DGANS：基于双重生成式对抗网络的稳健图像隐写模型

竺乐庆[1]，郭钰[1]，莫凌强[1]，张大兴[2]
（1. 浙江工商大学计算机与信息工程学院，浙江 杭州 310018；2. 杭州电子科技大学计算机学院，浙江 杭州 310018）

摘 要：深度卷积神经网络可有效地应用于大容量图像信息隐写，然而其稳健性研究却鲜有报道。双重生成式对抗网络（DGANS）模型对深度学习框架应用于图像隐写时，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。DGANS 由 2 个串联的生成式对抗网络构成，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。通过对生成的含密图像进行数据增强并进一步强化训练提取网络，使提取网络对输入图像的几何变换具有适应性。实验结果表明，DGAN 不仅可以实现高容量的图像信息隐写，而且可以对抗一定范围内的几何攻击，比同类模型有更好的稳健性。

关键词：图像隐写；稳健性；双重生成式对抗网络；深度学习
中图分类号：TP391
文献标识码：A
doi: 10.11959/j.issn.1000−436x.2020019

DGANS: robustness image steganography model based on double GAN

ZHU Leqing[1], GUO Yu[1], MO Lingqiang[1], ZHANG Daxing[2]
1. School of Computer and Information Engineering, Zhejiang Gongshang University, Hangzhou 310018, China
2. School of Computer Science and Technology, Hangzhou Dianzi University, Hangzhou 310018, China

Abstract: Deep convolutional neural networks can be effectively applied to large-capacity image steganography, but the research on their robustness is rarely reported. The DGANS (double-GAN-based steganography) applies the deep learning framework in image steganography, which is optimized to resist small geometric distortions so as to improve the model's robustness. DGANS is made up of two consecutive generative adversarial networks that can hide a grayscale image into another color or grayscale image of the same size and can restore it later. The generated stego-images are augmented and used to further train and strengthen the reveal network so as to make it adaptive to small geometric distortion of input images. Experimental results suggest that DGANS can not only realize high-capacity image steganography, but also can resist geometric attacks within certain range, which demonstrates better robustness than similar models.

Key words: image steganography, robustness, double GAN, deep learning

1 引言

图像信息隐藏利用图像数据的统计冗余和人类感知冗余，将有意义的秘密信息隐藏到图像中，且非授权者无法确认该载体中是否隐藏了信息，达到隐蔽通信、版权保护等目的[1]。从最初的最低有效位方法，到基于离散傅里叶变换、离散余弦变换[2]、离散小波变换[3]等变换域方法，以及提高安全性的高度不可检测隐写（HUGO, highly undetectable stego）算法[4]、空域通用小波相对失真（S-UNIWARD, spatial universal wavelet relative distortion）方法[5]、小波获得权重（WOW, wavelet

收稿日期：2019–08–16；**修回日期**：2019–12–05
基金项目：国家自然科学基金资助项目（No.61572160）；浙江省自然科学基金资助项目（No.LY20F020002）
Foundation Items: The National Natural Science Foundation of China (No.61572160), The Natural Science Foundation of Zhejiang Province (No.LY20F020002)

obtained weight）方法[6]等内容自适应隐写术，图像信息隐藏技术呈现多元化发展。这些传统隐写方法虽然在透明性、安全性方面已逐渐完善，但是在隐写容量及稳健性方面仍存在提升空间。近年来，随着深度学习技术的发展和推广，深度学习框架同样也被引入图像隐写术中。Shi 等[7]基于生成式对抗网络（GAN, generative adversarial network）[8]，结合高斯–神经元卷积神经网络提出了一种名为安全隐写 GAN（SSGAN, secure steganography based on GAN）的模型用于隐写术，SSGAN 生成的图像用 HUGO 隐写后更难于检测。Hayes 等[9]直接用 GAN 嵌入隐写信息并提取，可以在 32×32 大小的图像中隐写 100~400 位二进制位，取得了优于 HUGO、WOW 和 S-UNIWARD 的性能。Rehman 等[10]采用编码器–解码器结构深度学习框架在彩色图像中隐写灰度图像并提取，然而含密图像在色彩上有失真。Baluja[11]提出的深度隐写框架包括准备网络、隐藏网络、显现网络三部分，可以在彩色图像中隐藏小于等于原图的彩色图像。Chu 等[12]探索了使用 CycleGAN 在图像中隐藏信息并还原信息的可能性。Tang 等[13]提出了自动隐写失真学习框架，GAN 的产生器用于寻找图像中适合嵌入或隐藏信息的像素，区分器则训练为隐写分析器。Zhang 等[14]提出的不可见隐写 GAN（ISGAN, invisible steganography via GAN）可以在发送端隐藏灰度图像到彩色图像中，在接收端提取出所隐藏的灰度图，使用 GAN 提高隐写安全性和隐蔽性。Wu 等[15]提出的 StegNet 采用可分离卷积残差块，能在 64×64 大小的彩色图像中隐藏另一彩色图像，然而含密图像有明显色彩失真，隐蔽性不够理想。Duan 等[16]用 UNet 在彩图中隐藏彩图，最终的性能要优于前几种方法。上述框架大大提高了图像隐写的容量，但均未对隐写模型的稳健性进行测试和评估。本文提出的 DGANS 模型对基于深度学习的图像隐写模型的稳健性进行研究，主要贡献如下。

1) 在编码解码网络结构中，采用双重 GAN，第一个 GAN 用生成器生成含密图像，鉴别器作为隐写分析网络来增强隐写术的安全性；第二个 GAN 用生成器提取秘密图像，鉴别器用来增强网络的稳健性，使整体网络同时拥有较高安全性和稳健性。

2) 对训练集生成的含密图像集进行几何变换（平移、旋转、缩放）增强，用增强的数据集对第二个 GAN 进行单独强化训练，使该网络对上述变换具有适应性，进一步增强模型的稳健性。

2 DGANS 隐写模型

2.1 DGANS 模型总体结构

本文提出的 DGANS 隐写模型总体结构如图 1 所示，灰度秘密图像通过 DGANS 模型隐藏到彩色封面图中并能从中还原。为了不破坏原始封面的颜色信息，图像隐写在 YUV 颜色空间的 Y 通道。网络的基础模型的组成包括隐写网络 GAN_1 和提取网络 GAN_2。GAN_1 生成器输入为封面图像和秘密图像，判别器输入为封面及含密图像；GAN_2 生成器输入为含密图像，判别器输入为原秘密图像 $secret_1$

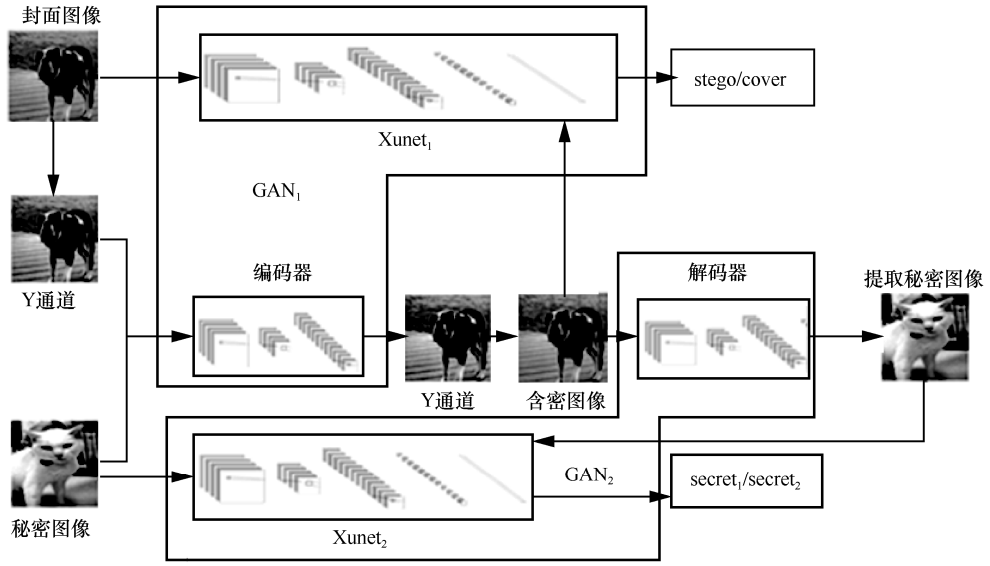


图 1 DGANS 隐写模型总体结构

以及提取出来的秘密图像 $secret_2$。2 个 GAN 分别用对抗训练来提高图像隐写的安全性和稳健性。其中 GAN_1 的判别器为隐写分析网络 $XuNet_1$[17]，接收三通道输入；GAN_2 的判别器为 $XuNet_2$，其输入为单通道。

2.2　基于 Inception 结构的编码器和解码器网络

在图 1 所示的编码器–解码器神经网络中，本文利用图 2 所示的 Inception[18]模块作为基础结构，Inception 结构将 1×1、3×3、5×5 的卷积和 3×3 的最大值池化堆叠在一起，一方面增加了网络的宽度，另一方面增加了网络对尺度的适应性，改善了图像嵌入和提取的效果。基于 Inception 结构的编码器网络如表 1 所示，解码器网络如表 2 所示。编码器网络的输入为封面图像的 Y 通道与秘密图像的通道连接，输出为含密图像的 Y 通道。解码器网络的输入为含密图像的 Y 通道，输出为提取的秘密图像。批归一化（BN, batch normalization）[19]对输入进行归一化处理，解决了训练偏移的影响，同时加快了训练的速度。其中 LeakyReLU、Tanh 和 Sigmoid 为 3 种不同的激活函数。

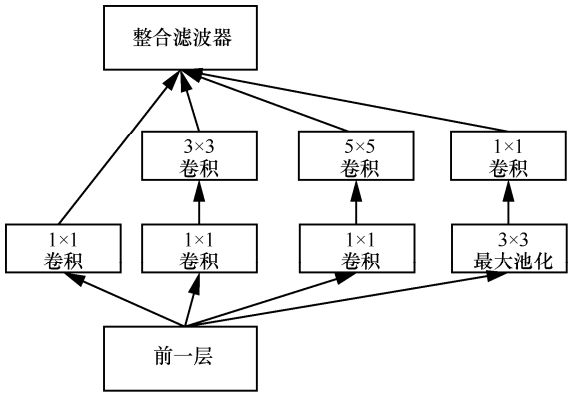


图 2　Inception v1 结构

表 1　　编码器网络结构

层结构	输入尺寸	输出尺寸
3×3×16 卷积+BN+LeakyReLU	256×256×2	256×256×16
Inception 模块	256×256×16	256×256×32
Inception 模块	256×256×32	256×256×64
Inception 模块	256×256×64	256×256×128
Inception 模块	256×256×128	256×256×256
Inception 模块	256×256×256	256×256×128
Inception 模块	256×256×128	256×256×64
Inception 模块	256×256×64	256×256×32
3×3×16 卷积+BN+LeakyReLU	256×256×32	256×256×16
1×1×1 卷积+Tanh	256×256×16	256×256×1

表 2　　解码器网络结构

层结构	输入尺寸	输出尺寸
3×3×16 卷积+BN+LeakyReLU	256×256×2	256×256×16
3×3×32 卷积+BN+LeakyReLU	256×256×16	256×256×32
3×3×64 卷积+BN+LeakyReLU	256×256×32	256×256×64
Inception 模块	256×256×64	256×256×128
Inception 模块	256×256×128	256×256×64
Inception 模块	256×256×64	256×256×32
3×3×16 卷积+BN+LeakyReLU	256×256×32	256×256×16
1×1×1 卷积+Sigmoid	256×256×16	256×256×1

2.3　通过对抗训练增强安全性

图像隐写的安全性极为重要，安全性表现为用通常的隐写分析方法难以检测到图像中是否包含隐藏信息。本文通过对抗训练来达到安全隐写的目的。Goodfellow 等[20]提出的 GAN 由一个生成器 G 和一个判别器 D 组成。生成器 G 努力让生成的图像更加真实，而判别器 D 则努力去识别出图像的真假，通过对抗训练使生成器 G 生成的图像无限逼近真实的图像，从而使判别器 D 对真假图像的辨别正确率降到 0.5 左右。将本文使用的隐写分析网络 $XuNet_1$ 作为判别器 D，编码器网络作为生成器 G，通过对抗训练使生成的含密图像难以被隐写分析网络检测到，从而提高生成算法的安全性。$XuNet_1$ 的网络结构如表 3 所示，其输入为原始封面图像或含密彩色图像，输出为隐写检测结果。

表 3　对抗训练判别器 $XuNet_1$ 及 $XuNet_2$ 结构

层结构	输入尺寸		输出尺寸
	$XuNet_1$	$XuNet_2$	
3×3 卷积+BN+LeakyReLU+Avgpool	256×256×3	256×256×1	128×128×8
3×3 卷积+BN+LeakyReLU+Avgpool	128×128×8		64×64×16
1×1 卷积+BN	64×64×16		32×32×32
1×1 卷积+BN	32×32×32		16×16×64
3×3 卷积+BN+LeakyReLU	16×16×64		8×8×128
空域金字塔池化	8×8×128		1×2 688
2 688×128 全连接层	1×2688		1×128
128×2 全连接层	1×128		1×2

2.4　通过对抗训练和数据增强提高稳健性

图像隐写的稳健性是一项十分重要的属性，反映了图像隐写技术的抗干扰能力。现有的基于深度

学习的图像隐写模型大多注重的是隐蔽性和容量，但对稳健性关注较少，本文研究的重点就是通过对抗训练来提高图像隐写的稳健性。

将本文中的解码器网络作为 GAN_2 的生成器，$XuNet_2$ 作为判别器，$XuNet_2$ 的网络结构如表 3 所示。由表 3 可知，本文将第二个判别器设计成与第一个判别器类似的结构，两者的区别是输入通道数不同，$XuNet_1$ 接收三通道输入，而 $XuNet_2$ 接收单通道的秘密图像。本文期望通过这样的设计让第二个判别器能分辨出秘密图像细微的变化。将解码器提取得到的秘密图像 $secret_2$ 作为负例，原秘密图像 $secret_1$ 作为正例输入进判别器 $XuNet_2$ 辨别，通过两者相互博弈促进，使 $secret_2$ 与 $secret_1$ 无限接近。同时单独对包含解码器网络的 GAN_2 进行增强训练，即对 GAN_1 生成的含密图像数据集进行旋转、裁剪、缩放数据增强后，再用增强数据单独对 GAN_2 进行强化训练，进一步提高了模型的稳健性。

2.5 损失函数

本文的损失函数主要包括 4 个部分，编码器的损失、解码器的损失以及 2 个生成式对抗网络的判别器损失。GAN 的训练过程为

$$\min_G \max_D E_{x\sim P_{\text{data}}}[\log D(x)] + E_{z\sim P_z}[\log(1-D(G(Z)))] \quad (1)$$

其中，$D(x)$表示对真实的样本进行判别，其判别结果越接近 1 越好，所以损失函数为 $\log(D(x))$；$G(z)$ 表示生成器编码网络输出的含密图像以及解码网络提取的秘密图像，本文希望判别器的判别结果 $D(G(z))$越接近 0 越好。对抗训练的目的是优化 D 使式(1)的期望最大化，同时优化生成器 G 使式(1)期望最小化。因此判别器 D 的损失会在生成器 G 和判别器 D 中反向传播，但生成器 G 的损失仅在生成器内反向传播。GAN_2 解码网络的损失会同时在编码网络和解码网络中反向传播，而编码网络的损失只在编码网络内反向传播。

2.5.1 编码器–解码器损失

编码器与解码器的损失主要由图像间像素值和结构上的统计差异计算得到，编码器损失为含密图像与封面的差异度，解码器损失为提取的秘密图像与原秘密图像的差异度。损失由均方误差（MSE, mean square error）和结构相似度（SSIM, structural similarity）[21]联合计算得到，MSE 表示图像 x 和图像 y 的均方误差，如式(2)所示。

$$\text{MSE} = \frac{1}{HW}\sum_{i=1}^{H}\sum_{j=1}^{W}(x(i,j)-y(i,j))^2 \quad (2)$$

SSIM 分别从亮度、对比度、结构 3 个方面衡量图像的相似性，如式(3)~式(6)所示。

$$\text{SSIM}(x,y) = L(x,y)C(x,y)S(x,y) \quad (3)$$

$$L(x,y) = \frac{2\mu_x\mu_y + \tau_1}{\mu_x^2 + \mu_y^2 + \tau_1} \quad (4)$$

$$C(x,y) = \frac{2\sigma_x\sigma_y + \tau_2}{\sigma_x^2 + \sigma_y^2 + \tau_2} \quad (5)$$

$$S(x,y) = \frac{\sigma_{xy} + \tau_3}{\sigma_x + \sigma_y + \tau_3} \quad (6)$$

其中，μ_x、μ_y 分别表示图像 x 与图像 y 的均值，σ_x 与 σ_y 分别表示图像 x 与图像 y 的方差，σ_{xy} 表示图像 x 与图像 y 的协方差，τ_1、τ_2、τ_3 为 3 个取值较小的正常量，用于避免除数为 0 出现计算异常，L、S、C 分别表示亮度、结构和对比度相似性。式(7)计算的多尺度结构相似度（MS-SSIM, multi-scale SSIM）可以对不同尺度进行结构相似度评判。

$$\text{MSSSIM}(x,y) = [L_N(x,y)]^{l_N}\prod_{j=1}^{N}[C_j(x,y)^{\gamma_j}][S_j(x,y)^{\rho_j}] \quad (7)$$

其中，N 表示多尺度下采样的级数，l_N、γ_j、ρ_j 为 0~1 之间的参数，用于控制各成分在相似度衡量时的重要性，L、S、C 的下标表示所在的尺度。

由 MSE、SSIM 和 MS-SSIM 联合计算编码器和解码器的损失，编码器损失如式(8)所示。

$$\text{en\_loss}(c,s) = \chi(1-\text{SSIM}(c,s)) + (1-\chi)(1-\text{MSSSIM}(c,s)) + \delta\text{MSE}(c,s) \quad (8)$$

其中，c 为封面图像，s 为含密图像。

解码器损失如式(9)所示。

$$\text{de\_loss}(s_1,s_2) = \chi(1-\text{SSIM}(s_1,s_2)) + (1-\chi)(1-\text{MSSSIM}(s_1,s_2)) + \delta\text{MSE}(s_1,s_2) \quad (9)$$

其中，s_1 为原始秘密图像，s_2 为提取出来的秘密图像。

编码器–解码器损失如式(10)所示。

$$\text{LOSS}(c_1,c_2,s_1,s_2) = \text{en\_loss}(c_1,c_2) + \varepsilon\text{de\_loss}(s_1,s_2) \quad (10)$$

其中，超参数值 χ=0.5，δ=0.85，ε=0.3。

2.5.2 GAN 判别器损失

DGANS 的 2 个生成式对抗网络的生成器分别

为上述的编码网络和解码网络，损失即为 2.5.1 节描述的损失。另外，判别器的损失使用二值交叉熵（BCE, binary cross entropy）损失，BCE 如式(11)所示。

$$\text{BCE\_LOSS}(x,y)=-\frac{1}{n}\sum_i(x_i\log(y_i)+(1-x_i)\log(1-y_i)) \tag{11}$$

其中，x 和 y 分别为判别器目标和预测输出。DGANS 包含 2 个判别器，其中一个判别器为隐写分析网络 XuNet，其损失 $\text{dis\_loss}(c,s)$ 如式(12)所示；另一个判别器为 XuNet_2，其损失函数 $\text{dis\_loss}(s_1,s_2)$ 如式(13)所示。

$$\text{dis\_loss}(c,s)=\text{BCE\_LOSS}(c,s\eta) \tag{12}$$

$$\text{dis\_loss}(s_1,s_2)=\text{BCE\_LOSS}(s_1,s_2\eta) \tag{13}$$

其中，η 为 0.8~1.2 之间的随机数。

3　实验结果

本节主要介绍实验使用的数据集、参数的设置以及实验过程和结果。实验数据集采用 PASCAL VOC2012[22]，使用其中 11 540 张图片作为训练集，前 5 770 张图片作为秘密图像，剩下的 5 770 张图片作为封面图像。随机选取 5 000 张图片用作验证集，前 2 500 张作为秘密图像，后 2 500 张作为封面图像，测试实验所使用的数据均为 5 000 张验证集图片得出的结果。所有图片的尺寸都归一化为 256×256 的大小。

在实验的参数设置上，模型所有的参数都使用 Xavier 初始化，本模型使用的实验环境为一台安装有 GTX1080Ti 显卡的服务器，操作系统为 Ubuntu16.04，显卡驱动的版本为 CUDA9.0+cuDNN7.0，程序用 python3.5 版本的 pytorch 深度学习框架实现，使用的集成开发环境为 pycharm。批大小设置为 5，初始学习率设置为 10^{-4}，训练了 80 轮后网络收敛。整个模型的损失为编码器–解码器损失以及 2 个 GAN 判别器的损失之和，如式(14)所示。

$$\text{LOSS\_TOTAL}=\text{LOSS}(c_1,c_2,s_1,s_2)+\text{dis\_loss}(c,s)+\text{dis\_loss}(s_1,s_2) \tag{14}$$

训练过程中总损失变化情况如图 3 所示。

3.1　隐写容量分析

一个好的隐写模型，应具备良好的隐蔽性、较大的隐写容量以及较高的稳健性，DGANS 模型嵌入的图像尺寸都是 256×256，嵌入容量为 8 bpp (bit per pixel)。不同模型隐写容量对比结果如表 4 所示。由表 4 可知，DGANS 具有较大的隐写容量。

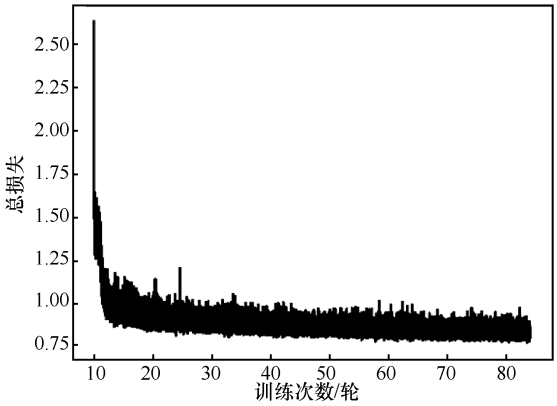


图 3　训练过程中总损失变化曲线

表 4　不同模型隐写容量对比结果

模型	秘密图像尺寸	封面图像尺寸	容量/bpp
HUGO[4]	32×32	32×32	0.1~0.4
SteGANography[9]	32×32	50×50	0.1~0.4
SSGAN[7]	64×64	204×204	0.4
S-UNIWARD[5]	64×64	64×64	0.4
ISGAN[14]	256×256	256×256	8
Rehman 等[10]	300×300	300×300	8
DGANS	256×256	256×256	8

3.2　隐蔽性测试结果

对于隐写模型的隐蔽性，首先用封面与含密图像之间以及原秘密和提取的秘密图像之间的峰值信噪比（PSNR, peak signal to noise ratio）以及 SSIM 值来衡量，图像的 PSNR 计算式为

$$\text{PSNR}=\frac{1}{10}\lg\left(\frac{(2^n-1)^2}{\text{MSE}}\right) \tag{15}$$

PSNR 数值越大表示差异越小。PSNR 使用广泛的图像差异度评价指标，但是 PSNR 不能反映人类的视觉差异，SSIM 更接近人类视觉感知。完全训练的 DGANS 在整个验证集上的 PSNR 和 SSIM 统计结果表 5 所示。图 4 给出了 DGANS、Rehman 等[10]模型及 ISGAN[14]的可视化结果展示。由表 5 和图 4 可知，本文隐写模型有很好的隐蔽性，肉眼观察不到含密图像与封面图像区别，同时也有很好的还原能力，从含密图像提取的秘密图像与原图也没有明显区别。由图 4 知，在未受攻击情况下，Rehman 等[10]及 ISGAN[14]等类似模型都表现出相当的性能。

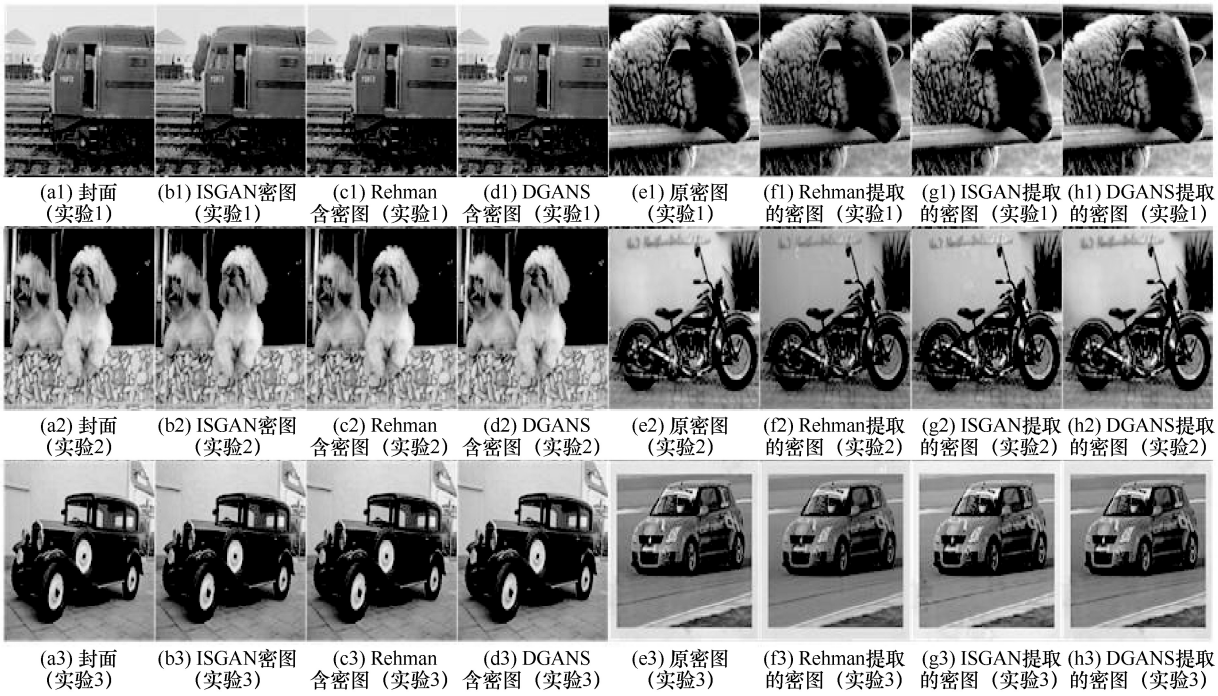


图 4　可视化结果比较

表 5　DGANS 的 PSNR 和 SSIM 统计结果

测试指标	取值
编码器 PSNR	24.002 8
解码器 PSNR	24.691 0
编码器 SSIM	0.907 2
解码器 SSIM	0.903 5
区分器 LOSS	0.659 7

使用 Ye 等[23]的隐写分析模型 YeNet 对 DGANS 进行隐写分析测试，YeNet 使用 BOSSBase 数据集采用 S-UNIWARD[5]隐写方法生成的数据集训练得到，检测结果所得的受试者工作特征（ROC, receiver operating characteristic）曲线以及曲线下面积（AUC, area under curve）的值如图 5 所示。与 S-UNIWARD 的 ROC 曲线相比，DGANS 的 AUC 值要低于 S-UNIWARD，不易手动检测，这进一步证明了 DGANS 有良好的隐蔽性。

3.3　稳健性测试及对比实验

本文分别对旋转、平移、缩放 3 种几何攻击进行了稳健性测试。当对含密图像进行低小角度旋转、小幅度平移、缩放后，用训练好的解码网络提取秘密图像，用 SSIM 值来评估从受攻击图像中提取的秘密图像与原始秘密图像的相似度，并与 Rehman 等[10]模型及 ISGAN [14]等其他类似模型进行对比，所有模型均用相同训练集训练至收敛。

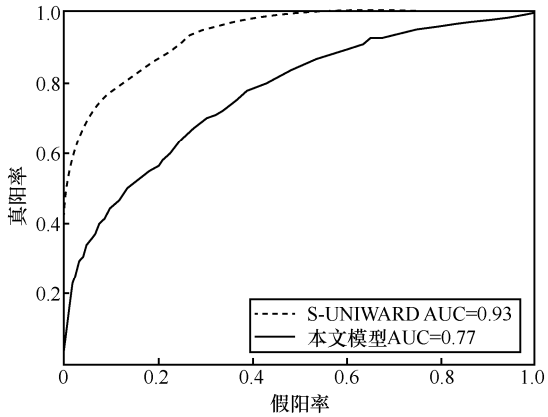


图 5　隐写检测 ROC 曲线

3.3.1　旋转攻击测试

旋转攻击测试包括逆时针旋转 2°、逆时针旋转 5° 共 2 组实验，具体实验结果及与其他模型对比结果如表 6 所示。由表 6 可知，在旋转攻击下，本文的 DGANS 模型提取的秘密图质量要明显优于其他模型，旋转 5° 时提取的秘密图 SSIM 值甚至是另外 2 个模型提取的秘密图 SSIM 值的 2 倍左右。

表 6　旋转攻击下不同模型提取的秘密图 SSIM 值对比

模型	旋转 5°	旋转 2°
Rehman 等[10]	0.338 5	0.447 3
ISGAN[14]	0.358 0	0.641 7
DGANS	0.685 4	0.789 1

图 6 为旋转攻击测试的可视化结果。从图 6 可以看出，经旋转之后 Rehman 等的模型基本是失效的，ISGAN 的提取比较差，出现了明显的图片裂化现象，而 DGANS 模型提取效果受旋转攻击影响较小。

3.3.2　平移攻击测试

平移攻击测试进行了如下实验：水平平移 2 个、5 个像素，垂直平移 2 个、5 个像素，随机水平平移 0~10 个像素，随机垂直平移 0~10 个像素，水平垂直同时随机平移 0~10 个像素。平移攻击下不同模型提取的秘密图 SSIM 值对比如表 7 所示。从表 7 可以看出，平移攻击后提取的秘密图像中，DGANS 要比其他模型好 15%~20%。

图 7 为平移攻击实验的可视化结果，其中第一行为水平随机平移 0~10 个像素，第二行为垂直随

(a1) 封面（旋转2°）　(b1) ISGAN含密图（旋转2°）　(c1) Rehman含密图（旋转2°）　(d1) DGANS含密图（旋转2°）　(e1) 原密图（旋转2°）　(f1) Rehman提取的密图（旋转2°）　(g1) ISGAN提取密图（旋转2°）　(h1) DGANS提取密图（旋转2°）

(a2) 封面（旋转5°）　(b2) ISGAN含密图（旋转5°）　(c2) Rehman含密图（旋转5°）　(d2) DGANS含密图（旋转5°）　(e2) 原密图（旋转5°）　(f2) Rehman提取的密图（旋转5°）　(g2) ISGAN提取密图（旋转5°）　(h2) DGANS提取密图（旋转5°）

图 6　旋转攻击实验

表 7　　平移攻击下不同模型提取的秘密图 SSIM 值对比

模型	水平平移/像素		垂直平移/像素		随机平移 0~10 个像素		水平垂直随机平移 0~10 个像素
	2	5	2	5	水平	垂直	
Rehman 等[10]	0.544 5	0.407 5	0.538 4	0.404 5	0.404 6	0.406 0	0.322 1
ISGAN[14]	0.650 4	0.630 7	0.660 9	0.650 5	0.389 9	0.405 7	0.297 7
DGANS	0.792 0	0.783 1	0.803 8	0.785 8	0.575 2	0.540 9	0.526 3

(a1) 封面（水平）　(b1) ISGAN含密图（水平）　(c1) Rehman含密图（水平）　(d1) DGANS含密图（水平）　(e1) 原密图（水平）　(f1) Rehman提取的密图（水平）　(g1) ISGAN提取密图（水平）　(h1) DGANS提取密图（水平）

(a2) 封面（垂直）　(b2) ISGAN含密图（垂直）　(c2) Rehman含密图（垂直）　(d2) DGANS含密图（垂直）　(e2) 原密图（垂直）　(f2) Rehman提取的密图（垂直）　(g2) ISGAN提取密图（垂直）　(h2) DGANS提取密图（垂直）

(a3) 封面（水平垂直）　(b3) ISGAN含密图（水平垂直）　(c3) Rehman含密图（水平垂直）　(d3) DGANS含密图（水平垂直）　(e3) 原密图（水平垂直）　(f3) Rehman提取的密图（水平垂直）　(g3) ISGAN提取密图（水平垂直）　(h3) DGANS提取密图（水平垂直）

图 7　平移攻击实验

机平移 0~10 个像素，第三行为水平垂直同时进行 0~10 像素的随机平移的可视化结果。由图 7 可知，本文 DGANS 模型稳健性较佳，并明显优于另外 2 个模型，Rehman 等的模型出现密图提取失效的情况，ISGAN 提取的结果出现大量噪点，DGANS 则受影响较小。

3.3.3 缩放攻击测试

缩放攻击实验主要是对含密图像随机缩放 80%~120%后再进行提取，观察提取效果。得出的提取结果与 ISGAN 的结果相当，都具有很好的抗缩放攻击的提取效果，而 Rehman 等的模型则对尺度变化较为敏感。具体对比结果如表 8 所示，可视化结果如图 8 所示。由图 8 知，DGANS 和 ISGAN 基本对尺度变化有较好的适应性，Rehman 等的模型则出现提取失效的情况。

表 8 缩放攻击下不同模型提取的秘密图 SSIM 值对比

模型	随机缩放 80%~120%
Rehman 等[10]	0.791 8
ISGAN[14]	0.890 1
DGANS	0.929 5

4 结束语

本文在使用深度学习实现图像信息隐藏的过程中，提出了 DGANS 图像隐写模型，该模型具有高容量、良好的隐蔽性和稳健性等特性。DGANS 可以有效地将单通道的灰度秘密图像嵌入隐藏进封面图像中，并从中提取出来。本文在保证高隐蔽性的基础上，对基于深度学习的隐写模型的稳健性进行了研究，使隐写图像在受到一定的几何攻击后，仍能以较高的保真度将秘密图像提取出来，这是大多数基于深度学习的图像隐写模型未涉及的。本文在 PACAL VOC2007 数据集上进行了实验验证，实验结果表明，本文在提高稳健性方面的模型设计是有效的。

参考文献：

[1] 沈昌祥, 张焕国, 冯登国, 等. 信息安全综述[J]. 中国科学(信息科学), 2007, 37(1):129-150.
SHEN C X, ZHANG H G, FENG D G, et al. Survey on information security[J]. Science in China Series (Information Sciences), 2007, 37(1): 129-150.

[2] 王向阳, 杨红颖. DCT 域自适应彩色图像二维数字水印算法研究[J]. 计算机辅助设计与图形学学报, 2004, 16(2): 243-247.
WANG X Y, YANG H Y. Adaptive 2-D color image watermarking based on DCT[J]. Journal of Computer-Aided Design and Computer Graphics, 2004, 16(2): 243-247.

[3] 王向阳, 杨红颖. 基于视觉掩蔽特性的小波域彩色数字水印技术[J]. 计算机辅助设计与图形学学报, 2004, 16(9): 1240-1243.
WANG X Y, YANG H Y. Color digital watermarking based on integer lifting wavelet transform and visual masking[J]. Journal of Computer-Aided Design and Computer Graphics, 2004, 16(9): 1240-1243.

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

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

[6] HOLUB V, FRIDRICH J. Designing steganographic distortion using directional filters[C]// IEEE International Workshop on Information Forensics & Security. IEEE, 2012: 234-239.

[7] SHI H C, DONG J, WANG W, et al. SSGAN: secure steganography based on generative adversarial networks[C]//18th Pacific-Rim Conference on Multimedia. Springer, 2017: 534-544.

[8] ARJOVSKY M, CHINTALA S, BOTTOU L. Wasserstein GAN[J]. arXiv Preprint, arXiv: 1701.07875, 2017: 1-32.

[9] HAYES J, DANEZIS G. Ste-GAN-ography: generating steganographic images via adversarial training[J]. arXiv Preprint, arXiv:170300371v2, 2017: 1-9.

[10] REHMAN A U, RAHIM R, NADEEM S, et al. End-to-end trained CNN encode-decoder networks for image steganography[J]. arXiv Preprint, arXiv:1711. 07201, 2017: 1-5.

[11] BALUJA S. Hiding images in plain sight: deep steganography[C]//Advances in Neural Information Processing Systems. 2017:

(a) 封面 (b) ISGAN含密图 (c) Rehman含密图 (d) DGANS含密图 (e) 原密图 (f) Rehman提取的密图 (g) ISGAN提取密图 (h) DGANS提取密图

图 8 缩放攻击实验

2069-2079.

[12] CHU C, ZHMOGINOV A, SANDLER M. CycleGAN, a master of steganography[C]// Thirty-first Conference on Neural Information Processing Systems. 2017: 1-6.

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

[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] WU P, YANG Y, LI X. StegNet: MEGA image steganography capacity with deep convolutional network[J]. Future Internet, 2018,10(6): 54-68.

[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(1): 9314-9323.

[17] XU G. Deep convolutional neural network to detect J-UNIWARD[C]// The 5th ACM Workshop on Information Hiding and Multimedia Security. 2017: 67-73.

[18] SZEGEDY C, VANHOUCKE V, IOFFE S, et al. Rethinking the inception architecture for computer vision[C]// IEEE Conference on Computer Vision and Pattern Recognition. 2016: 2818-2826.

[19] IOFFE S, SZEGEDY C. Batch normalization: accelerating deep network training by reducing internal covariate shift[J]. arXiv Preprint, arXiv: 1502.03167, 2015.

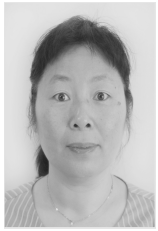
[20] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial nets[C]//International Conference on Neural Information Processing Systems. 2014: 2672-2680.

[21] WANG Z, BOVIK A, SHEIKH H R, et al. Image quality assessment: from error visibility to structural similarity[J]. IEEE Transactions on Image Processing, 2004, 13(4): 600-612.

[22] EVERINGHAM M. The PASCAL visual object classes challenge (VOC2007) Results[J]. Lecture Notes in Computer Science, 2007, 111(1): 98-136.

[23] YE J, NI J, YI Y. Deep learning hierarchical representations for image steganalysis[J]. IEEE Transactions on Information Forensics and Security, 2017, 12(11): 2545-2557.

[作者简介]



竺乐庆（1972– ），女，浙江嵊州人，博士，浙江工商大学副教授、硕士生导师，主要研究方向为图像处理、模式识别、视频处理、信息隐藏等。

郭钰（1995– ），男，安徽宿州人，浙江工商大学硕士生，主要研究方向为图像处理、模式识别、图像信息隐藏等。

莫凌强（1994– ），男，浙江嘉兴人，浙江工商大学硕士生，主要研究方向为模式识别、图像处理、图像信息隐藏等。

张大兴（1971– ），男，浙江嵊州人，博士，杭州电子科技大学副教授、硕士生导师，主要研究方向为信息安全、多媒体技术、软件工程等。