软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2023,34(7):3385−3407 [doi: 10.13328/j.cnki.jos.006537]

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

基于 U-Net 结构的生成式多重对抗隐写算法*

马 宾[1,2], 韩作伟[1,2], 徐 健[3], 王春鹏[1,2], 李 健[1,2], 王玉立[1,2]

[1](齐鲁工业大学 (山东省科学院) 网络空间安全学院, 山东 济南 250353)

[2](山东省计算机网络重点实验室, 山东 济南 250098)

[3](山东财经大学 计算机科学与技术学院, 山东 济南 250014)

通信作者: 徐健, E-mail: sdfixj@126.com

摘 要: 人工智能的发展为信息隐藏技术带来越来越多的挑战, 提高现有隐写方法的安全性迫在眉睫. 为提高图像的信息隐藏能力, 提出一种基于 U-Net 结构的生成式多重对抗隐写算法. 所提算法通过生成对抗网络与隐写分析器优化网络、隐写分析对抗网络间的多重对抗训练, 构建生成式多重对抗隐写网络模型, 生成适合信息隐写的载体图像, 提高隐写图像抗隐写分析能力; 同时, 针对现有生成对抗网络只能生成随机图像, 且图像质量不高的问题, 设计基于 U-Net 结构的生成式网络模型, 将参考图像的细节信息传递到生成载体图像中, 可控地生成高质量目标载体图像, 增强信息隐藏能力; 其次, 采用图像判别损失、均方误差 (*MSE*) 损失和隐写分析损失动态加权组合作为网络迭代优化总损失, 保障生成式多重对抗隐写网络快速稳定收敛. 实验表明, 基于 U-Net 结构的生成式多重对抗隐写算法生成的载体图像 *PSNR* 最高可达到 48.60 dB, 隐写分析器对生成载体图像及其隐写图像的判别率为 50.02%, 所提算法能够生成适合信息嵌入的高质量载体图像, 保障隐写网络快速稳定收敛, 提高了图像隐写安全性, 可以有效抵御当前优秀的隐写分析算法的检测.

关键词: 隐写; 隐写分析; 生成对抗网络; 多重对抗; U-Net

中图法分类号: TP309

中文引用格式: 马宾, 韩作伟, 徐健, 王春鹏, 李健, 王玉立. 基于U-Net结构的生成式多重对抗隐写算法. 软件学报, 2023, 34(7): 3385–3407. http://www.jos.org.cn/1000-9825/6537.htm

英文引用格式: Ma B, Han ZW, Xu J, Wang CP, Li J, Wang YL. Generative Multiple Adversarial Steganography Algorithm Based on U-Net Structure. Ruan Jian Xue Bao/Journal of Software, 2023, 34(7): 3385–3407 (in Chinese). http://www.jos.org.cn/1000-9825/6537.htm

Generative Multiple Adversarial Steganography Algorithm Based on U-Net Structure

MA Bin[1,2], HAN Zuo-Wei[1,2], XU Jian[3], WANG Chun-Peng[1,2], LI Jian[1,2], WANG Yu-Li[1,2]

[1](School of Cyber Security, Qilu University of Technology (Shandong Academy of Sciences), Jinan 250353, China)

[2](Shandong Provincial Key Laboratory of Computer Networks, Jinan 250098, China)

[3](School of Computer Science and Technology, Shandong University of Finance and Economics, Jinan 250014, China)

Abstract: The development of artificial intelligence brings more and more challenges to data hiding technology, and it is urgent to improve the security of existing steganography methods. In this study, a generative multiple adversarial steganography algorithm based on U-Net network structure is proposed to improve the image data hiding ability. A generative multiple adversarial steganography network (GMASN), including the generative adversarial network, the steganalyzer optimization network and the steganalysis network, is firstly constructed, and the anti steganalysis ability of the steganography image is improved through the competition of the networks in the

* 基金项目: 国家自然科学基金 (61802212, 61872203); 山东省重大科技创新工程 (2019JZZY010127, 2019JZZY010132, 2019JZZY010201); 山东省自然科学基金 (ZR2019BF017, ZR2020MF054); 山东省高校科研计划 (J18KA331); 山东省高等学校青创人才引育计划 (S019-161); 济南市“高校 20 条”引进创新团队 (2019GXRC031)
收稿时间: 2021-08-10; 修改时间: 2021-09-02, 2021-10-10; 采用时间: 2021-11-16; jos 在线出版时间: 2023-01-13
CNKI 网络首发时间: 2023-01-19

GMASN. At the same time, aiming at the problem that the existing generative adversarial network can only generate low-quality images randomly, a generative network based on U-Net structure is designed to transfer the details of the reference image to the generated carrier image, by which the image can be generated objectively with high visual quality. Moreover, the image discrimination loss, mean square error (*MSE*) loss, and steganalysis loss are dynamically combined in the proposed scheme to enable the GMASN to converge rapidly and stably. Experimental results show that the *PSNR* of the generated carrier image can reach 48.60 dB, and the discrimination rate between the generated carrier image and the steganographic image is 50.02%. The proposed algorithm can generate high-quality carrier images suitable for data hiding, enable the steganographic network to converge rapidly and stably, and improve the security of image steganography effectively.

Key words: steganography; steganalysis; generative adversarial network (GAN); multiple adversarial; U-Net

1 引 言

信息隐藏通过将秘密信息嵌入到载体冗余信息中实现隐蔽通信功能[1]. 隐写术作为信息隐藏领域中的一项重要技术, 其不仅隐藏了信息的内容, 还隐藏了信息的存在. 典型的隐写系统利用人类视觉特征, 将秘密信息不可见的隐藏到载体图像中, 使之不会引起检测者的怀疑, 从而保证秘密信息安全地传递.

隐写术要求尽可能少地改变载体的统计信息. 由于图像包含丰富的内容信息, 目前主流的隐写方法大多选择图像作为载体. 基于图像的隐写算法大致可以划分为空域隐写、变换域隐写和无载体隐写 3 种类型. 空域图像隐写算法主要通过修改图像像素值嵌入秘密信息, 典型的空域隐写算法包括最低有效位 (LSB) 算法[2]、空间通用小波相对失真算法 (S-UNIWARD)[3]、小波求权算法 (WOW)[4]、高难度检测隐写算法 (HUGO)[5]等, 其中 S-UNIWARD、WOW、HUGO 算法通过选择合适的修改位置, 从而最小化嵌入失真函数, 降低隐写图像的统计可检测性; 变换域隐写算法通过修改载体图像的频域系数实现秘密信息的隐藏, 如: 离散傅里叶变换 (DFT) 隐藏算法[6]、离散余弦变换 (DCT) 隐藏算法[7]、离散小波变换 (DWT) 隐藏算法[8]等; 无载体隐写算法[9−11]不直接改变载体图像数据, 而是通过建立图像的特征属性与秘密信息间的映射关系, 实现秘密信息隐蔽传输.

另一方面, 作为针对隐写技术的一种检测判别手段, 隐写分析根据观测到的图像特征信息的变化, 判断其是否含有秘密信息. 经典的空间富模型 (SRM)[12]隐写分析算法等通过分析图像的统计特性来检测秘密信息. 近年来, 研究人员将深度学习 (deep learning) 应用到隐写分析领域, 并得到了迅速的发展. 基于深度学习的隐写分析模型通过设计不同网络结构来学习图像的深层特征, 以区分载体图像和隐写图像. 基于深度学习的隐写分析网络 (例如 GNCNN[13]、Xu'Net[14]、Ye'Net[15]、SRNet[16]) 检测准确率逐渐超过了传统的 SRM 隐写分析算法. 隐写分析技术的快速发展严重威胁了隐写算法的安全, 提高隐写算法的安全性迫在眉睫.

近年来, 随着深度学习技术的发展, 基于深度学习框架的各种信息隐藏模型不断出现. 其中, 基于生成对抗网络 (generative adversarial network, GAN)[17]的信息隐写技术取得了较快发展. 生成对抗网络是 Goodfellow 等人于 2014 年提出的一种深度生成模型. 该模型由一个生成器 G 和一个判别器 D 构成, 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, 通过建立生成器和判别器的极小/极大博弈优化其生成样本的能力, 将一个随机变量的分布逐步逼近真实数据样本的分布. 基于生成对抗网络的隐写算法通过生成图像实现信息隐藏, 从而提高隐写算法的安全性. 然而, 生成对抗网络存在以下两个方面问题, 影响了其在信息隐藏领域的进一步应用. 一方面, 生成对抗网络只能学习训练数据集的分布特征, 随机生成与原始图像分布尽可能一致的载体图像; 因而, 生成对抗网络难以可控的生成目标载体图像, 实现基于特定载体图像的信息隐藏. 另一方面, 生成对抗网络采用张量运算的方式生成类似目标图像分布特征的载体图像, 当生成目标图像尺寸较大时, 张量的长度增加, 张量空间变大, 网络运算代价迅速升高, 网络收敛困难, 难以生成高质量载体图像. 然而, 在信息隐藏领域, 一方面需要按照需求生成特定的载体图像, 以实现基于特定目标图像的数据隐写; 另一方面需要载体图像具有较大的尺寸和较高的分辨率, 以提高载体图像的隐写性能. U-Net 是 Ronneberger 等人[18]提出的一种可以实现二维图像特征提取的增强型全卷积神经网络 (FCN) 结构. 其解决了 FCN 模型在训练时丢失较多细节, 难以实现原始图像特征完全重构的问题. U-Net 网络模型是在 FCN 基础上采用跳跃连接的方式对 FCN 网络的上、下采样部分进行关联, 并在图像特征重构时叠加融合原始图像深层和浅层特征, 从而有效提升原始图像特征重构的能力, 生成和原始图像尽可

能一致的图像. 为应对隐写分析技术的快速发展, 提升载体图像的隐写能力, 克服生成对抗网络在信息隐藏领域所面临的问题. 本文提出一种基于 U-Net 结构的生成式多重对抗隐写网络模型, 按需生成适合隐写的特定载体图像, 对抗提升载体图像质量, 增强信息隐藏能力. 本网络模型由生成对抗网络、隐写器网络、隐写分析器优化网络、隐写分析对抗网络构建而成, 通过网络间对抗提高隐写图像抗隐写分析检测的能力, 生成更适合信息隐写的载体图像. 同时, 本算法将 U-Net 结构应用于图像生成网络, 通过跳接层协议传递真实图像的细节信息, 融合图像的深层和浅层内容分布特征, 提高生成图像质量, 并生成既定目标图像, 解决传统生成对抗网络仅能生成随机图像, 且生成图像质量不高的问题. 其次, 本方案将判别器损失 (*D_loss*), 均方差损失 (*MSE_loss*) 和隐写网络损失 (*SDO_loss*) 动态加权构成生成网络迭代优化的总损失, 保障多重对抗隐写网络快速稳定收敛. 实验中, 采用真实图像及其隐写图像, 生成图像及其隐写图像分别作为训练数据优化隐写分析器网络的性能, 并利用优化后的隐写分析网络对生成图像和隐写图像进行平行、交叉和再训练隐写分析检测, 验证基于 U-Net 结构的生成式多重对抗隐写分析算法的安全性. 本文的主要贡献如下.

(1) 提出一种基于多重对抗的生成式隐写算法模型. 首先对隐写分析器网络进行性能优化, 采用优化后的隐写分析器构建隐写分析对抗网络, 通过生成网络与判别网络、隐写分析对抗网络间的多重对抗迭代训练, 生成更适合信息嵌入的载体图像, 提升生成载体图像的信息隐写能力, 图像抗隐写分析能力的鲁棒性.

(2) 提出一种基于 U-Net 结构的载体图像生成对抗网络模型. 在生成网络中采用 U-Net 网络架构, 基于跳层连接方法将真实图像的细节信息传递到生成图像中, 可控地生成高视觉质量的目标载体图像, 增强信息隐藏能力, 解决了传统生成对抗网络无法生成既定图像, 且图像质量不高的问题.

(3) 将生成对抗网络判别损失、均方误差 (*MSE*) 损失和隐写分析损失加权组合构成多重对抗隐写总损失. 根据网络训练过程动态选取权值参数, 解决训练过程中生成器网络迭代梯度消失问题, 保障生成式多重对抗隐写网络快速稳定收敛.

本文第 1 节介绍了生成式对抗隐写算法的发展和面临的问题. 第 2 节介绍了隐写技术的发展状况以及所面临的问题. 第 3 节详细分析了本文所提出的基于 U-Net 结构的生成式多重对抗隐写算法方案. 第 4 节详细描述了实验的实现过程、数据集、参数设置和评估指标, 并对实验结果进行讨论和分析. 最后, 第 5 节给出全文结论.

2 相关工作

生成对抗网络 (GAN) 的发展将深度学习在图像处理领域中的应用推上了新高度, 其对复杂数据强大的建模能力, 为隐写算法与深度学习的结合提供了契机. 生成对抗网络可以通过网络之间的相互竞争生成所需的数据分布, 有效提高隐写术的安全性. 根据不同的信息隐藏方法, 基于深度学习的隐写算法可以划分为: 载体图像生成式隐写、真实图像嵌入式隐写和无载体隐写 3 种类型. 2016 年, Volkhonskiy 等人[19]首次提出生成式隐写模型 (SGAN), 其通过在 DCGAN[20]的基础上添加隐写分析网络, 使生成的载体图像嵌入秘密信息后能够在一定程度上抵抗隐写分析模型的检测. 其中, 生成网络负责生成载体图像; 判别网络评估生成图像的视觉质量; 隐写分析网络用来评估生成隐写图像的安全性. SGAN 利用 GAN 对抗训练的思想, 在训练阶段生成器分别与判别器、隐写分析器对抗, 优化生成的载体图像, 使其隐写图像能够抵抗隐写分析的检测, 但该方法生成图像质量和抗隐写分析能力不足. 在 SGAN 的基础上 Shi 等人[21]提出 SSGAN 生成式隐写方案, 其使用 WGAN[22]代替 SGAN 中的 DCGAN 网络, 并且使用基于卷积神经网络的隐写分析网络 (GNCNN) 重新设计了判别网络与隐写分析网络. SSGAN 与 SGAN 相比生成载体图像的视觉质量得到提升, 收敛速度和训练速度加快, 模型稳定性也得到了提高. 生成式对抗隐写方案, 一定程度上提高了隐写图像抗隐写分析能力, 但由于这类方案都是基于正态噪声生成的载体图像, 无法生成既定图像, 且视觉质量不高、语义上不够真实自然、抗隐写分析的能力也难以达到理想的效果. 2018 年, Zhang 等人[23]提出了一种基于对抗样本[24]的图像隐写算法并使相同的对抗图像获得相同的分类标签, 优化了损失函数. 用快速梯度下降模型 (fast gradient sign model)[25]生成对抗载体图像, 并采用经典的自适应隐写算法嵌入秘密信息. 由于隐写分析网络可以作为一个用于目标分类的二元分类网络, 对抗样本通过在输入图像中添加对抗性噪声来欺骗目标分类网络, 产生错误输出. 因而, 加入了对抗噪声的隐写图像获得主动欺骗隐写分析器的能力, 从而实现信息

隐藏. 然而, Zhang 等人的隐写模型需要在信息嵌入时需要根据添加噪声对目标分类网络的影响进行重复训练, 以生成适合信息嵌入的载体图像, 当面对大量需要嵌入秘密信息的图像时, 耗费时间巨大, 因而该模型只适用于少量载体图像的情况. 针对此问题, 2019 年, Zhou 等人[26]采用全卷积神经网络 (FCN) 作为载体图像生成器, 构建了一种可以快速生成对抗载体图像的网络模型. 并且, 设计了新的损失函数, 使得对抗载体图像和隐写图像能够欺骗隐写网络的分析. 该隐写模型一定程度上提高了载体图像的生成速度和质量, 增强了隐写图像的安全性, 但其抗隐写分析能力鲁棒性不够理想. 2020 年, Li 等人[27]采用对抗样本的思想, 将载体图像划分为两个部分, 在一部分中隐藏秘密信息, 而对剩余部分添加对抗样本, 使整幅图像一定程度上具有欺骗隐写分析模型的能力. 然而, 这种方法只对部分图像进行增强, 与另一部分图像在视觉上形成了一定差异, 降低了隐写的安全性.

另一方面, 相较于采用 GAN 生成更适合信息嵌入的载体图像, 研究人员还提出了自适应嵌入隐写框架, 通过优化失真函数, 减小隐写图像的总隐写失真, 降低因信息嵌入引发的图像统计异常. 其中一个重要的途径就是采用隐写分析网络与生成网络对抗训练, 得到信息嵌入的最小失真代价. 2018 年, Tang 等人[28]首次提出了一种自动学习失真函数的嵌入式隐写模型 (ASDL-GAN), 该模型在生成器与隐写分析器的对抗训练中提高了隐写图像的安全性. Yang 等人[29]在 ASDL-GAN 的基础上提出了 UT-SCA-GAN, 使用更紧凑的 U-Net 生成器, 利用 tanh-simulator 函数替代 ASDL-GAN 中的 TES 网络. 其模型安全性超过了经典的 S-UNIWARD 隐写算法. Meng 等人[30]利用快速目标识别网络 (faster RCNN) 识别图像中的纹理复杂区域, 通过在不同复杂程度的区域中选择最优自适应隐写算法嵌入信息, 提升隐写图像的安全性和抗检测性. 此外, 2018 年, Tang 等人[31]提出了一种对抗嵌入方法 (ADV-EMB), 该方法在传统的最小化失真函数框架下工作, 通过对攻击隐写分析器的梯度进行反向传播来调整图像元素的修改成本, 提高隐写的安全性. 相对于最小化失真函数的隐写方式, 编码-解码网络在隐写中的应用也非常广泛. 2017 年, Hayes 等人[32]提出基于 SteGAN 的隐写模型将编码-解码网络运用到隐写领域, 但该隐写方法所产生的隐写图像的视觉质量不高. Wang 等人[33]在此基础上提出了 SsteGAN 模型, 采用隐写器与判别器的对抗训练降低隐写图像与载体图像之间的差异, 提高隐写图像质量. 2019 年, Zhang 等人[34]年提出了 SteganoGAN 隐写模型, 可以在载体图像中隐藏最高可达到 4.4 bpp 的二进制比特数据. 编码-解码网络不仅可以将二进制比特数据或者文本信息隐藏到载体图像中, 甚至可以隐藏具有较高分辨率的彩色或灰度图像. 2017 年, Baluja 等人[35]首次提出以图藏图的深度学习隐写网络, 实现在彩色图像中隐藏尺寸相同大小的彩色图像, 但隐写图像存在颜色失真, 并且可以从载体图像的残差中提取秘密图像的残影, 算法安全性较低. Wu 等人[36]提出了一种新的隐写网络模型 StegNet, 该模型在编码网络中加入残差连接[37], 并在损失函数中增加了均方差损失, 提高了隐写图像质量, 但仍不能解决隐写图像颜色失真的问题. 2019 年, Duan 等人[38]利用全卷积网络构建了一个类似 U-Net 结构的编码网络. 将两幅尺寸大小相同的彩色图像拼接后输入编码网络生成隐写图像, 而通过解码网络可以提取出秘密图像, 并恢复与原始图像相似的载体图像. Baluja 等人则以增加隐写容量为目标, 在 2019 年提出了“一图藏多图”的图像隐写改进模型[39]. 该模型不仅能够隐藏、提取出两幅秘密图像, 而且能够混淆隐写图像与载体图像的残差信息, 提升了隐写图像的视觉效果, 但该方法抵抗隐写分析检测的能力较差. Atique 等人[40]提出了一种单通道灰度图像隐写模型, 其利用编码-解码网络将秘密图像嵌入到载体图像中, 但该模型生成的隐写图像仍然存在显示失真的问题, 视觉质量不高. 随后, Zhang 等人[41]提出 ISGAN 模型隐写模型, 将灰度图像仅隐藏在彩色载体图像的 Y 通道中. 与 Atique 等人相比, 隐写图像没有明显的颜色失真且 PSNR 值提升了 2 dB 左右, 但提取出的灰度图像 PSNR 值却下降了 3 dB 左右. 2020 年, Fu 等人[42]提出了一种称为 HIGAN 的隐写模型, 其中编码器基于 ResNet 的网络构建, 并在模型架构中加入隐写分析模型进行对抗训练, 但其隐写图像与提取的秘密图像的 PSNR、SSIM 都比较低. 总之, 以图藏图隐写模型通过牺牲隐写的部分安全性以达到隐写容量的提升, 但在具体应用中仍存在隐写图像颜色失真、细节信息损失较大、解码网络无法完全实现秘密图像的提取等问题.

综上, 载体图像生成式隐写和真实图像嵌入式隐写在秘密信息嵌入的过程中都需要对载体图像进行一定的修改, 使得图像的视觉质量、统计特性不可避免地受到影响, 让隐写分析检测有迹可循. 无载体图像隐写技术因不需要对图像修改, 因而具有天然的抗隐写分析能力. 2018 年, Hu 等人[43]提出了采用 DCGAN 的无载体隐写网络模型. 首先构建随机噪声与二进制秘密信息间的映射关系, 然后, 通过生成器将噪声作为输入生成隐写图像, 接收方

采用提取网络恢复原始噪声序列, 根据映射关系恢复出秘密信息. 该模型生成的隐写图像没有嵌入和修改的痕迹, 不会被攻击方发现, 但存在隐写容量较小、生成的隐写图像不够真实以及秘密消息不能完全正确提取等问题. 受 Hu 等人[43]方法的启发, Li 等人[44]提出了一种同时训练隐写图像生成器和秘密信息提取器的无载体隐写模型, 并使用 WGAN-GP 代替 DCGAN 提高隐写图像的视觉质量. 2019 年, Zhu 等人[45]在 Hu 等人[43]的基础上提出基于正交生成对抗网络 (O-GAN) 的无载体图像隐写方法, 通过对隐写图像的特征添加约束, 建立了噪声向量与特征之间的映射, 提高了模型训练和秘密信息提取的稳定性. Zhang 等人[46]利用训练好的 CycleGAN 网络先对噪声生成的图像风格化, 然后对风格化后的图像进行恢复, 并将噪声向量作为输出, 最后接收方通过相同的映射关系实现噪声向量与秘密信息的转换. 尽管风格迁移后的图像具有更高的安全性, 但此方法提取秘密信息的准确度有待提高. Meng 等人[47]采用目标检测方法搜索图像中的安全区域实现自适应无载体隐写, 进一步提升了无载体隐写的安全性和鲁棒性. 同时, Liu 等人[48]提出了一种基于生成对抗网络 (ACGAN) 的无载体图像隐写方法, 该方法利用隐写信息代替生成对抗网络的类标签作, 将隐写信息和噪声作为 ACGAN 网络的输入生成隐写图像. 提取秘密信息时, 通过 ACGAN 的判别器提取图像的类别标签, 最后根据类标签映射关系恢复秘密信息. 由于类别标签数量有限, 该方法的隐藏容量较低. 总体来讲, 基于无载体的隐写算法可以生成视觉质量较高的隐写图像, 但目前仍存在隐写信息恢复能力较弱, 嵌入容量较低的问题.

近年来, 基于深度学习的隐写分析模型也取得了巨大的进展. 与传统隐写分析方法相比, 基于深度学习的隐写分析方法通过采用不同网络结构学习图像的根本特征, 以区分载体图像和隐写图像. 2015 年, Qian 等人[13]首次将深度学习与隐写分析结合, 提出了基于卷积神经网络的隐写分析网络 (GNCNN), 其利用卷积神经网络提取高维特征, 并使用高斯函数作为激活函数构建隐写分析模型. 2016 年, Xu 等人[14]提出了 Xu'Net 隐写分析模型. 与 GNCNN 相比, 模型中增加了 ReLU 激活函数、ABS 函数和批量归一化层 (BN), 以提高隐写分析检测精度, 该模型可应用于空域和 JPEG 领域. 2017 年, Ye 等人[15]提出 Ye'Net 隐写分析模型, 其采用 SRM 中的 30 个高通滤波器初始化数输入信息, 并引入选择信道感知 (SCA) 来优化模型, 并取得了优于 SRM[12]隐写分析算法的性能. 2019 年, Boroumand 等人[16]提出了一种称为 SRNet 的隐写分析模型, 该模型采用具有快捷连接的残差层构建, 可以在训练过程中重新使用特征向量, 并取得更好的隐写分析效果. 然而, SRNet 存在梯度消失现象, 导致网络模型训练困难. 由以上分析可知, 隐写分析的快速发展, 为隐写技术的研究提出了新的课题.

3 基于 U-Net 结构的生成式多重对抗隐写算法

3.1 生成对抗网络基础模型

生成对抗网络的思想来源于二人零和博弈, 如图 1 所示, 网络由 1 个生成器和 1 个判别器组成, 生成器用来接收输入噪声生成图像, 判别器用来区分真实图像与生成图像. GAN 通过网络之间的竞争来生成与原始图像尽可能一致的数据分布. 在生成对抗网络中, 判别器和生成器同时被训练, 二者不断地进行对抗博弈, 通过网络之间的竞争来生成所需数据分布, 最终达到纳什平衡, 即判别器判断生成图像与真实图像的结果一致.

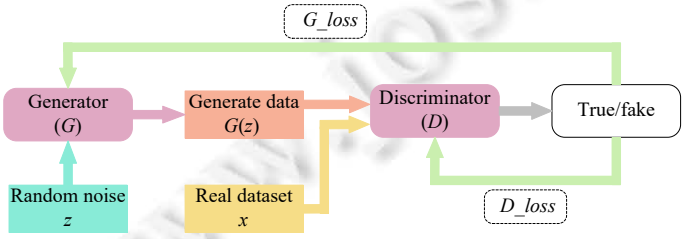

图 1 生成对抗网络模型

生成对抗网络 GAN 的训练过程可以看作是一个极小极大 (minimax) 问题的优化过程. 判别器 D 尽可能地正确判别输入的数据是来自真实数据样本 x 还是来自生成器生成样本 $G(z)$. 生成器 G 则去学习真实数据集样本的数据分布特征, 并尽可能使生成数据样本 $G(z)$ 在判别器 D 上有着真实数据 x 在 D 上一致的表现.

生成器和判别器相互对抗并迭代优化, 使得判别器 D 和生成器 G 的性能都能够不断地提升, 最终使 D 与 G 二者之间达到纳什平衡. 即, 当 D 无法正确判别输入样本数据是来源于真实数据还是生成数据时 (也即判别器判定来自真实数据集和生成数据集的概率都是 50%), 则认为生成器 G 已经学到了数据集样本的数据分布. 因此, GAN 的训练过程实际上是解决以下优化问题:

$$\min_G \max_D (D,G) = \sum\nolimits_{x\sim pata(x)} [\log D(x)] + \sum\nolimits_{x\sim pz(z)} [\log(1-D(G(z))] \tag{1}$$

其中, $D(x)$ 是真实图像的判别概率, $G(z)$ 为输入噪声 z 产生的生成图像, $D(G(z))$ 是生成图像的判别概率.

通过彼此交替训练, 生成对抗网络中 G 与 D 实现公式 (1) 中的优化: 在每个随机梯度优化的迭代过程中, 首先对 D 进行梯度上升训练, 然后对 G 进行梯度下降训练. 假定用 θ 表示神经网络 M 的参数, 那么更新规则为:

保持 G 不变, 通过 $\theta_D \leftarrow \theta_D + \gamma D \nabla_D L$ 更新 D:

$$\nabla_D L = \frac{\partial}{\partial \theta_D} \left\{ E_{x\sim pata(x)} [\log D(x,\theta_D)] + E_{x\sim pz(z)} [\log D(G(z,\theta_G))] \right\} \tag{2}$$

保持 D 不变, 通过 $\theta_G \leftarrow \theta_G + \gamma D \nabla_G L$ 更新 G:

$$\nabla_G L = \frac{\partial}{\partial \theta_G} \left\{ E_{z\sim pz(z)} [\log(1-D(G(z,\theta_G),\theta_D))] \right\} \tag{3}$$

3.2 生成式多重对抗隐写算法

Goodfellow 等人[17]提出的生成式对抗网络 (GAN) 模型, 为深度学习和信息隐藏的结合提供了契机. 本算法基于生成对抗网络基础模型, 在网络结构和损失函数两个方面进行改进, 提出一种基于 U-Net 结构的生成式多重对抗隐写算法. 区别于经典的 GAN 网络将噪声数据生成随机图像, 本算法以真实图像为基础, 可控的生成更适合信息嵌入的载体图像, 增强信息隐藏能力. 生成图像在保持和真实图像高度相似的同时, 实现更好的信息隐写性能. 本模型由生成对抗网络 GAN、隐写器网络 SN、隐写分析器优化网络 SON 和隐写分析对抗网络 SAN 这 4 个子网络构成. 其中, 生成对抗网络包含两个部分, 生成器通过前向传播将输入的真实图像转换为适合嵌入秘密信息的高质量载体图像; 判别器区分生成图像与真实图像, 对抗提升生成图像的视觉质量; 隐写器网络使用特定的隐写算法, 对载体图像嵌入秘密信息; 隐写分析器优化网络对当前最优的隐写分析器参数进行优化, 提高其隐写分析判别能力; 隐写分析对抗网络采用优化后的隐写分析器对生成载体图像的隐写性能进行评价, 并向生成对抗网络提供隐写分析损失, 促使生成对抗网络迭代生成更适合信息嵌入的载体图像.

本研究在生成对抗网络中设计基于 U-Net 结构的生成器网络模型, 将编码网络的降维层与解码网络的升维层之间进行连接, 充分利用真实图像各个维度的特征信息, 生成更近似于真实图像的载体图像, 增强信息隐藏能力. 如图 2 中生成对抗网络所示, 实验中首先通过基于 U-Net 结构的生成器 G 生成高质量的载体图像 X_G, 将 G 生成的载体图像 X_G 与真实图像 X 输入到判别网络 D 进行判别, 并计算真实图像 X 和生成图像 X_G 间的损失函数 D_loss, 同时计算生成图像 X_G 与真实图像 X 的均方误差, 使用像素空间最小均方差损失 MSE_loss 共同促进生成网络生成拥有更多语义信息的、高视觉质量的载体图像. 其次, 如图 2 中隐写分析器优化网络 (steganalysis optimization network, SON) 所示, 本方案将真实图像通过隐写器网络 (steganography network, SN) 生成隐写图像 X_S, 然后将真实图像 X 与其隐写图像 X_S 共同输入到隐写分析器优化网络 SON 优化其参数, 提高隐写分析器针对特定类型图像数据类型的判别能力. 隐写器网络 SN 分别采用 S-UNIWARD、ASDL-GAN、UT-SCA-GAN 等隐写嵌入算法对图像嵌入随机数据作为秘密信息. 再次, 在图 2 隐写分析对抗网络 (steganalysis adversarial network, SAN) 中, 采用隐写分析器优化网络 SON 中优化的隐写分析器 SD 构建隐写分析对抗网络, 并把由生成器 G 生成的载体图像 X_G 经过隐写器 SN 嵌入秘密信息后输出隐写图像 X_G_S, 将生成图像 X_G 和隐写生成图像 X_G_S 共同输入到优化后的隐写分析器中, 对抗提升生成图像的隐写性能. 隐写分析对抗网络 SAN 对生成载体图像信息隐藏的能力进行评估, 并为生成网络提供梯度损失 SDO_loss, 优化生成图像抵抗隐写分析的能力. 最后, 为提升生成图像的收敛速度和隐写能力, 将生成对抗网络模型的鉴别损失 D_loss、均方差损失 MSE_loss 和隐写分析对抗网络产生的损失 SDO_loss 加权叠加作为生成网络 G 的总损失, 循环优化其生成图像视觉质量和隐写能力, 保障网络快速稳定收敛.

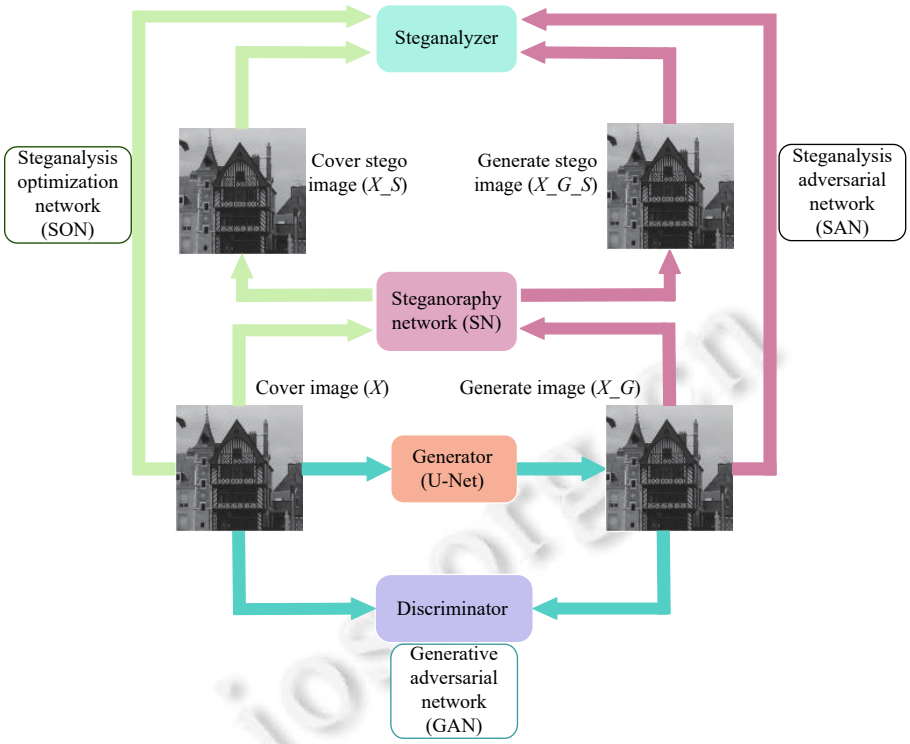

图 2 生成式多重对抗隐写算法结构图

本方案在实验过程中首先通过生成对抗网络生成具有较高视觉质量的图像并计算出损失函数 *D_loss*, 计算生成图像与真实图像的像素空间最小均方差损失 *MSE_loss*, 并采用不同的隐写算法嵌入随机秘密信息; 然后通过隐写分析器优化网络 SON 提升隐写分析器的性能, 再采用优化后的隐写分析器 *SD* 构建隐写分析对抗网络 SAN 对生成图像 *X_G* 及其隐写图像 *X_G_S* 进行判别, 计算出损失函数 *SDO_loss*; 最后, 将 *D_loss*、*MSE_loss* 和 *SDO_loss* 加权相加后作为生成网络 *G* 的损失函数. 通过生成对抗网络, 隐写分析器优化网络以及隐写分析对抗网络的连续迭代, 不断优化生成器、判别器和隐写分析网络的性能, 保障网络快速稳定收敛. 最终生成具有较高视觉图像质量的、适合信息隐写的生成图像, 提高隐蔽通信能力.

算法 1. 生成式多重对抗隐写算法训练步骤.

Require:

X: 真实图像

1. 采用基于 U-Net 的生成器, 生成与真实图像 *X* 一致分布的载体图像 *X_G*.
2. 将生成图像 *X_G* 与真实图像 *X* 输入到判别器 *D* 中, 对抗优化生成图像质量. 记录判别损失 *D_loss*, 并计算生成图像与真实图像的像素空间均方差损失 *MSE_loss*.
3. 将真实图像 *X* 输入到隐写器中输出其隐写图像 *X_S*, 将 *X* 和 *X_S* 分别输入到隐写分析器优化网络 SON 中, 训练优化隐写分析器网络模型参数.
4. 将载体图像 *X_G* 输入到隐写器中输出为隐写图像 *X_G_S*, 将 *X_G* 和 *X_G_S* 输入到采用优化后的隐写分析器构建的隐写分析对抗网络 SAN 中, 验证生成隐写图像的抗隐写分析能力, 并记录隐写分析损失 *SDO_loss*.
5. 动态加权叠加判别损失 *D_loss*、均方差损失 *MSE_loss* 和隐写分析损失 *SDO_loss*, 作为生成网络 *G* 的联合损失函数.
6. 循环优化更新生成器、判别器、隐写分析网络的参数, 分别最小化 *D_loss*、*MSE_loss* 和 *SDO_loss*, 保障生成式多重对抗隐写网络快速稳定收敛, 生成高质量适合隐写的载体图像.

本算法通过生成网络 G 与判别网络 D、隐写分析对抗网络 SAN 间的多重对抗, 实现生成式多重对抗隐写网络的快速稳定收敛, 并生成视觉质量更高、更适合嵌入秘密信息的载体图像. 生成对抗网络 GAN 通过区分生成图像与真实图像, 对抗提升生成图像的质量. 其使用真实图像 X 与生成图像 X_G 进行训练, 优化的目标函数为:

$$\max_{D} f = E_{x\sim pata(x)}[\log D(x)] + E_{x\sim pata(x)}[\log(1-D(G(x)))] \tag{4}$$

其中, $D(x)$ 代表的是真实图像的概率, $G(x)$ 为输入真实图像 X 产生的生成图像.

隐写分析器优化网络 SON 判断输入图像是否含有秘密信息, 使用真实图像 X 及其隐写图像 X_S 进行训练, 其优化的目标函数为:

$$\max_{SD} f = E_{x\sim pata(x)}[\log SD(x)] + E_{x\sim pata(x)}[\log(1-SD(S(x)))] \tag{5}$$

其中, $SD(x)$ 代表的是隐写分析网络判断为未嵌入秘密信息图像的概率, $S(x)$ 为在隐写器 SN 中输入真实图像 x 产生的隐写图像 X_S.

隐写分析对抗网络 SAN 对抗生成器网络, 产生隐写分析损失 SDO_loss, 促使生成器输入真实图像 X, 生成与真实图像分布一致的更适合秘密信息嵌入的载体图像 X_G, 使其能够欺骗判别模型与隐写分析模型, 其优化的目标函数为:

$$\min_{G} f = E_{x\sim pata(x)}[\log(1-D(G(x)))] + E_{x\sim pata(x)}[\log SD(G(x))] + E_{x\sim pata(x)}[\log(1-SD(S(G(x))))] \tag{6}$$

其中, $G(x)$ 为输入真实图像 x 产生的生成图像 X_G, $D(G(x))$ 代表的是生成图像判别为真实图像的概率, $SD(G(x))$ 代表的是隐写分析网络判断为未嵌入秘密信息图像的概率, $S(G(x))$ 为在隐写器 SN 中输入生成图像 X_G 产生的隐写图像 X_G_S.

模型优化总目标函数为:

$$\begin{aligned}\min_{G}\max_{D}\max_{SD} f =& \alpha\left\{E_{x\sim pata(x)}[\log D(x)] + E_{x\sim pata(x)}[\log(1-D(G(x)))]\right\}\\ &+(1-\alpha)\left\{E_{x\sim pata(x)}[\log SD(G(x))] + E_{x\sim pata(x)}[1-SD(S(G(x)))]\right\}\\ &+E_{x\sim pata(x)}[\log SD(x)] + E_{x\sim pata(x)}[\log(1-SD(S(x)))]\end{aligned} \tag{7}$$

3.2.1 基于 U-Net 结构的生成器网络

U-Net 是一种基于编码器-解码器结构的卷积神经网络模型, 其采用下采样-上采样的 U 型结构实现原始图像的重构. 在 U 型网络的左侧收缩路径 (constracting path) 上, 不断地堆叠卷积并进行最大池化操作, 感受野不断增强, 输出的特征图数目逐渐增多, 提取原始图像底层特征; 而在其右侧的展开路径 (expansive path) 上, 通过上采样后接卷积操作逐渐逼近原始图像. 同时, 在同级别的特征图之间使用跳跃连接融合深层和浅层的特征信息, 使整个网络可以利用编码器生成和原始图像一致的图像. 本方案采用 U-Net 结构将编码阶段不同感受野的细节信息传递到解码阶段, 协助生成器生成高质量的载体图像. 在基于 U-Net 结构的深度学习网络框架中, 下采样用来保存环境信息, 而上采样的过程是结合下采样所保存的各层信息, 并结合上采样的输入信息来还原细节信息, 逐步构建高精度真实图像. U-Net 结构通过对深层信息和浅层信息的拼接, 充分利用真实图像的局部和整体特征信息, 保障生成图像具有优秀的视觉表现. 同时, 相对于传统的生成对抗网络 (如: SGAN、SSGAN) 只能生成随机载体图像, 本方案通过设计基于 U-Net 结构的生成对抗网络可以有目的地生成具有特定内容的图像. 从而, 根据隐蔽通信内容与真实图像纹理结构特征, 生成更适合秘密信息嵌入的高质量载体图像, 增强信息隐藏能力.

本方案设计 U-Net 结构的生成器共包含 16 个数据处理单元, 前 8 组是下采样编码, 每组做包括步长为 2 的用于下采样的卷积层 (卷积核为 3×3)、批处理归一化层 (BN) 和 ReLU 激活函数, 9 到 15 组是扩展路径, 每组包括步长为 2 的反卷积层 (卷积核为 5×5)、批处理归一化层 (BN) 和 ReLU 激活函数, 第 16 组包括步长为 2 的反卷积层 (卷积核为 5) 和 ReLU 激活函数、Sigmoid 激活函数. 为了实现像素级学习并促进反向传播, 本方案在图像的解码阶段通过跳接方式将第 i 层和第 16–i 层的特征图拼接起来作为到 16–i+1 层的输入. 生成器的网络结构如图 3 所示, 网络模型具体参数如表 1 所示.

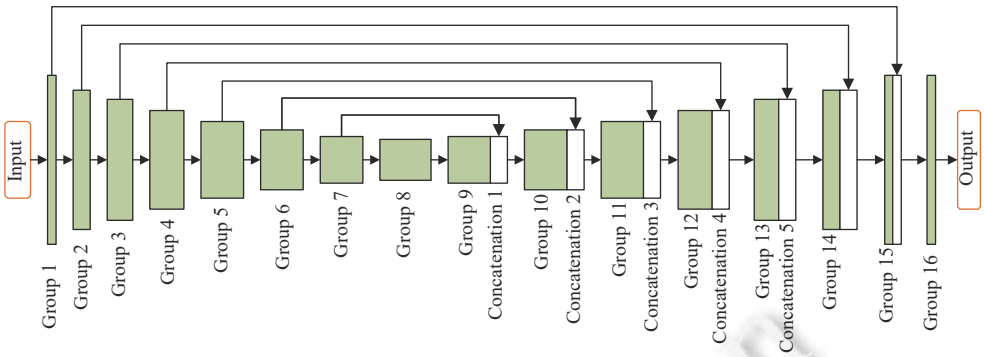

图 3 基于 U-Net 的生成器网络结构

表 1 生成器网络结构参数

Group/Layer	Process	Convolution/Deconvolution kernels	Output size
Input	—	—	1×(256×256)
Group 1	Convolution-BNormalization-ReLU	16×(3×3)	16×(128×128)
Group 2	Convolution-BatchNormalization-ReLU	32×(3×3)	32×(64×64)
Group 3	Convolution-BatchNormalization-ReLU	64×(3×3)	64×(32×32)
Group 4	Convolution-BatchNormalization-ReLU	128×(3×3)	128×(16×16)
Group 5	Convolution-BatchNormalization-ReLU	256×(3×3)	256×(8×8)
Group 6	Convolution-BatchNormalization-ReLU	256×(3×3)	256×(4×4)
Group 7	Convolution-BatchNormalization-ReLU	512×(3×3)	512×(2×2)
Group 8	Convolution-BatchNormalization-ReLU	512×(3×3)	512×(1×1)
Group 9	Deconvolution-BatchNormalization-ReLU	512×(5×5)	512×(2×2)
Concatenation 1	Concate the feature map from G7 and G9	—	1024×(2×2)
Group 10	Deconvolution-BatchNormalization-ReLU	256×(5×5)	256×(4×4)
Concatenation 2	Concate the feature map from G6 and G10	—	512×(4×4)
Group 11	Deconvolution-BatchNormalization-ReLU	256×(5×5)	256×(8×8)
Concatenation 3	Concate the feature map from G5 and G11	—	512×(8×8)
Group 12	Deconvolution-BatchNormalization-ReLU	128×(5×5)	256×(16×16)
Concatenation 4	Concate the feature map from G4and G12	—	256×(16×16)
Group 13	Deconvolution-BatchNormalization-ReLU	64×(5×5)	64×(32×32)
Concatenation 5	Concate the feature map from G3 and G13	—	128×(32×32)
Group 14	Deconvolution-BatchNormalization-ReLU	32×(5×5)	32×(64×64)
Concatenation 6	Concate the feature map from G6 and G10	—	64×(64×64)
Group 15	Deconvolution-BatchNormalization-ReLU	16×(5×5)	16×(128×128)
Concatenation 7	Concate the feature map from G5 and G11	—	32×(128×128)
Group 16	Deconvolution-Sigmoid-ReLU	1×(5×5)	1×(256×256)

3.2.2 判别器网络

本方案中判别器 D 用来区分生成图像与真实图像, 促进生成器生成的载体图像包含更多的语义信息, 从而生成更适合信息嵌入的高质量载体图像. 如图 4 所示, 本方案设计的判别器模型包含 8 个数据处理单元, 每个处理单元包括卷积层 (卷积核为 3×3) 和批处理归一化层 (BN), 并选用 Leaky-ReLU 作为激活函数, 加强网络对负信息的响应, 以避免整个网络中的最大池化. 随着网络层数加深, 特征图个数不断增加, 将特征输入到全连接层并得到特征向量, 选择 Sigmoid 作为激活函数, 输出检测为真实图像的概率. 判别器的网络结构如后文图 4 所示, 网络模型具体参数如后文表 2 所示.

3.2.3 隐写分析网络

隐写分析网络用来判别输入图像是否含有秘密信息, 随着深度学习的发展, 隐写分析与深度学习结合对隐写

算法构成了极大的威胁. 本研究采用优化后的隐写分析器构建隐写分析对抗网络 SAN, 实现生成图像隐写能力的提升. 如图 2 中隐写分析优化网络 SON 所示, 首先将真实图像 X 及其隐写图像 X_S 作为隐写分析器的输入, 优化隐写分析器 SD 的性能参数, 提升隐写分析器的判别能力. 然后, 利用优化后的隐写分析器 SD 构建隐写分析对抗网络 SAN, 对生成的载体图像 X_G 与其隐写图像 X_G_S 进行评价, 记录其判别损失 SDO_loss, 并作为生成器优化用总损失函数的组成部分, 对抗生成更适合隐写的载体图像, 提高隐写图像 X_G_S 抗隐写分析检测的能力. 本方案选择 Xu'Net 与 Ye'Net 两种当前优秀的基于深度学习的隐写分析器网络进行优化, 并构建隐写分析网络, 为生成对抗网络提供迭代损失 SDO_loss, 对抗提升生成的载体图像的信息隐藏能力.

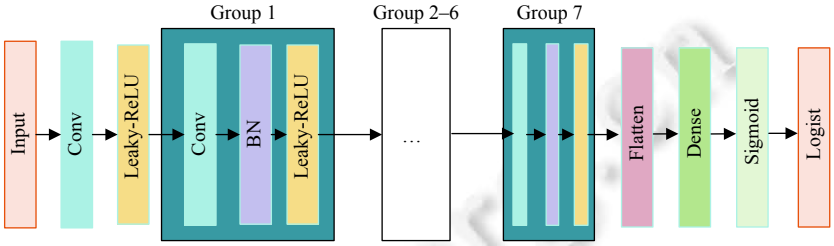

图 4 判别器网络结构

表 2 判别器网络结构参数

Group/Layer	Process	Convolution/Deconvolution Kernels	Output size
Input	—	—	1×(256×256)
Conv	Convolution-Leaky-ReLU	64×(4×4)	64×(128×128)
Group 1	Convolution-BatchNormalization-Leaky-ReLU	128×(4×4)	128×(64×64)
Group 2	Convolution-BatchNormalization-Leaky-ReLU	256×(4×4)	256×(32×32)
Group 3	Convolution-BatchNormalization-Leaky-ReLU	512×(4×4)	512×(16×16)
Group 4	Convolution-BatchNormalization-Leaky-ReLU	512×(4×4)	512×(8×8)
Group 5	Convolution-BatchNormalization-Leaky-ReLU	512×(4×4)	512×(4×4)
Group 6	Convolution-BatchNormalization-Leaky-ReLU	512×(4×4)	512×(2×2)
Group 7	Convolution-BatchNormalization-Leaky-ReLU	512×(4×4)	512×(1×1)
Group 8	Flatten-Dense-Sigmoid	—	(1, 1)

3.3 多重对抗网络损失

生成对抗网络 GAN 的主要思想来源于博弈论中零和博弈, 其通过生成器 G (generator) 和判别器 D (discriminator) 不断博弈, 进而使生成器 G 学习到数据的分布. 训练过程中, 生成器 G 通过梯度更新不断缩小生成图像和真实图像间的差异, 尽量生成和目标图像分布一致的图像, 而判别器 D 的目标就是尽量辨别出生成器 G 生成图像的真伪, 最终的平衡点即纳什平衡点. 生成对抗网络功能强大, 但有时会存在收敛速度慢、训练不稳定的问题.

生成对抗网络中生成器 G 的梯度是由判别器 D 通过损失函数产生的, 判别器通过损失函数计算生成图像和原始图像的一致性程度, 提供生成器迭代优化的梯度, 生成对抗网络的收敛速度和稳定性, 以及生成图像的质量很大程度上是由损失函数决定的. 设计性能优良的损失函数, 是保障生成对抗网络稳定收敛和生成高质量图像的前提. 因而, 本研究在生成式多重对抗隐写网络中动态叠加判别损失 D_loss、均方差损失 MSE_loss 和隐写分析损失 SDO_loss 作为生成网络 G 的总损失, 保障生成对抗网络快速稳定收敛, 并形成更适合信息隐写的高质量载体图像.

在基于 U-Net 结构的生成式多重对抗隐写网络中, 生成器 G 学习真实图像的数据分布特征, 并通过判别器 D 判别生成图像与真实图像一致的概率, 输出概率越大, 生成图像越倾向于被认定为真实图像. 判别器 D 的训练目标是: 当输入为生成图像时, 期望输出概率接近为 0; 当输入真实图像时, 期望输出概率接近为 1. 判别器 D 的损失函数可以表示为:

$$L_D = -\sum_{i=1}^{2} x_i' \log(x_i) \tag{8}$$

其中, x_1, x_2 分别是真实图像与生成载体图像的经过判别器 D 后 Softmax 层的输出, x'_1, x'_2 分别是输入的真实图像与生成图像对应的标签. 隐写分析对抗网络 SAN 通过检测输入图像是否含有秘密信息, 输出 2 个类标签上产生的分布. 隐写分析对抗网络 SAN 的训练目标是: 当输入图像为未隐写图像时, 输出为 [0, 1] 分布 (更靠近 0), 输入图像为隐写图像时, 输出为 [1, 0] 分布 (更靠近 1). 因此实验中隐写分析对抗网络 SAN 的损失函数为:

$$L_{SD} = -\sum_{i=1}^{2} z'_i \log(z_i) \tag{9}$$

其中, z_1, z_2 是原始图像与隐写图像经过隐写分析器 SD 后 Softmax 层的输出, z'_1, z'_2 分别是输入的真实图像与隐写图像对应的标签.

本研究中通过生成器 G 与判别器 D 和隐写分析对抗网络 SAN 进行对抗, 并试图迷惑判别器, 使其输出错误的判断. 生成器的损失为公式 (8)、公式 (9) 两个损失的反向加权叠加. 同时, 为了保证生成图像的视觉质量, 在总损失中引入像素空间均方差损失 *MSE_loss*, 促进生成的载体图像获得更高的 *PSNR* 数值. 因此生成器 G 的总损失为:

$$L_G = -\alpha \cdot L_D - \beta \cdot L_{SD} + \lambda \cdot \sum_{i=1}^{n} \frac{\left(y_i - y'_i\right)^2}{n} \tag{10}$$

其中, y_i 是真实图像像素点的值, y'_i 是对应位置的生成图像像素点的值, n 是图像像素点数量.

综上, 本算法通过构建基于判别器判别损失 *D_loss*、均方差损失 *MSE_loss* 和隐写分析损失 *SDO_loss* 的加权组合作为生成网络 G 的总损失 L_G, 并通过试验动态选择不同损失的权值, 保障生成式多重对抗网络快速稳定收敛, 并产生更适合信息嵌入的高质量载体图像.

4 实验及结果分析

4.1 实验设计

4.1.1 数据集

为验证基于 U-Net 结构的生成式多重对抗隐写网络的性能, 实验中选择隐写领域最常用的 BOSS Base 数据集开展隐写和隐写分析对抗实验研究. 为了提高训练效率, 实验中将 512×512 分辨率的 BOSS Base 图像采用 Matlab 中 Resize 指令 (默认参数) 降维为 256×256 分辨率的图像. 选取 10 000 张 BOSS Base 图像用于实验研究, 其中随机选取 8 000 张图像作为训练集参与训练, 剩余 2 000 张图像用于实验结果测试与性能验证.

4.1.2 实验参数设计

实验采用学习率为 0.000 1 的 Adam 优化器训练网络模型 (α=0.5, β=0.99). 在训练阶段, 每次迭代使用 32 张真实图像 X 作为输入, 生成相应的载体图像 *X_D*; 将生成图像 X 与真实图像 *X_G* 输入到判别器 D 中, 对抗提升生成图像的视觉质量; 然后, 将真实图像 X 采用隐写算法嵌入 0.4 bpp 秘密信息后生成隐写图像 *X_S*, 并将 X 和 *X_S* 同时输入隐写分析器优化网络 SON 中, 优化隐写分析器网络模型参数; 另一方面, 基于训练好的隐写分析器 SD 构建隐写分析对抗网络 SAN, 将生成的载体图像 *X_G* 与隐写生成图像 *X_G_S* 输入到隐写分析网络中进行检验, 并记录隐写分析损失 *SDO_loss*; 最后, 将对抗损失 *D_loss* 和隐写分析判别损失 *SDO_loss*, 以及生成图像与真实图像的均方差损失 *MSE_loss* 加权叠加作为新的损失, 进入下一轮生成图像优化过程, 通过对抗迭代优化生成图像的隐写能力和视觉质量. 训练结束后, 再从 BOSS Base 数据集中选择另外 2 000 张真实图像, 基于训练好的生成式多重对抗隐写网络产生适合信息隐写的高质量载体图像, 并分别采用 0.4 bpp 的 S-UNIWARD、ASDL-GAN、UT-SCA-GAN 等嵌入算法将随机信息嵌入到生成图像中, 构建基于 U-Net 结构的生成式多重对抗隐写算法测试集, 对生成隐写图像进行质量和隐写性能评估.

4.1.3 计算资源配置

由于生成对抗网络主要包含针对原始图像的卷积运算、反向传播所需的梯度运算、损失函数计算所需的浮点运算等运算, 以生成目标图像. 因而, 生成式多重对抗隐写网络对 CPU 和内存的占用不高, 网络通过 TensorFlow-GPU 框架把大量计算交给擅长浮点运算的 GPU 处理, 以加速网络的训练过程. 以本方案所提出的基于 U-Net 结

构的生成式多重对抗隐写算法为例, 模型本身包含生成对抗网络 GAN、隐写器网络 SN、隐写分析器优化网络 SON 和隐写分析对抗网络 SAN, 系统有 33 个卷积层, 16 个反卷积层, 2 个全连接层, 48 个 BatchNorm 层, 模型参数攻击 3 709 万个左右, 每次训练输入 8 000 幅 256×256 的图像, 整个网络的运算规模在数十亿次以上, 算法对浮点运算的需求较高. 因而, 实验设备采用戴尔 T630 4 核 i7 处理器, 主频 3.2 GHz, 内存 128 GB, GPU: RTX3090, 显存 24 GB, 实验环境使用 Python 3.7, TensorFlow 2.4.0. 本系统采用 RTX 3090 GPU 运算卡实现生成式多重对抗隐写算法, 主要是因为 RTX 3090 GPU 运算卡拥有 24 GB 显存, 10496 核心处理器, 384 bit 位宽, 显存速度 19.5 Gb/s 在神经网络运算等方面具有优异的性能. 实际试验过程中, 系统运行的平均 CPU 占用约为 20%, 占用内存峰值进位 4.1 GB, 平均 GPU 占用率达到 91.6%, 每个 epoch 运算的时间 100 s 左右, 可以实现本算法的稳定运行, 生成适合信息隐写的高质量载体图像, 增强信息隐藏能力.

4.1.4 评价指标

为客观评价生成图像的视觉质量, 实验中采用 *PSNR* 与 *SSIM* 两个经典指标对生成图像质量进行评价. *PSNR* 通过计算两图像对应像素之间的变化来评估生成图像与真实图像的相似度, *PSNR* 值越大, 生成图像与真实图像相似度越高、失真越小, 反之亦然. *PSNR* 计算公式如下:

$$MSE=\frac{1}{W\times H}\sum_{i=1}^{W}\sum_{j=1}^{H}\left(x_{i,j}-y_{i,j}\right)^2 \tag{11}$$

$$PSNR=10\cdot\log_{10}\left(\frac{(2^n-1)^2}{MSE}\right) \tag{12}$$

其中, W、H 是图像的宽与高, $x_{i,j}$ 与 $y_{i,j}$ 表示两幅图像相同空间位置的像素点, n 是像素点数量.

SSIM 也是一种衡量两幅图像相似度的指标, 它通过亮度、对比度和结构 3 个方面共同衡量图像的质量. *SSIM* 通过下述公式计算使用:

$$SSIM(X,Y)=l(X,Y)\cdot c(X,Y)\cdot s(X,Y) \tag{13}$$

$$l(X,Y)=\frac{2\mu_X\mu_Y+C_1}{(\mu_X)^2+(\mu_Y)^2+C_1} \tag{14}$$

$$c(X,Y)=\frac{2\sigma_X\sigma_Y+C_2}{(\sigma_X)^2+(\sigma_Y)^2+C_2} \tag{15}$$

$$s(X,Y)=\frac{\sigma_{XY}+C_3}{\sigma_X\sigma_Y+C_3} \tag{16}$$

其中, $l(\cdot)$, $c(\cdot)$, $s(\cdot)$ 和分别代表亮度, 对比度和结构. μ_X 和 μ_Y 是图像 X 和图像 Y 的像素平均值; σ_X 和 σ_Y 表示图像 X 和图像 Y 的标准差; $\sigma_X\sigma_Y$ 表示图像 X 和图像 Y 的协方差.

4.2 生成图像质量分析

实验中, 随着模型迭代次数的增加, 生成器生成的图像与真实图像相似度越来越高. 图 5 为实验中随机选择的 6 张生成图像和真实图像的对比, 第 1 列为真实图像、第 2 列为生成图像, 第 3 列和第 4 列分别是真实图像和生成图像的直方图分布. 由对比可知, 生成图像与真实图像直方图分布基本相同, 计算结果也表明生成图像和真实图像的 *PSNR* 值大于 42 dB 以上, 二者具有极高的相似度. 生成图像质量优于采用 SGAN[19]、SSGAN[21]等随机图像生成算法所生成的图像质量.

同时, 由于基于 U-Net 结构的生成式多重对抗隐写网络可以输出 256×256 的高质量图像, 本方案可用于生成高质量的大尺寸载体图像, 从而取得更大的信息嵌入容量和更好的视觉质量, 模型的实用性明显优于只能使用 CelebA 小尺寸图像集进行训练, 且只能生成随机图像的生成对抗网络算法 (如 SGAN[19]、SSGAN[21]等). 结果如图 6 所示.

图 7 为基于验证集所生成图像的 *PSNR* 与 *SSIM* 分布图. 其中横坐标为图像数量标签, 纵坐标为图像的 *PSNR* 与 *SSIM* 值. 由图 7 可知, 2 000 张生成图像的平均 *PSNR* 值大于 36.8 dB; 其中, 35.8% 以上的图像 *PSNR* 取值大于 40 dB, 11.2% 以上的生成图像 *PSNR* 取值大于 42 dB; 另一方面, 生成图像平均 *SSIM* 值大于 0.965; 其中, 35.2% 以

上的生成图像 *SSIM* 值大于 0.975, 10.7% 以上的图像 *SSIM* 值大于 0.980, 生成图像与真实图像相比具有很高的图像相似度和优秀的视觉质量. 实验证明结果表明, 基于 U-Net 结构的生成式多重对抗隐写网络可以输出高视觉质量的生成图像. 相较于基于原始噪声生成随机图像的生成对抗网络算法, 本算法具有生成目标可控, 生成图像尺寸大, 质量高等优点, 可以根据信息隐写需要生成内容可控的高质量载体图像, 增强信息隐藏能力, 有利于实现更优秀的隐蔽通信效果.

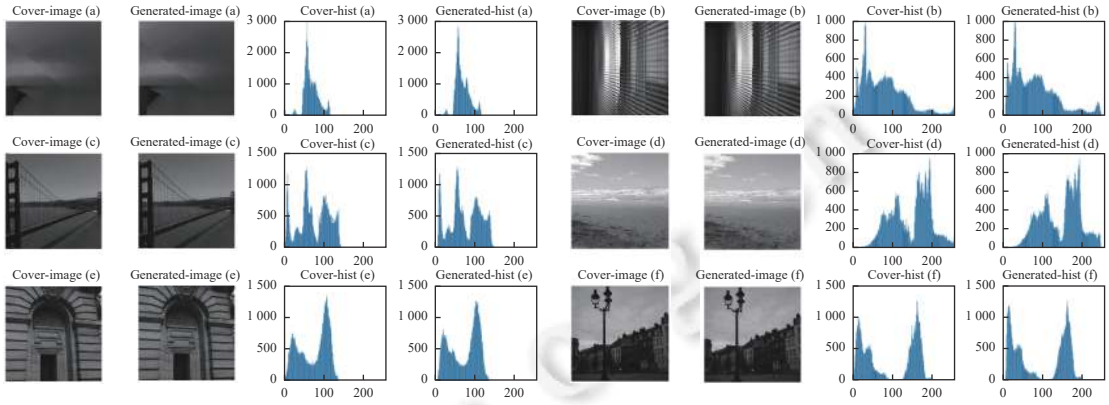

图 5 真实图像与生成图像及其直方图分布

(a) SGAN (CelebA 64×64) (b) SSGAN (CelebA 64×64) (c) 本文方案 (BOSS 256×256)

图 6 SGAN、SSGAN 基于 CelebA (64×64) 数据集的生成图像以及本方案基于 BOSS 数据集生成的图像

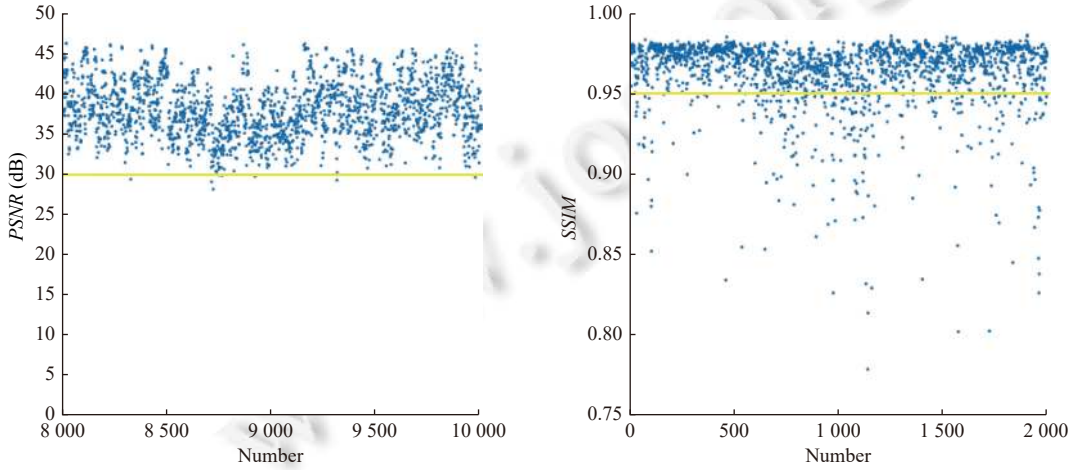

图 7 生成图像 *PSNR* 与 *SSIM* 分布图

4.2.1 不同权值对生成图像的影响

根据生成式多重对抗隐写网络的结构设计可知, 生成网络的总损失由生成网络判别损失 *D_loss*, 隐写分析网络损失 *SDO_loss*, 以及均方差损失 *MSE_loss* 这 3 部分组成, 3 种损失通过加权叠加作为下一轮载体图像对抗生成的损失函数. 因而, 3 种损失权重的分配对生成图像的质量也会产生较重要的影响.

表 3 展示了在公式 (10) 中不同 α, β 设置下, 运行 150 个 epoch 后达到的最优 *PSNR* 与 *SSIM* 值. 由表 3 可以看出, 在 α=0.4, β=0.6 时, 图像 *PSNR* 为 42.062、*SSIM* 为 0.9654; 在 α=0.3, β=0.7 时, 图像 *PSNR* 为 42.058、*SSIM* 为 0.9676; 在 α=0.2, β=0.8 时, 图像 *PSNR* 为 42.062、*SSIM* 为 0.9727; 而在 α=0.1, β=0.9 时, 图像 *PSNR* 达到了 42.821、*SSIM* 达到了 0.9744; Dynamic 列中最优 *PSNR* 取值可达 48.597、*SSIM* 取值达到 0.9873, 该列是基于 α, β 动态权重方案所取得的实验结果, 实验中首先令初始 α=0.4, β=0.6, 在模型训练的前 100 个 epoch 内参数 α 每个 epoch 降低 1%, 参数 β 每个 epoch 升高 1%, 100 个 epoch 后设定 α=0, β=1.0. 这样设计参数的原因是在生成对抗网络模型中训练 100 个 epoch 后, 判别器提供的梯度基本固定, 判别器损失 *D_loss* 对生成器优化的作用减弱. 为了促进生成更高质量的图像, 增强生成图像抗隐写分析检测能力, 在 100 个 epoch 后, 生成器不再与判别器对抗, 生成器损失函数中设置 α=0, β=1.0, 此时只通过隐写分析对抗网络 SAN 优化生成载体图像的隐写性能.

此外, 采用均方差 *MSE* 作为生成网络的损失参数时, *MSE_loss* 的权值也会对生成图像的质量产生影响. 表 4 为不同 λ 取值下生成图像的 *PSNR* 和 *SSIM* 评价结果, 实验中, 模型其他参数与表 3 中最优结果的参数设置相同. 由表 4 可以看出, 当 λ=0.005 时, 生成图像 *PSNR* 为 48.5978、*SSIM* 为 0.9873, 生成图像质量最优. 因而, 实验中将生成器损失函数中 *MSE_loss* 的权重设置为 0.005.

表 3 不同参数下的 *PSNR* 与 *SSIM*

指标	α = 0.4, β = 0.6	α = 0.3, β = 0.7	α = 0.2, β = 0.8	α = 0.1, β = 0.9	Dynamic
PSNR (dB)	42.062	42.058	42.062	42.821	48.597
SSIM	0.9654	0.9676	0.9727	0.9744	0.9873

注: Dynamic列为本方案设置的动态权值方案

表 4 *MSE_loss* 不同参数下的 *PSNR* 与 *SSIM*

指标	λ = 0.0001	λ = 0.005	λ = 0.001
PSNR (dB)	41.6073	48.5978	42.7324
SSIM	0.9691	0.9873	0.9704

4.2.2 不同迭代次数对生成图像的影响

实验中进一步研究了不同迭代次数对生成图像质量的影响, 图 8 展示了在不同 epoch 后生成图像 *PSNR* 与 *SSIM* 的最优值. 模型其他权重参数分别选择表 4 实验中的最优值, 令初始 α=0.4, β=0.6, 训练的前 100 个 epoch 内参数 α 每个 epoch 降低 1%, 参数 β 每个 epoch 升高 1%, *MSE_loss* 的权值 λ 设置为 0.005. 在第 50 个 epoch 后, 生成图像 *PSNR* 达到了 43.228、*SSIM* 达到了 0.9605; 第 150 个 epoch 后的最优 *PSNR* 值达到了 48.5978, *SSIM* 值达到了 0.9894; 然而, 在第 175 个 epoch 后, *PSNR* 下降到 48.5125, *SSIM* 下降到 0.9893. 实验表明, 生成图像的质量, 在训练 150 个 epoch 后, *PSNR* 与 *SSIM* 达到最优. 因而, 实验中设定模型训练 150 个 epoch 后结束训练.

4.3 算法隐写性能分析

为研究基于 U-Net 结构的生成式多重对抗隐写算法生成载体图像的信息隐写能力, 实验中选择 3 种不同的隐写方法 (S-UNIWARD, ASDL-GAN, UT-SCA-GAN), 并引入 Xu'Net 和 Ye'Net 两种当前性能优异的隐写分析器, 对抗验证本算法的性能.

4.3.1 实验采用的隐写方法

实验中分别采用 3 种性能优异的主流隐写方法实现秘密信息的嵌入. 其中, S-UNIWARD[3]作为经典自适应隐写算法之一, 其利用方向滤波器组对载体图像进行分解, 然后以像素嵌入前后滤波器中系数的相对改变量作为像素的嵌入失真; 基于方向滤波器的方向特性, 使得嵌入位置集中在各个方向都难以建模的纹理区域和噪声区域, 避免在图像中相对容易建模的平滑区域和干净的边缘区域隐藏信息, 保障了隐写信息的安全性. ASDL-GAN[28]是一种基于生成对抗网络的可自动学习失真函数嵌入式隐写模型, 其通过最小化加性失真函数的期望, 将载体图像输入到生成器 *G* 中得到图像变化概率矩阵 *P*, 概率矩阵 *P* 经过预训练好的三元嵌入式模拟器 (TES) 得到±1 的图像

修改矩阵 M. ASDL-GAN 可以自动寻找载体中更安全的像素点进行隐写, 但安全性低于传统的 S-UNIWARD 隐写算法. UT-SCA-GAN[29]的模型组件与 ASDL-GAN 相同, 但利用 tanh-simulator 函数替代 TES 激活函数, 并使用了更紧凑的 U-Net 生成器. 与 ASDL-GAN 相比, UT-SCA-GAN 由于引入 tanh-simulator 函数, 使得训练时间增加, 但其安全性优于传统的 S-UNIWARD 隐写算法.

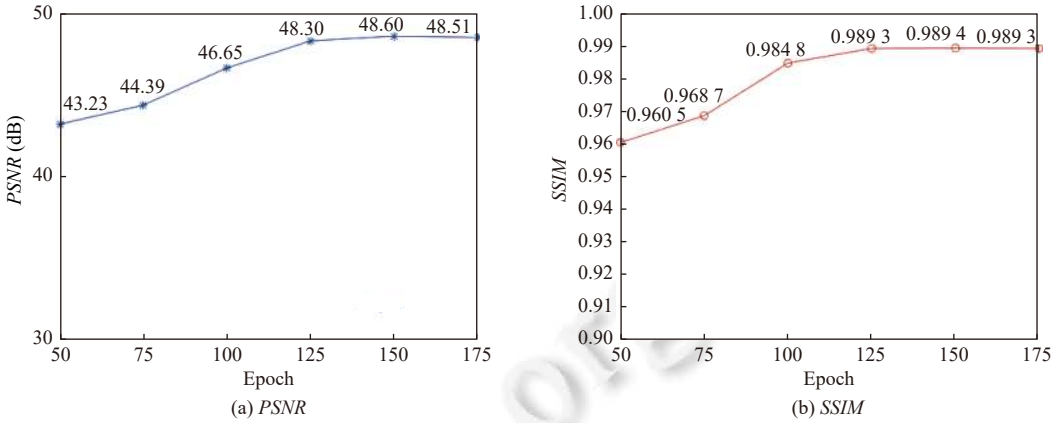

图 8 不同 epoch 的 *PSNR* 与 *SSIM*

4.3.2 实验采用的隐写分析模型

为优化基于 U-Net 结构的生成式多重对抗隐写网络, 实验中选择 Xu'Net 和 Ye'Net 两种高性能深度学习隐写分析器对抗提升模型隐写性能. Xu'Net[13]作为最经典的深度学习隐写分析器之一, 可以实现 90% 以上针对特定隐写样本的分辨能力. 本方案将 Xu'Net 的预处理层中的一个高通滤波器替换为 6 个 SRM 的高通滤波器[29], 从而更好地提取图像的细节信息, 提升了 Xu'Net 的隐写分析性能. 图 9 为 Xu'Net 隐写分析模型优化架构图.

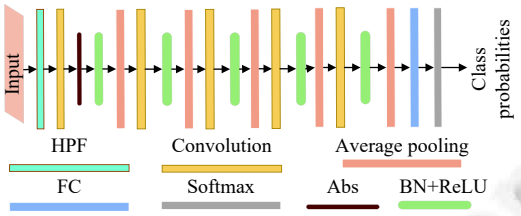

图 9 Xu'Net 隐写分析器模型结构

与 Xu'Net 相比, Ye'Net[14]在网络第 1 层采用了 30 个 SRM 卷积核来初始化参数, 并提出了新的激活函数 TLU, 因而能够更好地适应隐写噪声的分布, 收敛速度更快. 两个隐写分析模型输出都是在 2 个类标签上产生的分布. Ye'Net 模型结构如后文图 10 所示, 虚线框内的组件为选择通道感知的两个操作.

4.3.3 抗隐写分析能力验证

抗隐写分析能力是评价隐写算法性能的核心指标. 实验中采用 8 000 幅 256×256 分辨率的真实图像作为训练集, 2 000 幅图像作为测试集. 在实验过程中使用不同的隐写器网络嵌入信息, 并采用不同的隐写分析对抗网络 SAN 为生成网络提 GAN 供判别损失, 测试不同条件下生成图像在嵌入秘密信息后抵御不同隐写分析器 *SD* 检测的能力. 实验中采用不同隐写算法向载体图像嵌入容量为 0.4 bpp 的随机信息. 并选取相同数量的生成图像和隐写图像进行隐写分析验证. 由理论分析可知, 当隐写分析器的识别结果为 50% 时, 表明隐写分析器丧失对生成图像和生成隐写图像的分辨能力, 隐写分析器分辨不出目标图像是否包含隐写信息, 从而实现隐蔽通信的目的 (即隐写分析器检测的最好结果为 50%). 另一方面, 考虑到文献 [23] 和文献 [26] 分别采用对抗样本的方法对真实图像添加对抗样本合成对抗图像, 实现信息隐写, 相对于传统的生成对抗网络 (SGAN、SSGAN) 只能生成随机载体进行

隐写, 这两种方法都可以基于特定图像生成大尺寸载体图像, 并取得较好的隐蔽通信能力. 因而, 实验中将基于 U-Net 的生成式多重对抗隐写算法与这两种当前最优的生成式图像隐写方法进行比较.

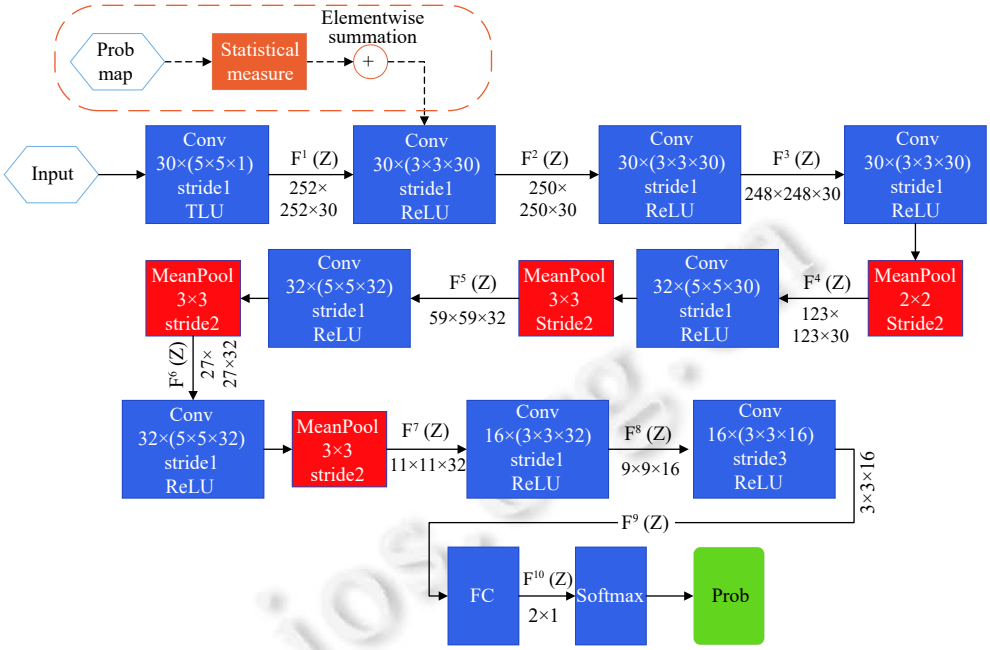

图 10 Ye'Net 隐写分析器模型结构

为了验证本模型的抗隐写分析能力, 实验中首先采用真实图像 X 及其隐写图像 X_G 作为训练集优化隐写分析器 SD 的性能, 并利用优化后的隐写分析器构建隐写分析对抗网络 SAN 为生成对抗网络 GAN 提供隐写损失, 并分别对生成图像 X_G 及其隐写图像 X_G_S 进行平行和交叉抗隐写分析能力验证; 进一步地, 实验中还采用生成图像及其隐写图像作为训练集再训练和优化隐写分析器 SD 的性能, 并利用优化后的隐写分析器构建隐写分析对抗网络 SAN 为生成对抗网络 GAN 提供隐写损失, 对生成图像 X 及其隐写图像 X_G_S 进行抗隐写分析能力验证, 评价基于 U-Net 结构的生成式多重对抗隐写算法的性能.

4.3.3.1 抗隐写分析能力平行验证

抗隐写分析能力平行验证通过在图像生成和隐写分析过程中选取相同的隐写分析器 SD 验证本算法的抗隐写分析能力. 首先采用数据库中 8 000 幅真实图像及其嵌入图像作为训练集送入隐写分析器优化网络 SON, 训练隐写分析器 SD 并构建隐写分析对抗网络 SAN, 对抗提升生成隐写图像的抗隐写分析能力, 优化生成式多重对抗隐写网络参数. 然后, 采用训练好的隐写分析器构建隐写分析网络, 联合生成对抗网络共同生成适合信息隐写的高质量载体图像. 实验中将测试集中 2 000 幅原始图像基于训练好的多重对抗隐写网络生成载体图像 X_G 和隐写图像 X_G_S, 将生成图像及其隐写图像输入到与图像生成阶段相同的隐写分析器 SD 中, 测试生成式隐写图像的抗隐写分析能力.

实验结果表明, 基于 U-Net 结构的生成式多重对抗网络可以生成更适合信息隐写的载体图像. 表 5 中展示了真实图像 X 和生成图像 X_G 分别经过 S-UNIWARD、ASDL-GAN、UT-SCA-GAN 这 3 种经典的隐写方法嵌入随机秘密信息后, 采用优化后的 Xu'Net 和 Ye'Net 隐写分析器进行隐写性能检测的结果. 其中真实图像列是隐写分析器对真实图像 X 及其隐写图像 X_S 的隐写判别准确率; "Zhang"列是使用 Zhang 等人[23]的方法为每个真实图像 X 采用快速梯度下降法生成对抗样本, 并将对抗样本添加到真实图像 X 上得到增强图像的 X_G, 然后在隐写分析器中输入 X_G 与其隐写图像 X_G_S, 所取得的隐写判别准确率. "Zhou"列是按照 Zhou 等人[26]的方法使用全卷积神经网络 (FCN) 生成对抗样本, 并将对抗样本与真实图像 X 叠加合成对抗图像 X_G, 然后在隐写分析器中输

入 *X_G* 与其隐写图像 *X_G_S* 所得到的隐写判别准确率. “所提方案”列是采用基于 U-Net 结构的生成式多重对抗隐写网络生成载体图像, 并采用与图像生成阶段相同的隐写分析器对生成图像 *X_G* 与其隐写图像 *X_G_S* 检测所得到的隐写判别准确率.

表 5 不同隐写方案的隐写分析器的检测准确率 (%)

分析器	ASDL-GAN				UT-SCA-GAN				S-UNIWARD			
	真实图像	Zhang[23]	Zhou[26]	所提方案	真实图像	Zhang[23]	Zhou[26]	所提方案	真实图像	Zhang[23]	Zhou[26]	所提方案
Xu’Net	90.2	50.3	50.1	50.0	85.2	50.2	50.1	50.1	80.5	50.3	50.1	50.0
Ye’Net	92.2	50.2	49.9	49.9	87.1	50.4	50.0	50.1	88.7	50.2	49.9	50.0

由实验结果可知, 基于 U-Net 结构的生成式多重对抗隐写网络生成的载体图像, 在秘密信息嵌入后可以有效欺骗 Xu’Net 和 Ye’Net 等高性能隐写分析网络的鉴别. 基于本算法所生成的载体图像隐写能力优于 Zhang 等人[23]和 Zhou 等人[26]所提出的基于对抗样本算法所生成的隐写载体, 可以实现高性能信息隐写. 其原因在于, 本算法在对抗模型中加入了共同训练的隐写分析器优化网络 SON, 并基于优化后的隐写分析器构建隐写分析对抗网络 SAN, 将隐写分析器网络输出的损失作为生成器损失的一部分, 设计了生成对抗网络动态加权联合损失, 从而在提升生成图像视觉质量的同时, 形成更适合信息隐写的载体图像, 增强了生成图像隐写能力.

由表 5 可以看出, 在 ASDL-GAN 隐写方案中, Xu’Net 与 Ye’Net 对生成图像 *X_G* 及其隐写图像 *X_G_S* 的检测准确率分别为 90.2%、92.2%, 但是对采用本方案的生成图像 *X_G* 及其隐写图像 *X_G_S* 的检测准确率分别只有 50%、49.9%. 在 UT-SCA-GAN 隐写方案中, Xu’Ne 与 Ye’Net 对生成图像及其隐写图像的检测准确率分别为 85.2%、87.1%, 而对采用本方案的生成图像及其隐写图像的检测准确率分别为 50.1%、50.1%. 而在 S-UNIWARD 隐写方案中, Xu’Ne 与 Ye’Net 对生成图像及其隐写图像的检测准确率为 80.5%、88.7%, 对采用本方案的生成图像及其隐写图像的检测准确率都为 50.0%. Zhang 等人[23]的网络和 Zhou 等人[26]的网络在生成载体图像上进行隐写后, 隐写图像的抗隐写分析能力也有较大提高, 采用这两种方法所生成的图像在嵌入秘密信息后, Xu’Net 与 Ye’Net 的隐写分析能力出现大幅下降, 隐写分析器的识别能力集中在 50% 左右, 即难以识别目标图像是否嵌入了秘密信息. 总体来看, 基于 U-Net 结构的生成式多重对抗隐写网络取得了较好的信息隐藏性能, 两种高性能隐写分析器对生成图像 *X_G* 及其隐写图像 *X_G_S* 的判别结果都是 50%, 隐写分析器无法区分目标图像是否嵌入了秘密信息, 实验结果表明基于 U-Net 结构的生成式多重对抗隐写算法提高了生成图像的隐写能力, 能够生成更适合信息隐写的载体图像, 与原始图像相比取得更好的抗隐写分析检测能力.

4.3.3.2 抗隐写分析能力交叉验证

抗隐写分析能力交叉验证通过在图像生成和隐写检测过程中选取不同的隐写分析器, 验证本算法的抗隐写分析能力及其鲁棒性. 在训练过程中分别采用 Xu’Net 和 Ye’Net 隐写分析器进行优化并构建隐写分析对抗网络 SAN, 而在测试过程中采用 Xu’Net/Ye’Net 测试基于 Ye’Net/Xu’Net 隐写分析器 *SD* 构建隐写分析对抗网络 SAN 所形成的生成图像 *X_G* 及其隐写图像 *X_G_S*, 以验证基于 U-Net 结构的生成式多重对抗隐写算法抗的抗隐写分析能力及其鲁棒性. 实验中首先选取数据库中 8000 幅真实图像及其嵌入图像作为训练集送入隐写分析器优化网络 SON, 训练隐写分析器, 然后选取经过 150 个 epoch 优化训练的 Xu’Net 和 Ye’Net 隐写分析器构建隐写分析对抗网络 SAN, 联合生成对抗网络共同生成适合信息隐写的高质量载体图像 *X_G*, 并对抗提升生成隐写图像 *X_G_S* 的抗隐写分析能力, 优化生成式多重对抗隐写网络参数. 然后, 将测试集中 2000 幅原始图像基于训练好的网络生成载体图像和隐写图像, 将生成图像及其隐写图像输入到与图像生成网络不同的隐写分析器 *SD* 中, 测试生成图像及其隐写图像抗不同隐写分析器交叉检验的能力. 实验结果如表 6 所示.

表 6 为图像经过 UT-SCA-GAN、ASDL-GAN 和 S_UNIWARD 这 3 种隐写方法嵌入随机秘密信息后, 分别采用不同的隐写分析网络进行检测的实验结果. 在模型训练过程中使用 Ye’Net 隐写分析器作为判别网络时, 采用 Xu’Net 隐写分析器构建多重对抗隐写网络, 并生成载体图像 *X_G* 及其隐藏图像 *X_G_S*; 而使用 Xu’Net 隐写分析

器作为判别网络时, 采用 Ye'Net 隐写分析器构建多重对抗隐写网络, 并生成载体图像 *X_G* 及其隐藏图像 *X_G_S*. 由此来验证基于 U-Net 结构的生成式多重对抗隐写算法抗隐写分析性能的鲁棒性.

表 6　隐写分析器交叉验证检测准确率 (%)

分析器	ASDL-GAN			UT-SCA-GAN			S-UNIWARD		
	Zhang[23]	Zhou[26]	所提方案	Zhang[23]	Zhou[26]	所提方案	Zhang[23]	Zhou[26]	所提方案
Ye'Net	90.1	66.3	50.1	85.7	61.8	50.2	84.5	64.5	50.2
Xu'Net	89.7	56.2	49.8	82.4	54.0	49.9	79.4	54.4	50.1

实验结果表明证明, 基于 U-Net 结构的生成式多重对抗隐写网络可以有效抵抗不同隐写分析网络的检测, 算法具有较好的鲁棒性. 一方面, 因为本算法与文献 [23] 和文献 [26] 的方案不同, 在训练过程中不只是对抗已经训练好的隐写分析网络, 而是在模型训练过程中同步优化生成网络和隐写分析网络. 因而, 在生成载体图像的过程中同步增强其抗隐写分析能力, 针对性地优化隐写分析网络所"关注"的某些敏感区域, 生成高视觉质量和较强鲁棒性的载体图像. 另一方面, 本文提出的基于 U-Net 结构的生成网络通过跳接层传递真实图像的细节信息, 尽可能多地保留了真实图像特征, 生成包含更多细节信息的载体图像, 增强信息隐藏能力; 而对抗样本则是通过在真实图像中添加噪声, 使隐写分析网络做出错误判断, 而在对抗过程中通过调节添加对抗噪声的大小改善图像的对抗性能. 相比而言, 基于 U-Net 结构的多重对抗隐写网络模型保留了更多的真实图像特征, 因而取得较强的抗隐写分析鲁棒性.

如表 6 所示, 在交叉验证生成图像 *X_G* 及其隐写图像 *X_G_S* 抗隐写分析性能时, Zhang 等人[23]的方法所生成的载体图像及其隐写图像, 在采用不同的隐写判别器进行识别时, 其识别准确率可以达到 80% 以上, 图像的抗隐写分析能力鲁棒性较差. 采用 Zhou 等人[26]的方案所生成的载体图像及其隐写图像对抗不同类型隐写分析器的判别性能有所增强, 但识别准确率仍然在 60% 以上. 而采用基于 U-Net 结构的生成式多重对抗隐写网络生成载体图像 *X_G* 及其隐写图像 *X_G_S*, 即使在载体图像生成阶段和隐写图像判别阶段采用不同的隐写分析器, 仍然能够取得较为理想的抗隐写分析能力. 从表 6 还可以看出示, 同样使用 UT-SCA-GAN 隐写方案, 在图像生成阶段采用 Ye'Net 网络构建隐写分析网络, 在测试阶段采用 Xu'Net 网络对生成图像 *X_G* 及其隐写图像 *X_G_S* 进行隐写分析, 采用 Zhang 等人[23]算法的隐写分析准确率为 82.41%, 采用 Zhou 等人[26]算法的隐写分析准确率为 54.03%, 而本模型的准确率为 49.97%. 因而, 基于 U-Net 结构的生成式多重对抗隐写算法具有更强的抗交叉隐写分析能力和鲁棒性.

4.3.3.3　基于隐写分析器再训练的抗隐写分析能力验证

更进一步地, 我们将生成的载体图像 *X_G* 与其隐写图像 *X_G_S* 作为训练集, 重新训练隐写分析器, 并采用再训练的隐写分析器构建隐写分析对抗网络 SAN, 联合生成对抗网络生成载体图像 *X_G* 及其隐写图像 *X_G_S*, 验证基于 U-Net 结构的生成式多重对抗网络经过再训练后的隐写性能. 实验中首先选取 8 000 幅原始及其嵌入图像作为训练集送入隐写分析器优化网络 SON, 训练隐写分析器, 选取经过 150 个 epoch 优化训练的 Xu'Net 和 Ye'Net 隐写分析器构建隐写分析对抗网络 SAN, 联合生成对抗网络 GAN, 共同生成适合信息隐写的高质量载体图像. 然后, 将生成的载体图像 *X_G* 及其隐写图像 *X_G_S* 作为训练集再次输入隐写分析器网络, 训练优化隐写分析器性能 *SD*, 并基于再训练后的隐写分析器构建新的隐写分析对抗网络 SAN, 联合生成对抗网络重新生成载体图像 *X_G* 及其隐写图像 *X_G_S*, 并对抗提升生成隐写图像的抗隐写分析能力, 优化生成式多重对抗隐写网络参数. 然后, 将测试集中 2 000 幅原始图像基于训练好的网络生成载体图像 *X_G* 及其隐写图像 *X_G_S*, 将生成图像及生成隐写图像输入到隐写分析器中, 测试生成隐写图像抗隐写分析能力. 实验结果如表 7 所示.

表 7　生成图像再训练后的隐写分析器检测准确率 (%)

分析器	ASDL-GAN			UT-SCA-GAN			S-UNIWARD		
	Zhang[23]	Zhou[26]	所提方案	Zhang[23]	Zhou[26]	所提方案	Zhang[23]	Zhou[26]	所提方案
Ye'Net	68.5	55.8	55.2	65.7	53.9	55.0	63.4	52.8	53.1
Xu'Net	65.3	55.1	54.6	62.5	54.2	52.8	60.7	53.1	53.0

由实验结果可知, 采用生成图像及其隐写图像对隐写分析器进行再训练后, 基于 U-Net 结构的多重对抗隐写网络仍具有很高的安全性. 生成图像 *X_G* 及其隐写图像 *X_G_S* 仍然可以有效欺骗隐写分析网络. 这是因为采用对抗样本的方法生成载体图像过程中, 通过添加噪声的方式对图像的敏感区域进行修改, 使得隐写分析器对嵌密后的载体图像进行错误的分类, 从而提高载体图像的隐写性能. 噪声信息的添加影响了生成载体图像的高频区域, 也不可避免地增添了冗余的“特征”信息, 导致隐写分析再训练时更容易在图像中检测到这些信息, 一定程度上降低了生成图像及其隐写图像的抗再训练隐写分析的能力. 而基于 U-Net 结构的生成式多重对抗隐写网络通过学习图像的关键特征生成更适合信息隐写的高质量载体图像, 增强信息隐藏能力; 在对抗生成载体图像的过程中通过添加图像对抗损失 *D_loss*、均方误差 (*MSE*) 损失 *MSE_loss* 和隐写分析损失 *SDO_loss* 这 3 种联合损失, 保障了生成图像在具有更高抗隐写分析能力的同时, 还具有更好的图像质量.

由表 7 可以看出, 在采用生成图像 *X_G* 及其隐写图像 *X_G_S* 进行对抗训练后, 基于 U-Net 的生成式多重对抗网络生成的图像仍具有较强的隐写能力, 不同隐写分析网络对再训练后的生成图像 *X_G* 及其隐写图像 *X_G_S* 的判别能力不大于 55.2%; 而采用 Zhang 等人的方法生成的载体图像及其隐写图像对判别其进行再训练后 (基于训练集图像), 判别器对生成图像 *X_G* 及其隐写图像 *X_G_S* 的判别准确率达到了 68.5% (基于测试集图像); 采用 Zhou 等人的方法生成的载体图像及其隐写图像对判别其进行再训练后 (基于训练集图像), 判别器对生成图像 *X_G* 及其隐写图像 *X_G_S* 的判别精度也达到了 55.8% (基于测试集图像). 因此, 可以看出基于 U-Net 结构的生成式多重对抗隐写网络, 即使在采用生成图像及其隐写图像对隐写分析器再训练后, 仍然可以生成具有很强抗隐写分析能力的高质量图像, 从而为图像隐写提供更高的安全性.

4.4 算法消融实验

为了进一步验证基于 U-Net 的生成式多重对抗隐写算法的性能, 实验中设计消融实验对比检验本文所提出的方法对生成载体图像隐写能力的影响. 考虑到本文模型由生成对抗网络、隐写器网络、隐写分析其优化网络、隐写分析对抗网络 4 部分构成, 生成器的损失函数由生成对抗损失 *D_loss*、均方差损失 *MSE_loss*、隐写分析损失 *SDO_loss* 加权组合而成. 实验中设计了不同网络结构的消融实验, 进一步研究不同模块对模型生成图像视觉质量和抗隐写分析能力的影响. 如表 8 所示, 分别采用基础 GAN 网络&*D_loss*、Encoder-Decoder 结构 GAN 网络& *D_loss*、U-Net 结构 GAN 网络& *D_loss*、U-Net 结构 GAN 网络&*MSE_loss*、U-Net 结构 GAN 网络&*D_loss*+*MSE_loss*、U-Net 结构 GAN 网络&*D_loss*+*SDO_loss*、U-Net 结构 GAN 网络&*MSE _loss*+*SDO_loss*、U-Net 结构 GAN 网络& *D_loss*+*MSE_loss*+*SDO_loss* 对本文所提出的生成式多重对抗隐写网络模型开展性能验证消融实验.

表 8　不同网络结构与损失函数的隐写算法性能

网络结构	*PSNR* (dB)	准确率 (%)	Epoch	训练时间 (min)
基础GAN网络&*D_loss*	—	—	—	—
Encoder-Decoder结构GAN网络&*D_loss*	10.9	100	149	128
U-Net结构GAN网络&*D_loss*	20.3	99.87	144	176
U-Net结构GAN网络&*MSE_loss*	49.7	91.97	135	168
U-Net结构GAN网络&*D_loss*+*MSE_loss*	49.9	89.80	130	193
U-Net结构GAN网络&*D_loss*+*SDO_loss*	16.8	100	150	221
U-Net结构GAN网络&*MSE _loss*+*SDO_loss*	48.4	51	137	203
U-Net结构GAN网络&*D_loss*+*MSE_loss*+*SDO_loss*	48.6	50	122	230

如表 8 所示, 基于 U-Net 结构的生成式多重对抗隐写网络在采用 3 种损失加权组合后取得最好的信息隐藏性能. 表 8 中横行为实验中采用的网络结构、模型训练时间、生成图像的 *PSNR* 以及采用 Ye’Net 隐写分析模型对生成隐写图像的判别准确率; 竖列代表消融实验中不同网络架构与损失的组合. 由表 8 可知, 基础 GAN 网络无法可控地生成 256×256 大小的既定目标图片, 而直接采用编解码器结构的 GAN 网络所生成的图像质量很差, 在网络采用 U-Net 结构后, 生成图像质量明显提升. 其中采用基于 U-Net 结构的生成式多重对抗隐写网络在结合

D_loss 和 *MSE_loss* 两种损失后, 生成图像的质量最优, *PSNR* 值达到 49.9 dB. 而在基于 U-Net 结构的生成式多重对抗隐写网络中采用 *D_loss*、*MSE_loss* 和 *SDO_loss* 这 3 种损失后, 生成图像的质量有所降低, 但生成隐写图像的抗隐写能力明显增加, 隐写分析器的识别能力为 50%. 这是因为在添加隐写分析对抗网络后, 生成对抗网络在保障生成更高质量载体图像的同时, 还需要与隐写分析网络进行对抗以提高生成图像的信息隐藏能力, 导致生成图像的视觉质量略微下降, 但其抗隐写分析能力得到了显著的提高. 同时, 由表 8 还可以看出, 随着在基于 U-Net 结构的多重对抗生成网络中所采用的损失数量增加, 隐写分析损失的生成需要增加计算成本, 延长了网络的运算时间, 但系统的学习效率提高, 网络收敛所需要的 epoch 大大缩短, 网络收敛的稳定性更好. 实验结果证明基于 U-Net 结构的生成式多重对抗隐写网络在采用 3 种损失加权组合后, 不但可以实现多重对抗网络的稳定收敛, 而且可以增加载体图像的信息隐藏能力, 生成更适合信息嵌入的高质量载体图像.

4.5 网络运行时效比较

为了验证基于 U-Net 结构的生成式多重对抗隐写网络的时效性, 实验中分别选取 Zhang 等人[23]与 Zhou 等人[26]的隐写网络模型与本文模型开展时效性对比实验. Zhang 等人[23]采用基于对抗样本的载体图像生成算法, 利用快速梯度下降法生成对抗载体图像, 并采用经典的自适应隐写算法嵌入秘密信息. Zhou 等人[26]采用全卷积神经网络 (FCN) 生成与原始图像近似的载体图像, 提高了载体图像的生成速度, 并且设计了新的损失函数保障载体图像和隐写图像能够欺骗隐写网络的分析. 以上两种方法都取得了非常优秀的信息隐藏能力. 因而, 实验中选取这两种优秀的隐写网络与本文中所提出的生成式多重对抗隐写网络模型进行运行效率比较. 实验中分别选取了基于不同网络架构与损失函数的生成式多重对抗隐写网络 (包括基础 GAN 网络&*D_loss*、Encoder-Decoder 结构 GAN 网络&*D_loss*、U-Net 结构 GAN 网络&*D_loss*、U-Net 结构 GAN 网络&*MSE_loss*、U-Net 结构 GAN 网络&*D_loss*+*MSE_loss*、U-Net 结构 GAN 网络&*D_loss*+*SDO_loss*、U-Net 结构 GAN 网络&*MSE_loss*+*SDO_loss* 以及 U-Net 结构 GAN 网络&*D_loss*+*MSE_loss*+*SDO_loss* 生成式多重对抗隐写网络模型) 与这两种隐写网络模型开展时效性对比试验. 由于实验中不同网络所采用的隐写算法和隐写分析网络一致, 所以只需要对不同网络生成载体图像所需要的时间进行比较. 实验结果如表 9 所示 (其中, 表中所列时间是生成 2 000 张载体图像所需要的时间).

表 9 基于不同网络结构的载体图像生成时间

网络结构	生成时间 (s)
Zhang等人[23]	6 815.12
Zhou等人[26]	41.59
基础GAN网络&*D_loss*	—
Encoder-Decoder结构GAN网络&*D_loss*	45.35
U-Net结构GAN网络&*D_loss*	50.45
U-Net结构GAN网络&*MSE_loss*	46.35
U-Net结构GAN网络&*D_loss*+*MSE_loss*	47.34
U-Net结构GAN网络&*D_loss*+*SDO_loss*	47.26
U-Net结构GAN网络&*MSE_loss*+*SDO_loss*	46.49
U-Net结构GAN网络&*D_loss*+*MSE_loss*+*SDO_loss*	48.61

由表 9 可知, 基于 U-Net 结构的生成式多重对抗隐写网络生成载体图像所需的时间不超过采用 Zhang 等人[23]的方法生成载体图像所需时间的 1%, 在运行效率方面取得了优异的性能. 这是因为 Zhang 等人[23]的方法需要根据嵌入噪声对载体图像判别结果的影响进行重训练, 以生成适合隐写的对抗样本图像, 导致其生成载体图像所需的时间较长. 另一方面, 采用本文所提出的生成式多重对抗隐写算法生成载体图像所需时间略高于 Zhou 等人[26]的网络所需的时间. 然而, 由前述实验可知, 基于本算法所生成的载体图像具有更好的图像质量和信息隐藏能力, 所形成的隐写图像也具有更强的抗隐写分析能力的鲁棒性. 由表 9 还可以看出, 基于 U-Net 结构的生成式多重对抗隐写网络在采用不同的损失函数组合后, 生成载体图像的时间并没有发生明显的变化. 其中, 采用基于 U-Net

结构的生成式多重对抗隐写网络在采用 3 种损失加权组合后实现了载体图像生成质量、隐写能力和生成时间之间的最好平衡. 实验结果表明基于 U-Net 结构的生成式多重对抗隐写网络在采用 3 种损失加权组合后, 不但可以生成更适合信息嵌入的高质量载体图像, 增加载体图像的信息隐藏能力, 而且还具有很高的载体图像生成效率.

5 结 语

本文提出了一种基于 U-Net 结构的多重对抗生成式隐写算法, 基于生成对抗网络与隐写分析网络的多重对抗, 生成比真实图像更适合信息隐藏的载体图像, 使得隐写图像取得更好的抗隐写分析能力. 算法在生成对抗网络中加入 U-Net 网络架构, 通过跳接层将真实图像的细节信息传递到生成图像中, 提高目标图像的生成质量, 增强信息隐藏能力. 同时, 在载体图像生成过程中引入多重对抗隐写网络, 分别采用真实图像及其隐写图像, 生成图像及其隐写图像训练隐写分析器并构建隐写分析网络, 对抗提升生成对抗网络的性能, 生成更适合信息嵌入的高质量载体图像; 并针对不同隐写分析器采用平行、交叉和再训练方案验证生成载体隐写图像的抗隐写分析能力. 其次, 本算法采用图像对抗损失、均方误差 (MSE) 损失和隐写分析损失的加权组合作为生成网络的总损失, 迭代优化生成对抗网络模型, 保障多重对抗生成式隐写网络快速平稳收敛. 实验结果表明, 基于 U-Net 结构的生成式多重对抗隐写算法与其他方法相比具有更好的载体图像生成能力和更强的抗隐写分析能力. 在以后的工作中, 我们将继续关注生成式对抗网络的优化模型, 进一步提高适合信息隐写载体图像的生成效率.

References:

[1] Petitcolas FAP, Anderson RJ, Kuhn MG. Information hiding—A survey. Proc. of the IEEE, 1999, 87(7): 1062−1078. [doi: 10.1109/5.771065]

[2] Mielikainen J. LSB matching revisited. IEEE Signal Processing Letters, 2006, 13(5): 285–287. [doi: 10.1109/LSP.2006.870357]

[3] 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]

[4] Holub V, Fridrich J. Designing steganographic distortion using directional filters. In: Proc. of the 2012 IEEE Int'l Workshop on Information Forensics and Security (WIFS). Costa Adeje: IEEE, 2013. 234–239. [doi: 10.1109/WIFS.2012.6412655]

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

[6] Ruanaidh JJKO, Dowling WJ, Boland FM. Phase watermarking of digital images. In: Proc. of the 3rd IEEE Int'l Conf. on Image Processing. Lausanne: IEEE, 1996. 239–242. [doi: 10.1109/ICIP.1996.560428]

[7] Cox IJ, Kilian J, Leighton FT, Shamoon T. Secure spread spectrum watermarking for multimedia. IEEE Trans. on Image Processing, 1997, 6(12): 1673–1687. [doi: 10.1109/83.650120]

[8] Lin WH, Horng SJ, Kao TW, Fan PZ, Lee CL, Pan Y. An efficient watermarking method based on significant difference of wavelet coefficient quantization. IEEE Trans. on Multimedia, 2008, 10(5): 746–757. [doi: 10.1109/TMM.2008.922795]

[9] Zhou ZL, Sun HY, Harit R, Chen XY, Sun XM. Coverless image steganography without embedding. In: Proc. of the 2015 Int'l Conf. on Cloud Computing and Security. Nanjing: Springer, 2015. 123–132. [doi: 10.1007/978-3-319-27051-7_11]

[10] Ruan SH, Qin ZC. Coverless covert communication based on GIF image. Communications Technology, 2017, 50(7): 1506−1510 (in Chinese with English abstract). [doi: 10.3969/j.issn.1002-0802.2017.07.031]

[11] Duan XT, Song HX, Qin C, Khan MK. Coverless steganography for digital images based on a generative model. Computers, Materials & Continua, 2018, 55(3): 483–493. [doi: 10.3970/cmc.2018.01798]

[12] Fridrich J, Kodovsky J. Rich models for steganalysis of digital images. IEEE Trans. on Information Forensics and Security, 2012, 7(3): 868–882. [doi: 10.1109/TIFS.2012.2190402]

[13] Qian YL, Dong J, Wang W, Tan TN. Deep learning for steganalysis via convolutional neural networks. In: Proc. of the 2015 Media Watermarking, Security, and Forensics. San Francisco: SPIE, 2015. 94090J. [doi: 10.1117/12.2083479]

[14] 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]

[15] Ye J, Ni JQ, Yi Y. Deep learning hierarchical representations for image steganalysis. IEEE Trans. on Information Forensics and Security, 2017, 12(11): 2545–2557. [doi: 10.1109/TIFS.2017.2710946]

[16] 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]

[17] Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y. Generative adversarial networks. arXiv:1406.2661, 2014.

[18] Ronneberger O, Fischer P, Brox T. U-Net: Convolutional networks for biomedical image segmentation. In: Proc. of the 18th Int'l Conf. on Medical Image Computing and Computer-assisted Intervention. Munich: Springer, 2015. 234–241. [doi: 10.1007/978-3-319-24574-4_28]

[19] Vplkonskiy D, Borisenko B, Burnaev E. Generative adversarial networks for image steganography. In: Proc. of the 5th Int'l Conf. on Learning Representations. Toulon: ICLR, 2016.

[20] Radford A, Metz L, Chintala S. Unsupervised representation learning with deep convolutional generative adversarial networks. In: Proc. of the 4th Int'l Conf. on Learning Representations. San Juan: ICLR, 2015. 534–544.

[21] Shi HC, Dong J, Wang W, Qian YL, Zhang XY. SSGAN: Secure steganography based on generative adversarial networks. In: Proc. of the 18th Pacific Rim Conf. on Multimedia. Harbin: Springer, 2017. 534–544. [doi: 10.1007/978-3-319-77380-3_51]

[22] Arjovsky M, Chintala S, Bottou L. Wasserstein GAN. arXiv:1701.07875, 2017.

[23] 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 Information Hiding Multimedia Security. New York: Association for Computing Machinery, 2018. 67–72. [doi: 10.1145/3206004.3206012]

[24] Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow I, Fergus R. Intriguing properties of neural networks. arXiv: 1312.6199, 2013.

[25] Goodfellow IJ, Shlens J, Szegedy C. Explaining and harnessing adversarial examples. arXiv:1412.6572, 2014.

[26] 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]

[27] Li SY, Ye DP, Jiang SZ, Liu CR, Niu XG, Luo XY. Anti-steganalysis for image on convolutional neural networks. Multimedia Tools and Applications, 2020, 79(7): 4315–4331. [doi: 10.1007/s11042-018-7046-6]

[28] Tang WX, Tan SQ, Li B, Huang JW. Automatic steganographic distortion learning using a generative adversarial network. IEEE Signal Processing Letters, 2017, 24(10): 1547–1551. [doi: 10.1109/LSP.2017.2745572]

[29] Yang JH, Liu K, Kang XQ, Wong EK, Shi YQ. Spatial image steganography based on generative adversarial network. arXiv:1804.07939, 2018.

[30] Meng RH, Rice SG, Wang J, Sun XM. A fusion steganographic algorithm based on faster R-CNN. Computers, Materials and Continua, 2018, 55(1): 1–16. [doi: 10.3970/cmc.2018.055.001]

[31] Tang WX, Li B, Tan SQ, Barni M, Huang JW. CNN-based adversarial embedding for image steganography. IEEE Trans. on Information Forensics and Security, 2018, 14(8): 2074–2087. [doi: 10.1109/TIFS.2019.2891237]

[32] Hayes J, Danezis G. Generating steganographic images via adversarial training. In: Proc. of the 31st Int'l Conf. on Neural Information Processing Systems. Red Hook: Curran Associates Inc., 2017. 1951–1960.

[33] Wang ZH, Gao N, Wang X, Qu XX, Li LH. SSteGAN: Self-learning steganography based on generative adversarial networks. In: Proc. of the 25th Int'l Conf. on Neural Information Processing. Siem Reap: Springer, 2018. 253–264. [doi: 10.1007/978-3-030-04179-3_22]

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

[35] Baluja S. Hiding images in plain sight: Deep steganography. In: Proc. of the 31st Int'l Conf. on Neural Information Processing Systems. Red Hook: Curran Associates Inc., 2017. 2069–2079.

[36] Wu P, Yang Y, Li XQ. StegNet: Mega image steganography capacity with deep convolutional network. Future Internet, 2018, 10(6): 54. [doi: 10.3390/fi10060054]

[37] He KM, Zhang XY, Ren SQ, Sun J. Deep residual learning for image recognition. In: Proc. of the 2016 IEEE Conf. on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 770–778. [doi: 10.1109/CVPR.2016.90]

[38] Duan XT, Jia K, Li BX, Guo DD, Qin C. Reversible image steganography scheme based on a U-Net structure. IEEE Access, 2019, 7: 9314–9323. [doi: 10.1109/ACCESS.2019.2891247]

[39] Baluja S. Hiding images within images. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2020, 42(7): 1685–1697. [doi: 10.1109/TPAMI.2019.2901877]

[40] 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]

[41] 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]
[42] Fu ZJ, Wang F, Cheng X. The secure steganography for hiding images via GAN. EURASIP Journal on Image and Video Processing, 2020, 2020: 46. [doi: 10.1186/s13640-020-00534-2]
[43] Hu DH, Wang L, Jiang WJ, Zheng SL, Li B. A novel image steganography method via deep convolutional generative adversarial networks. IEEE Access, 2018, 6: 38303–38314. [doi: 10.1109/ACCESS.2018.2852771]
[44] Li J, Niu K, Liao LW, Wang LJ, Liu L, Yu L, Zhang MQ. A generative steganography method based on WGAN-GP. In: Proc. of the 6th Int'l Conf. on Artificial Intelligence and Security. Hohhot: Springer, 2020. 386–397. [doi: 10.1007/978-981-15-8083-3_34]
[45] Zhu YM, Chen F, He HJ, Chen H. Orthogonal GAN information hiding model based on secret information driven. Journal of Applied Sciences, 2019, 33(5): 721–732 (in Chinese with English abstract). [doi: 10.3969/j.issn.0255-8297.2019.05.013]
[46] Di FQ, Liu J, Zhang Z, Yu L, Li J, Zhang MQ. Somewhat reversible data hiding by image to image translation. arXiv:1905.02872, 2019.
[47] Meng RH, Zhou ZL, Cui Q, Sun XM, Yuan CS. A novel steganography scheme combining coverless information hiding and steganography. Journal of Information Hiding and Privacy Protection, 2019, 1(1): 43–48. [doi: 10.32604/jihpp.2019.05797]
[48] Liu MM, Zhang MQ, Liu J, Zhang YN, Ke Y. Coverless information hiding based on generative adversarial networks. arXiv:1712.06951, 2017.

附中文参考文献:

[10] 阮书涵, 秦正才. 一种基于GIF图像的无载体隐蔽通信方法. 通信技术, 2017, 50(7): 1506–1510. [doi: 10.3969/j.issn.1002-0802.2017.07.031]
[45] 朱翌明, 陈帆, 和红杰, 陈鸿佑. 基于秘密信息驱动的正交GAN信息隐藏模型. 应用科学学报, 2019, 33(5): 721–732. [doi: 10.3969/j.issn.0255-8297.2019.05.013]

马宾(1973－), 男, 博士, 教授, 主要研究领域为信息隐藏与多媒体安全, 数字图像处理, 隐私计算.

王春鹏(1989－), 男, 博士, 副教授, CCF 专业会员, 主要研究领域为信息隐藏与多媒体安全, 数字图像处理.

韩作伟(1996－), 男, 硕士, 主要研究领域为深度学习, 隐写术, 信息隐藏.

李健(1982－), 男, 博士, 副教授, 主要研究领域为多媒体信息安全.

徐健(1972－), 女, 副教授, 主要研究领域为信息隐藏与多媒体安全, 数字图像处理.

王玉立(1973－), 男, 教授, 主要研究领域为网络空间安全.