软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2024,35(3):1502−1514 [doi: 10.13328/j.cnki.jos.006815]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

基于注意力机制的高容量通用图像隐写模型*

袁 超, 王宏霞, 何沛松

(四川大学 网络空间安全学院, 四川 成都 610065)

通信作者: 王宏霞, E-mail: hxwang@scu.edu.cn

摘 要: 随着深度学习与隐写技术的发展, 深度神经网络在图像隐写领域的应用越发广泛, 尤其是图像嵌入图像这一新兴的研究方向. 主流的基于深度神经网络的图像嵌入图像隐写方法需要将载体图像和秘密图像一起输入隐写模型生成含密图像, 而最近的研究表明, 隐写模型仅需要秘密图像作为输入, 然后将模型输出的含密扰动添加到载体图像上, 即可完成秘密图像的嵌入过程. 这种不依赖载体图像的嵌入方式极大地扩展了隐写的应用场景, 实现了隐写的通用性. 但这种嵌入方式目前仅验证了秘密图像嵌入和恢复的可行性, 而对隐写更重要的评价标准, 即隐蔽性, 未进行考虑和验证. 提出一种基于注意力机制的高容量通用图像隐写模型 USGAN, 利用注意力模块, USGAN 的编码器可以在通道维度上对秘密图像中像素位置的扰动强度分布进行调整, 从而减小含密扰动对载体图像的影响. 此外, 利用基于 CNN 的隐写分析模型作为 USGAN 的目标模型, 通过与目标模型进行对抗训练促使编码器学习生成含密对抗扰动, 从而使含密图像同时成为攻击隐写分析模型的对抗样本. 实验结果表明, 所提模型不仅可以实现不依赖载体图像的通用嵌入方式, 还进一步提高了隐写的隐蔽性.

关键词: 图像隐写; 注意力机制; 生成对抗网络; 对抗样本

中图法分类号: TP309

中文引用格式: 袁超, 王宏霞, 何沛松. 基于注意力机制的高容量通用图像隐写模型. 软件学报, 2024, 35(3): 1502–1514. http://www.jos.org.cn/1000-9825/6815.htm

英文引用格式: Yuan C, Wang HX, He PS. High-capacity Universal Image Steganographic Model Based on Attention Mechanism. Ruan Jian Xue Bao/Journal of Software, 2024, 35(3): 1502–1514 (in Chinese). http://www.jos.org.cn/1000-9825/6815.htm

High-capacity Universal Image Steganographic Model Based on Attention Mechanism

YUAN Chao, WANG Hong-Xia, HE Pei-Song

(School of Cyber Science and Engineering, Sichuan University, Chengdu 610065, China)

Abstract: With the development of deep learning and steganography, deep neural networks are widely used in image steganography, especially in a new research direction, namely embedding an image message in an image. The mainstream steganography of embedding an image message in an image based on deep neural networks requires cover images and secret images to be input into a steganographic model to generate stego-images. But recent studies have demonstrated that the steganographic model only needs secret images as input, and then the output secret perturbation is added to cover images, so as to embed secret images. This novel embedding method that does not rely on cover images greatly expands the application scenarios of steganography and realizes the universality of steganography. However, this method currently only verifies the feasibility of embedding and recovering secret images, and the more important evaluation criterion for steganography, namely concealment, has not been considered and verified. This study proposes a high-capacity universal steganography generative adversarial network (USGAN) model based on an attention mechanism. By using the attention module, the USGAN encoder can adjust the perturbation intensity distribution of the pixel position on the channel dimension in the secret image, thereby reducing the influence of the secret perturbation on the cover images. In addition, in this study, the CNN-based steganalyzer is

* 基金项目: 国家自然科学基金 (61972269, 61902263); 四川省科技计划 (2022YFG0320); 中国博士后科学基金 (2020M673276)
收稿时间: 2021-09-15; 修改时间: 2022-08-19; 采用时间: 2022-10-12; jos 在线出版时间: 2023-05-10
CNKI 网络首发时间: 2023-05-11

used as the target model of USGAN, and the encoder learns to generate a secret adversarial perturbation through adversarial training with the target model so that the stego-image can become an adversarial example for attacking the steganalyzer at the same time. The experimental results show that the proposed model can not only realize a universal embedding method that does not rely on cover images but also further improves the concealment of steganography.

Key words: image steganography; attention mechanism; generative adversarial network (GAN); adversarial example

图像隐写是一种将秘密信息隐藏在载体图像中从而获得含密图像, 然后通过含密图像恢复秘密信息的信息隐藏技术, 常用于隐蔽通信等用途. 评价图像隐写算法的基本标准是隐蔽性和嵌入容量, 隐蔽性要求含密图像的失真尽可能小, 且难以被隐写分析检测到, 嵌入容量则代表了载体图像中可以隐藏的秘密信息量. 因此, 如何在保证隐蔽性的前提下进一步提高隐写算法的嵌入容量是图像隐写发展的一个重要方向. 从最低有效位 (least significant bit, LSB) 隐写算法, 发展到基于最小失真和综合网格编码 (syndrome-trellis code, STC) 框架[1]的自适应隐写算法, 如 HUGO[2]、WOW[3]、S-UNIWARD[4]和 HILL[5]等, 图像隐写算法的隐蔽性不断提高, 但嵌入容量往往都在 0.5 bpp (bits per pixel) 以下, 并无明显改变. 直到基于深度学习的图像隐写算法的出现, 大大提高了隐写的嵌入容量, 一张 RGB 三通道的彩色图像作为秘密信息嵌入载体图像时的嵌入容量可达 24 bpp. 主流的基于深度学习的图像嵌入图像 (图嵌图) 隐写模型往往包含一对隐藏网络和提取网络用来嵌入和恢复秘密图像[6], 嵌入时需要将载体图像和秘密图像一起输入隐藏网络来生成含密图像, 因此在嵌入过程中载体图像和秘密图像是耦合的, 一次嵌入过程只能将秘密图像隐藏到一张载体图像中, 每生成一张新的含密图像都需要重新进行一次嵌入过程, 效率较低. 此外, 一张载体图像能隐藏的信息量也是根据训练过程中的设置决定的, 训练完成后便无法更改, 如训练时设置一张灰度图像作为秘密图像, 则测试时也只能将一张灰度图像隐藏进载体图像.

造成上述局限性的主要原因是目前主流的基于深度学习的图嵌图隐写方法的载体图像和秘密图像在生成含密图像的过程中具有耦合关系. 如图 1(a) 所示, 利用隐藏网络隐藏秘密图像时需要依赖载体图像的信息 (dependent deep hiding, DDH), 因此可以考虑从载体图像和秘密图像解耦的角度来克服这些局限. 在对抗攻击领域中, 通用对抗扰动 (universal adversarial perturbation, UAP) 的生成便满足解耦的特性[7], 即对抗扰动的生成不依赖某个具体的干净样本. 通过输入一段噪声, 生成模型可以生成一个微小的扰动添加到干净样本上, 从而生成对抗样本, 而 UAP 可以将扰动添加到多个不同的干净样本上同时生成多个对抗样本, 而无需根据干净样本重新生成对抗扰动[8]. 受到 UAP 的启发, 通用信息隐藏模型 (universal deep hiding, UDH)[9]实现了载体图像和秘密图像的解耦, 若将需要隐藏的秘密图像看作扰动, UDH 的图像隐写过程便与对抗样本的生成过程具有一定的相似性, 通过在原始图像上添加微小的扰动, 即可生成含密图像, 而这种微小的含密扰动通过 UDH 的隐藏网络生成. 如图 1(b) 所示, 将秘密图像输入 UDH 隐藏网络生成含密扰动, 这个过程无需载体图像的参与, 从而实现不依赖载体图像的嵌入方式. 通过输入一张秘密图像, 隐藏网络可以生成一个通用的含密扰动添加到不同的载体图像上, 便可同时生成多张含密图像, 实现通用图像隐写. 此外, 若想同时隐藏多张秘密图像, 可以生成多个含密扰动并添加到载体图像上.

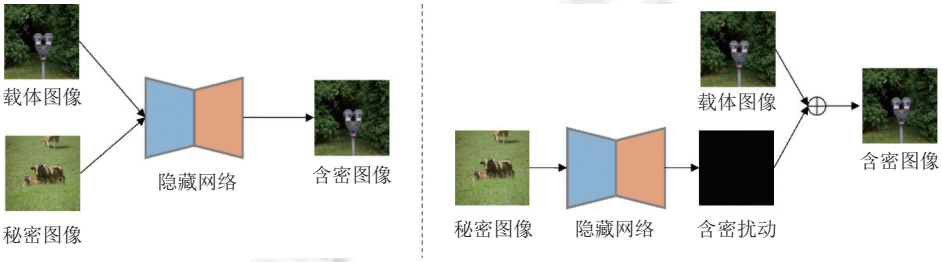


图 1 DDH 与 UDH 嵌入方式示意图

虽然 UDH 实现了不依赖载体的嵌入方式, 但并未考虑隐写的隐蔽性, 由于其较高的嵌入容量, 因此会不可避免地使隐写效果有所下降, 且基本不具备抵抗隐写分析检测的能力. 事实上, 得益于 UDH 与 UAP 在实现过程中

的相似性, 在嵌入过程中隐藏网络生成的扰动既包含了秘密信息, 也可以看作是一种对抗扰动, 因此, 可以通过对抗攻击策略促使隐藏网络学会如何将含密扰动转换成对抗扰动, 从而生成含密对抗扰动并添加到载体图像上, 使得含密图像成为可以攻击隐写分析模型的对抗样本. AdvGAN[10]是一种基于生成对抗网络 (generative adversarial network, GAN)[11]的对抗攻击方法, 其结构包含了生成器, 判别器和目标模型. 生成器用于生成对抗扰动, 判别器用于判断干净样本和对抗样本的差异, 通过对抗训练提高对抗样本的质量, 而目标模型则负责提供对抗攻击效果的反馈, 从而促使生成器生成更有效的对抗扰动. 受到 AdvGAN 的启发, 若将隐藏网络看作基于 GAN 的对抗攻击的生成器, 将隐写分析模型看作基于 GAN 的对抗攻击的目标模型 (隐写分析器用于区分载体图像和含密图像, 与判别器的作用一致, 因此可同时看作 GAN 的判别器), 利用对抗训练来促使隐藏网络学习如何使含密扰动拥有对抗攻击的效果, 便可同时达到图像隐写与对抗攻击的目的. 此外, 受到注意力机制 (attention mechanism, AM)[12]的启发, 本文在隐写模型中加入注意力模块, 用来提高本文模型的隐写效果. 注意力机制可以使用人类感知系统进行直观解释, 例如, 我们的视觉处理系统倾向于选择性地聚焦于图像的某些部分, 而以一种有助于感知的方式忽略其他不相关的信息[13]. 通过生成一个注意力概率图, 注意力模块允许隐写模型使用特定位置的秘密图像内容, 以确定需要注意的像素位置. 利用上述特性, 隐写模型可以在通道维度上将可能引起较高注意力的含密扰动进行抑制, 从而提高隐写的隐蔽性.

基于上述分析, 本文提出了一种基于注意力机制的通用图像隐写模型 (universal steganography GAN, USGAN). USGAN 包含 3 个主要部分, 分别是编码器, 解码器和目标模型. 编码器仅需秘密图像作为输入, 经过注意力模块生成注意力概率图, 促使编码器在秘密图像的不同位置根据注意力分布生成具有不同强度分布的扰动, 同时生成的含密对抗扰动既包含秘密信息又作为对抗扰动, 可直接添加到载体图像从而生成含密图像. 此外, 使用基于卷积神经网络 (convolutional neural network, CNN) 的隐写分析模型作为 USGAN 的目标模型, 用来识别载体图像和生成的含密图像的差异, 通过与编码器进行对抗训练促使编码器学习如何生成含密对抗扰动, 缓解高嵌入容量带来的隐蔽性下降的问题. 最后, 通过本文模型的解码器可以从含密图像中恢复嵌入的秘密图像. 本文的主要贡献如下.

(1) 本文提出了一种基于注意力机制的高容量通用图像隐写模型, 该模型利用注意力模块和对抗攻击策略促使编码器生成一个通用的含密对抗扰动, 可以同时对多张不同的载体图像实现嵌入, 并保证含密图像的隐蔽性.

(2) 通过实验证明本文模型在不同的隐写场景下均可实现秘密图像的嵌入和恢复, 且生成的含密图像有着较高的视觉质量, 同时在恢复秘密图像的效果和抵抗隐写分析的能力上要优于 UDH 的结果.

本文第 1 节介绍图像隐写的相关工作和研究现状. 第 2 节介绍本文提出的基于注意力机制的高容量通用图像隐写模型. 第 3 节通过实验验证了本文所提模型的有效性. 最后总结全文.

1 图像隐写相关工作

1.1 图嵌图隐写

基于深度神经网络的图嵌图隐写是近几年随着深度学习的发展而兴起的, 根据嵌入方式的不同, 图嵌图隐写可分为依赖载体的嵌入 DDH 和不依赖载体的嵌入 UDH.

DDH 嵌入方式: Baluja[14]提出了一种基于编解码器结构的卷积神经网络用于图像隐写, 可以利用编码器将一张秘密图像隐藏到相同尺寸的载体图像中, 并根据解码器恢复出秘密图像. Rehman 等人[15]也提出了一个类似的编解码器网络, 可以将灰度图像隐藏到载体图像中. Zhang 等人[16]提出了一个图像隐写模型 ISGAN, 可以将灰度秘密图像隐藏到相同大小的彩色载体图像中, 生成与载体图像在语义和颜色上都非常相似的含密图像. Zhang 等人[17]提出了一个基于 GAN 的高容量图像隐写模型, 可以将二进制秘密信息转换成类似图像的矩阵隐藏到彩色图像中, 并保证含密图像的视觉质量和隐蔽性. Yu[12]提出了一个基于 GAN 的深度信息隐藏模型, 可以将一张彩色图像隐藏到相同尺寸的彩色图像中, 并利用注意力机制提高鲁棒性, 可以很好地抵御噪声、裁剪和压缩等攻击. 竺乐庆等人[18]提出了一个基于双判别器的生成对抗网络的稳健图像隐写模型, 该模型由两个串联的生成对抗网络构成, 可将灰度图像隐藏到相同大小的彩色或灰度图像中并还原, 同时针对小幅度几何变换攻击进行了优化设计, 从

而提高模型的稳健性. Liu 等人[19]设计了一种联合压缩自编码器 (joint compressive autoencoder, J-CAE) 框架的图像隐藏算法, 以及一种用来扩大隐藏容量、实现多图像隐藏的新策略, 从而实现了极高的图像隐藏容量和较小的秘密图像重建损失. 段新涛等人[20]提出的隐藏方法中包含一个隐藏网络和两个结构相同的提取网络, 并且在隐藏网络和提取网络中加入了改进的金字塔池化模块和预处理模块, 实现了在一幅载体图像上同时对两幅全尺寸秘密图像进行有效的隐藏和提取, 所提方法较现有的图像信息隐藏方法在视觉质量上有显著提升. 虽然 DDH 能够实现较高的嵌入容量和较好的隐写视觉质量, 但嵌入时需要依赖载体图像信息, 因此嵌入效率较低.

UDH 嵌入方式: Zhang 等人[9]提出了一个通用信息隐藏模型 UDH, 可应用于图像隐写和数字水印等领域, 该方法通过将秘密信息输入隐藏网络生成一个含密扰动, 然后将扰动添加到载体图像上, 从而实现秘密信息的嵌入, 在嵌入效率上有所提高, 但该方法没有考虑隐写的隐蔽性. 本文在 UDH 的基础上进行改进, 通过注意力机制和对抗攻击策略进一步提高了隐写的隐蔽性.

1.2 对抗攻击与图像隐写

基于对抗样本生成与图像隐写在实现上的相似性, 近几年, 将对抗攻击应用于图像隐写的研究也逐渐兴起. Zhang 等人[21]提出利用快速梯度下降法 (fast gradient sign method, FGSM)[22]生成对抗样本作为载体图像, 使得隐写分析无法区分载体和含密图像. Zhou 等人[23]提出了一种快速生成大量对抗载体图像的方法, 通过训练一个生成器, 可以在简单的前向传播中将载体图像转换成对抗样本, 然后用于隐写. 这些方法初步将对抗攻击方法引入图像隐写领域, 但仅是将载体图像转换为对抗样本, 未实现端到端的图像隐写. AdvSGAN[24] 借鉴了基于深度神经网络的对抗扰动生成模型 AdvGAN 的结构, 通过在受限神经编码器和两个对抗模型 (即与编码器一起训练的判别器和预训练的目标模型) 之间进行对抗, 从头学习由受限神经编码器表示的图像隐写方案.

2 USGAN 隐写模型

2.1 模型结构

本文提出的 USGAN 隐写模型的总体结构如图 2 所示, 秘密图像可以通过 USGAN 隐藏到载体图像中并从含密图像中恢复. USGAN 由编码器, 目标模型和解码器 3 部分组成, 编码器以秘密图像 M 为输入生成含密对抗扰动 M_e, 然后与载体图像 C_i ($i=1,2,\ldots,n$) 相加生成含密图像 S_i, 含密图像同时也是攻击隐写分析模型的对抗样本. 不同于单纯的对抗样本只是为了攻击某个目标模型, 作为对抗样本的含密图像同时还要求能够从中恢复隐藏的秘密图像, 而解码器以含密图像 S_i 为输入并恢复出秘密图像 M_i'. 目标模型以载体图像 C_i 与对应的含密图像 S_i 为输入并输出分类概率, 用来判断输入的图像是载体还是含密图像, 并与编码器进行对抗训练, 提高 USGAN 隐藏秘密图像的隐蔽性.

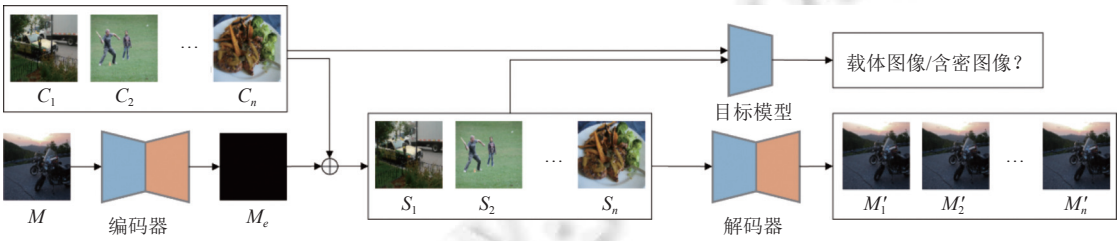


图 2 USGAN 模型结构图

USGAN 的编码器采用的是改进后的 UDH 的隐藏网络, 解码器采用的是 UDH 的恢复网络, 目标模型则是被攻击的目标隐写分析模型. 原始的 UDH 隐藏网络采用的是 CycleGAN[25]中的一个简化的 U-Net 网络, 而恢复网络则是一系列卷积层的堆叠. 本文在原始的 UDH 隐藏网络的基础上加入注意力模块, 并将改进后的隐藏网络作为本文模型的编码器, 从而提高本文模型的隐写效果. Simonyan 等人[26]和 Zhou 等人[27]的工作发现深度神经网络的激活特征图可以构建一个一般性的局部特征表达, 用以揭示深度神经网络在图像上的潜在注意力. 由于图像隐写

试图混淆载体图像和含密图像之间的视觉效果差异, 所以本文引入注意力机制可以帮助模型更明确地完成这一目标. 注意力模块会学习秘密图像中每个像素在通道维度上的信息的概率分布 (注意力分布), 并使用该分布来选择在嵌入过程中要注意的位置, 这使模型倾向于学习在不同对象和纹理中生成不同强度分布的扰动.

编码器中的注意力模块采用文献 [28] 中的网络结构, 该模块使用两个卷积块创建一个注意力概率图, 并鼓励编码器根据图像的内容关注像素的不同通道维度. 具体来说, 注意力模块以秘密图像作为输入, 输出注意力概率图 P_m 并与秘密图像相乘, 得到注意力特征图 M_a, 然后送进编码器剩下的网络结构中生成含密扰动 M_e, 过程如图 3 所示.

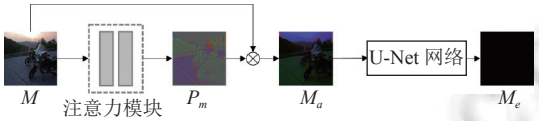


图 3 USGAN 编码器中的注意力机制示意图

如图 3 所示, 注意力模块的输出是一个注意力概率图, 其中每个像素上的概率向量可以解释为最终生成的含密对抗扰动在通道维度上的强度分布, 用来控制秘密图像中不同注意力的像素位置生成的扰动变化. 注意力机制可以被表示为:

$$\begin{cases} a = Conv_{d\to 32}(M) \\ b = Conv_{32\to d}(a) \\ P_m(b^{(i)}) = \dfrac{\mathrm{e}^{b_i^{(i)}}}{\sum_{k=1}^{d}\mathrm{e}^{b_k^{(i)}}} \\ M_a = M \times P_m \end{cases} \tag{1}$$

其中, $Conv$ 表示卷积层, a 和 b 表示卷积层输出的特征图, d 是秘密图像 M 的通道数, i 代表 b 中元素位置, j 和 k 表示 b 的通道维度.

2.2 含密对抗扰动

能否抵抗隐写分析的检测是图像隐写隐蔽性的一个重要的评价标准, 而 UDH 仅实现了秘密图像的嵌入和恢复, 没有考虑如何抵抗隐写分析的检测. 本文通过对抗攻击策略, 使得 USGAN 的编码器学会根据秘密图像生成含密对抗扰动, 并添加到载体图像上, 使生成的含密图像同时成为攻击隐写分析模型的对抗样本, 从而提高本文模型抵抗隐写分析检测的能力. 对抗攻击的目标模型是隐写分析模型, 因此可以通过和隐写分析模型进行对抗训练的方式使得编码器学习如何将含密扰动转换成对抗扰动, 同时又不破坏扰动中隐藏的秘密信息. 具体来说, 本文将目标隐写分析模型设置为目标模型, 编码器努力生成能够欺骗目标模型的含密对抗扰动, 而目标模型则努力去识别载体图像和含密图像的差异, 通过对抗训练使本文模型生成的含密图像具有更强的抵抗隐写分析的能力, 从而使隐写分析模型的识别准确率接近 0.5, 即相当于随机猜测. 经过迭代训练, 并通过目标模型的反馈更新编码器的参数, 使编码器最终可以学会生成含密对抗扰动. 本文中基于 GAN 的对抗训练可表示为:

$$\min_G \max_D V(D,G) = E_{x\sim p_{\text{data}}(x)}[\log D(x)] + E_{z\sim p_z(z)}[\log(1-D(x+G(z)))] \tag{2}$$

其中, D 表示目标模型, 目标是区分对抗样本 $x+G(z)$ 和原始样本 x. 在本文中, 原始样本 x 代表载体图像, z 代表秘密图像, 对抗样本 $x+G(z)$ 代表含密图像, x 是从载体图像的类别中进行采样的, 因此目标模型在对抗训练中会促使生成的含密图像更接近载体图像的类别.

对抗训练的目的是使编码器生成的含密扰动具有对抗扰动的特性, 因此需要给编码器设置一个目标, 使编码器向着目标进行参数的更新. 具体来说, 就是要编码器生成含密对抗扰动并添加到载体图像上得到含密图像, 使得目标模型将含密图像识别为载体图像. 整个目标的描述可表示为:

$$\arg\min_G E_{x\sim p_{\text{data}}(x), z\sim p_z(z)}\ell(D(x+G(z)), t) \tag{3}$$

其中, ℓ 表示用于训练原始目标模型的损失函数, t 表示目标类别. 若要欺骗目标模型, 则目标模型需要将含密图像的分类结果判定为错误的类别 (即无目标攻击), 或者判定为非原始类别的目标类别 (目标攻击), 从而误导目标模型的分类结果. 在本文中, 因为只有载体图像和含密图像两个类别, 所以无目标攻击和目标攻击的含义是等价的. 本文采用目标攻击的方式, 将载体图像类别标签设为 0, 含密图像的类别标签设为 1, 因此目标类别标签 $t=0$.

2.3 损失函数

训练 USGAN 的损失函数包括 3 部分, 分别是编码器的损失 ℓ_e, 目标模型的损失 ℓ_d, 以及解码器的损失 ℓ_r. ℓ_e 表示载体图像和含密图像之间的均方误差损失, 用来衡量含密图像的失真程度. ℓ_d 表示针对目标模型的目标攻击损失, 用来促使编码器学习生成含密对抗扰动. ℓ_r 表示恢复的秘密图像与原始秘密图像之间的信息损失. 三者的定义如下:

$$\begin{cases} \ell_e(c,s)=\dfrac{1}{n}\displaystyle\sum_{i=1}^{n}(s_i-c_i)^2 \\ \ell_d(y,\hat{y})=-\dfrac{1}{n}\displaystyle\sum_{i=1}^{n}[y\log\hat{y}+(1-y)\log(1-\hat{y})] \\ \ell_r(m,m')=\dfrac{1}{n}\displaystyle\sum_{i=1}^{n}(m'_i-m_i)^2 \end{cases} \tag{4}$$

其中, n 代表样本数量, c 和 s 分别代表载体图像和含密图像, y 和 $\hat{y}$ 分别代表目标模型的目标标签和预测标签, m 和 m' 分别代表原始秘密图像和恢复的秘密图像. 总的损失函数定义如下:

$$\ell_{\text{total}}=\ell_e+\ell_d+\beta\ell_r \tag{5}$$

其中, β 用于控制编码器损失和解码器损失的相对比例. 容易想到, 编码器和解码器之间存在博弈关系, 编码器希望生成的含密扰动强度尽可能小, 含密图像与载体图像尽可能接近, 而太小的含密扰动会影响秘密图像的恢复准确率. 与之相反, 解码器希望编码器生成的含密扰动尽可能大, 这样解码器才更容易从含密图像中恢复出秘密图像, 因此, 需要设置合理的 β 值来平衡二者的关系. 参考 UDH 的设置, 本文令 $\beta=0.75$, 可以让模型取得训练上的平衡.

3 实验分析

3.1 实验设置

本文使用 BOSSBASE[29]和 MSCOCO[30]作为实验的数据集. BOSSBASE 包含 10 000 张单通道的灰度图像, 将其按 8:2 划分为训练集和测试集, 训练集和测试集中又分别按 1:1 划分为载体图像和秘密图像, 为了提高训练效率, 所有的图像均归一化到 128×128 的尺寸. 从 MSCOCO 中取出 10 000 张三通道的 RGB 彩色图像, 并进行与 BOSSBASE 相同的数据集划分和图像预处理. 为了验证本文模型在不同隐写场景下的通用性, 根据载体和秘密图像来源的不同, 本文在 4 种嵌入模式下进行了实验, 具体设置如表 1 所示, 例如 SetA 表示在训练和测试时载体图像为 MSCOCO 中的三通道彩色图像, 秘密图像为 BOSSBASE 中的单通道灰度图像.

表 1 不同嵌入模式下载体图像和秘密图像的数据集

嵌入模式	载体图像	秘密图像
SetA	MSCOCO	BOSSBASE
SetB	BOSSBASE	BOSSBASE
SetC	BOSSBASE	MSCOCO
SetD	MSCOCO	MSCOCO

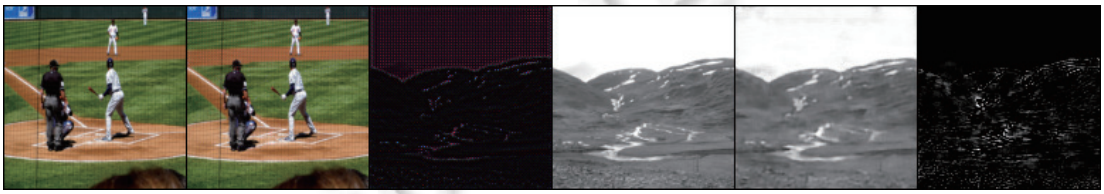
本文使用 ADAM (adaptive moment estimation) 优化器[31]训练 USGAN 隐写模型, 初始学习率为 0.001, 并随着训练轮次的增加逐渐衰减. UDH 的训练参照原论文中的设置. 本文使用 bpp 表示嵌入容量, 使用结构相似度 (structural similarity, SSIM) 和峰值信噪比 (peak signal-to-noise ratio, PSNR) 来评价含密图像质量, 使用 P_e 评价含密图像抵抗隐写分析的性能. P_e 表示隐写分析模型的检测错误率[24], 越高表示隐写模型的隐蔽性越好, 其定义如下:

$$P_e = \frac{1}{2}(P_{fa} + P_{md}) \tag{6}$$

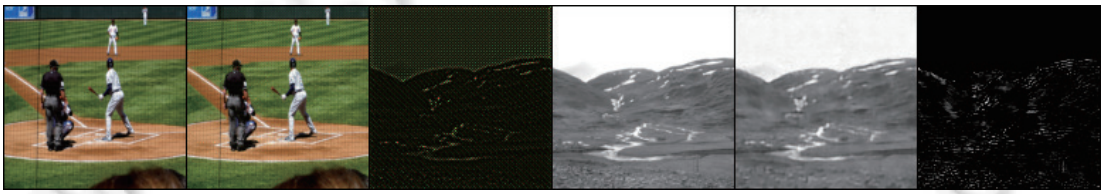
其中, P_{fa} 表示假正率, P_{md} 表示假负率.

3.2 不同嵌入模式下的隐写效果

本文在 4 种嵌入模式下进行了 UDH 和 USGAN 隐写效果的比较, 结果如图 4–图 7 以及表 2 所示. 图 4–图 7 给出了不同嵌入模式下 UDH 和 USGAN 在隐写前后载体和含密图像的示意图和残差图, 以及秘密图像在嵌入前和恢复后的示意图和残差图. 表 2 则给出了 UDH 和 USGAN 在不同嵌入模式下隐写性能指标的对比结果, 其中, s 代表含密图像, m' 代表恢复的秘密图像. 如图 4 所示, 在 SetA 嵌入模式下利用 UDH 和 USGAN 进行隐写得到的含密图像图 4(a2) 和图 4(b2) 的视觉效果没有明显差异, 从残差图像图 4(a3) 和图 4(b3) 也可以看出, 即使两个模型的载体和含密图像残差被放大了 5 倍, 其可视化结果依旧不够明显. 其他几种嵌入模式下也取得了类似的结果. 但从恢复的秘密图像的视觉质量来看, 本文的 USGAN 模型取得了比 UDH 更好的结果. 从图 4–图 7 这 4 种嵌入模式下的 UDH 和 USGAN 恢复的秘密图像的示意图 (a5) 和 (b5), 以及秘密图像残差图 (a6) 和 (b6) 可以看出, USGAN 恢复的秘密图像的细节更好, 和原始秘密图像的残差更小. 这表明 USGAN 在恢复秘密图像时取得了更好的效果.



(a1) 载体图 c (a2) UDH 含密图 s (a3) $5|s-c|$ (a4) 秘密图 m (a5) UDH 恢复秘密图 m' (a6) $5|m'-m|$



(b1) 载体图 c (b2) USGAN 含密图 s (b3) $5|s-c|$ (b4) 秘密图 m (b5) USGAN 恢复秘密图 m' (b6) $5|m'-m|$

图 4 SetA 嵌入模式下 UDH 和 USGAN 的隐写效果对比

(a1) 载体图 c (a2) UDH 含密图 s (a3) $5|s-c|$ (a4) 秘密图 m (a5) UDH 恢复秘密图 m' (a6) $5|m'-m|$

(b1) 载体图 c (b2) USGAN 含密图 s (b3) $5|s-c|$ (b4) 秘密图 m (b5) USGAN 恢复秘密图 m' (b6) $5|m'-m|$

图 5 SetB 嵌入模式下 UDH 和 USGAN 的隐写效果对比

(a1) 载体图 c (a2) UDH 含密图 s (a3) 5|$s-c$| (a4) 秘密图 m (a5) UDH 恢复秘密图 m' (a6) 5|$m'-m$|

(b1) 载体图 c (b2) USGAN 含密图 s (b3) 5|$s-c$| (b4) 秘密图 m (b5) USGAN 恢复秘密图 m' (b6) 5|$m'-m$|

图 6 SetC 嵌入模式下 UDH 和 USGAN 的隐写效果对比

(a1) 载体图 c (a2) UDH 含密图 s (a3) 5|$s-c$| (a4) 秘密图 m (a5) UDH 恢复秘密图 m' (a6) 5|$m'-m$|

(b1) 载体图 c (b2) USGAN 含密图 s (b3) 5|$s-c$| (b4) 秘密图 m (b5) USGAN 恢复秘密图 m' (b6) 5|$m'-m$|

图 7 SetD 嵌入模式下 UDH 和 USGAN 的隐写效果对比

表 2 UDH 和本文模型 USGAN 在 4 种嵌入模式下的性能指标

嵌入模式	隐写模型	嵌入容量 (bpp)	s		m'	
			SSIM	PSNR (dB)	SSIM	PSNR (dB)
SetA	UDH	8	0.9638	34.4921	0.9214	31.6334
	USGAN	8	0.9576	36.4534	0.9564	34.1907
SetB	UDH	8	0.8918	32.2914	0.9151	31.8812
	USGAN	8	0.8997	32.9694	0.9338	35.5412
SetC	UDH	24	0.8887	32.1209	0.8861	22.8740
	USGAN	24	0.8700	31.7936	0.8956	25.0946
SetD	UDH	24	0.9885	41.0615	0.8813	31.9077
	USGAN	24	0.9608	36.5715	0.9333	31.2558

此外, 从表 2 的结果可以看到, UDH 和 USGAN 的含密图像的 SSIM 和 PSNR 指标十分接近, 这也证明了两个模型具有相似的隐写性能. 而表 2 中 USGAN 恢复的秘密图像的 SSIM 和 PSNR 指标比 UDH 平均高出 5%, 这

证明了 USGAN 在恢复秘密图像的质量上性能更好.

值得一提的是, 实验结果表明 USGAN 在恢复秘密图像时取得了比 UDH 更好的结果. 我们推测产生这种现象的原因是 UDH 生成的含密图像上的秘密信息的分布很可能是均匀的或者是无规律的, 这限制了解码器学习一个固定的提取模式. 但基于注意力的 USGAN 生成的含密图像上的秘密信息分布可能具有一定的规律性, 即秘密图像中可能引起较高注意力的区域中的扰动强度受到一定的抑制, 导致大部分扰动都分布在秘密图像上难以被感知的位置, 从而使得生成的含密图像的秘密信息分布具有一定的统计规律, 而这有利于解码器学习一个固定的解码模式, 在含密图像上最有可能嵌入秘密信息的位置进行提取, 从而提高恢复秘密图像的准确率.

3.3 将多张秘密图像嵌入单张载体图像

UDH 和 USGAN 具有不依赖载体图像的嵌入方式, 这种灵活的嵌入方式极大扩展了图像隐写的应用场景, 例如可以通过编码器生成多个含密扰动并添加到载体图像上, 即可同时将多张秘密图像嵌入到载体图像中. 本文测试了 UDH 和 USGAN 将多张秘密图像嵌入到单张载体图像中时的隐写效果, 结果如图 8 和表 3 所示. 图 8 给出了 UDH 和 USGAN 在 3 张灰度图嵌入到单张彩色图时的载体和含密图像示意图, c 代表载体图像, s 代表含密图像, $m_i\ (i=1,2,3)$ 代表第 i 张秘密图像, $m_i'\ (i=1,2,3)$ 代表第 i 张恢复的秘密图像. 表 3 给出了对应的含密图像和恢复的秘密图像的 SSIM 和 PSNR 性能指标结果, s 代表含密图像, $m_i'\ (i=1,2,3)$ 代表第 i 张恢复的秘密图像. 如图 8 所示, 在嵌入多张秘密图像时, USGAN 生成的含密图像图 8(b2) 具有很好的视觉质量, 表 3 中对应的 SSIM 和 PSNR 值与 UDH 也非常接近. 而 USGAN 恢复出的 3 张秘密图像 (图 8(b4), 图 8(b6), 图 8(b8)) 在整体视觉质量和细节还原上要比 UDH 更好, 表 3 中对应的指标也更高. 这证明了 USGAN 在将多张秘密图像嵌入单张载体图像时的隐写性能要优于 UDH.

(a1) c (a2) s (UDH) (a3) m_1 (a4) m_1' (UDH) (a5) m_2 (a6) m_2' (UDH) (a7) m_3 (a8) m_3' (UDH)

(b1) c (b2) s (USGAN) (b3) m_1 (b4) m_1' (USGAN) (b5) m_2 (b6) m_2' (USGAN) (b7) m_3 (b8) m_3' (USGAN)

图 8 UDH 和 USGAN 在 3 张灰度图像嵌入单张彩色图像场景下的示意图

表 3 UDH 和本文模型 USGAN 在 3 张灰度图像嵌入单张彩色图像时的 SSIM 和 PSNR

隐写模型	s		m_1'		m_2'		m_3'	
	SSIM	PSNR (dB)	SSIM	PSNR (dB)	SSIM	PSNR (dB)	SSIM	PSNR (dB)
UDH	0.9626	34.4170	0.9156	31.1067	0.8724	24.7937	0.8617	27.6978
USGAN	0.9674	33.8730	0.9184	30.0377	0.8811	31.1767	0.9125	31.2368

3.4 抵抗隐写分析性能

UDH 虽然具有较高的嵌入容量和较好的隐写效果, 但并未提高隐写的隐蔽性, 而这是隐写性能最重要的评价标准之一. 本文利用对抗学习使得 USGAN 生成的含密图像同时作为抵抗隐写分析模型的对抗样本, 从而提高隐写的隐蔽性. 为了评价本文方法在提高隐蔽性上的效果, 本文采用基于 CNN 的隐写分析模型 XuNet[32]和 SRNet[33] 进行隐写模型隐蔽性的评估, 并分别与依赖载体和不依赖载体两种嵌入方式的图像隐写模型进行对比. ISGAN 和

SteganoGAN 是主流的采用依赖载体图像嵌入方式的图像隐写模型, ISGAN 可以将一张灰度图像嵌入一张灰度或彩色图像, 因此只能实现本文设置的 SetA 和 SetB 嵌入模式, 而 SteganoGAN 可以将灰度或彩色秘密图像嵌入灰度或彩色载体图像中, 因此可以实现 SetA–SetD 这 4 种嵌入模式. AdvSGAN 通过对秘密信息通道进行扩展, 也可实现将灰度图像嵌入载体图像的效果, 但隐写的隐蔽性会有所下降. 4 种嵌入模式下隐写分析的对比实验结果如表 4 所示.

表 4　4 种嵌入模式下不同图像隐写模型利用 XuNet 和 SRNet 进行隐写分析的检测错误率 P_e

隐写模型	SetA (8 bpp)		SetB (8 bpp)		SetC (24 bpp)		SetD (24 bpp)	
	XuNet	SRNet	XuNet	SRNet	XuNet	SRNet	XuNet	SRNet
ISGAN	0.0506	0.0558	0.0498	0.0692	—	—	—	—
SteganoGAN	0.0674	0.0512	0.0417	0.0631	0.0376	0.0203	0.0550	0.0204
AdvSGAN	0.1668	0.1475	0.1564	0.1426	—	—	—	—
UDH	0.0484	0.0348	0.0496	0.0587	0.0419	0.0363	0.0503	0.0441
USGAN	0.4995	0.4980	0.4787	0.4340	0.4722	0.4020	0.4646	0.4630

从表 4 可以看出, 不同嵌入模式下 ISGAN, SteganoGAN 和 UDH 在面对两种隐写分析模型 XuNet 和 SRNet 的检测时 P_e 均在 0.1 以下, 这说明这些图像隐写模型生成的含密图像几乎无法逃避隐写分析模型的检测. 由于 AdvSGAN 采用了两个对抗模型进行训练, 因此在面对隐写分析模型检测时具有一定的隐蔽性, 但因为较高的嵌入容量其 P_e 降低到了 0.15 左右. 而隐写分析模型在检测 USGAN 生成的含密图像时其 P_e 却已经接近 0.5, 这已经近似于随机猜测, 即隐写分析模型已经无法对载体图像和 USGAN 生成的含密图像进行准确的区分. 为了进一步评估隐写分析模型在检测 USGAN 生成的隐写图像上的性能, 本文利用隐写分析的实验结果绘出 4 种嵌入模式下 XuNet 和 SRNet 检测 USGAN 生成的隐写图像时的受试者工作特征 (receiver operating characteristic, ROC) 曲线并计算 ROC 曲线下的面积 (area under curve, AUC), 结果如图 9 所示.

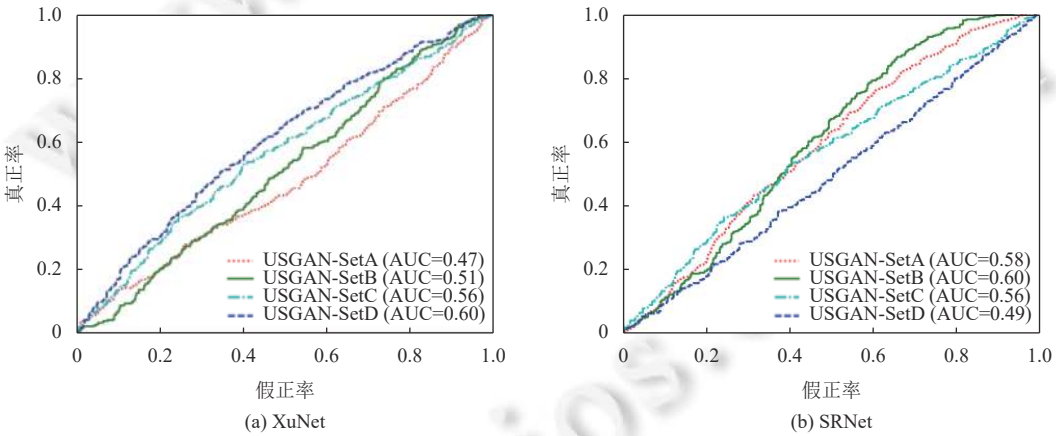


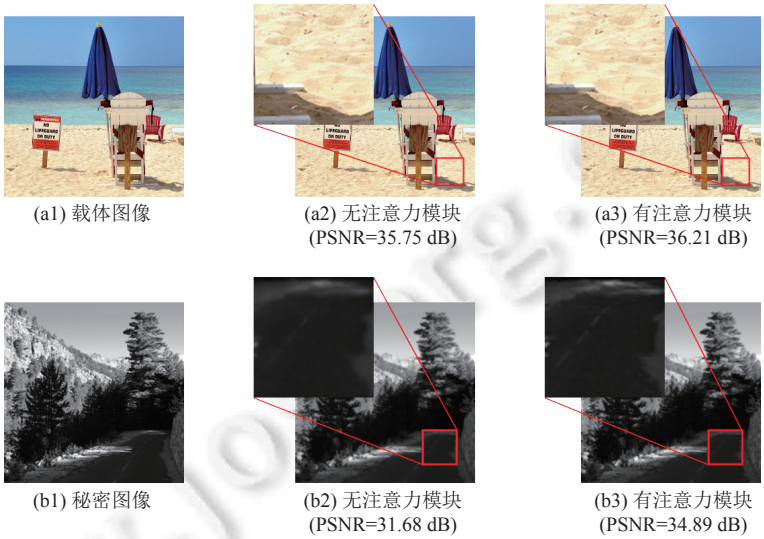
图 9　利用 XuNet 和 SRNet 检测 USGAN 生成的含密图像的 ROC 曲线

从图 9 可以看出, 4 种嵌入模式下 XuNet 和 SRNet 的 ROC 曲线都接近代表随机猜测的对角线, 且 AUC 都接近 0.5, 这证明这两种隐写分析模型在检测 USGAN 生成的含密图像时已经近似无效. 造成这种现象的原因是 USGAN 针对这两种隐写分析模型进行了对抗训练, 即利用 XuNet 和 SRNet 的检测结果优化编码器的训练, 从而使编码器生成可以欺骗隐写分析模型的含密对抗扰动.

3.5　消融实验

为了验证注意力模块在 USGAN 中的贡献, 本文设计了对应的消融实验. 具体来说, 评估在没有注意力模块的情况下 USGAN 的性能, 并将测试结果与含有注意力模块的完整模型进行比较. 图 10 给出了在没有注意力模块和

含有注意力模块的情况下, 编码器在 SetA 嵌入模式下生成的含密图像和解码器恢复的秘密图像的视觉质量及局部放大后的细节. 从图 10(a2) 和图 10(a3) 中可以看出, 含有注意力模块的编码器生成的含密图像局部的纹理细节与无注意力模块下没有明显差异, 视觉质量的评价指标 PSNR 也提升不大, 说明注意力模块对提高含密图像的生成质量效果不显著, 仅取得了少许提升. 但从图 10(b2) 和图 10(b3) 中可以看出, 解码器在恢复秘密图像时, 从具有注意力模块的编码器生成的含密图像中恢复的秘密图像的局部细节更加清晰丰富, PSNR 也更高, 证明注意力模块对提高恢复的秘密图像的视觉质量具有一定效果.



(a1) 载体图像　(a2) 无注意力模块 (PSNR=35.75 dB)　(a3) 有注意力模块 (PSNR=36.21 dB)

(b1) 秘密图像　(b2) 无注意力模块 (PSNR=31.68 dB)　(b3) 有注意力模块 (PSNR=34.89 dB)

图 10　USGAN 在有或无注意力模块下生成的含密图像和恢复的秘密图像及其局部放大图

此外, 表 5 给出了注意力模块在 4 种嵌入模式下抵抗隐写分析器 XuNet 的效果, 从检测错误率可以看出, 含有注意力模块时的检测错误率 P_e 相比无注意力模块时可提高 2.52%–3.31%, 说明注意力模块对抵抗隐写分析具有一定效果. 由于 USGAN 中的注意力模块略微提高了含密图像的视觉质量, 使得含密图像与载体图像之间的视觉差异更小, 因此具有一定的欺骗隐写分析器的作用, 但 USGAN 主要还是依赖对抗攻击策略来提高抵抗隐写分析检测的性能.

表 5　USGAN 在有或无注意力模块下利用 XuNet 进行隐写分析的检测错误率 P_e

模块	SetA	SetB	SetC	SetD
无注意力模块	0.4664	0.4535	0.4428	0.4389
有注意力模块	0.4995	0.4787	0.4722	0.4646

4 总　结

本文提出了一种基于注意力机制的高容量通用图像隐写模型 USGAN, 可以实现灰度图像嵌入彩色图像, 彩色图像嵌入灰度图像等多种嵌入模式, 并在多张秘密图像嵌入到单张载体图像的场景下进行了扩展, 提高了隐写模型的实用性. USGAN 在 UDH 的基础上进行改进, 通过增加注意力模块促使编码器学习如何在通道维度上根据秘密图像上具有不同注意力分布的像素获得生成扰动的强度分布, 对可能引起较高注意力的像素位置进行扰动的抑制, 提高隐写的隐蔽性. USGAN 倾向于在秘密图像上不易被感知的位置保留更多的信息, 这些位置的扰动对载体图像的影响更小, 从而使含密图像获得更好的隐蔽性, 并在恢复秘密图像的效果上具有比 UDH 更好的性能. 此外, 通过添加基于 CNN 的隐写分析模型作为目标模型进行对抗训练, USGAN 可以学会将含密扰动转换成对抗扰动, 从而使含密图像成为攻击隐写分析模型的对抗样本, 在抵抗隐写分析的性能上取得比 UDH 更好的结果. 实验

在 MSCOCO 和 BOSSBASE 两个数据集上进行了测试, 证明了本文模型在隐写性能上的优势.

本文通过实验给出了注意力机制作用于图像隐写模型的效果, 并对注意力机制产生效果的原因进行了分析. 下一步工作我们会着重探索注意力机制在图像隐写上的理论依据和作用机制.

References:

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

[2] Pevný T, Filler T, Bas P. Using high-dimensional image models to perform highly undetectable steganography. In: Proc. of the 12th Int'l Workshop on Information Hiding. Calgary: Springer, 2010. 161–177. [doi: 10.1007/978-3-642-16435-4_13]

[3] Holub V, Fridrich J. Designing steganographic distortion using directional filters. In: Proc. of the 2012 IEEE Int'l Workshop on Information Forensics and Security. Costa Adeje: IEEE, 2012. 234–239. [doi: 10.1109/WIFS.2012.6412655]

[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. [doi: 10.1186/1687-417X-2014-1]

[5] Li B, Wang M, Huang JW, Li XL. A new cost function for spatial image steganography. In: Proc. of the 2014 IEEE Int'l Conf. on Image Processing. Paris: IEEE, 2014. 4206–4210. [doi: 10.1109/ICIP.2014.7025854]

[6] Hayes J, Danezis G. Generating steganographic images via adversarial training. In: Proc. of the 31st Int'l Conf. on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 1951–1960. [doi: 10.5555/3294771.3294957]

[7] Moosavi-Dezfooli SM, Fawzi A, Fawzi O, Frossard P. Universal adversarial perturbations. In: Proc. of the 2017 IEEE Conf. on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 86–94. [doi: 10.1109/CVPR.2017.17]

[8] Hayes J, Danezis G. Learning universal adversarial perturbations with generative models. In: Proc. of the 2018 IEEE Security and Privacy Workshops. San Francisco: IEEE, 2018. 43–49. [doi: 10.1109/SPW.2018.00015]

[9] Zhang CN, Benz P, Karjauv A, Sun G, Kweon IS. UDH: Universal deep hiding for steganography, watermarking, and light field messaging. In: Proc. of the 34th Int'l Conf. on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2020. 857. [doi: 10.5555/3495724.3496581]

[10] Xiao CW, Li B, Zhu JY, He W, Liu MY, Song D. Generating adversarial examples with adversarial networks. In: Proc. of the 27th Int'l Joint Conf. on Artificial Intelligence. 2018. 3905–3911. [doi: 10.24963/ijcai.2018/543]

[11] Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y. Generative adversarial nets. In: Proc. of the 27th Int'l Conf. on Neural Information Processing Systems. Montreal: MIT Press, 2014. 2672–2680. [doi: 10.5555/2969033.2969125]

[12] Yu C. Attention based data hiding with generative adversarial networks. In: Proc. of the 34th AAAI Conf. on Artificial Intelligence. Palo Alto: AAAI, 2020. 1120–1128. [DOI: 10.1609/aaai.v34i01.5463] [doi: 10.1609/aaai.v34i01.5463]

[13] Xu K, Ba J, Kiros R, Cho K, Courville A, Salakhudinov R, Zemel R, Bengio Y. Show, attend and tell: Neural image caption generation with visual attention. In: Proc. of the 32nd Int'l Conf. on Machine Learning. 2015. 2048–2057.

[14] Baluja S. Hiding images in plain sight: Deep steganography. In: Proc. of the 31st Int'l Conf. on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 2066–2076. [doi: 10.5555/3294771.3294968]

[15] ur Rehman A, Rahim R, Nadeem S, ul Hussain S. End-to-end trained CNN encoder-decoder networks for image steganography. In: Proc. of the 2018 European Conf. on Computer Vision. Munich: Springer, 2018. 723–729. [doi: 10.1007/978-3-030-11018-5_64]

[16] Zhang R, Dong SQ, Liu JY. Invisible steganography via generative adversarial networks. Multimedia Tools and Applications, 2019, 78(7): 8559–8575. [doi: 10.1007/s11042-018-6951-z]

[17] Zhang KA, Cuesta-Infante A, Xu L, Veeramachaneni K. SteganoGAN: High capacity image steganography with GANs. arXiv: 1901.03892, 2019.

[18] Zhu LQ, Guo Y, Mo LQ, Zhang DX. DGANS: Robustness image steganography model based on double GAN. Journal on Communications, 2020, 41(1): 125–133 (in Chinese with English abstract). [doi: 10.11959/j.issn.1000-436x.2020019]

[19] Liu XY, Ma ZP, Chen ZH, Li FF, Jiang M, Schaefer C, Fang H. Hiding multiple images into a single image via joint compressive autoencoders. Pattern Recognition, 2022, 131: 108842. [doi: 10.1016/j.patcog.2022.108842]

[20] Duan XT, Wang WX, Li L, Shao ZQ, Wang XF, Qin C. Image hiding method based on two-channel deep convolutional neural network. Journal of Electronics & Information Technology, 2022, 44(5): 1782–1791 (in Chinese with English abstract). [doi: 10.11999/JEIT210280]

[21] Zhang YW, Zhang WM, Chen KJ, Liu JY, Liu YJ, Yu NH. Adversarial examples against deep neural network based steganalysis. In:

Proc. of the 6th ACM Workshop on Information Hiding and Multimedia Security. Innsbruck: ACM, 2018. 67–72. [doi: 10.1145/3206004.3206012]

[22] Goodfellow IJ, Shlens J, Szegedy C. Explaining and harnessing adversarial examples. In: Proc. of the 3rd Int'l Conf. on Learning Representations. 2015. 1–11.

[23] Zhou LC, Feng GR, Shen LQ, Zhang XP. On security enhancement of steganography via generative adversarial image. IEEE Signal Processing Letters, 2019, 27: 166–170. [doi: 10.1109/LSP.2019.2963180]

[24] Li L, Fan MY, Liu DF. AdvSGAN: Adversarial image steganography with adversarial networks. Multimedia Tools and Applications, 2021, 80(17): 25539–25555. [doi: 10.1007/s11042-021-10904-1]

[25] Zhu JY, Park T, Isola P, Efros AA. Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proc. of the 2017 IEEE Int'l Conf. on Computer Vision. Venice: IEEE, 2017. 2242–2251. [doi: 10.1109/ICCV.2017.244]

[26] Simonyan K, Vedaldi A, Zisserman A. Deep inside convolutional networks: Visualising image classification models and saliency maps. In: Proc. of the 2nd Int'l Conf. on Learning Representations. 2013.

[27] Zhou BL, Khosla A, Lapedriza A, Oliva A, Torralba A. Learning deep features for discriminative localization. In: Proc. of the 2016 IEEE Conf. on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 2921–2929. [doi: 10.1109/CVPR.2016.319]

[28] Zhang KA, Xu L, Cuesta-Infante A, Veeramachaneni K. Robust invisible video watermarking with attention. arXiv:1909.01285, 2019.

[29] Bas P, Filler T, Pevný T. "Break our steganographic system": The ins and outs of organizing BOSS. In: Proc. of the 13th Int'l Workshop on Information Hiding. Prague: Springer, 2011. 59–70. [doi: 10.1007/978-3-642-24178-9_5]

[30] Lin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P, Zitnick CL. Microsoft COCO: Common objects in context. In: Proc. of the 13th European Conf. on Computer Vision. Zurich: Springer, 2014. 740–755. [doi: 10.1007/978-3-319-10602-1_48]

[31] Kingma DP, Ba J. Adam: A method for stochastic optimization. In: Proc. of the 3rd Int'l Conf. on Learning Representations. San Diego: ICLR, 2015.

[32] Xu GS, Wu HZ, Shi YQ. Structural design of convolutional neural networks for steganalysis. IEEE Signal Processing Letters, 2016, 23(5): 708–712. [doi: 10.1109/LSP.2016.2548421]

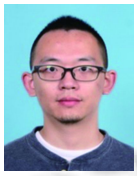
[33] Boroumand M, Chen M, Fridrich J. Deep residual network for steganalysis of digital images. IEEE Trans. on Information Forensics and Security, 2019, 14(5): 1181–1193. [doi: 10.1109/TIFS.2018.2871749]

附中文参考文献:

[18] 竺乐庆, 郭钰, 莫凌强, 张大兴. DGANS: 基于双重生成式对抗网络的稳健图像隐写模型. 通信学报, 2020, 41(1): 125–133. [doi: 10.11959/j.issn.1000-436x.2020019]

[20] 段新涛, 王文鑫, 李磊, 邵志强, 王鲜芳, 秦川. 基于两通道深度卷积神经网络的图像隐藏方法. 电子与信息学报, 2022, 44(5): 1782–1791. [doi: 10.11999/JEIT210280]

袁超(1996－), 男, 硕士生, 主要研究领域为图像隐写, 深度学习.



何沛松(1991－), 男, 博士, 讲师, 主要研究领域为多媒体安全, 深度学习.

王宏霞(1973－), 女, 博士, 教授, 博士生导师, CCF 高级会员, 主要研究领域为多媒体信息安全, 信息隐藏与数字水印, 数字取证, 智能信息处理.