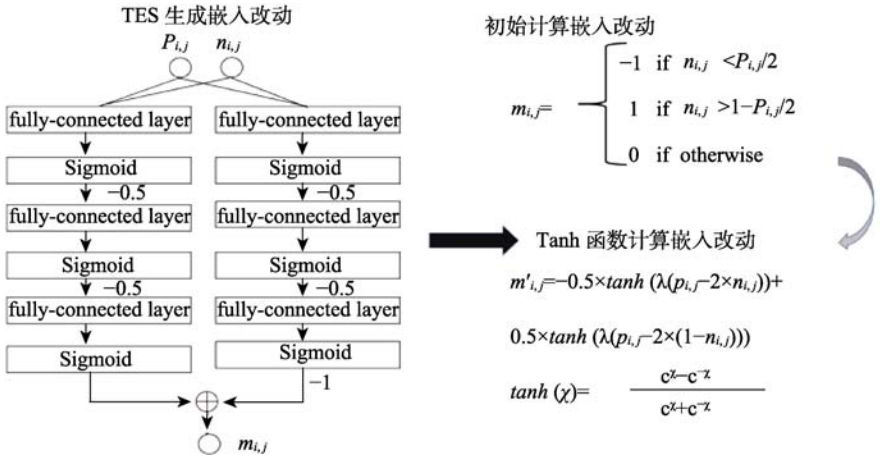

图 4　嵌入模拟器（左：ASDL-GAN 右：UT-GAN）

$$L_{Bob}(\theta_A,\theta_B,M,C)=d(M,B(\theta_B,C'))=d(M,B(\theta_B,A\theta_A,C,M))=d(M,M') \tag{2}$$

$$L_{eve}(\theta_E,C,C')=-\mathrm{y}.\log(\theta_E,x)-(1-y).\log(1-E(\theta_E,x)) \tag{3}$$

$$L_{\mathrm{Alice}}(\theta_A,C,M)=\lambda_A.d(C,C')+\lambda_B.L_{Bob}+\lambda_E.L_{Eve}(\theta_E,C,C') \tag{4}$$

其中，$\theta_A,\theta_B,\theta_C$ 分别代表 Alice，Bob，Eve 的参数，C 为载体图像，C' 为含密图像. 当 $x=C$ 时，$\mathrm{y}=0$；当 $x=C'$ 时，$\mathrm{y}=1$. $d=(M,M')$ 代表秘密信息与重构秘密信息之间的 EL 距离(又称为 Euclidean distance). $\lambda_A,\lambda_B,\lambda_C$ 分别代表三个损失所占权重. Alice 由全连接网络构成，它将载体图像和随机 n 位二进制序列的秘密消息作为输入生成含密图像，并传递给 Bob 提取出秘密信息. 而 Eve 则作为窃听方进行窃听并确认图像中是否有秘密消息的存在. 但是该隐写方法并未充分考虑到含密图像的图像质量以及与真实载体图像之间存在的差距. 于是，中科院 Wang 等人[31]提出了 SsteGAN 模型，该模型在 SteGAN 的基础上添加了一个 Dev 方，Alice 方通过与 Dev 方的对抗训练缩减含密图像与原载体图像之间的距离，然后生成更加真实的含密图像.

而 Zhu 等人则从另一角度出发，提出了可以抵抗噪声等攻击的隐写模型 HiDDeN[18]. 该模型中的编码-解码网络由卷积网络构成，考虑到含密载体在通信传输过程中的安全问题，在编码和解码网络之间加入了噪声层进行噪声建模. 编码网络经由噪声层“模拟”生成带有噪声的含密图像，解码网络将其作为输入进行秘密信息的提取. 训练好的 HiDDeN 模型具有很好的鲁棒性，即使含密图像受到一些常用噪声的攻击，提取网络仍然可以准确提取出秘密信息. 但是，该模型将秘密信息转换成数据张量时存在维数过大的问题，从而导致隐藏容量最高只能达到 0.2bpp 左右.

除了提高隐写算法的安全性和抗检测性，隐藏容量大小也是衡量隐写算法性能的重要指标. Zhang 等人[20]在 2019 年提出了 SteganoGAN，该隐写模型可以在载体图像中隐藏任意二进制比特数据，具体网络结构如图 5 所示. SteganoGAN 模型由三部分组成：编码器，解码器以及评价方. 编码器将秘密信息 M 转换成张量大小为 $D\times W\times H$ 的二进制数据，然后编码到尺寸大小为 $W\times H$ 的载体图像 X 中，其中 D 为秘密信息转换成的二进制比特数量，最后由解码网络从含密图像中重构出秘密信息 M'.

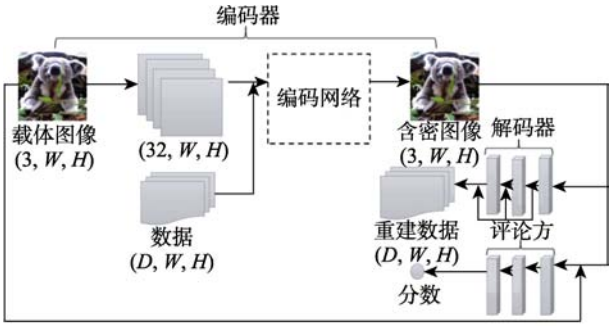

图 5　以图藏比特数据：SteganoGAN 隐写模型结构

与上述其他基于编码-解码的隐写方法不同，SteganoGAN 还构建了一个评估网络用来评价编码网络的表现，以生成更加真实的含密图像. 因此，该模型的损失函数可写为

$$L_d=\mathrm{E}_{x\sim P_c}Crossentroy(D(\varepsilon(X,M),M)) \tag{5}$$

$$L_S=\mathrm{E}_{X\sim P_c}\frac{1}{3\times W\times H}\left\|X-\varepsilon(X-M)\right\|_2^2 \tag{6}$$

$$L_{\mathrm{r}}=\mathrm{E}_{X\sim P_c}C(\varepsilon(X,M)) \tag{7}$$

其中，L_d 用于衡量解码准确性，L_s 用于衡量含密图像与原载体图像的相似性，L_r 用于衡量含密图像真实性. $\varepsilon(\cdot)$为编码器，$D(\cdot)$为解码器，$C(\cdot)$为评价方.

此外，该模型提出了一个新的隐写评估指标—RSBPP（Reed Solomon Bits Per Pixel），用于评估图像中嵌入的信息比特量（相对有效载荷），数学表达式可写为

$$RSBPP = n \times (1 - 2p) \tag{8}$$

其中，n 为尝试隐藏的比特数量，p 为隐写网络的误码率. 该隐写方法的隐写容量最高可达到 4.4bpp，相比于现有基于深度学习的以图藏二进制比特信息类型的隐写算法，其隐写容量高出 10 倍.

编码-解码网络不仅可以将二进制比特数据或者文字信息编码到载体图像中，还可以将彩色或灰度图像编码到同尺寸载体图像中. Baluja 等人[17]在 2017 年首次提出以图藏图的深度学习隐藏网络 Deep Steganography，该网络结构如图 6 所示，由预处理器，编码器和解码器三部分组成. 该模型将秘密图像压缩并分布到载体图像的所有可用位中，预处理器最重要的作用是将基于色彩的像素转换成可以成功编码到载体图像中的有用特征.

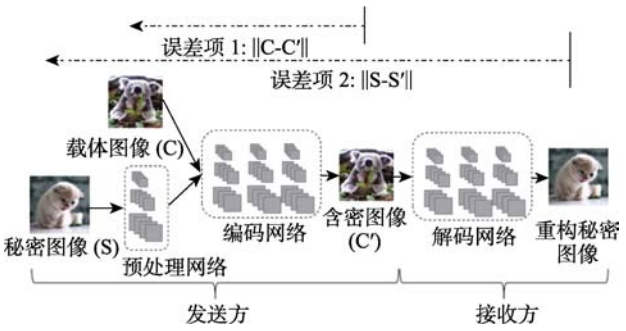

图 6 以图藏图：Deep Steganography 隐写模型结构

但 Deep Steganography 并未实现秘密图像的无损提取，而是在含密图像生成真实性与秘密图像提取准确率之间做了权衡考虑，网络具体的损失函数可表示为

$$L(C, C', S, S') = \|C - C'\| + \|S - S'\| \tag{9}$$

以图藏图隐写模型虽然扩大了隐写容量，但是过多的秘密信息会导致隐写的安全性降低以及含密图像失真较大. Deep Steganography 模型生成的含密图像会出现明显的颜色失真. 并且若原始载体图像被泄露，那么，生成的含密图像与载体图像构建出的残差图像可以明显看到秘密图像的残影.

之后，Wu 等人[32]提出了一个不包含预处理模块的端对端隐写网络模型 Stegnet. 该模型在编码网络的初始化模块（Inception module）[33]中加入残差连接[34]，用于加快网络的训练速度，并增加了一个方差损失函数用来解决含密图像在非纹理区域中掺杂噪声像素点的问题. 但是，Wu 等人[32]指出 Stegnet 隐写模型在嵌入完成后对载体图像的平均改变率仅有 0.76%，而且模型对于含密图像的颜色失真问题并没有提出有效解决方法.

从 Wu 等人给出的含密图像结果来看，添加方差损失函数的隐写模型生成的含密图像包含的噪声点明显减少，也从侧面反映出通过调整损失函数以及编码网络的结构能够提升含密图像质量. 因此，Duan 等人[35]利用全卷积网络构建了一个类似 U-Net 结构的编码网络，将两幅同尺寸大小的彩色图像 C，S 当做编码网络的输入然后生成含密图像，接着使用由卷积网络构成的解码网路提取出秘密图像. U-Net 结构型的编码网络与普通的编码网络相比，它能够将浅层特征与深层特征进行更好的融合，从而生成与原载体图像更加相似、视觉质量更好并且不存在颜色失真问题的含密图像.

而 Baluja 等人则以突破隐写容量为目标不断研究，在 2019 年又提出了“一图藏多图”的改进模型. 该模型不仅能够提取出两幅秘密图像，而且从含密图像与载体图像的残差来看，该模型能够混淆两幅秘密图像的具体信息. 混淆后的秘密图像虽然保证了隐写的安全性但仍然无法抵抗隐写分析模型的检测[36].

Atique 等人[37]从另一角度出发，将秘密图像由 3 通道的彩色图像更换为单通道的灰度图像，然后利用编码-解码网络将秘密图像嵌入到载体图像中，同时构建一个新型损失函数用于编码-解码网络的训练，该损失函数的表达式可写为

$$L(I_g - I_h) = \alpha \|I_h - O_e\|^2 + \beta \|I_g - O_d\|^2 + \lambda(\|W_e\|^2 + \|W_d\|^2) \tag{10}$$

其中，I_g，I_h 代表载体彩色图像、秘密灰度图像，O_e，O_d 分别代表含密图像与重建灰色图像，W_e，W_d 分别代表已学习到的编码器和解码器的权重. 但是该模型生成的含密图像质量较差并且仍然存在颜色失真问题. 之后，Zhang 等人[38] 提出的 ISGAN 模型同样将灰度图像做为秘密信息，但该隐写模型仅在彩色载体图像的 Y 通道嵌入灰度图像，之后将含密的 Y 通道图像与 U,V 通道图像进行拼接，恢复成一张彩色的含密图像，因此有效避免了 U，V 通道中亮度，

彩色信息的丢失. 同时，该模型在损失函数中考虑到含密图像的真实性，添加了结构相似性（Structual Similarity index，简称 SSIM）和 MSSIM 损失，用于提升含密图像的生成质量. 与 Atique 等人的隐写模型相比，ISGAN 生成的含密图像没有明显的颜色失真且峰值信噪比（Peak Signal to Noise Ratio,简称 PSNR）[39]值提升了 2db 左右，但是解码网络提取出的秘密图像的 PSNR 值却下降了大约 3db.

基于以图藏图的深度学习隐写模型存在明显的优势和缺点，它牺牲了隐写的部分安全性以达到隐写容量的大范围提升，不可避免的会引起图像质量下降，颜色失真等问题. 同时，利用编码-解码网络自动学习隐写算法，可以缩减人为耗费的时间和精力，但是解码网络无法百分百的实现秘密消息的重新提取. 除此以外，编码-解码网络这种端对端隐写模型并不适用于现实情况，现有编码网络大多直接将输出的浮点型张量作为解码网络的输入，从而实现秘密图像的重建. 但是，在现实世界中发送方必须要将编码网络输出的张量转换成图像后才能发送给接收方进行提取. 因此，这个过程必然会造成张量信息的丢失从而造成解码网络提取准确度的下降.

2.2.3　对抗样本

对抗样本是指在原始数据集中故意添加细微扰动导致分类器对其进行错误分类的样本[34]. 自 2014 年 Christian Szegedy 等人提出对抗样本概念之后，对抗样本就成为深度学习中不可避免的一个难题. 机器学习的显著弱点之一就是易受对抗样本的攻击，而基于深度学习的隐写分析模型则可被看作判断图像是否含有秘密消息的二分类判别器. 因此，学者们开始思考利用对抗攻击或者对抗噪声去欺骗基于深度学习的隐写分析二分类器，从而提高隐写算法安全性及抗检测性.

Ma 等人[40]在 2018 年提出了基于对抗样本的隐写网络模型，该模型有效提升传统空域自适应隐写算法的安全性. 该模型首先利用对抗样本的生成模型—快速梯度下降模型(Fast Gradirnt Sign Model，简称 FGSM)[41]，更改它在前向传播过程中 softmax 函数的输出向量$[P_c, P_s]=[0,1]$，具体流程如图 7 所示. 接着将 Softmax 的输出向量更改为$[P_c, P_s]=[0.5, 0.5]$，经过反向传播生成对抗梯度图，该梯度图中每个元素都是梯度值，使得隐写分析器倾向于具有错误的分类结果. 最后利用 STC 编码寻找到最小失真嵌入位置再结合之前生成的对抗梯度图完成秘密信息的嵌入. 该模型在保证载体图像最小嵌入失真的同时，利用对抗样本去提升模型的抗检测能力，但是，最后生成的含密图像抗检测能力并没有得到明显增强.

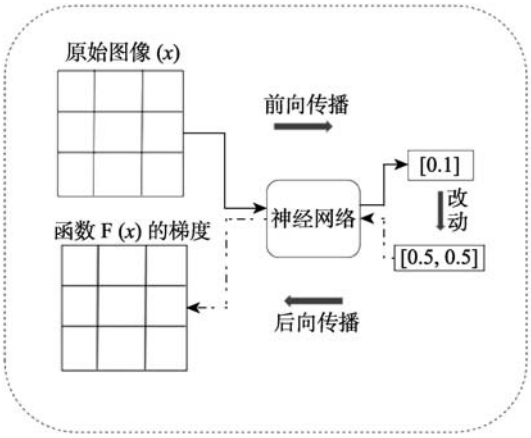

图 7　对抗梯度图生成过程

之后，Zhang 等人[42]直接利用 FGSM 生成“增强型”载体抵抗基于深度学习的隐写分析模型，该模型中增强型载体图像的生成过程如图 8 所示，它更改了对抗噪声与信息嵌入之间的先后顺序，第一步首先在载体图像中加入模拟嵌入的随机噪声生成“含密”图像，然后通过与目标隐写分析模型进行对抗训练生成对抗噪声，从而生成增强型可抵抗隐写分析模型检测的对抗载体. 第二步是在增强型的载体图像上利用自适应隐写算法完成秘密信息的嵌入. 因此，该隐写模型不仅提升了隐写安全性，同时还保证了秘密消息能够准确提取.

对抗样本不仅可以用来增强载体图像的抗检测能力还能用来调整自适应隐写算法的嵌入代价，因此，Li 等人[21]又提出对抗嵌入隐写方案 ADV-EMB. 该隐写模型根据隐写分析网络反向传播回来的梯度值非对称地调整部分原始嵌入成本. ADV-EMB 首先利用 J-UNIWARD 算法获得初始嵌入成本，然后采用对像素的一个方向除以 2，另一个方向乘以 2 的方法对 β%的初始嵌入成本进行更新，最终生成的含密图像不仅能够保证图像的最小嵌入失真还能抵抗基于深度学习模型的检测. 因为该方法仅对初始成本图造成少量修改，所以初始嵌入方法可以保留并且不会添加过多被“非目标”隐写分析模型检测到的痕迹. 但是该隐写模型对载体图像造成的改动仍会略微高于普通自适应隐写算法.

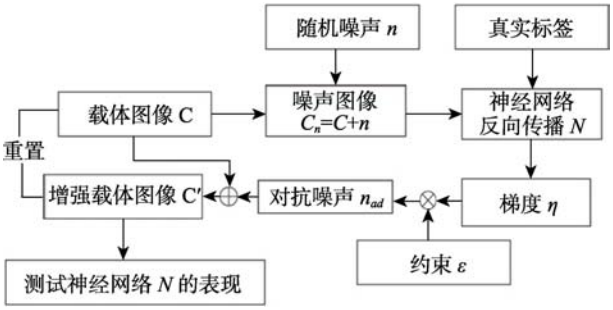

图 8 基于对抗样本的增强载体生成过程

基于嵌入载体式的深度学习隐写模型是信息隐藏领域主要隐写方式之一. 目前分别有三种嵌入式隐写方案，各类方案的优缺点对比如表 3 所示. 其中，基于最小嵌入失真代价的隐写模型能够有效提升传统自适应隐写算法的安全性，但是模型训练缓慢且抗检测能力相较于传统自适应隐写算法并没有提升. 基于编码-解码网络的隐写模型能够代替传统隐写算法并提升隐写容量，但是隐写安全性低并且无法实现秘密信息的无损提取. 基于对抗样本的隐写模型能够有效降低隐写分析网络模型的检测准确率，但是模型结构庞大且耗时久，除此以外生成的抗检测含密图像泛化性差，当隐写分析器为非目标隐写分析器时，它的抗隐写分析检测能力就会大大降低.

除了利用上述三种方法完成秘密消息的嵌入以外，Meng 等人在文献[43]中还提出利用目标识别网络 faster RCNN 识别图像中的纹理复杂区域，然后在识别出的不同区域中选择合适的自适应隐写算法，以此提升含密图像的安全性和抗检测性.

2.3 基于合成载体式深度学习隐写模型

基于合成载体式深度学习隐写模型是指在现有载体图像基础上利用 GAN 网络进行图像合成(比如图像修复、图像拼接、前景生成等) 的同时完成信息隐藏，而重新生成的载体图像不仅包含了秘密信息还对原始载体图像进行语义修复或者语义更新. 与基于载体生成式深度学习隐写方法不同，它仍需要依赖原始载体图像的部分语义信息以及结构信息. 该类隐写模型的优缺点对比如表 4 所示.

表 3 基于嵌入载体式深度学习隐写模型优缺点对比

方法	实现细节	优点	缺点
ASDL-GAN[19] UT-SCA-GAN[29]	依据与隐写分析对抗网络自动学习嵌入失真代价和嵌入改变概率，实现信息隐藏.	自动学习最小嵌入失真代价	相较于传统自适应隐写算法,抗隐写分析检测能力没有很大提升
Deep Steganography[17] Stegnet[32] Atique's model[37] ISGAN[38]	利用编码-解码网络实现以图藏图的隐写模型	扩大了隐写容量	含密图像颜色失真; 重构秘密图像质量有损
SteGAN[30] SsteGAN[31] HiDDeN[18] SteganoGAN[20]	利用编码-解码网络实秘密二进制数据及文字信息的隐藏和提取,并添加第三方网络进行对抗训练，增强隐写的安全性	提升隐写安全性和鲁棒性	难以抵抗基于深度学习隐写分析模型的检测
Ma's model[40] Zhang's model[42] ADV-EMB[21]	利用与隐写分析网络在对抗训练过程中,生成对抗扰动或对抗梯度图 ，干扰隐写分析的判别结果	抵抗基于深度学习的隐写分析模型的检测	模型结构庞大，训练耗时久; 含密图像泛化性差

表 4 基于合成载体式深度学习隐写模型优缺点对比

方法	实现细节	优点	缺点
Liu's model[22]	利用卡丹格将秘密信息隐藏到图像破损区域，利用 DCGAN 进行图像修复	提供一种新型的隐写方法，在图像修复过程中完成隐写扩大了可用作隐写的载体图像类型	信息提取准确度不稳定
SGSRGAN[23]	利用 MC-GAN 生成适合隐写的复杂前景区域，LSBM 隐写后与前景与载体图像进行拼接合成	LSBM 隐写算法安全性提升且含密图像质量高	信息提取需要依赖额外提供 mask，容易引起起攻击方的怀疑

Liu 等人[22]在 2018 年提出一种基于卡丹格掩码(Cardan Grille)的载体修复式信息隐藏框架，利用卡丹格进行秘密消息的嵌入和提取，然后再利用 DCGAN 进行受损图像的语义修复，该隐写模型如

图 9 所示，首先将秘密消息隐藏到卡丹格掩码对应的图像受损位置上，并对含有秘密消息的受损图像放入 DCGAN 网络进行重建，最后，在信息提取过程中，接收方在修复好的图像上依据卡丹格完成秘密信息的提取.

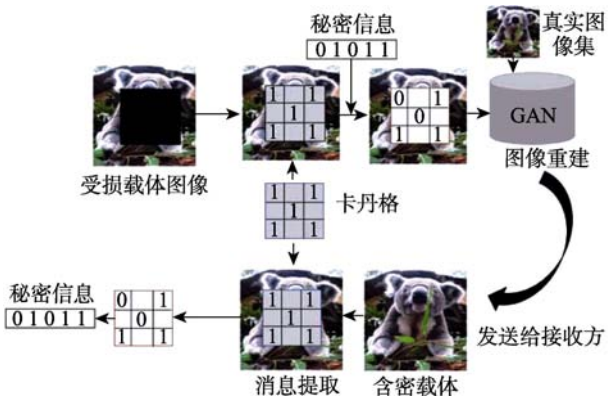

图 9　基于卡丹格的深度学习隐写模型结构

该隐写模型在训练过程中考虑了三种损失：感知损失、情境损失以及秘密信息损失. 其中，感知损失就是恢复后图像与自然图像的像素位置之间的差距，而情境损失代表的是恢复后图像与真实自然图像之间的差距. 当秘密信息损失与情境损失二者关系设置为紧密时，生成的图像内容会变差；当设置为毫无关系时，提取准确率会变低，但生成图像内容会更好. 因此，该模型需要在含密图像与重构秘密图像之间寻找一个折中点，除此以外，含密图像质量还受隐写容量（即卡丹格尺寸）的影响. 当嵌入比特过大时，导致卡丹格尺寸超过受损区域的 50%，生成的含密图像会出现肉眼可见的严重失真.

生成对抗网络（GAN）除了可以修复受损图像，还可以根据背景图像生成前景图像. Cui 等人[23]利用 MC-GAN[44]在给定载体图像上生成前景区域，并在生成过程中将秘密消息利用 LSBM 隐写算法[45]嵌入到前景区域中. 与在原始载体图中进行 LSBM 嵌入算法相比，该隐写方法增加了载体图像的区域纹理复杂度，提升了隐写算法安全性.

基于合成载体式深度学习隐写方法目前有两种实现途径：一种是先利用生成对抗网络对原始载体图像进行二次改动，生成全新的图像，并在新生成的图像区域利用传统隐写算法完成秘密信息的嵌入；另一种是在原始载体图像受损区域中利用传统隐写算法嵌入秘密信息，然后将图像放入生成对抗网络完成对图像的修复. 但是，该类隐写方法在秘密信息提取时都需要依靠额外的 mask（比如前景图像轮廓、卡丹格等）才能完成，这对隐写模型的安全性造成了一个潜在的威胁. 同时，该类隐写模型中生成对抗网络性能的好坏对于含密图像质量有着很大影响，因此，选择合适的生成对抗网络对该类隐写模型性能的提升有很大帮助.

2.4　基于映射关系式深度学习隐写模型

“无载体”信息隐藏技术相较于嵌入式信息隐藏具有天然抗隐写分析检测的优势，该技术通过在载体图像与秘密信息之间构建映射关系完成秘密信息的隐藏和提取，并且对载体图像本身不做任何改动[46]. 因此，将无载体信息隐藏技术与深度学习网络相结合，既可以保证隐写模型的安全性，又能提升秘密信息的嵌入和提取效率. 该类隐写模型结构优缺点对比如表 5 所示.

Hu 等人[24]依据 SWE（无嵌入隐写）又称作无载体隐写思想提出了一个基于 DCGAN 的隐写网络模型，其模型结构如图 10 所示. 首先，将秘密信息 m 转换成二进制序列，与[–1,1]范围的 100 维随机噪

表 5　基于映射关系式深度学习隐写模型优缺点对比

方法	实现细节	优点	缺点
Hu’s model[24]	将随机噪声与二进制秘密信息之间构建映射关系，再通过 DCGAN 将噪声作为输入生成载体图像（含密图像），并构建提取网络恢复噪声序列，最后通过映射关系恢复处秘密信息	生成的载体图像中没有嵌入和修改痕迹	隐写容量小 生成的图像不够真实和自然 秘密消息不能完全正确提取
Zhang’s model[25]	在 Hu 的基础上，利用预训练好的 CycleGAN 对生成图像进行风格迁移，并在恢复原始生成图像过程时，直接将噪声向量作为输出	风格迁移后的图像具有更高的安全性	秘密信息的提取准确度低
Meng’s model[26]	利用 faster RCNN 检测在载体图像上检测目标物体的颜色，类型等属性与秘密消息构建映射关系	相较于传统无载体隐藏方法，能够更加自动化的获取映射关系对应的秘密信息	需要建立大型数据集构建寻找与映射关系相对应的载体图像

声向量 $\boldsymbol{z}$ 构建映射关系，然后将该随机噪声作为 DCGAN 网络的输入，经过对抗训练输出尽量符合自然图像统计分布的生成图像，该图像也可看作含密载体图像. 接收方利用由反卷积层组成的提取网络提取出秘密信息. Zhang 等人[25]在 Hu 等人的基础上，利用 CycleGAN 对噪声生成的图像 C 进行了风格迁移，并对风格迁移后的图像 C'进行恢复. 在恢复原始生成图像 C 的过程中，该隐写模型直接将恢复的噪声向量 $\boldsymbol{z}$ 作为输出，最后接收方通过映射关系完成噪声向量到秘密信息的转换.

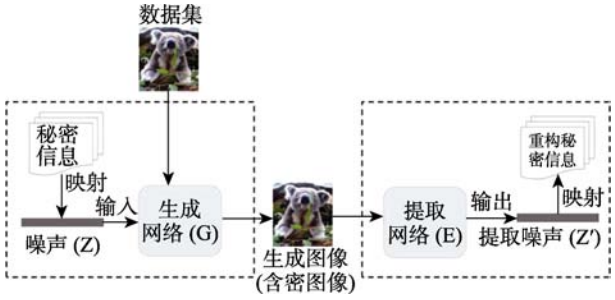

图 10 基于 SWE 的深度学习隐写模型结构

此外，Meng 等人[26]利用 VGG-19 目标检测网络以及映射关系，将秘密信息与图像中多个目标的类别、颜色等特征形成映射关系. 该隐写方法首先选择符合秘密消息对应映射关系的自然载体图像，再使用 VGG-19 检测出目标所在位置以及适合隐写的安全区域，最后使用隐写算法完成对秘密消息的隐藏和提取.

基于映射关系式深度学习隐写模型大多都是将秘密信息与深度学习网络的输入建立某种特定映射关系，Hu、Meng 等人将秘密信息分别与生成器的输入和 VGG-19 网络的输入图像之间构建映射关系. 但前者采用的映射方法是将随机噪声分段，后者采用的映射方法是在合适的图像上寻找能够构建映射关系的多个属性和对象. 该类隐写方法不会对载体图像造成二次改动，但是因为受映射多样性限制，隐写容量仍然很低.

3 实验对比与分析

该章节将包含不同类型的深度学习隐写模型的实验结果，以及不同隐写方法之间安全性的比较分析.

3.1 生成载体式深度学习隐写模型的性能表现

基于生成载体式深度学习隐写方法的实验数据集使用 CelebA 数据集. 在预处理过程中将图像裁剪成 64×64 像素，然后利用 ±1 隐写算法分别在 20 万张 CelebA 图像上嵌入 0.4bpp 的秘密信息，其中，18 万对自然图像和含密图像被当作隐写分析模型(GNCNN)[47]的训练集，另外 2 万对则被当作隐写分析网络的测试集. SGAN[15] 与 SSGAN[16]的性能表现对比如表 6 所示. SSGAN 将 SGAN 中的载体图像生成模型由 DCGAN 替换成 WGAN 来生成更加自然并适合隐写的载体图像，使得隐写分析模型（GNCNN）的检测错误率增加了 7%.

表 6 0.4 bpp 嵌入容量下，GNCNN 隐写分析模型检测下，生成载体式深度学习隐写模型的性能表现

模型	提出者	嵌入算法	检测错误率（GNCNN）
SGAN[15]	Volkhonskiy	±1	18%
SSGAN[16]	Dong Jing	±1	11%

3.2 嵌入载体式深度学习隐写方法的性能表现

在 0.4bpp 嵌入容量下，将文献[19, 21, 29]中提出的利用深度学习进行自适应嵌入的隐写模型 ASDL-GAN、ADV-EMB 以及 UT-SCA-GAN 与传统空域隐写算法 S-UNIWARD 的安全性进行比较，实验结果如表 7 所示. 该实验设定 10000 张 512×512 尺寸的 BOSSbase 图片作为实验数据集. 经过 18000 次迭代训练后的 ASDL-GAN 使得隐写分析模型 SRM[8]以及 Xu'Net [48]的检测错误率相较于 S-UNIWARD 算法[4]降低了 3%左右，Yang 等人[29]在 ASDL-GAN 的基础上针对生成模型与嵌入模拟器做出改进，使其隐写分析模型（SRM）的检测错误率相较于 S-UNIWARD 算法提升 2%左右.

表 7 0.4 bpp 嵌入容量下，自动学习隐写失真框架和嵌入成本的深度学习隐写模型与传统自适应隐写算法的性能比较

隐写模型及算法	隐写分析模型	
	SRM	Xu'Net
ASDL-GAN[19](18000 次迭代后)	17.4%	16.20%
UT-SCA-GAN[29]	22.36%	\
S-UNIWARD[4]	20.50%	20.10%

基于编码-解码网络的隐写模型相比于其他隐写算法及模型，它在扩大隐写容量的同时提升了含密图像质量. 为了评估编码-解码网络隐写模型隐藏二进制比特信息的性能表现，我们将其与 HUGO、S-UNIWARD 等空域隐写算法进行比较，并用隐写分析模型 Eve（4Conv+FC+Sigmoid）和 ATS[49]评价隐写算法的安全性，实验结果如表 8 所示. 该实验

的数据集使用 10000 张 32×32 像素的 CelebA 数据集，隐写容量最高设置为 0.4bpp. 其中，HiDDeN 模型适用于任意尺寸大小的实验数据集，但由于秘密信息扩展成的数据张量维数过大导致最大隐写容量仅达到 0.2bpp 左右.

SsteGAN[31]在 SteGAN[30]的基础上添加 Dev 方即判别器方与 Alice 方即编码器方进行 5K 次迭代对抗训练，最终使得隐写分析模型（EVE 方）检测错误率增加了 30%. 但是，当嵌入容量为 0.1bpp 时 SteGAN 的解码准确率为 99.8%，而 SsteGAN 的解码准确率只有 98.8%. 除此以外，HiDDeN 模型的隐写容量低于其他基于编码-解码的隐写模型，但是生成的含密图像具有更高的鲁棒性. 即使含密图像加入噪声后，解码网络所提取的秘密信息准确率仍能达到 99.9%. 而 SteganoGAN 模型的隐写容量上限相比于上述的 SsteGAN 和 Stegan 扩大了 10 倍，最高可达到 4.4bpp[20]. 具体实验结果如表 9 所示，实验数据集使用种类更加多样性的 ImageNet[50]. 在隐写分析模型 Ye'Net[12]固定 20%的检测错误率下，SteganoGAN 的隐写容量上限仍能达到 2bpp. 从侧面说明 SteganoGAN 的隐写安全性比传统自适应隐写算法更高.

表 8　基于编码-解码网络的深度学习隐写模型与自适应隐写算法的比较

隐写分析模型	隐写模型及算法								
	0.4 bpp		0.4 bpp/0.2 bpp		0.4 bpp/0.2 bpp		0.4 bpp/0.2 bpp		0.2 bpp
	SteGAN[30]	SsteGAN[31]	S-UNIWARD[4]		HUGO[7]		WOW[6]		HiDDeN[18]
EVE	10%	40.8%	23.4%	\	21.8%	\	22.7%	\	\
ATS	5%	\	10%	32%	6%	30%	11%	32%	50%（模型权重未知）

利用编码-解码网络进行以图藏图将隐写算法的隐写容量提高一个量级，Deep Steganography[17]与 Stegnet[32]两种隐写模型的秘密信息皆为彩色图像. 为了比较两种隐写模型的安全性，我们利用 ImageNet 作为数据集，并且使用 StegExpose[51]作为评估模型，具体实验结果如图 11 所示. 从图中可以明显看出 Stegnet 隐写模型在 StegExpose 工具的检测性只比随机猜测好一点，随着检测预知的随机差值不断上升，隐写分析的检测准确性也在不断增强. 但从整体效果来看，Stegnet 的隐写安全性明显高于 Baluja 等人提出的 Deep Steganography 模型，由此说明，针对含密图像进行方差损失函数的构建不仅能够提升含密图像的质量，还能增强模型的抗检测能力.

表 9　当 Ye'Net 隐写分析模型固定为 20%的检测错误率，各隐写算法的隐写容量

隐写模型及算法	嵌入容量（bpp）
SteganoGAN[20]	2bpp
HUGO[7]	0.5bpp
S-UNIWARD[7]	0.4bpp
WOW[6]	0.3 bpp

而文献[37, 38]中提出的以彩图藏灰度图的隐写模型，隐写容量皆为 8bpp. 其中，Zhang 等人[38]在 Atique 等人提出的模型基础上提出的 ISGAN，针对秘密图像的不可见性做出了改进. 为了验证秘密图像的不可见性，我们使用 256×256 像素的 ImageNet 和 LFW 数据集，并根据 PSNR[39]和 SSIM[52]两种最常用的图像质量评估指标，分别对两种模型在两种数据集上生成的 1K 张含密图像进行比较，实验结果如表 10 所示.

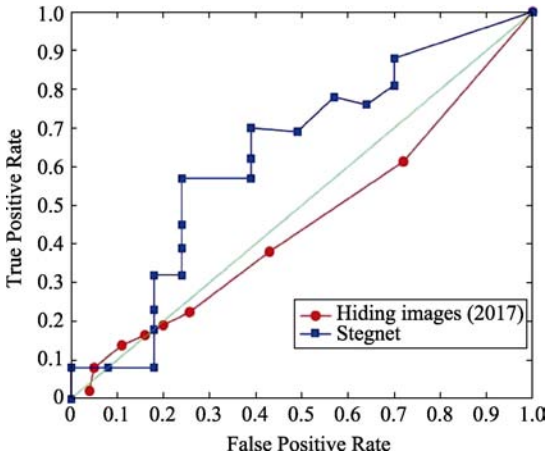

图 11　ROC 曲线：StegExpose 隐写分析模型下，以彩图藏彩图的隐写模型性能比较

从实验结果来看，相比于 Atique 模型，ISGAN 模型的改进使得含密图像质量得到了提升，但是 ISGAN 提取的秘密图像质量却出现了下降. 由此说明 ISGAN 将秘密图像隐藏在单通道（Y 通道）中能够有效减少图像亮度以及色彩信息的丢失，但是解码网络仍需进一步改进. 利用对抗样本进行信息隐藏的主要目标是使得隐写分析网络模型对于含

表 10 以彩图藏灰度图的隐写模型生成的含密图像与重构图像质量比较

隐写模型	数据集	含密图像（PSNR）	重构秘密图像(SSIM)	秘密图像(SSIM)	重构秘密图像(SSIM)
ISGAN[31]	LFW	34.6	33.6	0.957	0.94
Atique's model[30]	LEW	33.7	39.9	0.95	0.95
ISGAN[31]	ImageNet	34.9	33.4	0.97	0.95
Atique's model[30]	ImageNet	32.9	36.6	0.96	0.96

密图像的分类出现错误，从而抵抗隐写分析网络模型的检测. 该实验数据集为 ImageNet，在 0.4bpp 隐写容量下，由文献[37]提出的模型生成增强型载体图像使得目标分析模型（Xu'Net）的检测错误率可以达到 56%，SRM 的检测错误率可以达到 36%.

而文献[21]提出的模型则根据与目标隐写分析模型（Xu'Net）在对抗训练过程中生成的对抗梯度图，来修改由 J-UNIWARD 算法选取的待改变像素的像素翻转方向. 该隐写模型在 0.5bpnzAC 隐写容量下，使得未使用对抗含密图像训练的隐写分析模型 Xu'Net 检测错误率达到了 29%,使得使用对抗含密图像训练的隐写分析模型 Xu'Net 检测错误率达到了 24%左右. 相较于普通的 J-UNIWARD 隐写算法,两种不同方式训练的隐写分析网络 Xu'Net 检测错误率分别增加了大约 10%和 5%. 由此说明,通过与隐写分析网络对抗训练生成的对抗含密图像能够有效干扰基于深度学习的隐写分析模型的判断.

3.3 合成载体式深度学习隐写模型的性能表现

基于合成载体式深度学习隐写模型在嵌入信息的过程中利用生成对抗网络将已有载体图像变成一张“新”的图像，且该图像与原始载体图像并不完全相同,因此该类隐写方法所生成的图像的质量好坏对隐写安全性有很大影响. 文献[22]提出的隐写模型将中心剪切后的 64×64 像素 LFW 数据集作为受损图像，其中，受损区域大小为 32×32 像素. 最终，该模型对于秘密消息的提取准确率最高仅能达到 42%左右. 并且当待嵌入的卡丹格尺寸大于 32×32 像素时，生成的含密图像质量肉眼可见的下降，隐写模型不再具有实际意义. 而文献[23]提出的 SGSR-GAN 模型，实验数据集为 CUB bird image 数据集，最终生成的含密图像的 SSIM 取值范围大多都在[0.996,1]之间. 在相同嵌入容量下，经过 SGSR-GAN 进行 LSB 隐写后的含密图像相较于普通 LSB 隐写后生成的含密图像，欺骗 SPA 隐写分析器的成功率提升了大约 50%. 因此，基于合成载体式的深度学习隐写模型方法无论是秘密信息的提取准确度还是隐写安全性，都还存在着较大提升空间.

3.4 映射关系式深度学习隐写模型的性能表现

基于映射关系式隐写模型优缺点十分明显，隐写安全性得到保证的同时，隐写容量十分局限. Hu 等人在文献[24]中将任意二进制比特信息映射成均匀分布的[–1,1]的高斯噪声作为 DCGAN 的输入生成 64×64 像素的合成图像,其隐写容量仅达到 9.16×10^{-3}bpp，秘密信息的提取准确率能达到 90%以上. 隐写分析模型（Ye'Net）的训练集由文献[24]的隐写模型生成的含密图像构成，模型的检测准确率可以达到 98%；当训练集不包含由文献[24]的隐写模型生成的含密图像时，Ye'Net 的检测准确率仅有 47%. 因此，该类隐写方法安全性取决于由隐写模型生成的含密图像数据集是否被泄露和恶意采集.

4 存在的问题

尽管基于深度学习的隐写模型在近几年的时间里已经获得长足发展并取得了优异表现，不仅扩大了隐写容量还提升了隐写安全性，但是目前各类隐写模型都还存在一些未能解决的问题，在未来需要去进一步改善：

（1）生成的载体图像不够真实自然. 目前基于生成载体式深度学习隐写模型通过生成更加适合隐写的合成载体图像，以提高待嵌入秘密信息的安全性. 但是，文献[15]和文献[16]所使用的生成网络分别是 DCGAN 和 WGAN. 其中，DCGAN 自身存在训练不平衡问题,使得 GAN 网络在生成过程中极其不稳定. WGAN 虽然从理论上证明了 DCGAN 训练不稳定的原因—损失函数（交叉熵）不适合衡量不相交分布间的距离，而在实际训练过程中它仅解决了 DCGAN 的模式崩溃问题，并未很好提升生成图像的质量. 因此，该类模型的安全性极大地依赖于生成图像的质量及语义合理性. 在未来，模型可以结合更加先进的 BEGAN 或者 PG-GAN 等网络，生成高清、纹理细节丰富的载体图像.

（2）利用 FGSM 攻击算法生成对抗含密图像耗时久，泛化性差. 目前隐写模型中生成对抗样本采用的方法都是—FGSM，而 FGSM 是通过求出模

型对输入的导数，然后用符号函数得到具体的梯度方向，接着再乘以一个步长，得到的“扰动”再添加到原本的输入，最终得到一个对抗样本. 换句话说，隐写分析模型是固定不变的，唯一改变的是它的输入也就是含密图像，因此生成的对抗含密图像仅对目标隐写分析模型的抗检测能力比较强，而对其他隐写分析模型的抗检测能力就会削弱很多. 除此以外，FGSM 是单次梯度更新，所以有些时候含密图像利用 FGSM 算法得到的扰动不足以被误分类，这样就导致生成对抗含密图像的过程耗时过久.

（3）隐写模型安全性差，难以抵抗隐写分析模型的检测. 这个问题是针对基于编码-解码网络大容量隐写模型提出的，该类隐写模型可以在载体图像中成功隐藏同尺寸大小的秘密图像，但是隐写容量的扩大意味着对于载体图像的统计分布改动较大，因此该类方法难以抵抗隐写分析模型的检测.

（4）提取网络无法实现秘密信息的无损恢复. 该问题不仅存在于基于映射关系式的隐写模型中，还存在于基于编码-解码网络的隐写模型中. 因为深度学习网络会经过一系列的池化和归一化等操作，所以导致图像内部很多细节信息的丢失，从而使得提取网络难以无损地恢复出原始秘密图像或者原始输入噪声.

（5）秘密信息的提取需要依靠额外信息以及含密图像语义性不合理. 该问题是针对基于合成载体式深度学习隐写模型提出的，该类隐写模型目前无论是将含有秘密信息的卡丹格藏于载体图像受损区域，或是将秘密信息藏于根据载体图像生成的前景区域，最终接收方若想成功提取秘密信息都要依靠额外信息（比如卡丹格，前景区域 MASK 等）才能实现. 而这些额外信息在传递过程中容易招到窃听方的怀疑，导致秘密信息不安全. 不仅如此，该类隐写方法生成的改动图像存在质量较差，语义不合理等问题，当卡丹格的尺寸大于原载体图像受损区域的 50%，最终生成的图像会出现肉眼可见的失真和扭曲.

5　基于对抗样本的大容量隐写方法

针对上一节中提到的基于编码-解码网络的大容量隐写模型存在的安全性较低问题，本文提出了如下具体改进方法.

在现实情况中，发送方需要利用编码网络将秘密图像编码到载体图像中并将含密图像发送给接收方，接收方需要利用解码网络从含密图像中提取秘密信息. 但是，在采用端对端方式的训练模型中，解码网络的输入是编码网络输出的一个没有丢失任何信息的浮点张量. 这样就使得当我们将编码网络的输出额外保存成秘密图像再传递给解码网络时，解码网络的提取准确率明显降低. 除此以外，发送方在载体图像编码的秘密信息量过多，换算下来相当于在每个像素中隐藏了 8Bit 的秘密信息（一个像素可以换算成 8 比特）[38]，所以导致生成的含密图像存在较大的安全问题，难以抵抗隐写分析网络模型的检测.

于是，本文提出在编码-解码网络的隐写模型中加入对抗样本来提升含密图像的安全性，并且模型不采用端对端的训练模式，从而能够使提取网络从含密图像中准确提取秘密图像. 该想法的具体框架结构如图 12 所示，该隐写模型由三部分组成：隐藏网络（编码网络），对抗噪声生成网络，提取网络（解码网络）. 首先利用隐藏网络将秘密图像 S 编码到载体图像 C 中，然后将生成的普通含密图像 C’放入对抗噪声生成网络，让它与目标隐写分析对抗来生成对抗噪声. 最后将叠加了对抗噪声的含密图像 C’’放入解码网络中提取秘密图像.

其中，对抗噪声生成网络也包含三个子网络：噪声生成网络、判别网络、目标网络. 噪声生成网络 G 将含密图像 Stego 作为输入，生成能够成功干

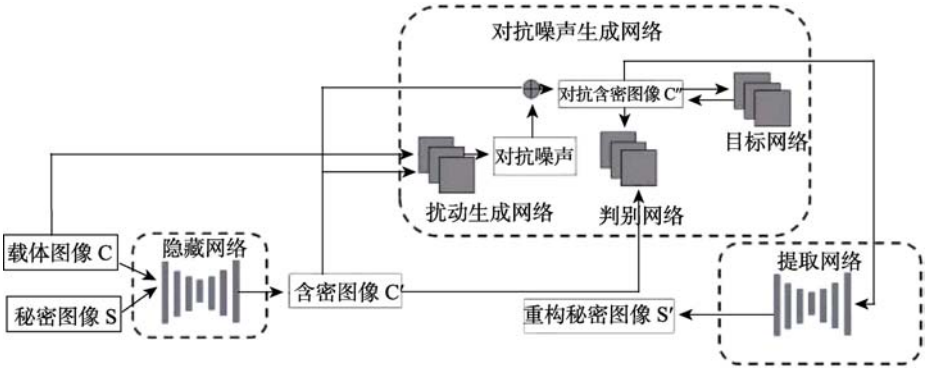

图 12　对抗隐写图像的隐藏和提取过程

扰隐写分析分类的对抗噪声. 判别网络 D 用于缩小对抗含密图像与含密图像之间的差别，减少对抗噪声对于含密图像质量造成的影响. 目标网络 f 即预训练好的隐写分析网络通过输出的概率值，引导对抗噪声的生成，最终使得叠加了对抗噪声的含密图像 C''尽可能被分类为载体图像.

我们将 50000 张 ImageNet 数据集预处理成 256×256 像素大小的图像，一半作为秘密图像集，一半作为载体图像集. 并将两个图像集输入到编码网络生成 25000 张含密图像，然后分别将含密图像与载体图像打上标签，其中载体图像标签为 1，含密图像标签为 0. 之后，将打上标签的两个图像集放入对抗噪声生成网络中添加对抗噪声. 最后，利用目标网络 f 进行分类测试，把错误分类成标签 1 的含密图像保存下来放入到提取网络中，进行秘密图像的提取. 初步实验结果显示，该模型目前能够使得目标网络的分类错误率达到 50%，其中，含密图像被错分为载体图像的比例在 20%左右，载体图像被分类为含密图像的比例在 30%左右. 并且，加入了对抗噪声的含密图像提取成功率较低，大约仅在 40%.

从初步实验结果来看，利用对抗样本能够有效提升基于编码-解码网络的大容量隐写模型的安全性. 另外，我们还对实验结果进行了分析：（1）针对对抗含密图像生成概率低的问题，我们认为是含密图像与载体图像的概率分布相差太大或者对抗噪声生成网络的损失函数设计不合理，所以导致有效的对抗噪声难以生成. （2）针对秘密图像提取成功率低的问题，我们认为是含密图像叠加对抗噪声幅度太强，导致提取网络无法从噪声干扰中完全恢复出秘密图像.

6 总结与展望

本文对目前各类基于深度学习的图像隐写模型进行了分析与总结，其中，基于载体嵌入式隐写模型是目前利用深度学习进行隐写所采用的主流方法. 深度学习网络在训练过程中不仅能够自动学习最小嵌入失真代价来降低载体图像的失真率，还可以替代普通隐写算法完成秘密信息的隐藏及提取. 因此，将深度学习与信息隐藏领域相结合，不仅增强了普通自适应隐写算法的安全性，还扩大了隐写容量. 另外，本文探讨了各类隐写模型存在的一些问题，比如基于生成载体式隐写模型生成的图像不够真实、基于编码-解码网络的大容量嵌入式隐写模型的隐写安全性低、基于合成载体式隐写模型提取信息时依赖额外信息且图像修复拼接后效果不自然、基于映射关系式隐写模型的嵌入容量小等. 本文针对基于编码-解码网络的大容量隐写模型的安全问题，提出了利用对抗样本提升隐写安全性的新思路，并进行了初步实验验证其有效性. 综上所述，深度学习能够有效提升隐写模型各方面的性能，基于深度学习的数字图像隐写术是信息隐藏领域未来发展的重要方向之一.

本文针对上述问题，结合目前深度学习网络的发展状况，对未来基于深度学习的隐写模型的发展方向进行了几点展望：

（1）隐写术的安全性是网络能够进行安全通信的重要保证. 近几年来，基于深度学习的各类隐写分析模型飞速发展，比如：SRNet[13]、Ye'Net[12]、Xu'Net[48]等. 目前，基于深度学习的隐写模型采用的方法大多是通过与隐写分析模型的对抗训练，根据反向传播回来的梯度，更新生成器的网络参数重新生成含密图像. 但是，该方法对于隐写模型的优化效果并不明显，隐写分析的检测准确率仍然能达到 70%甚至以上. Zhang 等人提出在载体图像中添加对抗扰动，提升含密图像的安全性和抵抗隐写分析模型检测的能力，并取得了优异的效果. 因此，如何利用对抗样本构建一个高效的端对端深度学习隐写模型，有效抵抗隐写分析网络模型的检测是未来需要继续研究的方向.

（2）基于深度学习的隐写模型大多是利用生成对抗网络中对抗学习的思想提升隐写算法安全性和含密图像质量. 但是，目前利用隐写分析网络进行对抗训练的隐写模型，其含密图像的抗隐写分析检测能力仍存在较大提升空间. 对抗学习存在寻找纳什平衡的过程，但是网络之间最终达到的相对平衡状态并不代表网络所能达到的最优状态. 因此，在未来可以考虑将强化学习应用到深度学习模型中，将预先训练好的隐写分析网络当做理想状态，然后利用强化学习将隐藏网络训练到优于理想状态的最优状态.

（3）基于编码-解码网络的大容量隐写模型近两年来不断涌现. 编码-解码网络可以被看做是端对端的隐藏和提取网络，秘密图像被编码网络编码到载体图像的可用位，再由解码网络解码出秘密图像. 秘密信息的提取准确度是衡量隐写方法好坏的重要衡量标准之一，但现有的该类隐写模型，比如：Deep Steganography[17]、Stegnet[32]、ISGAN[38]等都存在重构秘密图像质量较差的问题. 所以，如何构建一个能够更精确的提取出秘密图像的提取网络，或者通

过设计一个新的多元损失函数，缩小重构秘密图像和原秘密图像在特征分布之间的距离，从而更有效的训练提取网络. 这些都是基于编码-解码网络的隐写模型未来可以继续研究的方向.

参 考 文 献

[1] Feng Deng-Guo, Zhang Min, Li Hao, et al. Big data security and privacy protection. Chinese Joural of Computers, 2014, 37(1): 246-258 (in chinese)
(冯登国, 张敏, 李昊. 大数据安全与隐私保护. 计算机学报, 2014, 37(1): 246-258)

[2] Zhai Li-Ming, Jia Ju, Ren Wei-Xiang, et al. Progress in deep learning in the field of image steganography and steganalysis. Journal of Cyber Security, 2018, 3(6): 2-12 (in chinese)
(翟黎明, 嘉炬, 任魏翔, 等. 深度学习在图像隐写术与隐写分析领域中的研究进展. 信息安全学报, 2018, 3(6): 2-12)

[3] Liu Jia, Ke Yan, Lei Yu, et al. Recent advances of image steganography with generative adversarial networks. arXiv preprint arXiv:1907.01886, 2019

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

[5] Li Bin, Wang Ming, Huang Ji-Wu, et al. A new cost function for spatial image steganography//Proceedings of the 2014 IEEE International Conference on Image Processing. Paris, France, 2014: 4206-4210

[6] Holub V, Fridrich J. Designing steganographic distortion using directional filters//Proceedings of the 2012 IEEE International workshop on information forensics and security. Tenerife, Spain, 2012: 234-239

[7] Pevný T, Filler T, Bas P. Using high-dimensional image models to perform highly undetectable steganography//Proceedings of the International Workshop on Information Hiding. Calgary, Canada, 2010: 161-177

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

[9] Denemark T, Sedighi V, Holub V, et al. Selection-channel-aware rich model for steganalysis of digital images//Proceedings of the 2014 IEEE International Workshop on Information Forensics and Security (WIFS). Georgia, USA, 2014: 48-53

[10] Schmidhuber J. Deep learning in neural networks: an overview. neural networks, 2015, 61: 85-117

[11] Hinton G, Osindero S, Welling M, et al. Unsupervised discovery of nonlinear structure using contrastive backpropagation. Cognitive science, 2006, 30(4): 725-731

[12] Ni Jiang-Qun, Ye Jian, Yang Yi. Deep learning hierarchical representations for image steganalysis. IEEE Transactions on Information Forensics and Security, 2017, 12(11): 2545-2557

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

[14] Goodfellow I, Pouget A J, Mirza M, et al. Generative adversarial nets//Proceedings of the 2014 International Conference on Neural Information Processing Systems. Montreal, Canada, 2014: 2672-2680

[15] Vplkonskiy D, Borisenko B, Burnaev E. Generative adversarial netwoks for image steganography//Proceedings of the Open Review Conference on Learning Representations. Puerto Rico, USA, 2016

[16] Shi Hai-Chao, Dong Jing, Wang Wei, et al. SSGAN: secure steganography based on generative adversarial networks// Proceedings of the Pacific Rim Conference on Multimedia. Harbin, China, 2017: 534-544

[17] Baluja S. Hiding images in plain sight: deep steganography// Proceedings of the Advances in Neural Information Processing Systems. Long Beach, USA, 2017: 2069-2079

[18] Zhu Ji-Ren, Kaplan R, Johnson J, et al. HiDDeN: hiding data with deep networks//Proceedings of the European Conference on Computer Vision. Munich, Germany, 2018: 657-672

[19] Tang Wei-Xuan, Tan Shun-Quan, Li Bin, et al. Automatic steganographic distortion learning using a generative adversarial network. IEEE Signal Processing Letters, 2017, 24(10): 1547-1551

[20] Zhang K A, Cuesta-Infante A, Xu Lei, et al. SteganoGAN:High capacity image steganography with gans. arXiv:1901.03892, 2019

[21] Tang Wei-Xuan, Li Bin, Tan Shun-Quan, et al. Cnn-based adversarial embedding for image steganography. IEEE Transactions on Information Forensics and Security, 2019, 14(8): 2074-2087

[22] Liu Jia, Zhou Tan-Ping, Zhang Zhuo, et al. Digital cardan grille: a modern approach for information hiding//Proceedings of the 2nd International Conference on Computer Science and Artificial Intelligence. Shenzhen, China, 2018: 441-446

[23] Cui Qi, Zhou Zhi-Li, Fu Zhang-Jie, et al. Image steganography based on foreground object generation by generative adversarial networks in mobile edge computing with internet of things. IEEE Access, 2019, 7: 90815-90824

[24] Hu Dong-Hui, Wang Liang, Jiang Wen-Jie , et al. A novel image steganography method via deep convolutional generative adversarial networks. IEEE Access, 2018, 6: 38303-38314

[25] Di Fu-Qiang, Liu Jia, Zhang Zhuo et al. Somewhat Reversible data hiding by image to image translation. arXiv preprint arXiv:1905.02872, 2019

[26] Meng Ruo-Han, Zhou Zhi-Li, Cui Qi, et al. A novel steganography scheme combining coverless information hiding and steganography. Joural of Information Hiding and Privarcy Protection, 2019, 1(1): 43-48

[27] Radford A, Metz L, Chintala S. Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434, 2015

[28] Arjovsky M, Chintala S, Bottou L. Wasserstein generative adversarial networks//Proceedings of the International Conference on Machine Learning. Sydney, Australia, 2017: 214-223

[29] Yang Jian-Hua, Liu Kai, Kang Xian-Qui, et al. Spatial image steganography based on generative adversarial network. arXiv preprint arXiv:1804.07939, 2018

[30] Hayes J, Danezis G. Generating steganographic images via adversarial training//Proceedings of the Advances in Neural Information Processing Systems. Los Angeles, USA, 2017: 1954-1963

[31] Wang Zi-Han, Gao Neng, Wang Xin, et al. SsteGAN: Self-learning steganography based on generative adversarial networks//Proceedings of the International Conference on Neural

Information Processing. Siem Reap, Cambodia, 2018: 253-264
[32] Wu Ping, Yang Yang, Li Xiao-Qiang. Stegnet: mega image steganography capacity with deep convolutional network. Future Internet, 2018, 10(6): 54
[33] Zegedy C, Liu Wei, Jia Yang-Qing, et al. Going deeper with convolutions//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA, 2015: 1-9
[34] He Kai-Ming, Zhang Xiang-Yu, Ren Shao-Qing, et al. Deep residual learning for image recognition//Proceedings of the IEEE conference on computer vision and pattern recognition. Las Vegas, USA, 2016: 770-778
[35] Duan Xin-Tao, Jia Kai, Li Bao-Xia, et al. Reversible image steganography schemebased on a U-Net structure. IEEE Access, 2019, 7: 9314-9323
[36] Baluja S. Hiding images within images. IEEE transactions on pattern analysis and machine intelligence, DOI:10.1109/TPAMI. 2901877, 2019
[37] Rehman A , Rahim R, Nadeem S，et al. End-to-end trained cnn encoder-decoder networks for image steganography. arXiv: 1711.07201, 2017
[38] Zhang Ru, Dong Shi-Qi, Liu Jian-Yi. Invisible steganography via generative adversarial networks. Multimedia Tools and Applications, 2019, 78(7): 8559-8575
[39] Hore A, Ziou D. Image quality metrics: psnr vs. ssim// Proceedings of the 20th International Conference on Pattern Recognition. Istanbul, turkey, 2010: 2366-2369
[40] Ma Sai, Zhao Xian-Feng, Liu Ya-Qi. Adaptive spatial steganography based on adversarial examples. Multimedia Tools and Applications, 2019 78(22): 32503-32522
[41] Goodfellow I J, Shlens J, Szegedy C. Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572, 2014
[42] Zhang Yi-Wei, Zhang Wei-Ming, Chen Ke-Jiang, et al. Adversarial examples against deep neural network based steganalysis// Proceedings of the 6th ACM Workshop on Information Hiding and Multimedia Security. Innsbruck, Austria, 2018: 67-72
[43] Meng Ruo-Han, Rice S G, Wang Jin, et al. A fusion steganographic algorithm based on faster r-cnn. Computers, Materials&Continua, 2018, 55(1): 1-16
[44] Park H, Yoo Y J, Kwak N. MC-GAN: multi-conditional generative adversarial network for image synthesis. arXiv preprint arXiv:1805.01123, 2018
[45] Mielikainen J. LSB Matching revisited. IEEE Signal Processing Letters, 2006, 13(5): 285-287
[46] Zhou Zhi-Li, Sun Hui-Yun, Harit R, et al. Coverless image steganography without embedding//Proceedings of the International Conference on Cloud Computing and Security. Haikou, China, 2015: 123-132
[47] Qian Yin-Long, Dong Jing, Wang Wei, et al. Deep learning for steganalysis via convolutional neural networks//Proceedings of the Media Watermarking, Security, and Forensics. San Francisco, USA, 2015: 94090J
[48] Xu Guan-Shuo, Wu Han-Zhou, Shi Yun-Qing. Structural design of convolutional neural networks for steganalysis.IEEE Signal Processing Letters, 2016, 23(5): 708-712
[49] Lerch-Hostalot D, Megías D. Unsupervised steganalysis based on artificial training sets. Engineering Applications of Artificial Intelligence. Francisco, USA, 2016, 50: 45-59
[50] Deng Jia, Dong Wei, Socher R, et al. ImageNet: a large-scale hierarchical image database//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Miami, USA, 2009: 248-255
[51] Boehm B. Stegexpose-a tool for detecting lsb steganography. arXiv preprint arXiv:1410.6656, 2014
[52] Wang Zhou, Bovik A C, Sheikh H R, et al. Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing, 2004, 13(4): 600-612

FU Zhang-Jie Ph.D., professor. His research interests include information hiding, digital forensics, network and information security.

WANG Fan B.S. Her research interests include information hiding and deep learning.

SUN Xing-Ming Ph.D., professor. His research interests include network and information security.

WANG Yan B.S. His research interests include information hiding and deep learning.

Background

Steganography is an important means to ensure the security of network information. It not only guarantees the security of the information itself, but also ensures the security of the information transmission process. The advent of the era of big data makes data security easy to become one of the key issues of concern, so it has achieved amazing results in the field of information hiding. This paper will classify and analyze image steganography based on deep learning in recent years. According to different steganography methods, the steganographic models based on deep learning are divided into four categories. At the same time, the paper also analyzes the advantages and disadvantages of various steganographic models. At last, the paper proposes a new steganographic method and gives the future development directions.

This paper is supported by National Natural Science Foundation of China (U1836110, U1836208), and National Key Research and Development Project Natural Science Foundation(2018YFB1003205). These projects aim to promote the development of information hiding technology enrich the theoretical knowledge in the field of information hiding and ensure information security in cyberspace.