引用格式: 苏海, 余松森, 杨珊. 基于分离训练与图像去噪的频率域彩色图像隐写方法[J]. 微电子学与计算机, 2024, 41(2): 28-36.

SU H, YU S S, YANG S. The color image steganography in frequency domain based on separation training and image denoising[J]. Microelectronics & Computer, 2024, 41(2): 28-36.

DOI: 10.19304/J.ISSN1000-7180.2023.0034

基于分离训练与图像去噪的频率域彩色图像隐写方法

苏 海, 余松森, 杨 珊

(华南师范大学 软件学院, 广东 佛山 528225)

摘 要: 彩色图像隐写方法具有秘密传输、不易察觉的特性。其中, 基于频率域的彩色图像隐写方法不论在传统图像隐写方法还是深度学习图像隐写方法中都取得了更好的隐写性能。然而, 当前大多基于自编码器结构的彩色图像隐写模型在提升重构秘密图像能力方面均存在局限性。针对这一问题, 本文基于频率域彩色图像隐写方法的现有优势, 提出了一种基于分离训练与图像去噪的频率域彩色图像隐写方法, 并构建了相应的隐写模型。面对自编码器的编码网络与解码网络在训练过程中的性能权衡问题, 本文的隐写方法采用分离训练对默认的神经网络训练方式进行优化。除此之外, 为了进一步提升重构秘密图像的质量, 模型还添加了去噪卷积神经网络(Denoising Convolutional Neural Network, DnCNN)结构的图像去噪模块。经实验验证, 本文模型生成的彩色载密图像与重构秘密图像的峰值信噪比(Peak Signal to Noise Ratio, PSNR)高达 82.31 dB 和 39.27 dB, 结构相似度(Structural Similarity Index Measure, SSIM)均达到 0.99。与同类型的深度学习彩色图像隐写模型相比, 提出的隐写模型不仅具有更强的不可察觉性, 而且具有更好的重构秘密图像的能力。

关键词: 图像隐写; 信息隐藏; 离散小波变换; 深度学习; 图像去噪

中图分类号: TP391 **文献标识码:** A **文章编号:** 1000-7180(2024)02-0028-09

The color image steganography in frequency domain based on separation training and image denoising

SU Hai, YU Songsen, YANG Shan

(School of Software, South China Normal University, Foshan 528225, China)

Abstract: Color image steganography attracts the attention of scholars because of its secretive and imperceptibility. The color image steganography based on frequency domain has achieved better performance in both traditional steganography and deep learning steganography. However, most current steganographic models based on auto-encoder have limitations in improving the ability of reconstructing secret images. Based on this problem and the existing advantages of steganography in the frequency domain, a color image steganographic method based on separation training and image denoising is proposed. In the face of the performance trade-off between the encoder and the decoder, the proposed method uses the separation training to optimize the model training. In addition, the proposed model adds an image-denoising module which is a Denoising Convolutional Neural Network(DnCNN). Experimental results show that the Peak Signal to Noise Ratio(PSNR) of the stego image and the reconstructed secret image reached 82.31 dB and 39.27 dB respectively, and the Structural Similarity Index Measure(SSIM) reached 0.99. Compared with other models, the proposed model not only has stronger imperceptibility but also has better ability to reconstruct secret images.

Key words: image steganography; information hiding; discrete wavelet transform; deep learning; image denoising

收稿日期: 2023-01-27; **修回日期:** 2023-03-13

基金项目: 广东省基础与应用基础研究基金(2021A1515110673)

1　引言

随着互联网技术的飞速发展，信息传输变得愈发高效。依托于开放的网络平台，图片、视频等形式的多媒体信息更容易进行大量的复制和传播。因此，信息泄露、侵权滥用或是恶意篡改等信息安全问题层出不穷。 为了解决上述信息安全问题，许多学者将目光聚焦于彩色图像隐写技术。彩色图像隐写技术是信息隐藏领域的重要研究分支之一，其可被广泛应用于军事、医疗或知识产权保护等领域，实现秘密信息的安全传输。彩色图像隐写技术是一种以难以察觉的方式将需要传输的信息隐藏至彩色图像中，并且在公开的网络环境中进行定向传输的一种信息隐藏技术。近几年，为了进一步提升彩色图像隐写技术的可用性，越来越多的学者尝试将深度学习技术与彩色图像隐写技术相结合，以实现隐写容量、信息安全性或是其他隐写性能的提升。借助深度神经网络强大的学习能力，彩色图像隐写技术已进入飞速发展的阶段。众多深度学习彩色图像隐写方案在隐写容量、安全性或是不可察觉性等方面表现出了优秀的性能。这些隐写方案的提出为图像的安全传输提供了更多可用的隐写方案，而且还进一步丰富了图像隐写技术的应用场景。

按照网络结构，基于深度学习的彩色图像隐写模型可分为 3 类：基于自编码器的模型、基于生成对抗网络(Generative Adversarial Networks, GAN)的模型和基于可逆神经网络(Invertible Neural Networks, INN)的模型。其中，编码网络-解码网络结构 的自编码器是最常见的深度学习隐写模型架构之一。自编码器中，编码网络用于隐写流程中秘密信息的嵌入任务，解码网络用于完成秘密信息的提取任务。由于自编码器网络结构与隐写在不同流程阶段的目标具有极高的贴合度，因此，该结构成为了深度学习彩色图像隐写模型的主流网络框架之一。然而，目前基于自编码器网络构建隐写模型在提升重构秘密图像能力方面仍然存在一定的局限性。

自编码器结构的图像隐写模型默认使用的是端到端的联合训练方式，并且当前大多数基于深度学习的彩色图像隐写模型以不可察觉性为目标进行损失函数的设计。然而在训练过程中，自编码器中的编码网络与解码网络需考虑到生成的彩色载密图像与重构秘密信息的质量权衡问题。这往往会导致当编码网络和解码网络中一方性能表现优越时，另一方网络的性能受限。GAN 结构的隐写模型可看作是自编码器结构的隐写模型与隐写分析模型结合、从而提升抗隐写分析能力的一种模型结构。INN 结构的隐写模型则更专注于提升隐写容量和鲁棒性。因此，目前仍无能够有效解决自编码器隐写模型的性能权衡问题的方法。

除了深度神经网络训练模式的问题，解码网络有限的学习能力同样是限制深度学习彩色图像隐写模型性能的重要原因之一。当前，许多深度学习彩色图像隐写模型以图像作为需要隐藏的秘密信息，从而达到大容量隐写的目的。然而，经过卷积、下采样等操作，输入的秘密图像信息矩阵难免丢失部分信息。因此，深度学习图像隐写模型生成的重构秘密图像仍存在一定程度的图像失真问题。

为了解决当前自编码器结构的彩色图像隐写模型的性能权衡问题并进一步减小重构秘密图像的失真，本文设计了一种基于分离训练与图像去噪模块的频率域彩色图像隐写方法，并基于自编码器构建了相应的隐写模型。经实验，该模型可在以图藏图的大隐写容量的前提下，生成具有高图像质量的彩色载密图像并成功提取灰度秘密图像。相比于其他同类模型，本文提出模型在生成的载密图像质量以及重构秘密图像能力方面均具有更好的表现。本文的主要贡献可以归结为以下 3 点：

(1)本文提出了一种基于分离训练与去噪模块的频率域彩色图像隐写方法，并以隐写模型中常见的自编码器结构构建了相应模型。相比于其他同类的隐写模型，该模型在不可察觉性以及重构秘密图像能力方面都表现出了最好的性能。

(2)本文使用分离训练方法对模型进行训练。在端到端联合训练方式下，本文提出模型生成的彩色载密图像已经具有十分接近原图的质量，然而重构的秘密图像仍有一定程度的图像失真。因此本文利用分离训练进一步挖掘解码网络的潜在学习能力，从而有效缓解了自编码器隐写模型的性能权衡问题。经实验，分离训练可有效提升自编码器中解码网络的性能。

(3)本文使用去噪卷积神经网络(Denoising Convolutional Neural Network, DnCNN)作为模型的图像去噪模块，从而有效缓解了重构秘密图像的失真问题。模型将重构秘密图像中的隐写痕迹看作一种噪声，使用图像去噪的思路，对解码网络输出的秘密图像进行图像去噪。添加去噪模块后，模型生成的重构秘密图像的质量明显提升。

2 基于深度学习的彩色图像隐写方法研究现状

隐写术(Steganography)是一种可实现信息隐藏与秘密传递的技术，起源于古希腊并且历史悠久。自 2016 年起，通过与深度学习技术相融合，许多新提出的彩色图像隐写方法在各种隐写性能方面都发生了质的飞跃。

根据模型结构，深度学习彩色图像隐写模型可分为基于自编码器的隐写模型、基于生成对抗网络(GAN)的隐写模型和基于可逆神经网络 (INN) 的隐写模型。早期的深度学习彩色图像隐写模型多为基于自编码器的隐写模型。例如第一个可隐藏同尺寸图像的隐写模型是谷歌的 Baluja 提出的基于自编码器结构的深度隐写模型[1-2]。在这之后，StegNet[3]、Rehman 的端到端隐写模型[4] 和基于 ResNet 的隐写模型[5] 等大容量的隐写模型，还有鲁棒性明显提升的 StegaStamp[6] 和王彦设计的基于编码-解码网络的鲁棒彩色图像隐写方案[7] 等基于自编码器的隐写模型被陆续提出。不同于基于自编码器的隐写模型，SteganoGAN[8]、ISGAN[9]、TISGAN[10]、UMC-GAN[11]、CHAT-GAN[12] 和 CIS-GAN[13] 等基于 GAN 的隐写模型以深度学习隐写分析模型作为判别器进行训练，从而有效抵御深度学习隐写分析模型的检测，提升了隐写的安全性；可逆隐写网络(ISN)[14]、RIIS[15] 和 Steg-cINN[16] 等基于 INN 的隐写模型利用 INN 的可逆性进行隐写模型的构建，在隐写容量、鲁棒性或是提取准确率等方面取得了显著的提升。

在上述 3 种基于深度学习的图像隐写模型中，自编码器结构是最常见的结构之一。相比于基于 GAN 的隐写模型和基于 INN 的隐写模型，基于自编码器的隐写模型可使用更加轻量的网络结构完成隐写[17]。然而，端到端的联合训练模式和网络自身有限的学习能力仍然是限制自编码器隐写模型性能的重要原因。考虑到基于频率域的深度学习隐写模型往往能够表现出更好的隐写性能(例如 ISN)，本文提出的模型选择在载体图像的频率域中隐写，并利用分离训练优化基于自编码器的隐写模型的训练方式，再通过添加 DnCNN 去噪模块解决自编码器隐写模型的学习能力有限的问题，从而构建基于分离训练与图像去噪的频率域彩色图像隐写模型。

3 基于分离训练与图像去噪的频率域彩色图像隐写模型

基于分离训练与图像去噪的彩色图像隐写模型的完整网络结构如图 1 所示。

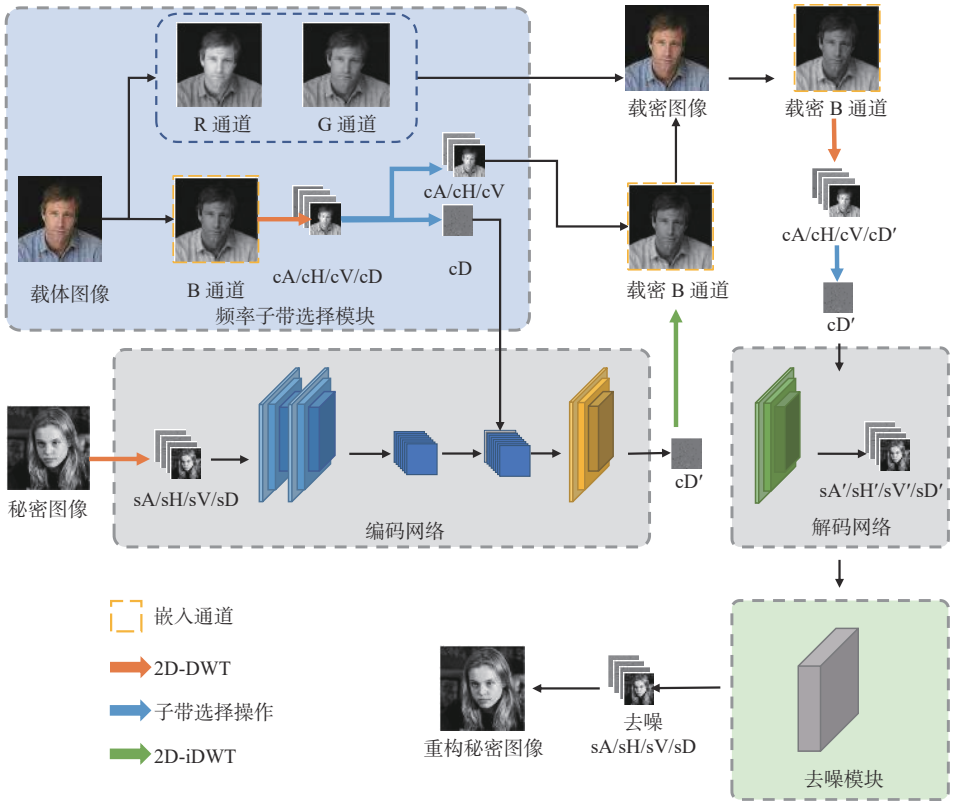

图 1 基于分离训练与图像去噪的频率域彩色图像隐写模型结构图

Fig. 1 Structure of the color image steganographic model in frequency domain based on separation training and image denoising

由图 1 可知，模型主要由频率子带选择模块、编码网络、解码网络和去噪模块这 4 部分构成。频率子带选择模块用于对载体数据进行预处理；编码网络与解码网络分别用于秘密图像信息的隐藏与提取；而去噪模块则用于对解码网络的输出结果进行去噪，优化输出结果。该模型使用的载体信息为尺寸 250*250 的 RGB 彩色图像，秘密信息为同尺寸的灰度图像。图中：cA/cH/cV/cD 分别代表彩色载体图像嵌入通道的通过离散小波变换(Discrete Wavelet Transform, DWT)得到的 4 个频率分量(对角低频分量/水平高频分量/垂直高频分量/对角高频分量)；sA/sH/sV/sD 分别代表灰度秘密图像的 4 个 DWT 频率分量；cD′表示载密图像的对角高频分量；sA′/sH′/sV′/sD′分别代表重构秘密图像的 4 个 DWT 频率分量。

根据各个模块的作用，模型的隐写流程主要分为 4 个阶段：第一个阶段，频率子带选择模块将对载体图像进行数据预处理，该模块选择了 RGB 图像的 B 通道的 cD 作为秘密信息的嵌入域，这主要是因为相比于另外两个通道，人眼对 B 通道的修改更不敏感[18]，并且视觉对图像中具有复杂纹理、高对比度或非常亮/暗的区域修改相对不敏感[19]；完成载体图像的预处理后，隐写流程的第二阶段和第三阶段将通过编码网络与解码网络完成秘密图像的嵌入和提取，并生成相应的载密图像和重构秘密图像的 DWT 频率分量；第四阶段，DnCNN 去噪模块将对解码网络的输出进行图像去噪操作。该模块的输入与输出均为 DWT 频率分量。因此，去噪模块的输出结果需要经过二维离散小波逆变换(2D inverse Discrete Wavelet Transform, 2D-iDWT)操作转换为最终的重构秘密图像。

3.1　分离训练

在端到端的联合训练方式下，编码网络已能够生成具备十分接近原始载体图像的载密图像，但解码网络生成的重构秘密图像仍然存在一定程度的失真。自编码器的训练模式是重构秘密图像出现失真的重要原因之一。隐写模型要完成的任务可分为生成载密图像与提取秘密信息。然而网络的学习能力是有限的，因此，在联合训练的过程中，如果模型太过专注于学习如何提升编码网络的性能，解码网络的性能就会受到抑制，反之亦然。由此可知，模型的重构秘密图像能力难以提升的原因之一是在编码-解码网络的联合训练模式下，解码网络的能力在一定程度上受到了抑制。

为了解决自编码器默认的端到端训练方式所导致的性能权衡问题，本文尝试根据隐写在不同阶段的目标对训练流程进行拆解，对模型进行分离训练，具体流程如图 2 所示，只有编码网络与解码网络会发生权重的更新。分离训练按照以图藏图隐写方法的两个主要目标分为两个训练阶段。

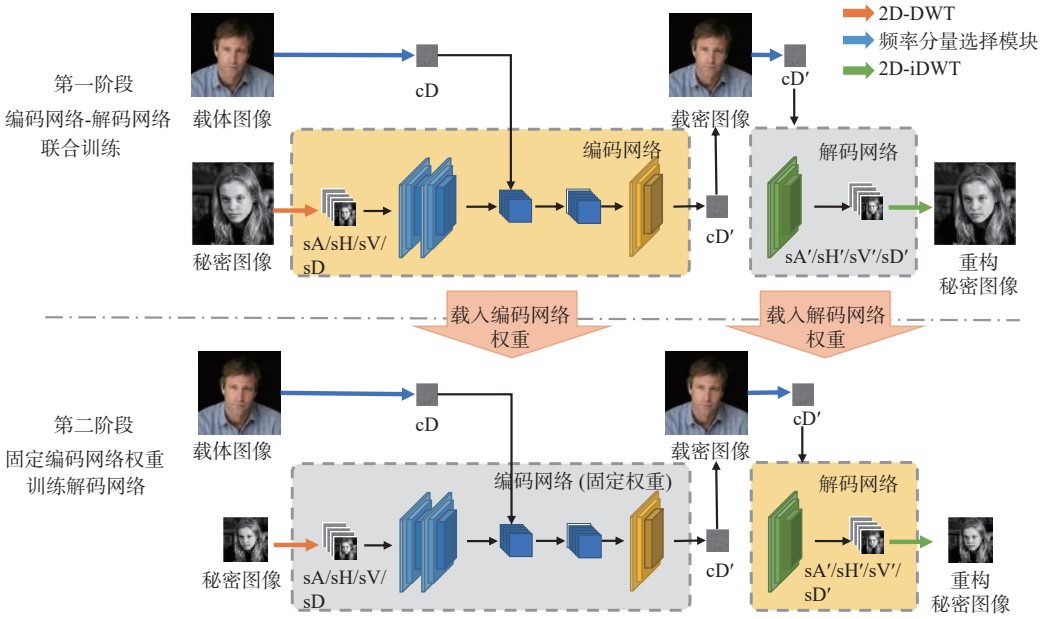

图 2　分离训练流程示意图

Fig. 2 Flow-process of the separation trainin

在第一阶段，模型按照原本的端到端联合训练方式同时训练编码网络与解码网络，并重点关注编

码网络的训练情况。此时，模型的损失函数如式(1)所示，两个损失项分别表示载密图像与载体图像之间的距离和秘密图像与重构秘密图像之间的距离。第二阶段的训练则更加关注于解码网络的训练。为了避免编码网络影响解码网络，编码网络的权重将被固定，并且不断生成载密图像来辅助解码网络进行独立训练。因为不再需要对编码网络进行训练，因此，第二阶段的损失函数(见式(2))不再使用载密图像与载体图像之间的距离损失函数项。

$$\mathrm{loss}(c,c',s,s') = \|c-c'\|^2 + \beta \times \|s-s'\|^2 \quad (1)$$

$$\mathrm{loss}(s,s') = \|s-s'\|^2 \quad (2)$$

参与分离训练的编码网络与解码网络均为全卷积神经网络，主要由卷积层构成，并通过连接层将各个卷积组进行连接。图 3 和图 4 分别给出了编码网络和解码网路的详细架构。在第二阶段，编码网络将接收两组输入数据：输入层 1 接收灰度秘密图像完成 DWT 操作之后获得的频率分量[sA, sH, sV, sD]；输入层 2 则通过频率分量选择模块接收 cD。编码网络的前两组卷积层可看作预处理网络，预处理网络输入 [sA, sH, sV, sD]，并输出提取的秘密图像特征图。特征图与 cD 连接后，由最后一组卷积层，即隐藏网络对连接的数据进行编码，最后输出含有秘密图像特征的载密图像对角高频分量 cD′。在第三阶段，接收者接收到载密图像之后，解码网络将会把 cD′输入至解码网络中，并使用一组卷积组提取秘密图像的 DWT 频率信息(用 [sA′, sH′, sV′, sD′] 表示)。最后，模型使用 2D-iDWT 操作将 [sA′, sH′, sV′, sD′] 转换为空间域信息，即重构的一通道灰度秘密图像。

使用分离训练方法，自编码器结构模型的训练可以更加贴合隐写在不同阶段的目标，从而达到更好的训练效果。基于分离训练的方式对本文提出的模型进行训练，既保证了彩色载密图像的质量不受影响，又可有效提升解码网络重构秘密图像的能力。

3.2 DnCNN 去噪模块

在隐写流程中，频率域-空间域变换、卷积操作都会不可避免地造成秘密图像的精度损失。而在公开的网络信道中，常见的 JPEG 图像压缩、高斯噪声和下采样等网络环境因素同样会导致传输的图像出现不同程度的失真。因此，本论文认为图像隐写方法可视为秘密图像在公开的网络信道中传输时可能遭遇的一种无意攻击。

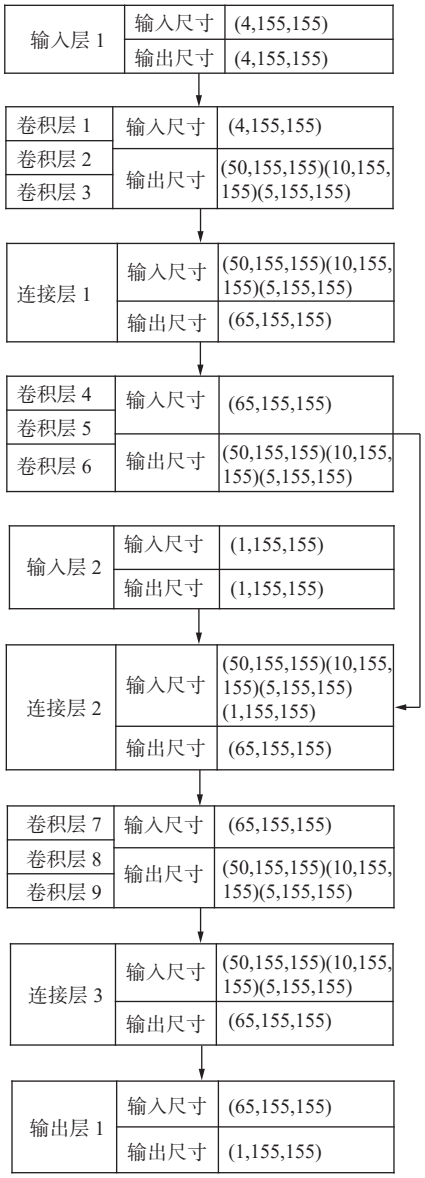

图3 编码网络结构图

Fig. 3 The structure of encoder

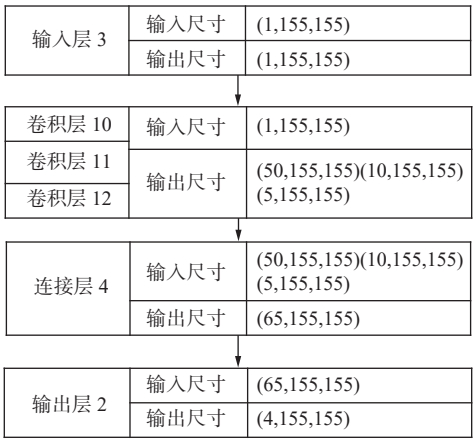

图4 解码网络架构图

Fig. 4 The structure of decoder

隐写技术可以保护秘密图像以一种隐蔽的方式被安全地传输到接收者处，但会在秘密图像中添加一些噪声，从而影响到重构秘密图像的质量。为了解决隐写带入的修改痕迹，进一步提升重构秘密图像质量，本文在解码网络后添加了图像去噪模块，以减轻秘密图像信息中的隐写痕迹。

本文使用的图像去噪模块基于 DnCNN。DnCNN[20] 是一种网络结构简明、训练时间成本低、去噪效果明显的图像去噪网络，主要由卷积层、BN 层、ReLU 激活函数级联组成。虽然原始的 DnCNN 在模型复杂度上远低于其他深度学习图像去噪模型，但其难以直接应用于失真程度比较微小的重构秘密图像。因此，本文的图像去噪模块参考了 TISGAN[10] 中去噪器的结构，通过添加跳跃连接、减少网络层数的方式降低 DnCNN 的结构复杂度，在保证模型效果的同时进一步降低了训练成本。图 5 给出了 DnCNN 去噪模块的具体网络结构。去噪模块的第一层和最后一层分别为卷积+ReLU 层和卷积层，模块中间部分为 15 层卷积+BN+ReLU 层，且每 5 层卷积+BN+ReLU 层间添加了跳跃连接，以避免梯度爆炸/消失的问题。不同于人眼可见的高斯噪

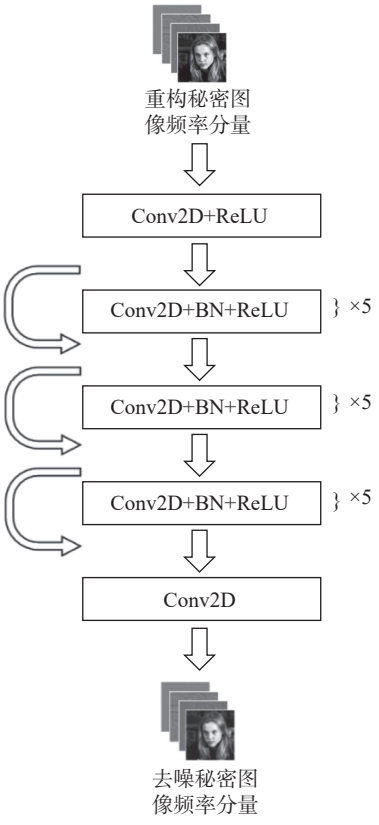

图5　DnCNN 图像去噪模块结构图

Fig. 5 Structure of DnCNN image-denoising module

声或是 JPEG 造成的块效应，解码网络生成的重构秘密图像与原始图像之间的差异十分微小，往往需要通过细微的观察才能够察觉到隐写对重构秘密图像造成的失真。因此，对于 DnCNN 来说，生成隐写噪声图像的残差图像是比较困难的任务。为了让 DnCNN 能够达到去噪的目的，本节设计的去噪模块的训练目标并不是输出原始秘密图像与重构秘密图像之间的残差图像，而是直接输出优化后的去噪秘密图像。

从隐写流程的角度分析，去噪模块可看作秘密信息提取阶段的一部分，但是该模块并不影响隐写的完整流程，只是对重构秘密图像信息进行了进一步的优化。因此，DnCNN 的网络训练可独自进行。DnCNN 去噪模块的训练目标是生成去噪后的重构秘密图像信息，因此，损失函数设计为原始灰度秘密图像与重构秘密图像的距离。DnCNN 去噪模块的损失函数如式(3)所示：

$$\text{loss}_{\text{DnCNN}} = \|s - s'\|^2 \tag{3}$$

4　实验与分析

本章将对本文提出的彩色图像隐写模型进行实验以及分析。实验中的模型主要使用了 the Labeled Faces in the Wild(LFW)数据集。训练集与测试集的数量比例为 4 000∶400。训练集和测试集将被平均分为载体图像和秘密图像，并且秘密图像将被统一处理成灰度图像。所有图像的大小重新裁剪为 250×250 的尺寸。训练过程中统一使用 Adam 优化器进行模型的权重参数的更新。分离训练时，第一阶段的训练迭代周期为 100，第二阶段的迭代周期为 300，联合训练的总训练迭代周期数为 400，相当于分离训练迭代数的总和。分离训练与联合训练的学习率均随着训练迭代周期增加而下降。模型的初始学习率为 0.001，第 150 个迭代后下降至 0.000 3，第 300 个迭代后将下降到 0.000 1。DnCNN 去噪模块的超参设置与隐写模型基本相同。除此之外，去噪模块的训练集和测试集均由原始的灰度秘密图像和解码网络生成的重构秘密图像组成。

在实验结果评价工作上，将使用隐写技术中较为常见的峰值信噪比(Peak Signal to Noise Ratio, PSNR)和结构相似度(Structural Similarity, SSIM)衡量生成图像的质量。PSNR 可用于衡量图像的失真程度，取值区间在 [0, 100]，值越大，则失真程度越小。PSNR 的计算表达式为

$$\mathrm{MSE}=\frac{1}{mn}\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}\|I(i,j)-K(i,j)\|^2 \tag{4}$$

$$\mathrm{PSNR}=10\times\lg\left(\frac{\mathrm{Max}_I^2}{\mathrm{MSE}}\right) \tag{5}$$

式中：m 和 n 为图像的长和宽；$I(i,j)$和 $K(i,j)$为原始图像 I 和对比图像 K 在(i,j)位置的像素值；MSE 为 I 和 K 的均方误差；Max_I 的默认值为 255。

SSIM 可衡量两幅图像的相似程度，取值区间为 [0, 1]，值越大则相似度越高。图像 x 和图像 y 的 SSIM 计算公式如式(6)所示：

$$\mathrm{SSIM}(x,y)=\frac{(2\mu_x\mu_y+c_1)(2\sigma_{xy}+c_2)}{(\mu_x^2+\mu_y^2+c_1)(\sigma_x^2+\sigma_y^2+c_2)} \tag{6}$$

式中：μ_x、μ_y 为平均值；σ_x、σ_y 为标准差；σ_{xy} 为协方差；c_1、c_2 为正值常数，用于防止除 0 异常。

4.1 隐写效果对比实验

对比实验将使用同样为以彩图藏灰图的深度学习隐写模型：Antique 提出的端到端模型[4]、ISGAN[9]和 TISGAN[10]。表 1 给出了各个模型生成的载密图像与重构秘密图像的质量情况。由表 1 可知：除秘密图像的 PSNR，本文提出模型的其他隐写指标均高于其他对比模型的隐写指标。此外，因本文提出模型仅在载体图像的 B 通道的对角高频分量进行修改，因此，在秘密信息嵌入后，R 通道与 G 通道的 PSNR 仍为 100，B 通道的 PSNR 为 46.93。另 3 组对比模型在生成的载密图像质量方面的表现并不出色，这主要是因为隐写模型进行的是大容量的隐写。目前，基于深度学习的隐写模型的主流隐写方式是修改载体图像的信息，从而完成秘密信息嵌入与提取[7]。然而，隐写模型的不可察觉性与隐写容量之间存在反关联性。这意味着隐写容量的增大往往会导致彩色载密图像出现人眼可见的隐写修改痕迹。但本文提出的模型仅在载体图像嵌入通道的高频分量进行信息隐藏，将修改控制在人眼不可察觉的范围内，从而保证了生成的彩色载密图像的质量。不仅如此，减少对载体图像的修改有利于提升载密图像的安全性。

表 1 各个隐写模型生成的载密图像质量与重构秘密图像质量对比表

Tab. 1 Comparison between the quality of the stego images and the reconstructed secret images generated by each steganographic model

模型	载密PSNR	秘密PSNR	载密SSIM	秘密SSIM
Antique's Model[4]	33.70	**39.90**	0.95	0.96
ISGAN[9]	34.63	33.63	0.95	0.94
TISGAN[10]	37.52	35.42	0.96	0.95
本文模型	**82.31**	39.27	**0.99**	**0.99**

综上可知，相比于同类的隐写模型，本文提出模型表现出了最好的不可察觉性与秘密图像的重构能力。

4.2 分离训练和去噪模块的优化效果验证实验

为了提升隐写模型的重构秘密图像的能力，本文主要进行了两点优化：将模型的联合训练方式更改为分离训练，并添加 DnCNN 去噪模块。本节将根据这两点改动进行消融实验，从而体现不同的优化点对模型的隐写性能提升程度。

表 2 给出了消融实验结果，通过表中数据可知：同时使用分离训练方法并添加了去噪模块的模型生成的重构秘密图像的质量指标结果最好；相比第一组实验结果，第四组模型的重构秘密图像的每像素平均误差有约 0.6 比特/像素的降低，并且 PSNR 和 SSIM 分别有约 1.8 dB 和 0.5 的提升；除此之外，将第二组、第三组的实验结果与第一组进行对比，可看出，不论是使用分离训练模式还是去噪模块，模型生成的重构秘密图像的质量均出现了小幅提升；第二组以及第三组的实验结果差异并不明显，由此可知，分离训练和去噪模块对最终模型的重构秘密图像能力的程度差别不大，并且不会相互影响。

表 2 隐写模型的分离训练与去噪模块消融实验结果

Tab. 2 Ablation experimental results of the separation training and the denoising module of the steganographic model

模型	基于最终模型的修改		重构秘密图像质量指标		
	分离训练	去噪模块	每像素误差	PSNR	SSIM
第一组	×	×	3.40	37.50	0.98
第二组	√	×	3.05	38.49	0.98
第三组	×	√	3.18	38.08	0.98
第四组	√	√	**2.79**	**39.27**	**0.99**

图 6(a) 为原始秘密图像；图 6(b) 为第一组未优化的模型重构秘密图像的 30 倍残差图，从而将图

像的残差放大至易于人眼观察的程度；图 6(c) 为第四组优化后模型的重构秘密图像的 30 倍残差图。图中边框表示出修改痕迹差异比较明显的区域。不难看出，第四组使用分离训练并添加去噪模块的模型的重构秘密图像更加接近于原始秘密图像。

(a) 原始秘密图像
(a) Original secret image

(b) 重构秘密图像 residual×30
(b) Reconstruct secret image, residual×30(first group)

(c) 重构秘密图像 residual×30
(c) Reconstruct secret image, residual×30(fourth group)

图 6　未优化模型与优化模型重构秘密图像残差对比图

Fig. 6 Comparison of reconstructed secret images and residual images between the unoptimized model and the optimized model

在分离训练与去噪模块的作用下，优化模型的重构秘密图像的能力实现了明显的提升，但是编码网络并没有因为解码网络性能的提升而受到抑制。经实验，经分离训练与去噪模块的模型生成的载密图像的每像素平均误差为 1.02，PSNR 高达 81.05 dB（嵌入通道的 PSNR 达到了 43.162 2 dB），SSIM 高达 0.998 1。相比之下，优化的模型与未优化模型分别生成的载密图像的质量指标结果之间仅存在细微差异，但是各项质量指标的具体差异数值均在正常波动范围内。综上，通过分离训练与添加去噪模块可有效避免训练过程中编码网络与解码网络的性能权衡问题。

5　结束语

(1)本文提出了基于分离训练与图像去噪的彩色图像隐写方法，并使用自编码器网络结构组建了对应的彩色图像隐写模型。

(2)针对自编码器结构的彩色图像隐写方法所面临的端对端训练受限、重构秘密图像质量难以提升等问题，本文提出了使用分离训练与添加 DnCNN 图像去噪模块两个优化点加以解决。其中，分离训练使模型在后期的训练中避免了不同模块性能相互影响。添加去噪模块则是将重构秘密图像中的隐写痕迹视作无意攻击带入的噪声，以图像去噪的思路进一步提升了重构秘密图像的质量。实验证明了所提出的模型在同类隐写模型中具有最好的不可察觉性和重构秘密图像的能力，而且分别验证了分离训练和去噪模块对于最终模型的优化效果。

参考文献：

[1] BALUJA S. Hiding images in plain sight: deep steganography[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. Vancouver: Neural Information Processing Systems Foundation, 2017.

[2] BALUJA S. Hiding images within images[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 42(7): 1685-1697. DOI: 10.1109/TPAMI.2019.2901877.

[3] WU P, YANG Y, LI X Q. StegNet: mega image stegano-

graphy capacity with deep convolutional network[J]. Future Internet, 2018, 10(6): 54. DOI: 10.3390/fi10060054.

[4] RAHIM A U, RAHIM R, NADEEM S, et al. End-to-end trained CNN encoder-decoder networks for image steganography[C]//Proceedings of the European Conference on Computer Vision. Heidelberg: Springer, 2018: 723-729.

[5] LIU L S, MENG L Z, WANG X L, et al. An image steganography scheme based on ResNet[J]. Multimedia Tools and Applications, 2022, 81(27): 39803-39820. DOI: 10.1007/s11042-022-13206-2.

[6] TANCIK M, MILDENHALL B, REN N. StegaStamp: invisible hyperlinks in physical photographs[C]//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2020. DOI: 10.1109/cvpr42600.2020.00219.

[7] 王彦. 基于 Encoder-Decoder 的图像隐写技术研究[D]. 南京: 南京信息工程大学, 2021. DOI: 10.27248/d.cnki.gnjqc.2021.000450.
WANG Y. Image steganography based on encoder-decoder model[D]. Nanjing: Nanjing University of Information Science and Technology, 2021. DOI: 10.27248/d.cnki.gnjqc.2021.000450.

[8] ZHANG K A, CUESTA-INFANTE A, XU L, et al. SteganoGAN: high capacity image steganography with GANs[J]. arXiv, 2019, 1901.03892. https://arxiv.org/pdf/1901.03892.pdf.

[9] ZHANG R, DONG S Q, LIU J Y. Invisible steganography via generative adversarial networks[J]. Multimedia Tools and Applications, 2018, 78(7): 8559-8575. DOI: 10.1007/S11042-018-6951-Z.

[10] WU G Z, YU X Y, LIANG H, et al. Two-step image-in-image steganography via GAN[J]. International Journal of Digital Crime and Forensics (IJDCF), 2021, 13(6): 1-12. DOI: 10.4018/IJDCF.295814.

[11] ZHAO J F, WANG S. A stable GAN for image steganography with multi-order feature fusion[J]. Neural Computing and Applications, 2022, 34(18): 16073-16088. DOI: 10.1007/s00521-022-07270-w.

[12] TAN J X, LIAO X, LIU J T, et al. Channel attention image steganography with generative adversarial networks[J]. IEEE Transactions on Network Science and Engineering, 2022, 9(2): 888-903. DOI: 10.1109/TNSE.2021.3139671.

[13] 廖鑫, 唐志强, 曹纭. 基于生成对抗网络的空域彩色图像隐写失真函数设计方法[J]. 软件学报, 2022, 33(9): 3470-3484. DOI: 10.13328/j.cnki.jos.006290.
LIAO X, TANG Z Q, CAO Y. Steganographic distortion function design method for spatial color image based on GAN[J]. Journal of Software, 2022, 33(9): 3470-3484. DOI: 10.13328/j.cnki.jos.006290.

[14] LU S P, WANG R, ZHONG T, et al. Large-capacity image steganography based on invertible neural networks[C]//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2021: 10811-10820. DOI: 10.1109/cvpr46437.2021.01067.

[15] XU Y M, MOU C, HU Y J, et al. Robust invertible image steganography[C]//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2022: 7875-7884. DOI: 10.1109/cvpr52688.2022.00772.

[16] REN Y Z, LIU T, ZHAI L M, et al. Hiding data in colors: secure and lossless deep image steganography via conditional invertible neural networks[J]. arXiv, 2022, 2201.07444. https://arxiv.org/abs/2201.07444.

[17] 付章杰, 王帆, 孙星明, 等. 基于深度学习的图像隐写方法研究[J]. 计算机学报, 2020, 43(9): 1656-1672. DOI: 10.11897/SP.J.1016.2020.01656.
FU Z J, WANG F, SUN X M, et al. Research on steganography of digital images based on deep learning[J]. Chinese Journal of Computers, 2020, 43(9): 1656-1672. DOI: 10.11897/SP.J.1016.2020.01656.

[18] ALMAZAYDEH L. Secure RGB image steganography based on modified LSB substitution[J]. International Journal of Embedded Systems, 2020, 12(4): 453. DOI: 10.1504/ijes.2020.107644.

[19] 陈孟华, 刘嘉勇, 何沛松. 基于注意力机制与生成对抗网络的彩色图像隐写算法[J]. 现代信息科技, 2022, 6(7): 70-76. DOI: 10.19850/j.cnki.2096-4706.2022.07.018.
CHEN M H, LIU J Y, HE P S. A color image steganography algorithm based on attention mechanism and generative adversarial network[J]. Modern Information Technology, 2022, 6(7): 70-76. DOI: 10.19850/j.cnki.2096-4706.2022.07.018.

[20] ZHANG K, ZUO W M, CHEN Y J, et al. Beyond a Gaussian denoiser: residual learning of deep CNN for image denoising[J]. IEEE Transactions on Image Processing, 2017, 26(7): 3142-3155. DOI: 10.1109/TIP.2017.2662206.

作者简介:

苏 海 博士, suhai@m.scnu.edu.cn

余松森(通信作者) 博士, yss8109@163.com