

中图法分类号: TP309 文献标识码: A 文章编号: 1006-8961(2023)03-0749-11

论文引用格式: Wang J, Chen K J, Zhang W M and Yu N H. 2023. Image processing network-inverted identifiable secure natural steganography. Journal of Image and Graphics, 28(03): 0749-0759 (王健, 陈可江, 张卫明, 俞能海. 2023. 面向可逆图像处理网络的可证安全自然隐写. 中国图象图形学报, 28(03): 0749-0759) [DOI: 10.11834/jig.220529]

面向可逆图像处理网络的可证安全自然隐写

王健, 陈可江*, 张卫明, 俞能海

1. 中国科学技术大学网络空间安全学院, 合肥 230027; 2. 中国科学院电磁空间信息重点实验室, 合肥 230027;
3. 网络空间安全态势感知与评估安徽省重点实验室, 合肥 230027

摘要: **目的** 自然隐写是一种基于载体源转换的图像隐写方法, 基本思想是使隐写后的图像具有另一种载体的特征, 从而增强隐写安全性。但现有的自然隐写方法局限于对图像 ISO (International Standardization Organization) 感光度进行载体源转换, 不仅复杂度高, 而且无法达到可证安全性。为了提高安全性, 本文结合基于标准化流的可逆图像处理模型, 在隐空间完成载体源转换, 同时通过消息映射的设计做到了可证安全的自然隐写。**方法** 利用目前发展迅速的基于可逆网络的图像处理方法将图像可逆地映射到隐空间, 通过替换使用的隐变量完成载体源的转换, 从而避免对原始图像复杂的建模。同时, 改进了基于拒绝采样的消息映射方法, 简单地从均匀分布中采样以获得需要的条件分布, 高效地将消息嵌入到隐变量中, 并且保证了嵌入消息后的分布与原本使用的分布一致, 从而实现了可证安全的自然隐写。**结果** 针对图像质量、隐写容量、消息提取准确率、隐写安全性和运行时间进行了实验验证, 结果表明在使用可逆缩放网络和可逆去噪网络时能够在每个像素值上平均嵌入 5.625 bit 消息, 且具有接近 99% 的提取准确率, 同时隐写分析网络 SRNet (steganalysis residual network) 和 Zhu-Net 的检测准确率都在 50% 附近, 即相当于随机猜测。**结论** 本文提出的隐写框架利用可逆图像处理网络实现了可证安全的自然隐写, 在隐写容量和安全性上都具有很大优势。

关键词: 隐写; 自然隐写; 可证安全隐写; 可逆神经网络 (INN); 图像处理

Image processing network-inverted identifiable secure natural steganography

Wang Jian, Chen Kejiang*, Zhang Weiming, Yu Nenghai

1. School of Cyber Science and Technology, University of Science and Technology of China, Hefei 230027, China;
2. Key Laboratory of Electromagnetic Space Information, Chinese Academy of Sciences, Hefei 230027, China;
3. Anhui Province Key Laboratory of Cyberspace Security Situation Awareness and Evaluation, Hefei 230027, China

Abstract: **Objective** Natural steganography is regarded as a cover-source switching based image steganography method. To enhance the steganographic security, its objective is focused on more steganographic image-related cover features. Natural steganography is originally designed for ISO (International Standardization Organization) sensitivity through adding noise to

收稿日期: 2022-06-10; 修回日期: 2022-09-22; 预印本日期: 2022-09-29

* 通信作者: 陈可江 chenkj@ustc.edu.cn

基金项目: 国家自然科学基金项目 (62102386, 62002334, 62072421, 62121002); 中国博士后科学基金项目 (2021M693091); 网络空间安全态势感知与评估安徽省重点实验室开放基金项目; 中央高校基本科研业务费专项资金资助 (WK2100000018)

Supported by: National Natural Science Foundation of China (62102386, 62002334, 62072421, 62121002); China Postdoctoral Science Foundation (2021M693091); Open Fund of Anhui Province Key Laboratory of Cyberspace Security Situation Awareness and Evaluation; Fundamental Research Funds for the Central Universities (WK2100000018)

a low ISO image to yield a high ISO image feature, and modeling this noise signal to complete the message embedding. This approach is required for modeling the generation of ISO sensor noise, the development pipeline from raw sensor data stored in RAW format is commonly used like portable network graphics (PNG) or joint photographic experts group (JPEG) format images, which is very complex and not precise enough and the existing natural steganography approaches cannot be identified for safety inefficiently. To make the stylized images generated by steganography indistinguishable from other stylized images, some existing approaches are employed to explore steganography on the basis of image style transformation. However, it is challenged that the steganography-generated stylized image has the same distribution as the stylized image from another source, and none of them is as secure as traditional natural steganography. Actually, it is possible to achieve clarified security via using generated images as the cover image. Steganography is tackled for stronger invisibility than cryptography, but it has been difficult to achieve identifiable security, and most of methods are constrained of empirical security. Due to existing identifiable secure methods is required to obtain the distribution of cover datasets or the ability to sample from the cover distribution accurately, it is not feasible for traditional cover datasets. However, datasets-generated are easy to exact sampling because generative models random variables are required to be introduced to manipulate data generation. Therefore, to accomplish cover-source switching in latent space, and achieve identifiable secure natural steganography, the invertible image processing network is coordinated in terms of normalization flow. **Method** First, the image is mapped to the latent space in reverse using invertible neural networks-based image processing method, and the distribution of the latent variables are used to determine the cover features. To avoid modeling the original image features like ISO sensitivity, latent variable is configured as cover-source to switch and it can optimize the complexity of steganography significantly. At the same time, the identifiable secure steganography is implemented in the invertible image processing. The source switching-after cover is exactly the same as the cover of the other source, instead of the traditional natural steganography method, which can only be approximated in maximize. Since most of the invertible neural networks-based image processing methods use normal distribution-oriented latent variables to participate in recovery, conditional probability distribution of the stego latent variables about uniformly distributed messages can be designed. To enable message embedding and extraction, the stego latent variables can meet the requirement of normal distribution and the messages-linked correlation is obtained at the same time. To obtain the stego latent variable that matches the target conditional probability distribution, the inverse transform sampling-based message-embedding can be done efficiently via simplified regular distribution sampling. By keeping the stego latent variables with the same normal distribution as the latent variables used in normal image processing, the images involved in recovering or generating have the same distribution as well. **Result** Experiments are carried out in terms of image quality, steganographic capacity, message extraction accuracy, steganographic security, and its runtime. Using the same quality evaluation method with the original image processing network, there is almost no difference in peak signal-to-noise ratio and structural similarity between our Steganography and non-steganography images. To optimize steganographic capability, invertible de-noising and invertible rescaling network-based steganography methods are able to extract the message with about 99% accuracy while embedding 5.625 bits per pixel on the stego image. And, invertible de-colorization network-based steganography method achieves more than 99% extraction accuracy while embedding 0.67 bits of message per pixel. For steganographic security, we demonstrate the identifiable security of the proposed method, and neither of the two deep learning steganalysis networks used can distinguish the stego image from the cover image with better accuracy than random guesses. Furthermore, to reduce the time from the exponential growth of the latter to a constant time, our improved message mapping algorithm can be used to improve the efficiency compared to the rejection sampling approach. **Conclusion** An identifiable secure natural steganographic system is facilitated based on invertible image processing network, and the experimental results show that our method has its potentials for steganographic-related capacity and security.

Key words: steganography; natural steganography; provably secure steganography; invertible neural network (INN); image processing

0 引言

隐写术是保密通信的重要技术,与密码技术只要求消息无法破解不同,隐写术进一步要求通信过程无法察觉。由于数字图像具有较大冗余信息且广泛使用,以图像为载体的图像隐写技术最为成熟。目前主流的图像隐写术基于最小化失真框架,基本流程为设计嵌入代价,再利用 STC (syndrome-trellis code) (Filler 等, 2011) 等隐写编码进行嵌入。与之相对的,隐写分析技术通过提取隐写修改的特征将其与正常图像区分开。由于深度学习技术的不断发展,基于深度学习的隐写分析技术对主流的最小化失真隐写取得了很高的检测率,发展全新的安全隐写术势在必行。

为了提高隐写安全性,Denemark 和 Fridrich (2015) 使用图像处理过程中的边信息对失真进行调整,获得了比初始失真更好的性能。赵鑫等人 (2022) 利用超分辨率下采样的边信息获得了某些情况下比原始边信息更好的安全性。同样是利用边信息,Bas (2016) 提出了一种基于载体源转换原则的全新隐写方法,称为自然隐写。这种方法的特点是添加特定噪声使载体图像从一个源转换到另一个源,使载密图像具有另一个源中载体的统计特性,例如从低 ISO (International Standardization Organization) 感光度图像转换到高 ISO 感光度的图像。后续针对 ISO 的自然隐写对 JPEG (joint photographic experts group) 域和更合理的图像处理管道设计了更有效的方法 (Denemark 等, 2018; Taburet 等, 2019, 2021), 获得了更高的嵌入容量和隐写安全性。但这一类方法都依赖于载体的原始传感器信息以及合理的图像处理管道的建模,因此实现复杂,需要从 RAW (raw image format) 格式图像开始对图像处理管道建模,隐写速度很慢。并且隐写分析实验证明它们无法达到 50% 检测错误率的理论安全。

这种自然隐写方法局限性的来源就是自然图像分布的复杂性,只能通过建模复杂的图像处理管道获取不同 ISO 的噪声分布。但是,随着深度学习技术的发展,生成模型已经广泛应用在图像处理领域,生成模型处理后图像也运用在各种网络场景。若以生成模型产生的图像为载体,使嵌入消息后生成的图像与正常生成模型产生的图像保持相同的分布,

则能够完成基于生成模型的自然隐写。

Zhong 等人 (2019) 通过失真调整方法在图像风格迁移后的图像上探索了这一方法。他们使用略有不同的风格迁移方法产生载体图像和参考图像,如使用同一风格迁移网络不同迭代次数下的模型。然后使用 HILL (high-pass, low-pass, low-pass) (Li 等, 2014) 在载体图像上生成初始嵌入代价,在这个基础上调整嵌入代价使载体图像倾向于往参考图像的方向修改,从而使隐写后的载密图像具有一定的参考图像的特征。尽管这种方法比针对 ISO 的自然隐写更简单,但需要使用最小化失真框架,且安全性不如传统的自然隐写。Wang 等人 (2019) 和 Bi 等人 (2021) 都尝试了直接修改图像风格迁移网络,在完成风格迁移的同时嵌入消息,而不是像前者一样分成两步。这种方法完全由网络完成,不需要借助最小化失真框架,但他们的消息嵌入操作是通过修改网络结构完成的,因此嵌入时的网络和未嵌入的网络结构有所不同,生成图像的分布也无法保持一致,只能依靠风格迁移过程本身的随机性掩盖隐写行为,无法抵抗针对该方法训练过的隐写分析网络的攻击。

上述方法虽然一定程度上用到了生成模型生成的图像,但都没有考虑利用生成图像的分布来完成可证安全隐写。Cachin (1998) 用载体数据与载密数据分布之间的相对熵来定义隐写的信息论安全。设载体数据的分布为 $P_c(x)$, 载密数据的分布为 $P_s(x)$, 它们之间的相对熵为

$$D(P_c \| P_s) = \sum_{x \in X} P_c(x) \log_2 \frac{P_c(x)}{P_s(x)} \leq \varepsilon \quad (1)$$

若 ε 为 0, 则称隐写是绝对安全的。但要达到这个目标是十分困难的,传统的图像隐写术通常采用自然图像作为载体,而自然图像的分布是难以获取的,要使载密图像保持与载体图像同分布更无从谈起。但生成图像的分布却通常是已知或者容易采样的,因此以生成图像为载体,设法保证载密图像与载体图像的分布一致,即可使它们的相对熵为 0, 从而做到可证安全的隐写。

近年来,基于标准化流的生成模型广泛运用在图像处理领域,包括可逆去噪 (Liu 等, 2021)、可逆图像缩放 (Xiao 等, 2020) 和可逆图像灰度化 (Zhao 等, 2021) 等。它们利用标准化流设计可逆神经网络,构造了一个可逆变换,将输入图像可逆地变换到

隐空间中,从而在隐空间中分离出待处理部分变量和剩下的隐变量,再进一步完成去噪或者恢复原始图像的功能。这类可逆图像处理网络通常将剩下的无用部分隐变量丢弃,然后从正态分布中重新采样隐变量。对于同一幅图像的处理,网络逆过程恢复的图像分布完全由隐变量决定,所有服从正态分布的隐变量生成的图像都有可能作为正常的图像输出。因此本文以这类可逆图像处理方法为基础,结合自然隐写的思想,通过设计消息映射,将消息嵌入到隐变量中,使嵌入后的图像与载体图像具有一致的分布,从而完成了可证安全的自然隐写。

本文的主要贡献为:1)结合自然隐写的思想和可逆图像处理网络,完成了在生成载体上的自然隐写;2)设计了新的采样算法进行消息映射,相比用拒绝采样的消息映射大幅提高了效率;3)从理论上证明了本文方法的可证安全性,并通过实验验证了该隐写框架的大容量和安全性。

1 相关工作

1.1 可证安全隐写

Hopper 等人(2002)给出了计算安全隐写的形式化定义,并基于拒绝采样构造了两类可证安全隐写方法。一类是利用伪随机函数族进行拒绝采样,这类方法将安全性归约于伪随机函数族的伪随机性;一类是假设存在载体分布上的无偏函数进行拒绝采样,这类方法将安全性归约于加密算法的安全性。这两类方法都需要一个完美的采样器获得载体分布。传统的修改式隐写无法做到可证安全,正是因为传统的自然图像不存在这种采样器。

但人工智能生成模型的发展为可证安全隐写提供了机会。生成模型产生的图像、音频或文字等生成数据的分布通常可以从生成模型中直接或间接获得,因此在生成数据广泛运用的前提下,以生成数据为载体可以做到可证安全的隐写。

以图像为载体,Yang 等人(2019)实现了基于图像自回归生成模型 PixelCNN (pixel convolutional neural networks)的可证安全隐写,利用自适应算数编码将消息嵌入到生成图像所需要的序列中。但这种图像生成模型的效果和应用场景都有限。Chen 等人(2022)探索了基于文本合成语言生成模型的可证安全隐写。对于自回归生成模型,分别使用算

数编码的译码算法和编码算法完成消息嵌入和提取;对于基于标准化流的可逆生成模型,使用拒绝采样将消息映射为服从标准正态分布的隐变量参与语音合成。Zhang 等人(2021)提出了可证安全的文本生成隐写,通过自适应动态分布(adaptive dynamic grouping, ADG)保证消息嵌入过程几乎不破坏条件分布。Meteor(Kapchuk 等,2021)同样利用文本生成模型设计了一个利用对称的可证安全隐写模型。

1.2 基于可逆神经网络的图像处理

可逆神经网络(invertible neural network, INN)即基于标准化流的生成网络,是生成模型的一种,能够学习从一个空间到隐空间的双射,从而将一个分布可逆地映射为另一个分布。由于其精确的密度估计特性(Dinh 等,2017),INN 广泛运用于图像处理领域,如可逆图像缩放(Xiao 等,2020)和可逆图像灰度化(Zhao 等,2021)。这类可逆图像处理网络的特点是进行有信息损失的图像处理后能够较好地恢复原始图像,基本框架如图1所示。其中, x 表示待处理的图像,如图像缩放中的高分辨率图像以及灰度化处理中的彩色图像; y 表示处理后生成的图像,即对应的低分辨率图像和灰度图像; z 则是一个隐变量,用于维持变换前后变量的总尺寸不变以保持可逆。通常来说,INN 由多种可逆的网络层堆叠而成,主要包括 Haar 小波映射层和仿射耦合层,在不考虑计算误差时都是完全可逆的。其中前者用于完成尺度变换和高低频分离,而后者则作为主要可学习部分。在图像处理任务中,研究者根据不同任务的需求设计不同的 INN 结构,但基本都满足图1所示的基本框架。可逆缩放和可逆灰度化都是利用可逆神经网络的正向过程完成图像处理,再利用逆向过程恢复原始图像。由于这两种操作都会损失信息,为了能够从信息量较少的处理后图像 y 中恢复原始图像 x ,他们设计损失函数在训练时将隐变量约束为正态分布,从而使隐变量中蕴含尽可能少的信息。在需要恢复原始图像时,从正态分布中采样一个新的隐变量 $\hat{z}$ 参与恢复过程。这种做法会导致

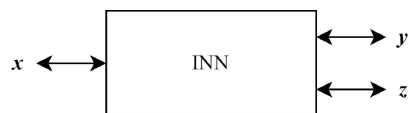


图1 可逆网络图像处理的基本框架

Fig. 1 Basic framework for INN-based image processing

恢复的图像与原始图像不完全一致,但能够获得可接受的图像质量。

Liu 等人(2021)则注意到了可逆神经网络能够用于图像去噪,使用了与可逆缩放网络类似的结构。首先将含噪声的图像 x 映射为干净低分辨率图像 y 和含噪隐变量 z , 在这个过程中将噪声信息与干净图像解耦。然后用从正态分布采样的隐变量 $\hat{z}$ 和 y 通过可逆缩放网络的逆过程得到目标干净图像。与前文提到的两种可逆图像处理方法不同,去噪网络使用可逆神经网络是为了通过在隐空间中替换潜在变量完成去噪任务,所以在去噪过程中同时使用了可逆神经网络的正向过程和逆向过程。该方法虽然称为可逆去噪网络,但事实上并不追求能够从去噪后的图像恢复原始的含噪声图像。

这两类使用可逆神经网络的图像处理方法具有一个共同特征,即在可逆神经网络的逆向过程都是从正态分布中采样新的隐变量,而不是保存

正向过程产生的隐变量。对于可逆去噪,这种设计的目的是将噪声集中在隐变量中丢弃,从而从正态分布中采样生成干净图像;对于可逆缩放和可逆灰度化,这种做法能够保证只利用处理后的图像能够恢复较高质量的原始图像。通过损失函数的设计,这两类方法都做到了使用从正态分布中采样不同的隐变量得到的图像恢复效果或去噪效果一致。

图2为不同隐变量 z 产生的图像质量。第1、2行分别由不同隐变量 z 恢复得来,第3行是前两幅图像绝对差放大40倍后取反的结果。图2(a)(b)中的可逆图像处理方法使用不同隐变量恢复的原始图像视觉上无法区分,峰值信噪比(peak signal to noise ratio, PSNR)和结构相似性(structural similarity index measure, SSIM)也相差无几。图2(c)中不同隐变量参与去噪的干净图像同样在视觉质量和定量指标上都无法区分。

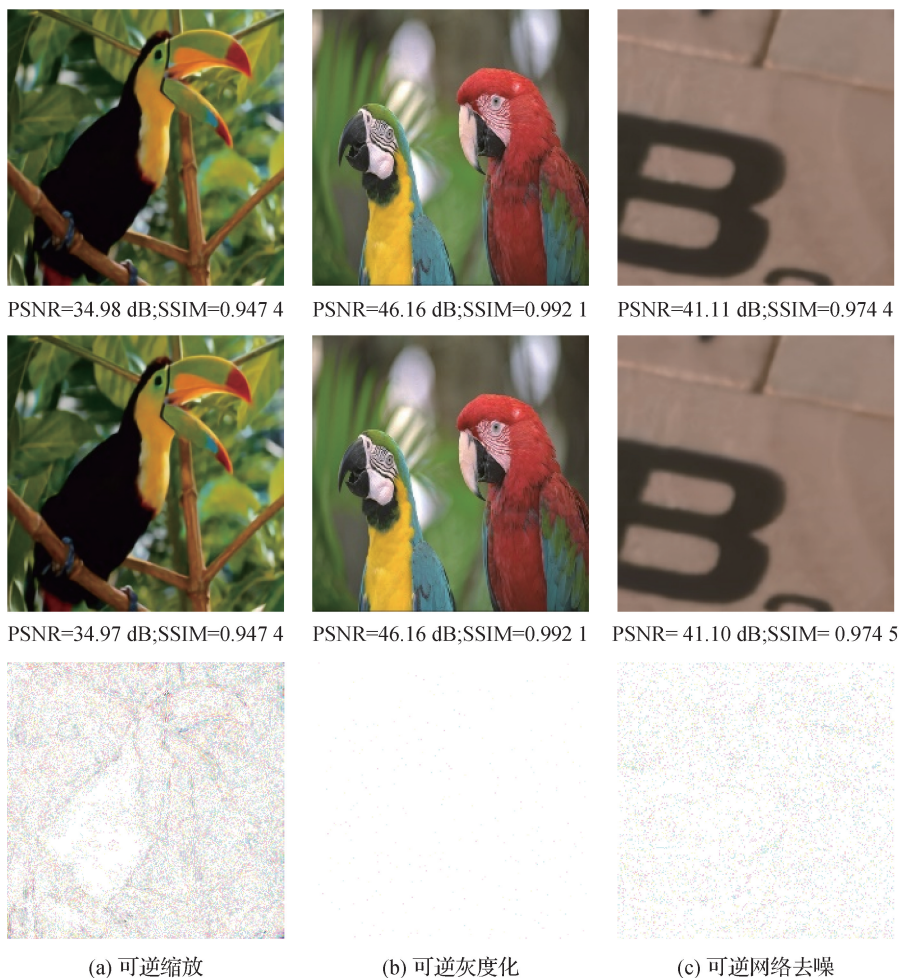


图2 不同隐变量 z 产生的图像质量

Fig. 2 Quality of images generated with different z ((a) invertible rescaling; (b) invertible decolorizing; (c) invertible denoising)

本文将这类基于可逆神经网络图像处理方法的逆过程看做一个生成模型,同一幅图像的生成完全由正态分布的隐变量决定。这种模型很适合使用基于生成模型的可证安全隐写方法,只需要设计合适的消息映射方法将消息映射为正态分布,即可将消息嵌入到恢复图像的随机性中。

另外,这两类基于可逆神经网络的图像方法都具有广泛的运用场景。例如对一幅高分辨率图像或者彩色图像,可以利用可逆图像处理网络将其处理为低分辨率图像或者灰度图像,用于不需要原始图像的场景。与传统图像处理方法相比,这种处理是可逆的,因此能在需要原始图像的时候恢复出原始图像,从而适合用在各种需要恢复原始图像的场景;而基于可逆网络的去噪方法,能够对真实相机拍摄得到的含噪声图像进行去噪,适用于各种使用含噪声真实图像的场景。因此,在这两类基于可逆神经网络的图像方法上进行隐写,能够利用正常的图像处理行为掩盖隐写行为,进一步达到行为安全的特点,即不会因为异常行为被识别为隐写行为。

2 本文方法

本文设计了一个基于可逆图像处理方法的可证安全自然隐写框架,如图3所示。该隐写框架接受加密过的消息 m 作为输入,经过消息映射模块得到服从正态分布的含密隐变量 z_m ,参与到正常可逆图像处理的反向恢复过程中,从而得到载密图像 x_m 。由于可逆网络和消息映射算法都是可逆的,可以简单地从载密图像中提取出原消息。这个过程与正常使用可逆网络图像处理过程唯一的区别在于反向经过网络使用的隐变量是由消息映射得来,而正常使用可逆网络图像处理过程是从正态分布中采样得到。可以证明,通过这种方式嵌入消息生成的图像具有与载体图像一致的分布,因此实现了可证安全的自然隐写。

2.1 消息嵌入与提取

由于可逆神经网络的可逆性,消息的嵌入和提取都能由一个网络实现,因此隐写的接收双方只需要共享同一个可逆图像处理网络以及消息加密使用的对称密钥。

本文在可逆神经网络逆向过程中嵌入消息,然后利用其正向过程提取消息。对于可逆缩放和可逆

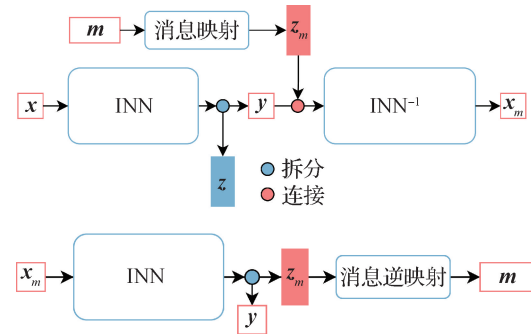


图3 本文方法的整体框架

Fig. 3 Framework of our method

灰度化这种使用正向过程处理图像的,在恢复原始图像的过程嵌入消息,生成与正常恢复的图像不可区分的载密图像;而对于基于可逆网络的去噪这类方法,由于它们同时使用了可逆网络的正向过程和逆向过程处理图像,就在其中的逆向过程中嵌入消息,生成与正常去噪后的图像不可区分的载密图像。

嵌入时,发送方首先对消息进行加密,使用消息映射算法获得待嵌入的隐变量。然后将该隐变量代替正常使用可逆图像处理网络时随机采样的隐变量参与后续过程得到生成图像。为了保证可逆性,得到的图像需要以标签图像存储格式(tag image file format, TIFF)存储。TIFF格式支持多种色彩空间和存储精度,能够避免以PNG(portable network graphics)格式存储时从浮点数量化到整数的信息损失。

提取时,直接将载密图像正向通过共享的可逆网络恢复隐变量,再通过消息映射的逆过程和解密过程即可得到明文消息。

2.2 消息映射

假设消息 m 在嵌入前经过加密,服从 $0 \sim 1$ 均匀分布。本文的目标是将消息 m 可逆地映射为服从正态分布的隐变量 z 。

设嵌入位数为 k ,每 k 比特(bit)消息可以看做一个整数 $y \in [0, 2^k - 1]$,在消息经过加密的前提下, y 服从离散均匀分布,即

$$p(y = i) = \frac{1}{2^k}$$

$$i = 0, 1, \dots, 2^k - 1 \quad (2)$$

设标准正态分布的密度函数为 $f(x) \sim N(0, 1)$,其累积分布函数的逆函数即分位数函数为 f_{pp} 。当消息 $y = i$ 时,希望采样得到的数 z 满足条件分布,即

$$p(z | y = i) =$$

$$\begin{cases} 2^k f(z) & f_{pp}\left(\frac{i}{2^k}\right) < z < f_{pp}\left(\frac{i+1}{2^k}\right) \\ 0 & \text{其他} \end{cases} \quad (3)$$

由全概率公式可得

$$p(z) = \sum_{i=0}^{2^k-1} p(z|y=i)p(y=i) = f(z) \quad (4)$$

即当加密后的消息服从离散均匀分布时,嵌入消息后的隐变量仍然服从正态分布。

Chen 等人(2022)的方法是使用拒绝采样。具体而言,假设隐变量中每个数嵌入 k bit 消息,即嵌入位数为 k ,可以通过正态分布的累计分布函数的逆函数(又称分位数函数)获得 2^k 个等概率区间,即从正态分布中采样落在这些区间内的概率是相等的,从而, k bit 消息的每一种情况都能唯一地与一个区间对应。对每组嵌入的 k bit 消息,重复地从正态分布中采样,直到采样得到的数落入消息对应的区间才接受这次采样。

但是,拒绝采样有一个严重的缺点。随着嵌入位数的增大,划分的区间数会指数增长,而从均匀分布中采样落入目标区间的概率为区间数的倒数,因此每次采样的期望时间将指数增长,这会极大地限制隐写容量和隐写效率。事实上,拒绝采样通常用于难以获得目标分布的显式表达的情况,而本文的目标分布是已知的,所以可以使用逆变换采样,构造均匀分布到目标分布的变换,从而通过从均匀分布中采样获得服从满足式(3)分布的结果。逆变换采样来自于这样一个事实:一个随机变量的累积分布函数将其自身变换到标准均匀分布,因此对标准均匀分布进行形如目标累积分布函数的逆函数的变换即可获得目标分布。

设正态分布的累积分布函数为 f_{cd} ,其逆函数即分位数函数为 f_{pp} ,通过式(3)很容易计算目标累积分布函数,具体为

$$F(z|y=i) = \begin{cases} 0 & z < f_{pp}\left(\frac{i}{2^k}\right) \\ 1 & z > f_{pp}\left(\frac{i+1}{2^k}\right) \\ 2^k f_{cd}(z) - i & \text{其他} \end{cases} \quad (5)$$

其逆函数为

$$F^{-1}(u|y=i) = f_{pp}\left(\frac{u+i}{2^k}\right) \quad (6)$$

则当 $u \sim \text{Uniform}(0,1)$ 时,进行以下变换

$$z = f_{pp}\left(\frac{u+i}{2^k}\right) \quad (7)$$

得到的随机变量 z 即服从式(3)的条件分布。

上述映射过程是可逆的,根据式(7)容易恢复隐变量中每个数对应的整数,具体为

$$i = \lfloor 2^k f_{cd}(z) \rfloor \quad (8)$$

式中, $\lfloor \cdot \rfloor$ 表示向下取整。再将整数 i 转换回 2 进制比特串后解密即可恢复出明文,从而完成从隐变量恢复消息的逆映射。

另外,式(3)的分布形式上是截断正态分布,可以利用截断正态分布直接采样获得目标结果。容易看出,这两种采样方法的时间成本都与嵌入位数无关,且没有拒绝采样的失败概率,因此能够大幅提高采样效率。但实验结果表明使用逆变换采样的效率更高,因此最终采用逆变换采样的方法完成消息映射。

图4演示了嵌入位数为2时的消息映射与逆映射的过程。其中 m 和 m' 分别为原消息和恢复的消息, y 和 y' 分别为原消息对应的数组和恢复的数组, z 为映射生成的隐变量, u 为从均匀分布采样得到的变量。

2.3 可证安全的自然隐写

根据可逆图像处理网络的可逆性,该网络构造

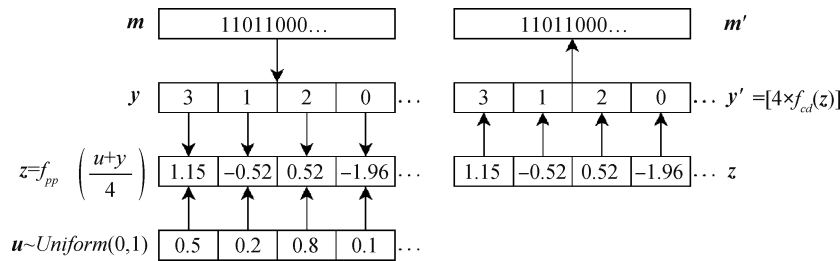


图4 消息映射与逆映射

Fig.4 Message mapping and inverse mapping

了一个可逆映射 f , 其逆过程将隐变量 z 和中间图像 y 映射到目标图像 x (如可逆去噪得到的干净图像、可逆缩放恢复的高分辨率图像和可逆灰度化恢复的彩色图像)。因此对同一幅中间图像 y , 正常使用该网络时直接从正态分布中采样得到隐变量 z , 生成目标图像 $x = f(y, z)$ 的所有可能取值完全由 z 决定。假设给定一幅中间图像 y , 使用正常采样的隐变量 z_c 生成的载体图像为 x_c , 而使用嵌有消息的隐变量 z_s 生成的载密图像为 x_s 。由式(4)可知, 消息映射得到的隐变量服从正态分布, 与正常从正态分布中采样得到的隐变量是同分布的。因此直接从正态分布中采样时一定存在 $z'_c = z_s$, 它对应的载体图像为 x'_c 。即载密图像与某个可能的载体图像完全一致, 这正是自然隐写的思想: 使隐写后的载密图像与另一种情况下的载体图像不可区分。本文方法更进一步地使载密图像与一种可能的载体图像完全一致, 从而达到了可证明的安全性。

事实上, 隐写的接收方和拥有可逆网络的攻击者能够从该网络逆向过程生成的任意图像中提取得到隐变量, 进一步提取出离散均匀分布的比特流, 但只有接收方拥有解密密钥, 从而通过能否解密出合理的信息分辨载体图像和载密图像。因此本文的安全性也依赖于使用的加密算法的安全性, 即加密后的消息具有密码安全的随机性。

3 实验与分析

本文基于可逆去噪网络、可逆缩放网络和可逆灰度化网络进行实验。由于需要保证网络的可逆性, 使用 TIFF 格式存储浮点数图像, 以避免转换为 8 bit 位图的取整误差。实验对生成图像的质量、隐写容量与消息提取准确率、隐写安全性以及隐写的时间成本进行了验证。

在图像质量方面, 本文的载密图像与载体图像相当于不同隐变量生成的图像, 理论上无论是视觉还是定量指标都相差无几。实验对 3 种方法分别在 SIDD (smartphone image denoising dataset) 验证集 (Abdelhamed 等, 2018)、Div2k (diverse 2K resolution high quality images) (Agustsson 和 Timofte, 2017) 和 Kodak24 (Franzen, 1999) 数据集上进行测试, 这些数据集都是原方法在测试时使用的。视觉效果如图 5 所示。对于可逆去噪网络, 嵌入消息生成的去噪图

像与正常去噪的图像视觉效果无法区分; 对于可逆缩放网络和可逆灰度化网络, 在恢复过程中嵌入消息, 可以看到隐写后的恢复效果与正常恢复效果基本一致。

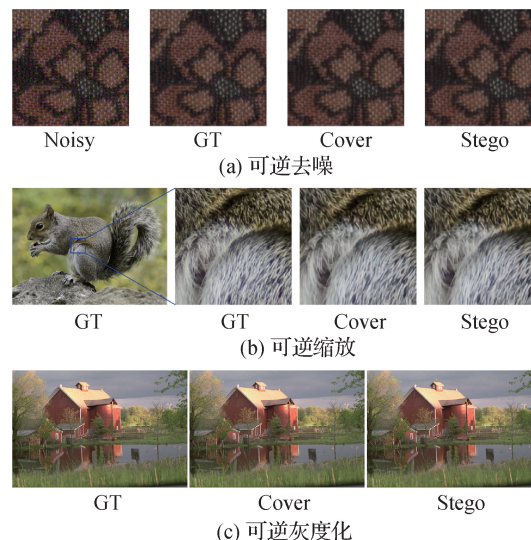


图 5 载体图像与载密图像的视觉质量

Fig. 5 Visualization results of cover and stego

((a) invertible denoising; (b) invertible rescaling;
(c) invertible decolorizing)

为了定量描述隐写对使用的可逆图像处理网络的影响, 使用对比各自真实数据的峰值信噪比 PSNR 和结构相似度 SSIM 描述载体图像与载密图像的视觉指标。从表 1 可以看出, 对各种方法, 载体图像和载密图像的指标的差距都可以忽略, 证实了隐写与否理论上不会影响原网络处理与恢复的结论。

表 1 载体图像与载密图像的定量指标

Table 1 Quantitative metrics of cover and stego

图像处理方法	载体图像		载密图像	
	PSNR/dB	SSIM	PSNR/dB	SSIM
可逆缩放	33.02	0.914 6	33.02	0.914 5
可逆灰度化	43.35	0.991 5	43.34	0.991 5
可逆网络去噪	39.09	0.913 6	39.09	0.913 6

对一种隐写方法而言, 隐写容量和安全性是最重要的两个方面。本文提出的隐写框架的隐写容量由使用可逆图像处理网络的隐变量的大小和嵌入位数决定。设嵌入位数为 k , 可逆去噪网络和可逆缩放网络的正向过程都是将图像的宽和高减少到初始

尺寸的1/4,因此隐变量的大小为初始图像像素值的15/16,即隐写容量都为0.937 5k bit/像素。而可逆灰度化的正向过程则是将RGB三通道的彩色图像转换为单通道灰度图,因此隐变量的大小为初始图像像素值的2/3,即隐写容量为2k/3 bit/像素。

虽然使用TIFF格式存储载密图像避免了浮点数到整数的取整误差,但由于可逆网络自身使用浮点数计算会引入部分误差,本文方法实际上不能完全准确地提取消息。并且嵌入位数越多,隐变量中不同消息对应的区间越窄,误差导致的提取错误率就越高。表2给出了本文方法在不同嵌入位数下的消息提取准确率。可以看到,对于可逆去噪网络和可逆缩放网络,即使每个隐变量嵌入6 bit消息,隐写容量高达5.625 bit/像素,仍然能获得接近99%的消息提取准确率,在使用纠错码后能够在保持较大隐写容量的同时完成准确提取。另外,可以观察到利用可逆灰度化网络进行隐写的消息提取准确率明显低于前两者,即使在相同嵌入位数下可逆灰度化隐写的嵌入容量远小于前两者。这是因为可逆灰度化网络对正向过程输出的隐变量和灰度图解耦更彻底导致的。如图2所示,不同的隐变量恢复的彩色图像之间的差值极小。因此相应地,彩色图像中存在的微小计算误差会导致消息提取时获得的隐变量具有较大差距,进而影响消息提取。

表2 不同嵌入位数下的消息提取准确率

Table 2 Message extraction accuracy with different embedding bits

	/%				
图像处理方法	1 bit	2 bit	4 bit	6 bit	8 bit
可逆缩放	99.94	99.90	99.71	99.16	97.48
可逆灰度化	99.18	98.54	95.87	88.98	80.58
可逆网络去噪	99.90	99.81	99.52	98.64	95.88

对于本文方法的安全性,用深度学习隐写分析方法SRNet(Boroumand等,2019)和Zhu-Net(Zhang等,2020)进行了实验验证。首先,在10 000幅彩色的Bossbase图像上使用了可逆缩放网络,产生了10 000幅由可逆缩放网络下采样并恢复的图像,作为载体图像。对这些载体图像使用HILL(Li等,2014)在0.2 bit/像素嵌入率下隐写生成载密图像,由于HILL只针对灰度图像,对彩色图像的每个通

道都使用了HILL进行隐写。在这10 000对载体—载密对上训练深度学习隐写分析器,检测准确率如表3所示,第1行结果表明,深度学习隐写分析器能够有效区分在使用的图像载体上进行的自适应隐写。并且,因为可逆缩放网络恢复的高分辨率图像会不可避免地丢失部分复杂区域的信息,而HILL需要在复杂区域嵌入消息才能获得安全性,所以这种分两步完成的隐写方式安全性很低,两种隐写分析网络都能以超过90%的准确率检测。

表3 深度学习隐写分析器对不同方法的检测准确率

Table 3 Detection accuracy of deep learning steganalyser for different methods

	/%	
隐写方法	SRNet	Zhu-Net
可逆缩放 + HILL(0.2 bpp)	99.00	94.55
可逆缩放 + 本文方法	50.05	50.10
可逆灰度化 + 本文方法	50.00	49.90
可逆网络去噪 + 本文方法	50.00	50.00

随后,测试这两种深度学习隐写分析方法能否检测使用本文框架的隐写。由于可逆缩放和可逆灰度化测试用的数据集太小,无法用于隐写分析网络的训练,本文在10 000幅彩色BossBase图像上运用可逆缩放和可逆灰度化,再分别使用正常采样得到的隐变量和消息映射得到的隐变量进行恢复,得到各自的载体图像和载密图像。而对于可逆网络去噪,本文则直接在SIDD验证集上分别使用正常采样得到的隐变量和消息映射得到的隐变量参与去噪,得到载体图像和载密图像。分别使用SRNet和Zhu-Net在3种方法上训练后,测试结果如表3所示。本文多次调整网络参数进行训练,结果网络计算的交叉熵损失都保持在0.693 1附近不下降,验证时的准确率保持50%附近,即相当于随机猜测。这表明网络无法从载体图像和载密图像中学习到任何能够区分它们的特征。

由于本文的隐写框架可以基于不同的可逆图像处理网络,隐写所需的时间为正常使用网络消耗的时间加上消息映射耗费的时间,因此对不同消息映射方法的效率进行了实验测试。实验时,固定隐变量长度为10 000,测试在不同嵌入位数下不同采样方法消耗的时间。如图6所示,拒绝采样所需要的

时间随嵌入位数增大指数上升,而直接由截断正态分布采样和由均匀分布采样映射消耗的时间都基本不变。这与本文 2.2 节的理论推导一致,证明了本文提出的逆映射采样能够大幅提高这类消息映射的效率,进而提高隐写算法的效率。

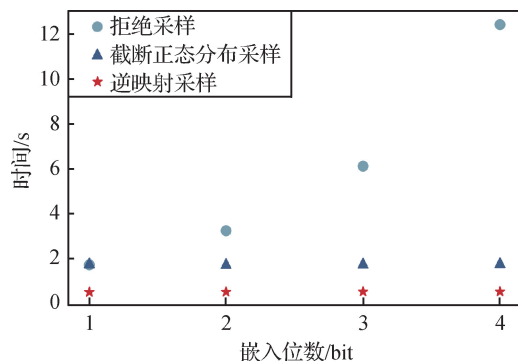


图6 不同消息映射方法的时间成本

Fig. 6 Time cost of different message mapping methods

4 结 论

可逆神经网络越来越多地用于图像处理领域,这类图像处理方法为隐写术提供了新的应用场景和方法。本文考虑到传统自然隐写只针对 ISO 感光度的缺点,将自然隐写的思想运用到可逆神经网络图像处理上,提出了一种可证安全的自然隐写算法。首先,分析了可逆网络图像处理的基本框架,使用这类网络常用的隐变量设计了消息嵌入算法;其次,设计了新的消息映射算法,大幅提高了消息映射的效率;最后,证明了本文方法的可证明安全性,在使用 TIFF 格式存储图像时能够与正常使用可逆网络图像处理方法产生不可区分的图像。

在可逆缩放、可逆灰度化和可逆网络去噪上运用本文提出的可证安全自然隐写框架进行了实验。结果表明,隐写后产生的图像与正常使用网络产生的图像从视觉上不可区分,且 PSNR 和 SSIM 都相差无几;对可逆缩放和可逆网络去噪,在隐写容量高达 5.625 bit/像素时仍能以 99% 左右的准确率提取消息;在安全性上,深度学习隐写分析方法 SRNet 和 Zhu-Net 对 3 种网络上的隐写检测准确率均为 50% 左右,即无法比随机猜测更好。与传统自然隐写方法相比,本文方法省去了复杂的处理过程,并且将安全性提高到了可证明安全;与其他可证安全隐写方

法相比,本文设计的消息映射算法大幅提高了效率,从拒绝采样随嵌入位数指数上升的时间成本降低到了常数成本,并且能够借助自然隐写的思想用可逆网络图像处理行为掩盖隐写行为。

但本文方法也存在一些不足。由于可逆网络本身较差的鲁棒性,在隐变量上嵌入消息需要整个过程高度可逆才能保证准确提取。因此需要使用 TIFF 格式存储图像,避免使用 PNG 格式存储导致的浮点数取整误差,这导致本文方法的使用范围受到限制。后续的研究工作中,将设法改善本文方法的鲁棒性问题,争取在适用范围更广的 PNG 格式上成功实现可证安全的自然隐写。

参考文献 (References)

- Abdelhamed A, Lin S and Brown M S. 2018. A high-quality denoising dataset for smartphone cameras//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA; IEEE: 1692-1700 [DOI: 10.1109/CVPR.2018.00182]
- Agustsson E and Timofte R. 2017. NTIRE 2017 Challenge on Single Image Super-Resolution: dataset and study//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Honolulu, USA; IEEE: 1122-1131 [DOI: 10.1109/CVPRW.2017.150]
- Bas P. 2016. Steganography via cover-source switching//Proceedings of 2016 IEEE International Workshop on Information Forensics and Security (WIFS). Abu Dhabi, United Arab Emirates; IEEE: 1-6 [DOI: 10.1109/WIFS.2016.7823905]
- Bi X L, Yang X Y, Wang C and Liu J. 2021. High-capacity image steganography algorithm based on image style transfer. Security and Communication Networks, 2021: 4179340-1-4179340-14 [DOI: 10.1155/2021/4179340]
- 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]
- Cachin C. 1998. An information-theoretic model for steganography//Proceedings of the 2nd International Workshop on Information Hiding. Portland, USA; Springer: 306-318 [DOI: 10.1007/3-540-49380-8_21]
- Chen K J, Zhou H, Zhao H Q, Chen D D, Zhang W M and Yu N H. 2022. Distribution-preserving steganography based on text-to-speech generative models. IEEE Transactions on Dependable and Secure Computing, 19(5): 3343-3356 [DOI: 10.1109/TDSC.2021.3095072]

- Denemark T, Bas P and Fridrich J. 2018. Natural steganography in JPEG compressed images. *Electronic Imaging*, 30 (7): #316 [DOI: 10.2352/ISSN.2470-1173.2018.07.MWSF-316]
- Denemark T and Fridrich J. 2015. Side-informed steganography with additive distortion//Proceedings of 2015 IEEE International Workshop on Information Forensics and Security (WIFS). Roma, Italy: IEEE: 1-6 [DOI: 10.1109/WIFS.2015.7368589]
- Dinh L, Sohl-Dickstein J and Bengio S. 2017. Density estimation using Real NVP//Proceedings of the 5th International Conference on Learning Representations. Toulon, France: OpenReview.net
- 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]
- Franzen R. 1999. Kodak lossless true color image suite [DB/OL]. [2022-05-20]. <http://r0k.us/graphics/kodak>
- Hopper N J, Langford J and Von Ahn L. 2002. Provably secure steganography//Proceedings of the 22nd Annual International Cryptology Conference. Santa Barbara, USA: Springer: 77-92 [DOI: 10.1007/3-540-45708-9_6]
- Kapchuk G, Jois T M, Green M and Rubin A D. 2021. Meteor: cryptographically secure steganography for realistic distributions//Proceedings of 2021 ACM SIGSAC Conference on Computer and Communications Security. Virtual Event, Korea(South): ACM: 1529-1548 [DOI: 10.1145/3460120.3484550]
- Li B, Wang M, Huang J W and Li X L. 2014. A new cost function for spatial image steganography//Proceedings of 2014 IEEE International Conference on Image Processing (ICIP). Paris, France: IEEE: 4206-4210 [DOI: 10.1109/ICIP.2014.7025854]
- Liu Y, Qin Z Y, Anwar S, Ji P, Kim D, Caldwell S and Gedeon T. 2021. Invertible denoising network: a light solution for real noise removal//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 13360-13369 [DOI: 10.1109/CVPR46437.2021.01316]
- Taburet T, Bas P, Sawaya W and Fridrich J. 2019. A natural steganography embedding scheme dedicated to color sensors in the JPEG domain. *Electronic Imaging*, 31 (5): #542 [DOI: 10.2352/ISSN.2470-1173.2019.5.MWSF-542]
- Taburet T, Bas P, Sawaya W and Fridrich J. 2021. Natural steganography in JPEG domain with a linear development pipeline. *IEEE Transactions on Information Forensics and Security*, 16: 173-186 [DOI: 10.1109/TIFS.2020.3007354]
- Wang Z H, Gao N, Wang X, Xiang J and Liu G Q. 2019. STNet: a style transformation network for deep image steganography//Proceedings of the 26th International Conference on Neural Information Processing. Sydney, Australia: Springer: 3-14 [DOI: 10.1007/978-3-030-36711-4_1]
- Xiao M Q, Zheng S X, Liu C, Wang Y L, He D, Ke G L, Bian J, Lin Z C and Liu T Y. 2020. Invertible image rescaling//Proceedings of 16th European Conference on Computer Vision. Glasgow, UK: Springer: 126-144 [DOI: 10.1007/978-3-030-58452-8_8]
- Yang K, Chen K J, Zhang W M and Yu N H. 2019. Provably secure generative steganography based on autoregressive model//Proceedings of the 17th International Workshop on Digital Watermarking. Jeju Island, Korea (South): Springer: 55-68 [DOI: 10.1007/978-3-030-11389-6_5]
- 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 S Y, Yang Z L, Yang J S and Huang Y F. 2021. Provably secure generative linguistic steganography//Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. Bangkok, Thailand: ACL: 3046-3055 [DOI: 10.18653/v1/2021.findings-acl.268]
- Zhao R, Liu T S, Xiao J, Lun D P K and Lam K M. 2021. Invertible image decolorization. *IEEE Transactions on Image Processing*, 30: 6081-6095 [DOI: 10.1109/TIP.2021.3091902]
- Zhao X, Wang Y F, Chen K J, Zhang W M and Yu N H. 2022. Spatial image steganography based on side information estimated by super resolution. *Journal of Image and Graphics*, 27 (1): 226-237 (赵鑫, 王垚飞, 陈可江, 张卫明, 俞能海. 2022. 引入超分辨率下采样误差的图像边信息估计隐写. *中国图象图形学报*, 27(1): 226-237) [DOI: 10.11834/jig.210433]
- Zhong N, Qian Z X, Wang Z C and Zhang X P. 2019. Steganography in stylized images. *Journal of Electronic Imaging*, 28 (3): #033005 [DOI: 10.1117/1.JEI.28.3.033005]

作者简介

王健,男,硕士研究生,主要研究方向为信息隐藏与人工智能安全。E-mail: wangjian1227@mail.ustc.edu.cn

陈可江,通信作者,男,副研究员,主要研究方向为信息隐藏与人工智能安全。E-mail: chenkj@ustc.edu.cn

张卫明,男,教授,主要研究方向为多媒体安全、信息隐藏与人工智能安全。E-mail: zhangwm@ustc.edu.cn

俞能海,男,教授,主要研究方向为图像处理与视频分析、媒体内容安全、网络通信与安全。

E-mail: ynh@ustc.edu.cn