结合图像加密与深度学习的高容量图像隐写算法

杨晓元[1,2]，毕新亮[1,2]，刘佳[1,2]，黄思远[1]

（1. 武警工程大学密码工程学院，陕西 西安 710086；2. 网络与信息安全武警部队重点实验室，陕西 西安 710086）

摘 要：针对基于深度学习的高容量图像隐写方案存在的载体图像和含密图像的残差图像会暴露秘密图像的问题，提出了结合图像加密和深度学习的高容量图像隐写算法。该算法设计使用了一种图像特征提取方法，使得从载体图像中提取的特征与从含密图像中提取的特征是一致的。发送方在图像隐写前，从载体图像中提取特征作为密钥，用来加密秘密图像。提取方提取加密过的秘密图像后，从含密图像中提取特征作为密钥，用来解密秘密图像。实验结果表明，攻击者无法从残差图像中发现秘密图像的信息，且密钥传递的频率更低，算法安全性得到了提升。

关键词：深度学习；高容量；图像加密；图像隐写

中图分类号：TP309.2

文献标识码：A

DOI: 10.11959/j.issn.1000−436x.2021134

High-capacity image steganography algorithm combining image encryption and deep learning

YANG Xiaoyuan[1,2], BI Xinliang[1,2], LIU Jia[1,2], HUANG Siyuan[1]

1. College of Cryptographic Engineering, Engineering University of PAP, Xi'an 710086, China
2. Key Laboratory of Network and Information Security of the PAP, Xi'an 710086, China

Abstract: Aiming at the problem that the residual image of cover image and carrier image in the high-capacity image steganography scheme based on deep learning will expose the secret image, a high-capacity image steganography scheme combining image encryption and deep learning was proposed. An image feature extraction method was used, so that the features extracted from the cover image were consistent with the features extracted from the carrier image. Before the image steganography, the sender extracted features from the cover image as a key to encrypt the secret image. After the extractor extracted the encrypted secret image, the features were extracted from the carrier image as a key to decrypt the secret image. The experimental results show that the attacker cannot find the information of the secret image from the residual image, and the frequency of key transmission is lower, and the security of the algorithm is improved.

Keywords: deep learning, high capacity, image encryption, image steganography

1 引言

隐写术是将秘密信息隐藏在公开通信载体中进行秘密传输的一种技术。与密码学相比，它更加注重隐藏通信过程本身。传统的隐写技术中，广泛使用的是基于载体修改的隐写技术[1]，它将秘密信息隐藏在人类感官难以察觉到的区域[2-3]。

随着深度学习技术的发展与成熟，它被广泛应用于图像翻译[4]、图像分割[5]、图像分类[6]、图像生成[7]等领域，相比于传统的图像处理技术，深度学习技术表现出了更大的潜力和更好的效果。2014 年，生成对抗网络（GAN, generative adversarial net-

收稿日期：2021−03−30；**修回日期**：2021−06−15

通信作者：刘佳，liujia1022@gmail.com

基金项目：国家自然科学基金资助项目（No.61872384）

Foundation Item: The National Natural Science Foundation of China (No.61872384)

work）[8]被提出，深度学习技术给图像隐写研究带来了新突破[9]。基于深度学习的隐写算法的主要特点是以大量数据为驱动，相较于人工设计，该类隐写算法更智能和自动化，它不需要隐写算法设计者制定一系列复杂的操作步骤，设计者只要根据经验设计好网络结构，给网络明确训练目标，并输入数据，网络就可以自动调整参数，以满足设计者的需求。

基于深度学习的高容量图像隐写技术较大地提升了隐写容量[10]。该技术在一张图像中隐藏一张和该图像尺寸大小相同的图像。但是这类技术存在一定的缺陷，即载体图像和含密图像的残差图像会暴露秘密图像。有学者提出使用图像加密的方式进行改进[11-12]，先对秘密图像进行加密，而后再将加密后的秘密图像嵌入载体图像中，但是这些算法都需要频繁地更新密钥以保证其安全性，这也就需要频繁地传递新的密钥。为了减少密钥传递的频率，本文算法提出了使用载体图像的特征作为密钥，通过更换载体图像，在不减少密钥更新频率的基础上，减少了密钥传递的频率，安全性更高。本文主要贡献如下。

1) 本文直接以载体图像的特征作为图像加密过程中使用的密钥，通过更换载体图像，可以直接更新密钥。

2) 本文设计使用的图像特征提取算法，可以从载体图像与含密图像中提取相同的特征，确保正确解密出秘密图像。

3) Sharma 等[12]的图像加密方法不同，本文没有将图像的某一行或某一列进行移动，而是给每一个图像块一个新的位置，更难被破解。

2　相关工作

随着深度学习技术的不断创新与发展，它的应用范围越来越广，也给图像隐写技术带来了新的发展。Baluja[10]将秘密图像隐藏在载体图像中，最小化载体图像和含密图像的差异，最终达到载体图像与含密图像的视觉不可分辨。与传统的隐写算法相比，该方案能够隐藏较多的信息。但是攻击者如果能够得到载体图像，就能从载体图像与含密图像的残差图像中发现秘密图像的信息。Rahim 等[13]提出在载体图像特征提取的过程中就不断将秘密图像特征与之融合，并且保证编码图像与载体图像的视觉和统计特征的一致性，以确保隐藏的图片不会被发现。但是这种方法仅能隐藏单通道的灰度图像，且含密图像颜色失真较明显，还会出现秘密图像的轮廓，被发现的可能性较大。Zhang 等[14]在此基础上进行改进，在 Y 通道中隐藏灰度图像，以确保含密图像不会产生颜色失真，且淡化了含密图像中秘密图像的轮廓，使用 inception module[15]构成了隐藏网络，添加了隐写分析网络，提升了方案的隐蔽性，但是该方案与文献[13]一样，仅能隐藏一张单通道的灰度图像。Duan 等[16]基于这两者的工作，再次改进，使用 U-net[17]作为编码器，U-net 能更好地捕获图像的细节信息，取得了较好的生成效果。但是上述这些算法仍然能从载体图像和含密图像的残差图像中发现秘密图像的信息。为了改善这个不足，Duan 等[11]提出了秘密图像预处理办法，首先对秘密图像进行离散余弦变换（DCT, discrete cosine transform），而后使用椭圆曲线加密算法对 DCT 后的秘密图像进行加密，把秘密图像处理成一个伪随机的噪声图像，并使用改进的 SegNet[18]，将加密的秘密图像隐藏进载体图像中。该方案通过 DCT 和椭圆曲线加密解决了含密图像中存在秘密图像轮廓的问题。Sharma 等[12]同样也使用了图像加密的方式来解决秘密图像暴露的问题，与 Duan 等提出的算法不同的是，Sharma 等首先将图像进行分块，而后根据参数以行和列为单位，对图像块进行移动。但是这 2 种算法都需要对密钥进行更新，如果长时间使用同一种密钥，有被攻击者攻破的风险，安全性有待提升。

本文同样也采用“先加密，再嵌入”的方式，先对秘密图像进行加密，再将加密后的秘密图像嵌入载体图像。但与上述 2 种算法不同的是，本文的特征提取方式可以分别从载体图像和含密图像中提取出相同的特征，再将特征映射为密钥，由于不同的载体图像提取的特征不一样，因此通过更换载体图像就可以完成密钥的更新，而不需要再单独传递密钥，当特征映射到密钥的映射方式使用一段时间后，传递新的映射方式，就可以完成对密钥整体的更新。文献[11-12]对比方案中都必须通过传递新的密钥信息才能完成密钥更新，本文只需使用不同的载体图像就可以做到，减少了密钥传递的次数。

3　本文方案

为了解决基于深度学习的高容量图像隐写算法存在的载体图像和含密图像的残差图像会暴露秘密图像的问题，本文提出了一种结合图像加密与深度学习的高容量图像隐写算法，该算法可以将一张

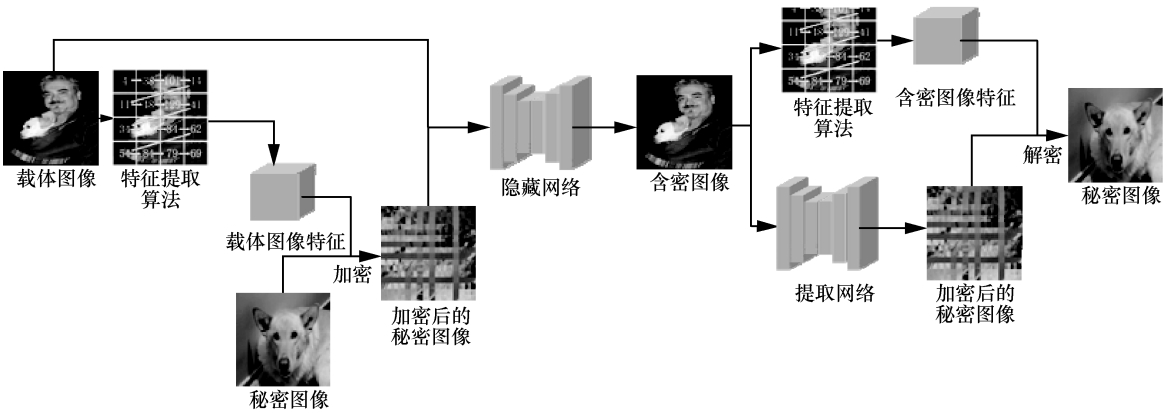


图 1　本文算法的总体结构

彩色图像隐藏在另一张彩色图像中，且在视觉上无法发现载体图像的修改，即使攻击者得到了载体图像，也无法读取秘密图像的语义信息。本文算法的总体结构如图 1 所示，发送方使用特征提取算法从载体图像中提取载体图像特征作为加密密钥对秘密图像进行加密，隐藏网络将加密后的秘密图像隐藏进载体图像中。接收方使用提取网络从含密图像中提取加密后的秘密图像，使用特征提取算法从含密图像中提取含密图像特征作为解密密钥对提取的图像进行解密，得到秘密图像。特征提取算法能够从载体图像和含密图像中得到相同的特征信息，所以分别将载体图像特征与含密图像特征作为加密密钥、解密密钥能够保证秘密图像被准确恢复。

本文算法共分为 4 个部分：特征提取部分、图像加密部分、隐藏部分、提取部分。

3.1　特征提取部分

目前特征提取使用的方法主要有基于深度学习[19-20]的方法和传统方法[21]，基于深度学习的方法需要大量的数据训练，且难以保证载体图像和含密图像特征提取的一致性与多样性。传统方法在无载体信息隐藏中的使用较广泛，比如通过特征提取算法提取图像特征并将特征映射为秘密信息，进而构建图像和秘密信息对应的关系库。

本文需要的特征提取算法要从载体图像与含密图像中提取出相同特征，而含密图像可以看作是对载体图像修改后的结果，所以本文选择了具有一定稳健性的特征提取算法，也就是文献[21]中的特征提取算法，其特征提取过程如下。

1) 图像灰度化。将彩色图像变为灰度图像。

2) 图像分块。本文将 256×256 尺寸大小的图像分为 16 块，每个块图像大小为 64×64。

3) 计算每一个图像块的像素平均值。

4) 按从左到右、自上而下的顺序对像素均值进行比较，其顺序如图 2 所示，如果第一个图像块的像素平均值小于或等于第二个，则映射为 0，如果第一个图像块的像素平均值大于第二个，则映射为 1，剩下的图像块依次类推。

通过对 16 个图像块的均值进行比较，可以得到一个 15 位的 0/1 比特串，在此比特串后添加 1，即可得到一个 16 位的 0/1 比特串，如图 2 所示，可以将该图像映射为 0011001101110111。

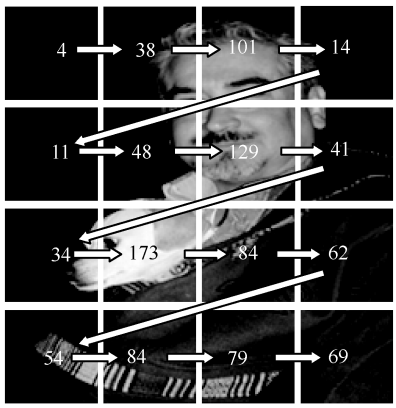


图 2　特征提取方式

该算法能够较好地应对图像修改造成的损失，即使图像块内的像素值遭到一定修改，也能够保持特征的稳定性。

3.2　图像加密部分

目前基于深度学习的高容量图像写算法中，使用的图像加密方式有 2 种类型，一种是将图像分成多个图像块，而后将图像块位置打乱，从而达到无法读取图像语义信息的效果；另一种是将图像进行 DCT，再使用椭圆曲线加密算法进行加密。

本文的图像加密方式是在第一种加密方式的基础上改进而来的。

本文加密过程如下。

1) 特征提取。利用 3.1 节的特征提取方法，提取载体图像的特征作为密钥。

2) 图像分块。将 256×256 尺寸大小的彩色图像分为 256 块，每个块图像尺寸大小为 16×16。

3) 制定置乱方法。将从载体图像中提取的 16 位特征比特串按从左到右的顺序编定位置，比特串第 1 位编定为 1，第 2 位编定为 2，依次类推。按从左到右的顺序，如果比特串第 2 位数值为 0，则将 2 放置在当前位置信息的最左边，如果数值为 1，则将其放置在当前位置信息的最右边，依次类推，不断更新位置信息，最终将 16 位的比特串转换为位置信息。其具体流程如图 3 所示。

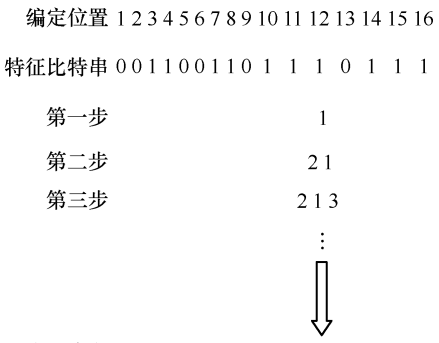


图 3　比特串与位置信息转换流程

4) 分块置乱。按从左到右、自上而下的顺序，以每 16 个图像块为一组，共分为 16 组，使用 3)中的置乱方法，对每组内的 16 个图像块进行位置置乱。

5) 总体置乱。对 16 组图像块使用 3)中的置乱方法进行置乱。

解密过程如下。

1) 特征提取。利用 3.1 节的特征提取方法，从含密图像中提取含密图像特征，作为解密过程中的密钥。由于特征提取算法可以从载体图像和含密图像中提取相同的密钥，因此该密钥与加密过程中 1)的密钥是一致的。

2) 图像分块。将 256×256 尺寸大小的彩色图像分为 256 块，每个块图像尺寸大小为 16×16。

3) 总体解密。加密过程中 5)的逆过程。

4) 分块解密。加密过程中 4)的逆过程。

相比前文提出的 2 种加密方法[11-12]，本文加密方法利用了载体图像的特征进行加密，攻击者如果想还原秘密图像需要进行 256!次运算。同时，本文算法通过更换载体图像，即可完成密钥的更换。图 4 和图 5 分别为分块置乱和总体置乱的结果。

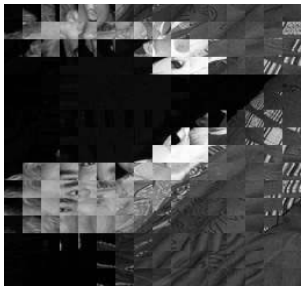


图 4　分块置乱结果

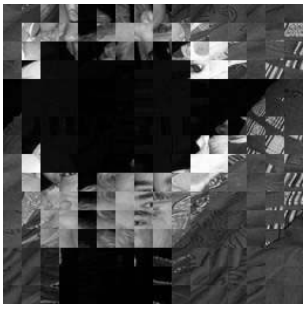


图 5　总体置乱结果

3.3　隐藏部分

隐藏部分网络类似于 U-net 网络。本文使用的隐藏网络是一个输入通道为 6、输出通道为 3 的 U 型网络，它的具体结构如表 1 所示。

表 1　　隐藏网络结构

层结构	输入尺寸	输出尺寸
4×4×64 卷积+LeakyReLU	256×256×6	128×128×64
4×4×128 卷积+BN+LeakyReLU	128×128×64	64×64×128
4×4×256 卷积+BN+LeakyReLU	64×64×128	32×32×256
4×4×512 卷积+BN+LeakyReLU	32×32×256	16×16×512
4×4×512 卷积+BN+LeakyReLU	16×16×512	8×8×512
4×4×512 卷积+BN+LeakyReLU	8×8×512	4×4×512
4×4×512 卷积+ReLU	4×4×512	2×2×512
4×4×512 反卷积+BN+ReLU	2×2×512	4×4×512
4×4×512 反卷积+BN+ReLU	4×4×1024	8×8×512
4×4×512 反卷积+BN+ReLU	8×8×1024	16×16×512
4×4×256 反卷积+BN+ReLU	16×16×1024	32×32×256
4×4×128 反卷积+BN+ReLU	32×32×512	64×64×128
4×4×64 反卷积+BN+ReLU	64×64×256	128×128×64
4×4×3 反卷积+ Sigmoid	128×128×128	256×256×3

隐藏网络中，除了第一层和第七层没有使用批量标准化（BN, batch normalization）层外，其余层都使用了 BN 层，因为第一层要尽可能保留图像的原始信息，减少特征信息的损失，第七层的特征尺寸为 2×2，此时已经提取了最终特征，也不加入 BN 层，保留最终的特征提取准确率。每层网络中，不论是卷积操作还是反卷积操作，它们的卷积核大小都为 4×4，stride 为 2，padding 为 1，与文献[11,16]在卷积操作后先使用激活函数再使用 BN 不同，本文先使用 BN，再使用激活函数，减少激活函数处理后的某些神经元失活对 BN 操作产生的影响。

该隐藏网络是一个编码与解码网络，它首先对输入的两张图像进行编码，转换成一个较深层的特征，再对深层特征进行解码，生成一个与载体图像相似的含密图像。

隐藏部分可表示为

$$I_{\text{ca}} = S\left(E(I_s, f_c), I_c\right) \tag{1}$$

发送方使用特征提取算法提取载体图像 I_c 的特征值 f_c，使用特征值 f_c 作为密钥，使用加密算法 $E(x,y)$ 对秘密图像 I_s 加密，将加密后的秘密图像 I_s^E 与载体图像 I_c 同时输入隐藏网络 $S(x,y)$ 中，得到含密图像 I_{ca}，将含密图像 I_{ca} 发送给接收方。

3.4 提取部分

提取部分使用的网络是一个多层卷积网络，网络结构如表 2 所示，它的输入通道和输出通道都为 3。

表 2 提取网络结构

层结构	输入尺寸	输出尺寸
3×3×64 卷积+BN+ReLU	256×256×3	256×256×64
3×3×128 卷积+BN+ReLU	256×256×64	256×256×128
3×3×256 卷积+BN+ReLU	256×256×128	256×256×256
3×3×128 卷积+BN+ReLU	256×256×256	256×256×128
3×3×64 卷积+BN+ReLU	256×256×128	256×256×64
3×3×3 卷积+sigmoid	256×256×64	256×256×3

提取网络中，除了最后一层，每一层都使用了 BN 和激活函数，加快网络训练速度，每一个卷积操作使用的卷积核大小都为 3×3，stride 为 1，padding 为 1，始终保持提取网络的输入特征尺寸和输出特征尺寸不变。

相较于隐藏网络，提取网络层数较少，不同于隐藏网络将两张图像映射为一张图像，提取网络只需将一张图像映射为一张图像，所以与隐藏网络相比，结构较简单。

提取部分可表示为

$$\begin{cases} I_s^E = R\left(I_{\text{ca}}\right) \\ I_s = D\left(f_{\text{ca}}, I_s^E\right) \end{cases} \tag{2}$$

接收方接收到含密图像 I_{ca} 后，先使用特征提取算法提取含密图像 I_{ca} 的特征值 f_{ca}，再使用解密算法 $D(x,y)$ 和 f_{ca} 对提取网络 $R(x)$ 提取的加密后的秘密图像 I_s^E 进行解密，得到 I_s。由于本文特征提取算法的特点，即图像的一定修改不影响特征的提取，因此 $f_{\text{ca}} = f_c$。

3.5 损失函数

本文的损失函数使用了均方误差（MSE, mean square error）度量损失。

$$\text{MSE}(\hat{y}_i, y_i) = \frac{1}{M}\sum_{m=1}^{M}(y_i - \hat{y}_i)^2 \tag{3}$$

损失函数包括 2 个部分。一是隐藏网络中，载体图像和含密图像的差异以及秘密图像和提取图像之间的差异的加权和 Loss_1。

$$\text{Loss}_1 = \text{MSE}(I_c, I_{\text{ca}}) + \omega\text{MSE}(I_s^E, R(I_{\text{ca}})) \tag{4}$$

隐藏网络在隐藏加密图像过程中，不仅要考虑隐藏图像后的载体图像损失，还要考虑提取加密后的秘密图像时的准确率，所以使用参数 ω 权衡隐藏效果和提取效果。

二是提取网络中秘密图像和提取图像之间的差异 Loss_2。

$$\text{Loss}_2 = \text{MSE}(I_s^E, R(I_{\text{ca}})) \tag{5}$$

提取网络仅需准确地提取出加密后的秘密图像，因此其损失函数仅包含加密后的秘密图像 I_s^E 与提取的加密后的秘密图像 $R(I_{\text{ca}})$ 的损失。

4 实验分析

本文在训练过程中使用的图像数据集为 ImageNet，使用 45 000 张图像训练，5 000 张图像测试，图像尺寸在训练前均被调整为 256×256，网络初始学习率为 0.001，并使用 Adam 算法对学习率进行调整，ω 设置为 0.75，batchsize 设置为 32，迭代次数为 100。GPU 为 NVIDIA RTX2080Ti，使用的深度学习框架为 Pytorch，使用的编程语言为 Python3.8。为了测试训练集在其他图像数据集的适配情况，本文

在 VOC2007 数据集上对方案进行了实验。

4.1　特征提取准确率

由于特征中的一个比特错误会对之后的位置映射产生严重的影响，造成图像解密错误，因此，本文使用的加密算法对特征提取的准确率要求较高。

为了对比基于深度学习的特征提取方法与本文特征提取方法的提取准确程度，本文分别对 ResNet50[22]、VggNet11[23]以及本文的特征提取算法的准确率进行了实验。在特征提取过程中仅使用 ResNet50 以及 VggNet11 的特征提取部分，舍弃其用于图像分类的全连接网络，并在网络最后加入新的全连接层，用于将特征调整至特定长度，并通过训练新加入的全连接层，使特征提取网络从载体图像和含密图像中提取的特征尽量一致。共训练 10 轮，使用 500 组载体图像、含密图像对作为训练集。其训练过程中的损失函数如下

$$\mathrm{Loss}_V = \mathrm{MSE}(V(I_c), V(I_{\mathrm{ca}})) \tag{6}$$

$$\mathrm{Loss}_R = \mathrm{MSE}(R_e(I_c), R_e(I_{\mathrm{ca}})) \tag{7}$$

其中，Loss_V 与 Loss_R 分别代表 VggNet11 与 ResNet50 训练过程中的损失函数，$V(x)$ 与 $R_e(x)$ 分别代表 VggNet11 与 ResNet50 的特征提取部分与新加入的全连接层构成的网络。

本文在提取特征数量为 16 bit、64 bit 以及 256 bit 3 种情况下对 3 个特征提取算法进行了 500 组实验，实验中，对特征提取准确率的计算如式(8)所示。

$$A_c = \frac{n_e}{N} \tag{8}$$

其中，A_c 是提取准确率，n_e 是从载体图像中提取的特征与从含密图像中提取的特征完全相同的组数，N 是实验总组数，即 500。

图 6 展示了 3 种算法在提取不同长度特征时的提取准确率对比。通过实验可以观察出，在提取特征长度相同时，本文算法能够更准确地提取特征，在特征数量为 16 bit、64 bit 和 256 bit 时，准确率能达到 0.964、0.832 和 0.422，与 ResNet 以及 VggNet 相比，准确率有一定提升。

本文算法的核心要求是载体图像特征与含密图像特征的一致。仅在 ResNet50 与 VggNet11 的特征提取网络后加入一个全连接层并训练全连接层，难以保证载体图像特征与含密图像特征的一致性。如果加入较多的全连接层，则会导致不同的输入被网络映射为相同的特征，影响特征的多样性。

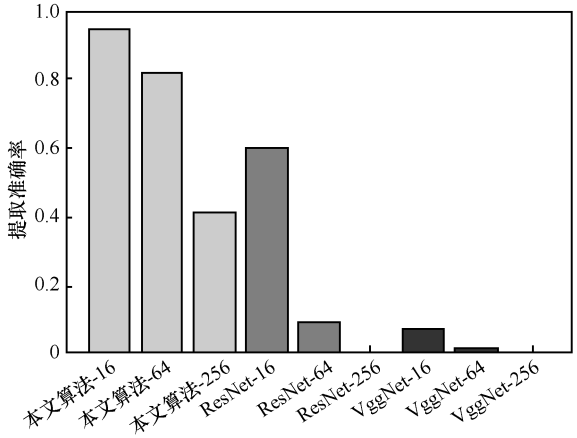


图 6　3 种算法在提取不同长度特征时的提取准确率对比

本文要求的特征提取需要有较高的准确率，因为提取过程中某一位的差错都会对接下来的秘密图像恢复产生影响，所以设计使用了传统的较稳定的特征提取算法。

4.2　安全性分析

4.2.1　加密方法对比

图 7 为本文加密算法与文献[12]的加密算法的结果对比。从图 7 中可以观察到，本文加密算法的图像块置乱效果更好。在加密中，文献[12]加密后的图像块的目标分布没有较大的改变，单独站立的男人主要部分仍在图像的右侧，而本文算法加密后的图像已无法读取具体的语义信息。

(a) 未加密图像　(b) 本文算法加密结果　(c) 文献[12]加密结果

图 7　加密结果图像对比

这主要是因为本文加密算法会给予每一个图像块一个具体的位置，而不是仅对列和行的位置进行调换。而文献[12]中的加密算法加密后，大部分原来相邻的列或行转换后仅隔一行或一列，比如在列变换中原来的第 1 列被移动到第 2 列，原来的第 2 列被移动到第 4 列，原来的第 3 列被移动到第 6 列，它们之间仅隔一列，而在下一步的行变换时，同一行内的这些情况没有得到改变。所以加密后仍然能从加密的结果中发现一些秘密图像的语义信息。同时，文献[12]也对加密的位移参数有一定的要求。比如说位移参数为 3，那么逐行或逐列的移动位置分别为 1，2，3，1，2，3…。当位移参数为

偶数 4 时，每一列或每一行都会移动到偶数列或行，就造成了奇数列或行的位置信息并没有被改变。本文以图像大小为 256×256，像素块大小为 16×16，位移参数为 3 和 4 为例进行比较，表 3 和表 4 为行或列的位置变化。

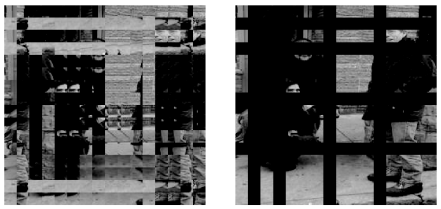
表 3　位移参数为 3 时的位置变化

移动前位置	移动后位置
1	2
4	5
7	8
10	11
13	14
16	16
2	4
5	7
8	10
11	13
14	1
3	6
6	9
9	12
12	15
15	3

表 4　位移参数为 4 时的位置变化

移动前	移动后
1	2
5	6
9	10
13	14
2	4
6	8
10	12
14	16
3	6
7	10
11	14
15	2
4	8
8	12
12	16
16	4

从表 3 与表 4 可以观察到，当位移参数为偶数时，所有的位置变换都只会变换到偶数行或列，而不会变换到奇数行或列。图 8(a)展示了位移参数为偶数时的加密效果，由于奇数行或列没有被改变，因此可以从图像中看到部分行或列的重复。同时，在移动过程中，移动顺序靠后的行或列会将移动顺序靠前的列或行替换掉，造成信息的丢失。如图 8(b)所示，图像中的黑色部分代表了信息的丢失。



(a) 加密后结果　(b) 缺失图

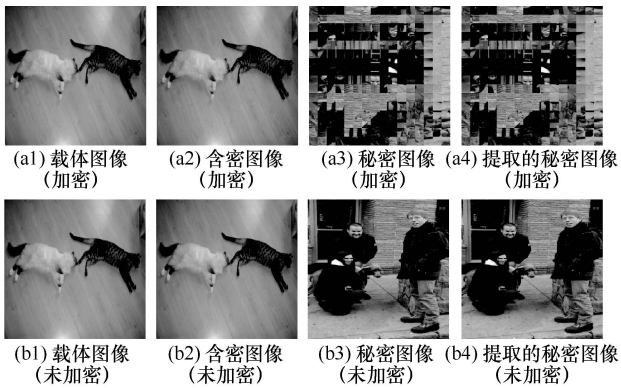
图 8　位移参数为偶数时的影响

4.2.2　穷举所需次数

文献[12]将秘密图像分为 196 块，通过将图像块的行、列移动一定距离实现加密，如果想要穷举出其结果，需要 $196!\approx5.08\times10^{365}$ 次计算。本文将图像块分为 256 块，而后根据载体图像的特征对图像块位置进行交换，进而完成加密。相比而言，本文的组合方式有 $256!\approx8.58\times10^{506}$ 种，更加难以穷举出正确的排列。

4.2.3　秘密图像不可见性分析

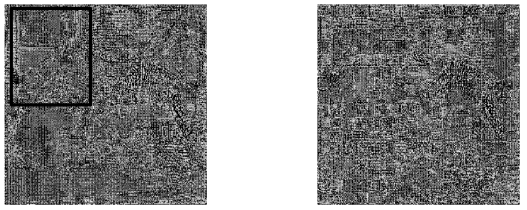
本节基于图 9 中的载体图像和秘密图像对秘密图像的不可见性进行分析。图 9(a1)～图 9(a4)为对秘密图像进行加密后再嵌入情况下的载体图像、含密图像、秘密图像与提取的秘密图像。图 9(b1)～图 9(b4)为未对秘密图像进行加密，直接进行嵌入的的载体图像、含密图像、秘密图像与提取的秘密图像。图 10 为载体图像与含密图像的残差图像数值扩大 5 倍的结果。



(a1) 载体图像（加密）　(a2) 含密图像（加密）　(a3) 秘密图像（加密）　(a4) 提取的秘密图像（加密）

(b1) 载体图像（未加密）　(b2) 含密图像（未加密）　(b3) 秘密图像（未加密）　(b4) 提取的秘密图像（未加密）

图 9　加密与未加密效果对比



(a) 无加密时的残差图（像素值×5）　(b) 加密后的残差图（像素值×5）

图 10　残差图像

从图 9 中可以观察到，图像加密并没有对含密图像产生较大的损失，人眼还是无法分辨载体图像与含密图像的区别。观察图 10 中的残差图像，可以发现，没有加密时，能够从载体图像与含密图像的残差部分观察到秘密图像的部分语义信息。从图 10(a)中可以看出，秘密图像中的透明门的轮廓。在这种情况下，如果攻击者可以同时得到载体图像与含密图像，就可能会造成秘密图像部分语义信息的泄露，无法保证秘密信息的安全性。而在加密后，已经无法观察到秘密图像的语义信息。

4.2.4　密钥更新频率对比

增加密钥更新频率也是提升安全性的有效方法之一，但是过于频繁地传递密钥也增加了隐写被发现的可能性。相比同为使用了图像加密的高容量图像隐写算法，本文的密钥产生依赖于载体图像，通过更换载体图像，就可以完成密钥的更新，如此一来，本文算法在保证密钥更新频率的同时，减少了密钥传递的次数，降低了隐写通信被发现的概率。

以特征长度为 16 bit、载体图像数量为 10 张为例，假设 10 张载体图像具有不同的特征信息，那么 10 张载体图像就分别代表了 10 个密钥，文献[11-12]的方案通过传递新的密钥进行一次密钥更新，本文算法则可通过改变载体图像进行一次密钥更新，当两者都进行了 10 次密钥更新时，文献[11-12]的算法需要传递 10 次密钥，而本文算法仅更改了 10 张载体图像，不需要向接收方发送密钥信息，图 11 展示了不同算法在更新 10 次密钥时需要传递密钥的次数对比。当本文算法的提取特征长度为 16 bit 时，最大密钥数量为 2^{16} 个；当特征长度为 64 bit 时，最大密钥数量为 2^{64} 个。

4.2.5　抗隐写分析对比

为了便于对比，本文同样使用开源的隐写分析工具 StegExpose[25]，对 250 张载体图像以及本方案产生的 250 张含密图像进行隐写分析。250 张载体图像中，有一张被错误判别为含密图像；250 张含密图像中，有 4 组被正确判别为含密图像，隐写分析准确率为 0.506，只略高于随机猜测准确率。其接收者操作特性（ROC, receiver operating characteristic）曲线如图 12(a)所示，图 12(b)、图 12(c)的 ROC 曲线分别是同为高容量隐写算法的文献[11,24]在 StegExpose 分析下的结果，其 threshold 参数设置与本文一致，为 0.2。通过观察 ROC 曲线，本文算法在面对 StegExpose 隐写分析工具分析时，抵抗隐写分析能力与文献[11,24]相当。

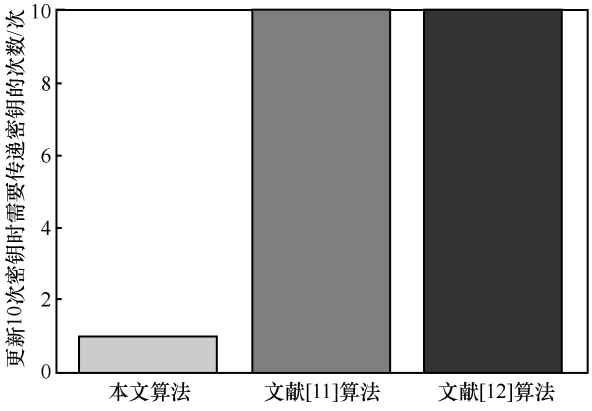


图 11　不同算法更新 10 次密钥时需要传递密钥的次数对比

4.3　隐写容量

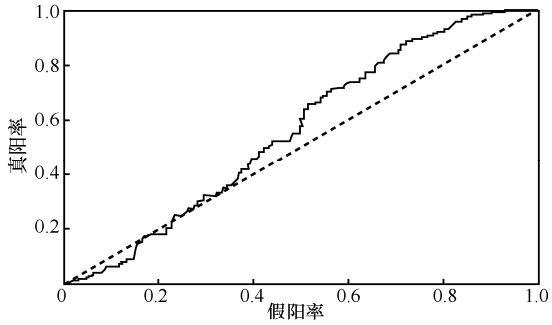
本文算法具有较高的隐写容量，能够在一张彩色图像中隐藏另一张彩色图像，丰富的色彩信息能够帮助接收方从秘密图像中获得更多的有用信息，方案[13,14,26]则是在一张彩色图像中隐藏一张与其尺寸相同的灰度图像。与方案[13,14,26]相比，本文算法的隐写容量是其 3 倍，表 5 是本文隐写容量与其他算法的隐写容量对比。

4.4　图像损失

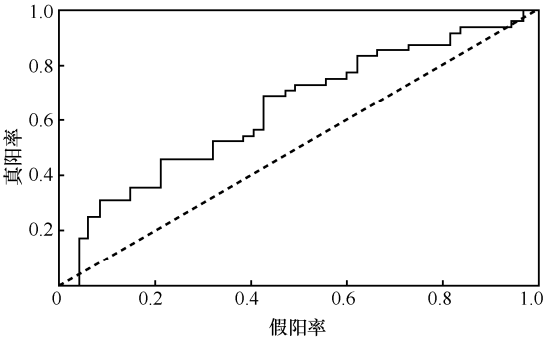
本文使用峰值信噪比（PSNR, peak signal to noise ratio）和结构相似性（SSIM, structural similarity）来评价载体图像和含密图像之间的损失，由于提取效果的好坏主要使用提取准确率来评估，这里不对秘密图像和提取的秘密图像之间的损失值做对比，仅对比载体图像与含密图像之间的损失，具体 PSNR 和 SSIM 值如图 13 所示，本文算法计算了 500 对秘密图像与含密图像对的 PSNR 和 SSIM 的平均值。

从图 13 数据可以观察到，相比同为高容量隐写的算法，本文算法能够保持较好的 PSNR 和 SSIM 值，载体图像到含密图像变换中的损失较少。本文算法在进行信息嵌入的同时，对图像的统计特征影响较小。图 14 为 3 对载体图像与含密图像的图像频率分布直

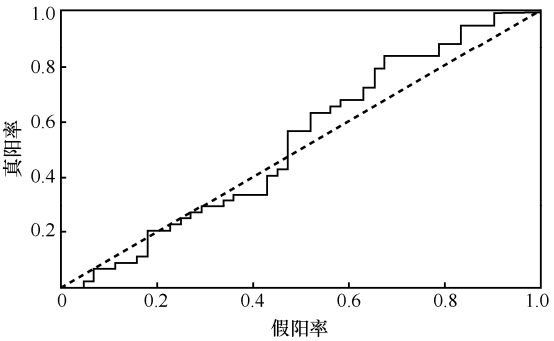
方图进行对比，其特征分布基本相同，通过观察图像直方图可以发现，本文算法是在保留了载体图像大部分统计特性的基础上进行的图像嵌入，对载体图像造成的损失较低，对于人眼而言更难以发现。



(a) StegExpose对本文方案分析的ROC曲线



(b) StegExpose对文献[11]方案分析的ROC曲线



(c) StegExpose对文献[24]方案分析的ROC曲线

图 12　本文算法与其他算法的 ROC 曲线

表 5　　不同算法隐写容量对比结果

算法	秘密图像尺寸	载体图像尺寸	容量/bpp
HUGO[1]	32×32	32×32	0.1~0.4
S-UNIWARD[3]	64×64	64×64	0.4
SSGAN[27]	64×64	204×204	0.4
ISGAN[14]	256×256	256×256	8
Rahim[13]	300×300	300×300	8
DGANS[26]	256×256	256×256	8
Duan[11]	256×256	256×256	24
Sharma[12]	256×256	256×256	24
本文算法	256×256	256×256	24

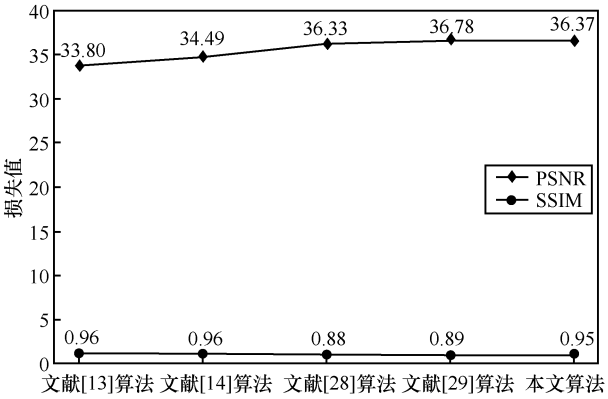
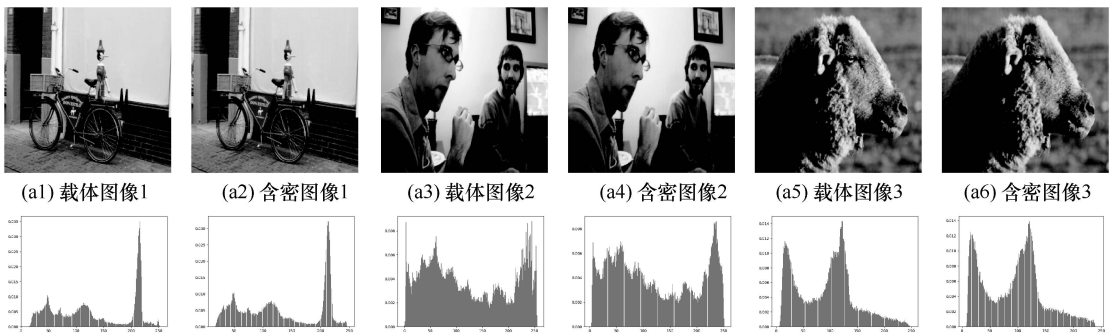


图 13　载体图像损失对比

5　结束语

本文算法可以有效地将一张彩色图像隐藏在另一张彩色图像中，且载体图像与含密图像的差异难以发现。本文算法在基于图像加密和深度学习的高容量图像隐写算法的基础上，通过设计使用特征提取算法，将载体图像特征作为密钥，通过更换载体图像即可实现对密钥的更换，能够在不减少密钥更新次数的同时，减少密钥传递的次数；通过将图



(a1) 载体图像1　(a2) 含密图像1　(a3) 载体图像2　(a4) 含密图像2　(a5) 载体图像3　(a6) 含密图像3

(b1) 载体图像1直方图 (b2) 含密图像1直方图(b3) 载体图像2直方图(b4) 含密图像2直方图(b5) 载体图像3直方图(b6) 含密图像3直方图

图 14　载体图像与含密图像的图像频率分布直方图

像块进行更细粒度的划分、加密，改善了载体图像与含密图像的残差图像暴露秘密图像的问题。实验结果表明，本文算法可以有效提高基于深度学习的高容量图像隐写算法的安全性。

参考文献：

[1] PEVNÝ T, FILLER T, BAS P. Using high-dimensional image models to perform highly undetectable steganography[C]//Information Hiding. Berlin: Springer, 2010: 161-177.

[2] HOLUB V, FRIDRICH J. Designing steganographic distortion using directional filters[C]//Proceedings of 2012 IEEE International Workshop onInformation Forensics and Security (WIFS). Piscataway: IEEE Press, 2012: 234-239.

[3] HOLUB V, FRIDRICH J, DENEMARK T. Universal distortion function for steganography in an arbitrary domain[J]. EURASIP Journal on Information Security, 2014(1): 1-13.

[4] HICSONMEZ S, SAMET N, AKBAS E, et al. GANILLA: generative adversarial networks for image to illustration translation[J]. Image and Vision Computing, 2020, 95: 103886.

[5] HE K, GKIOXARI G, DOLLÁR P, et al. Mask R-CNN[C]//Proceedings of 2017 IEEE International Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2017: 2980-2988.

[6] SZEGEDY C, IOFFE S, VANHOUCKE V, et al. Inception-v4, inception-ResNet and the impact of residual connections on learning[J]. arXiv Preprint, arXiv: 1602.07261, 2016.

[7] CHOI Y, CHOI M, KIM M, et al. StarGAN: unified generative adversarial networks for multi-domain image-to-image translation[C]//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 8789-8797.

[8] GOODFELLOW I J, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial nets[C]//Proceedings of the 27th International Conference on Neural Information Processing Systems. Cambridge: MIT Press, 2014: 2672-2680.

[9] TANG W X, TAN S Q, LI B, et al. Automatic steganographic distortion learning using a generative adversarial network[J]. IEEE Signal Processing Letters, 2017, 24(10): 1547-1551.

[10] BALUJA S. Hiding images in plain sight: deep steganography[C]//Advances in Neural Information Processing Systems 30 (NIPS 2017). [S.n.:s.l.], 2017: 2069-2079.

[11] DUAN X T, GUO D D, LIU N, et al. A new high capacity image steganography method combined with image elliptic curve cryptography and deep neural network[J]. IEEE Access, 2020, 8: 25777-25788.

[12] SHARMA K, AGGARWAL A, SINGHANIA T, et al. Hiding data in images using cryptography and deep neural network[J]. Journal of Artificial Intelligence and Systems, 2019, 1(1): 143-162.

[13] RAHIM R, NADEEM S. End-to-end trained CNN encoder-decoder networks for image steganography[C]//Proceedings of the European Conference on Computer Vision (ECCV). [S.n.:s.l.], 2018: 723-729.

[14] ZHANG R, DONG S Q, LIU J Y. Invisible steganography via generative adversarial networks[J]. Multimedia Tools and Applications, 2019, 78(7): 8559-8575.

[15] SZEGEDY C, VANHOUCKE V, IOFFE S, et al. Rethinking the inception architecture for computer vision[C]//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2016: 2818-2826.

[16] DUAN X T, JIA K, LI B X, et al. Reversible image steganography scheme based on a U-net structure[J]. IEEE Access, 2019, 7: 9314-9323.

[17] RONNEBERGER O, FISCHER P, BROX T. U-net: convolutional networks for biomedical image segmentation[C]//International Conference on Medical Image Computing and Computer-Assisted Intervention. Berlin: Springer, 2015: 234-241.

[18] BADRINARAYANAN V, KENDALL A, CIPOLLA R. SegNet: a deep convolutional encoder-decoder architecture for image segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(12): 2481-2495.

[19] LIU Q, XIANG X Y, QIN J H, et al. Coverless image steganography based on DenseNet feature mapping[J]. EURASIP Journal on Image and Video Processing, 2020, 2020(1): 39.

[20] ZHOU Z L, CAO Y, WANG M M, et al. Faster-RCNN based robust coverless information hiding system in cloud environment[J]. IEEE Access, 2019, 7: 179891-179897.

[21] ZHOU Z, SUN H, HARIT R, et al. Coverless image steganography without embedding[C]//International Conference on Cloud Computing and Security. Berlin: Springer, 2015: 123-132.

[22] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2016: 770-778.

[23] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[J]. arXiv Preprint, arXiv: 1409.1556, 2015.

[24] DUAN X T, NAO L, GOU M X, et al. High-capacity image steganography based on improved FC-DenseNet[J]. IEEE Access, 2020, 8: 170174-170182.

[25] BOEHM B. StegExpose - a tool for detecting LSB steganography[J]. arXiv Preprint, arXiv: 1410.6656, 2014.

[26] 竺乐庆, 郭钰, 莫凌强, 等. DGANS: 基于双重生成式对抗网络的稳健图像隐写模型[J]. 通信学报, 2020, 41(1): 125-133.
ZHU L Q, GUO Y, MO L Q, et al. DGANS: robustness image steganography model based on double GAN[J]. Journal on Communications, 2020, 41(1): 125-133.

[27] SHI H C, DONG J, WANG W, et al. SSGAN: secure steganography based on generative adversarial networks[C]//Advances in Multimedia Information Processing – PCM 2017. Berlin: Springer, 2018: 534-544.

[28] ZHANG K A, CUESTA-INFANTE A, XU L, et al. SteganoGAN: high capacity image steganography with GANs[J]. arXiv Preprint, arXiv: 1901.03892, 2019.

[29] QIN J H, WANG J, TAN Y, et al. Coverless image steganography based on generative adversarial network[J]. Mathematics, 2020, 8(9): 1394.

[作者简介]

杨晓元（1959− ），男，湖南湘潭人，博士，武警工程大学教授、博士生导师，主要研究方向为密码学、信息隐藏等。

毕新亮（1997− ），男，安徽合肥人，武警工程大学硕士生，主要研究方向为深度学习、图像隐写等。

刘佳（1982− ），男，河南汝州人，博士，武警工程大学副教授、硕士生导师，主要研究方向为信息隐藏、图像隐写、机器学习等。

黄思远（1997− ），男，陕西西安人，武警工程大学硕士生，主要研究方向为信息隐藏等。