基于双向生成对抗网络的感知哈希图像内容取证算法

马 宾[1),2)] 王一利[1),2)] 徐 健[3)] 王春鹏[1),2)] 李 健[1),2)]
周琳娜[4)] 施云庆[5)]

[1)](齐鲁工业大学(山东省科学院)计算机科学与技术学部 济南 250300)
[2)](山东省计算机网络重点实验室 济南 2500013)
[3)](山东财经大学计算机科学与技术学院 济南 250014)
[4)](北京邮电大学网络空间安全学院 北京 100876)
[5)](新泽西理工大学电气与计算机系 纽瓦克 07102 美国)

摘 要 传统的感知哈希算法通过提取图像特定属性生成感知哈希序列,难以充分利用原始图像全部特征信息,影响了基于感知哈希的图像内容认证与版权保护能力.本文提出一种基于双向生成对抗网络(Bidirectional Generative Adversarial Network, BiGAN)的无监督感知哈希图像内容取证算法,基于编码网络、生成网络和判别网络间的双向迭代对抗,生成具有较强图像语义特征表示能力的感知哈希码;并通过在编码网络和生成网络间添加跳接层网络结构,将原始图像不同维度的特征信息传递到生成网络,提高生成网络语义特征学习能力与网络收敛速度;同时,在对抗损失中添加MSE误差损失,增强生成图像的视觉质量与细节表示能力;最后,基于网络间的多重迭代与对抗训练,输出兼具相同内容图像认证鲁棒性和不同内容图像区分敏感性的高性能图像感知哈希码.本研究首次采用大型图像数据库进行算法性能评价,实验结果表明基于双向生成对抗网络的感知哈希图像内容取证算法与当前其他优秀研究方案相比具有更强的图像内容取证性能.

关键词 图像取证;生成对抗网络;感知哈希;跳接;均方误差
中图法分类号 TP391 **DOI号** 10.11897/SP.J.1016.2023.02551

A Bidirectional Generative Adversarial Network-Based Perceptual Hash Algorithm for Image Content Forensics

MA Bin[1),2)] WANG Yi-Li[1),2)] XU Jian[3)] WANG Chun-Peng[1),2)] LI Jian[1),2)]
ZHOU Lin-Na[4)] SHI Yun-Qing[5)]

[1)](*Department of Computer Science And Technology, Qilu University of Technology (Shandong Academy of Sciences), Jinan* 250300)
[2)](*Shandong Provincial Key Laboratory of Computer Network, Jinan* 250013)
[3)](*Department of Computer Science And Technology, Shandong University of Finance and Economics, Jinan* 250014)
[4)](*Department of Cyber Security, Beijing University of Posts and Telecommunications, Beijing* 100876)
[5)](*Department of Electrical and Computer Engineering, New Jersey Institute of Technology, Newark* 07102, *USA*)

收稿日期:2022-10-11;在线发布日期:2023-07-21.本课题得到国家自然科学基金(62272255, 61872203)、国家重点研发计划(2021YFC3340600)、山东省自然科学基金(ZR2019BF017, ZR2020MF054)、山东省重大科技创新工程项目(2019JZZY020127, 2019JZZY010132, 2019JZZY010201)、山东省高等学校青创人才引育计划(SD2019-161)、济南市"高校20条"引进创新团队(2019GXRC031)、济南市"高校20条"工作室带头人(2020GXRC056)、济南市市校融合发展战略工程项目(JNSX2021030)资助.
马 宾,博士,教授,中国计算机学会(CCF)会员,主要研究领域为可逆信息隐藏、多媒体安全、图像处理. E-mail: sddxmb@126.com. **王一利**,硕士研究生,主要研究领域为生成式对抗网络、感知图像哈希. **徐 健**(通信作者),博士,副教授,主要研究领域为信息隐藏、图像处理. E-mail: sdfixj@126.com. **王春鹏**,博士,副教授,中国计算机学会(CCF)会员,主要研究领域为图像处理、多媒体信息安全. **李 健**,博士,副教授,主要研究领域为数字图像与视频取证. **周琳娜**,博士,教授,主要研究领域为信息隐藏、多媒体内容取证. **施云庆**,博士,教授,主要研究领域为多媒体取证、多媒体安全.

Abstract The traditional perceptual hash algorithm creates image perceptual hash code by extracting image features with a pre-designed scheme. As it is hard to make full use of image inherent semantic characters, the performance of perceptual hash code on image content authentication and copyright protection is constrained. In this paper, an unsupervised perceptual hash algorithm for image forensics based on Bidirectional Generative Adversarial Network (BiGAN) is proposed. The main contributions of the paper are as follows: Firstly, depending on the bidirectional iterative adversary among the coding network, the generative network, and the discriminative network, the powerful learning ability of BiGAN on image inherent feature extraction is fully developed; so that the perceptual hash code that has strong image semantic feature representation capability can be created. As a result, both the identification robustness for images with identical content and the discrimination sensitivity for images with different contents are achieved. Hence, the capability of image forensics is improved. Secondly, a BiGAN optimization framework is constructed by adding a skip-connection structure between the coding and the generative network. By concatenating the shallow and deep layers′ features of the sampled image, different dimensional features are organically integrated to improve the learning efficiency and the convergence speed of the proposed scheme. Thereby, the semantic information representation ability of the perceptual hash code is enhanced, and the identification robustness for identical content images is heightened. Thirdly, a Mean Square Error (MSE) loss-based performance optimization strategy for BiGAN is investigated. By computing the difference between the output of the coding network and the generative network, not only the visual quality of the generated image but also the representation capability of the generated perceptual hash code is effectively improved. Consequently, the discrimination sensitivity for different content images is intensified. In the end, by virtue of multiple network iterations and adversarial training, a high-performance perceptual hash code for image forensics is obtained. Furthermore, a large image database CeleA Mask-HQ is employed for the first time to evaluate the performance of the perceptual hash algorithm in this study. The capability of the BiGAN-based perceptual hash algorithm for the identification of images with identical content and for the discrimination of images with different contents is discussed in detail. Meanwhile, both the influence of the skip-connection network structure and that of the mean square error (MSE) loss on the performance improvement of the BiGAN-based perceptual hash algorithm are explored at length. In addition, four excellent image perceptual hash algorithms are involved in the experiment to verify the performance of the proposed scheme in comparisons. Extensive experimental results indicate that the BiGAN-based perceptual hash algorithm gains higher image forensics ability than other state-of-the-art schemes.

Keywords image forensics; generative adversarial network; perceptual hash; skip-connection; mean square error

1 引 言

随着智能终端与数字图像处理技术[1]的快速发展,人们获取高精度图像的成本不断下降,互联网中存储的数字图像规模呈几何级数增长;同时,大量图像编辑软件的出现,使得图像修改成本不断降低,眼见不一定为实,数字图像的内容认证与版权保护成为图像领域的研究热点.

近年来,研究人员提出采用数字水印[2]的算法进行图像版权认证,然而,水印嵌入不可避免地破坏了图像内在结构,影响了图像的视觉效果.与此同时,图像感知哈希作为新兴的多媒体安全技术引起了众多学者关注[3-4].图像感知哈希是一种基于图像

视觉内容特征生成固定长度散列码的哈希算法.一方面,其继承了传统哈希函数单向性、抗碰撞性与摘要性的特点;另一方面,图像感知哈希还具有以下特性:(1)鲁棒性:不同于传统哈希算法对原始数据极度敏感的特征,原始图像经过内容保持性操作后(如压缩、加噪、旋转等)仍然能够产生相似的哈希序列码,任何感知上相似的图像都应具有相似的哈希值,从而实现相同内容图像的认证.(2)区分性:针对不同内容的图像,感知哈希算法应生成差异性较大的感知哈希码,从而实现不同内容图像的识别.(3)安全性:采用不同的密钥应产生完全不同的感知哈希码.即使图像感知哈希生成算法公开,不掌握密钥的攻击者也不能伪造出与目标感知哈希一致的散列码.基于上述特征,基于感知哈希的图像内容取证算法成为实现图像内容认证和版权保护的理想选择.由于鲁棒性与区分性之间是相互制约的,一个性能优良的感知哈希算法应能够实现二者之间的良好平衡性.

经典的图像感知哈希算法依赖预先设计的特征提取器与量化器生成图像感知哈希序列,这需要广泛的专家知识,而且难以捕捉数字图像内在或抽象的视觉特征,算法性能很大程度上受制于所提取的图像特定属性信息.近年来,伴随着计算机算力与深度学习理论的发展,许多研究开始采用深度学习网络模型生成图像感知哈希码,充分利用不同内容图像的深层特征信息,提升感知哈希算法的识别鲁棒性和区分敏感性.基于生成对抗网络的强大学习能力,本文提出了一种基于双向生成对抗网络(Bidirectional Generative Adversarial Networks,BiGAN)[5]的图像感知哈希生成框架实现图像内容取证.有别于采用生成对抗网络生成与原始图像分布一致的图像,本算法利用双向生成对抗网络对原始图象隐含特征的强大学习能力,输出对相同内容图像具有识别鲁棒性,而对不同内容图像具有区分敏感性的感知哈希码.同时,为了验证基于双向生成对抗网络生成图像感知哈希算法的性能,本研究选择采用图像数量更大、图像内容纹理更加复杂的CelebA Mask-HQ图像数据库,以验证本算法在海量图像环境下实现图像内容认证与版权保护的能力.本文工作的主要贡献如下:

(1)本文提出了一种基于双向生成对抗网络的无监督感知哈希图像内容取证算法,通过编码网络、生成网络与判别网络间的双向对抗优化,充分利用双向生成对抗网络对样本图像隐含特征的强大学习能力,生成具有较高图像语义特征表示能力的图像感知哈希码,实现相同内容图像识别鲁棒性与不同内容图像区分敏感性间的优化平衡,提高图像内容认证与版权保护能力.

(2)提出了一种在编码网络与生成网络之间添加跳接层结构的双向生成对抗网络优化模型,通过融合样本图像不同维度的特征,将浅层特征与深层特征有机结合,提升生成网络的学习能力与生成图像质量,并加速网络收敛,增强感知哈希码的图像语义信息表示能力,提高感知哈希码对相同内容图像的识别鲁棒性.

(3)提出了一种基于均方误差(Mean Sequare Error,MSE)损失的双向生成对抗网络性能优化策略,有效提升生成图像的细节表现能力,增强感知哈希码对图像内容的表示准确度,提高感知哈希码对不同内容图像的区分敏感性.

本文其余部分组织结构如下:第二部分讨论了图像感知哈希的相关研究方向与工作进展;第三部分探讨了生成对抗网络(GAN)与双向对抗生成网络(BiGAN)技术原理,提出基于BiGAN的感知哈希图像内容取证算法体系结构和训练流程,详细阐述了采用跳接层网络结构与MSE损失优化感知哈希网络的具体方法;第四部分对比分析了不同网络结构和网络参数对生成感知哈希码图像取证性能的影响,并对不同感知哈希算法性能进行对比验证;第五部分对全文研究进行了总结.

2　相关工作

图像感知哈希最早由Schneider和Chang[6]在1996年的国际图像处理会议(ICIP)上提出.其通过将图像分成若干子块,提取每个子块强度直方图的均值并量化为图像的哈希码.此后,针对图像感知哈希的研究大量出现.经典的图像感知哈希技术大致可划分为基于统计特征、基于局部特征点提取、基于频域变换、基于特征降维以及基于深度学习网络等几种不同类型的图像感知哈希算法.

2.1　基于统计特征的图像感知哈希算法

基于统计特征的感知哈希算法依据图像在经过失真操作之后,相邻像素的相对关系通常保持不变的特点,采用图像的均值、方差、偏度和峰度等统计属性产生感知哈希.Tang等人[7]通过计算图像的颜色向量角矩阵,提取该特征矩阵内接圆的直方图,并使用DCT压缩直方图生成哈希码;该算法能够抵抗

大角度旋转攻击，但时间复杂度较高. Zhao等人[8]采用Zernike矩对原始图像进行变换并提取生成全局特征，由图像显著区域的位置信息和纹理信息产生局部特征，将全局特征与局部特征结合起来生成哈希，取得了很好的旋转不变性. Chen等人[9]通过量化切比切夫(Tchebichef)矩来生成图像感知哈希，由于Tchebichef矩具有良好的正交性和鲁棒性，所提出的哈希算法取得较好的图像区分能力. 此后，Tang等人[10]又研究将归一化后的图像分成不同的环，从环中提取均值、方差、偏度、峰度四个统计特征向量，最后利用特征向量之间的不变距离形成紧凑的图像散列码. 基于统计量与图像像素的位置信息无关性，该方案对几何变换攻击具有较强鲁棒性. 最近，Hosny等人[11]通过计算灰度图的Gaussian-Hermite矩，然后对不同阶的矩系数进行量化构成感知哈希，从而在安全性和抵抗噪声攻击方面取得良好表现. Zhao等人[12]提出了一种基于图像三维颜色结构特征和亮度梯度特征的哈希算法，提高彩色图像感知哈希算法的分类性能；Tang等人[13]提出了一种基于颜色向量角度和Canny算子的鲁棒图像哈希. 首先通过插值和Gaussian低通滤波将输入图像转换为归一化图像. 然后从归一化图像中提取颜色向量角度和图像边缘，结合颜色向量角度和图像边缘计算统计特征以形成图像哈希. Abbas等人[14]将局部二元模式和反向局部二元图案直方图相结合，生成用于创建哈希向量的聚合双向局部二元模型特征. 该方法不但具有优越的计算效率，而且可以定位篡改位置，但图像认证精度仍需要进一步提升. 在该类方法中，图像感知哈希码的构造利用了内容保持操作下具有不变性的统计特征. 这类统计特征通常对噪声、压缩失真有较强鲁棒性，但对于图像的细节变化不敏感，使得基于统计特征方法产生的感知哈希码对于不同内容图像区分效果往往不够理想.

2.2 基于局部特征点提取的图像感知哈希算法

基于局部特征点提取的方法利用图像边缘、角、斑点等局部信息的鲁棒性特征生成图像哈希. Wang等人[15]提出了一种用于内容真实性分析的图像取证方法，采用Harris自适应角点检测算法提取图像特征点，然后利用特征点邻域的统计信息构造哈希码，该方法对JPEG压缩、加噪、滤波等内容保护操作具有较强的鲁棒性，但生成哈希码长度很大程度上受图像大小和纹理分布的影响. Tang等人[16]从预处理后的图像中提取颜色矢量角度和图像边缘特征形成图像感知哈希码. 由于颜色矢量角度对色相和饱和度的差异很敏感，因此其在评价图像颜色感知差异时效果显著. 然而，由于该算法仅在选取的同心圆上提取边缘特征信息，所生成的感知哈希忽略了图像圆环外的结构信息，降低了感知哈希码的代表性. Ouyang等人[17]使用四元数Zernike矩和尺度不变特征变换(Scale Invariant & Feature Transform，SIFT)相结合的方式来提取哈希码，通过Zernike矩计算图像全局特征保证模型的识别准确性，而采用SIFT变换模型提取图像的局部显著性特征保障算法对内容操作的敏感性. 此外，Vadlamudi等人[18]提出了一种将特征点与DWT结合的鲁棒哈希方法，首先利用SIFT算法从LAB彩色图像的L分量中计算不变特征点，然后对提取的特征内容进行离散小波变换(DWT)，并把其结果归一化为感知哈希码. 该算法能有效抵抗图像的几何失真，在拷贝检测上也具有较好的性能. Yuan等人[19]提出了一种结合三维全局特征和局部能量特征的新哈希算法. 该方法首先通过SVD分解对图像进行压缩以形成二次图像，然后提取二次图像在三维视角下的统计特征作为全局特征，通过使用来自不同三维视角的图像层的统计特征之间的关系来生成全局特征哈希. 不仅对传统的内容保持运算具有良好的鲁棒性，而且在判别能力和抗干扰性之间取得了良好的平衡. 基于局部特征点提取的感知哈希算法依靠图像关键信息点提取图像特征，这些局部特征的固有优势在于其在几何变换下的特征不变性. 但由于细节特征信息的敏感性，这种方法对滤波和噪声攻击的抵抗能力不强，算法鲁棒性受到一定限制.

2.3 基于频域变换的图像感知哈希算法

基于频域变换的图像感知哈希算法主要包括离散余弦变换(DCT)、离散小波变换(DWT)和离散傅里叶变换(DFT). Lin等人[20]利用不同图像子块之间相同位置DCT系数之间的关系不变性设计图像哈希算法，该方法对任意程度的JPEG压缩具有较强鲁棒性. Tang等人[21]通过把图像划分为多个子块，提取每个子块DCT系数中的主要特征值来构造特征矩阵，然后通过量化列向量间的距离进行矩阵压缩进而得到紧凑的哈希码. Qin等人[22]对原始图像进行归一化处理后提取图像显著结构特征，同时对纹理复杂的数据块进行离散余弦变换提取显著性DCT系数，对所提取的特征进行拼接和降维压缩后生成图像感知哈希码，提高生成感知哈希的图像认

证能力. 受Tang等人[21]工作的启发，Huang等人[23]结合灰度共生矩阵与DCT系数生成哈希序列，从而在鲁棒性与唯一性间取得良好平衡性. Venkatesan等人[24]通过量化三级DWT系数的均值和方差构造哈希函数，该算法对常见的内容保持操作具有鲁棒性，但是只能识别较低程度的攻击. Tang等人[25]还通过Gabor滤波和混沌映射从归一化图像中提取鲁棒和安全的图像特征，然后通过单层离散小波变换进行特征压缩，并将LL子带中的DWT系数级联得到图像哈希，提高了生成感知哈希序列的鲁棒性. Swaminathan等人[26]提出了一种基于傅里叶变换的图像感知哈希方法，该方法首先对归一化的图像进行傅里叶变换并转化为极坐标表示，然后提取图像统计特征设计成哈希码. 近期，Qin等人[27]通过从DFT变换后的图像中提取稳健的频率特征并进行非均匀采样，从频率分量中选取采样点作为图像的显著特征，提高生成感知哈希码的性能. 然而，该方案在抵抗大角度旋转攻击方面的能力有待提升. Yu等人[28]提出了一种具有互补彩色小波变换(CCWT)和压缩感知(CS)的感知哈希算法，并用于图像质量评估. 其首先利用CCWT将输入彩色图像分解为不同的子带，以保留颜色通道特征信息，然后基于块的CS变换从CCWT子频带中提取特征而生成哈希序列. 由于需要关注图像细节信息以提高图形质量评价性能，该算法的鲁棒性不强. 总体来讲，频域变换后所提取的特征依赖于变换空间中频率系数的相对稳定性构造图像感知哈希码，这类算法一般只针对特定的攻击方案有效，而对于多种组合攻击方案鲁棒性能不强.

2.4　基于特征降维的图像感知哈希算法

近年来，基于特征降维的图像感知哈希算法也取得很多进展，典型的方法包括奇异值分解(Singular Value Decomposition，SVD)、非负矩阵分解(Non-negative Matrix Factorization，NMF)、局部线性嵌入(Locally Linear Embedding，LLE)、主成分分析(Principal Component Analysis，PCA)、多维标度(Multi-Dimensional Scaling，MDS)等图像感知哈希码生成方案. Kozat等人[29]首先提出基于奇异值分解(SVD)的图像感知哈希码生成方案，采用低阶SVD系数的稳定性保障哈希算法的鲁棒性，为基于特征降维的图像哈希方案奠定了基础. Tang等人[30]利用环形分割技术构造具有旋转不变性的二级特征图像，并将NMF应用于特征图像，利用NMF系数以生成哈希码，从而能够抵抗大角度旋转攻击. 另外，Tang等人[31]还通过对随机子块组成的图像矩阵表示应用LLE降维操作后，取嵌入向量的方差生成感知哈希码，该方法对高斯滤波和JPEG压缩等操作取得较好的鲁棒性. 此后，Tang等人[32]在此基础又提出了一种DCT与LLE相结合的感知哈希生成算法，进一步提升了所生成感知哈希码的区分性能. Zhu等人[33]提出了一种基于PCA的图像感知哈希生成方法，该算法首先对数据集聚类，然后对其进行主成分分析并量化得到图像感知哈希码. 然而，该算法对图像内容变化比较敏感. Tang等人[34]利用对数极坐标和DFT变换提取归一化图像特征矩阵，通过MDS变换量化生成感知哈希码，该方案可以抵抗大角度的旋转攻击，但针对不同内容图像区分性表现不佳. Huang等人[35]设计了一种基于局部保持投影(Locality Preserving Projections，LPP)的感知图像哈希. 首先利用Gabor滤波自适应地提取方向和结构特征，然后采用LPP从最大Gabor滤波响应中学习有意义的低维信息，从而提高感知哈希码的图像识别能力. 该算法不仅对传统的内容保持操作具有良好的鲁棒性，也取得很好的图像篡改检测能力. 总体上讲，基于特征降维的感知哈希生成方案可以有效地消除高维特征的冗余信息，但是高强度的压缩降维使得哈希码区分能力较弱，因此在压缩生成感知哈希码的同时，有效提取图像关键特征是该类方法的亟待解决的问题.

2.5　基于深度学习的图像感知哈希算法

传统图像感知哈希算法大都依赖于预先设计的特征提取器与量化器生成图像感知哈希序列，算法性能很大程度上受制于所设计的特征提取算法. 由于不能充分利用相同内容和不同内容图像间的比较特征来对抗增强感知哈希算法的性能，因而难以对所提取特征进行优化和提升，导致经典感知哈希算法的性能优化能力不强. 近年来，基于深度学习网络的感知哈希算法充分利用不同类型图像(如：原始图像、相同内容受攻击图像和不同内容图像)的深层特征信息，有效提升了感知哈希算法的图像内容取证能力. Liong等人[36]利用神经网络多级非线性变换映射特性，搭建了基于深度神经网络生成图像感知哈希的算法，提高感知哈希序列的语义表示能力. Li等人[37]提出了一种能够同步实现特征提取与哈希码生成的神经网络，并通过实验证明该方案能够在鲁棒性与敏感性之间取得较好平衡. Deng等人[38]提出了一种跨模态感知哈希生成方案，采用三元组标签描述多模态数据间的相对关系，从而更好

地捕捉跨模态数据类内和类间的变化以形成更有代表性的感知哈希码. Liu等人[39]采用卷积神经网络(CNN)提取图像潜在特征生成感知哈希序列,用于学习和检测未经授权的可疑数字图像. Li等人[40]提出了一种数据驱动的图像感知哈希算法,训练神经网络自动搜索从图像到感知哈希的最佳映射. 为了改善训练的难度,其首先以分层方式训练指纹计算网络,逐步提高其对内容保持失真的鲁棒性,然后将网络重新训练为一个整体单元,最大化其内容识别准确性. Qin等人[41]通过权重分配策略将两对约束集成到一个整体约束函数中,构造了一种基于多约束卷积神经网络(CNN)的感知图像哈希生成方案,根据约束值的变化动态调整训练集结构,自动学习图像特征并生成感知哈希序列,提高感知哈希算法的鲁棒性和区分性. Sun等人[42]根据原始图像内容对训练图像进行分类,并使用Hadamard矩阵为Hamming空间中的每个类生成哈希中心;然后利用卷积神经网络自动学习图像的特征,使每类图像的哈希码收敛到其类的哈希中心,并生成最终的图像感知哈希码,从而在感知的鲁棒性和区分性之间取得平衡.

然而,上述基于深度学习的感知哈希算法仍需要人工标注标签,由于标签的标注往往依赖于人类的经验知识,限制了标注的准确性且费时费力,导致了网络对训练数据敏感且在实际应用中泛化性能不强. 不同于有监督图像感知哈希模型,无监督哈希模型不需要进行数据标注,而是通过网络自动学习图像的隐含特征,生成基于图像内容的感知哈希序列. Song等人[43]将生成对抗网络(Generative Adversarial Networks,GAN)的输入变量设置为二进制噪声,并以每个输入图像的特征为条件,生成对抗网络可以同时学习每个图像的二值化表示,生成图像的感知哈希表示. Lin等人[44]设计了一种无监督的深度神经网络,通过最小误差损失、哈希码元素分布均匀性和独立性三种约束提高感知哈希码生成质量. 由于没有语义标签的引导,自然图像在重建过程中通常包含很多变化的细节信息,如大小、斑点和形状等,导致模型对这些细节信息比较敏感,算法鲁棒性能有待提升. 总体上来讲,当前基于无监督深度学习网络实现感知图像哈希算法的研究处于起步阶段,算法的性能还有较大的提升空间.

因而,本文提出一种基于双向生成对抗网络的无监督感知哈希图像内容取证算法,基于双向生成网络对目标图像潜在特征空间的学习能力,并采用跳接层网络与MSE损失优化算法,增强双向生成对抗网络生成感知哈希码的图像内在特征表示能力,提高图像内容认证与版权保护性能.

3 基于BiGAN的图像感知哈希算法

生成式对抗网络(GAN)[45]最早由Goodfellow等人首先提出,并在图像合成、信息隐藏等领域得到了广泛应用[46-47]. 一个基本的生成对抗网络由生成网络G和判别网络D两部分组成,生成网络G负责学习数据的特征分布,判别网络D估计输入样本来自真实数据空间R还是来自生成网络G的概率. 生成对抗网络的优化是一个极小极大的博弈过程,其通过生成网络G和判别网络D的迭代对抗和误差反向传播,使得生成网络G具备生成与训练样本分布一致数据的能力.

生成对抗网络经过多轮迭代对抗可将预设的随机噪声信息映射成为任意复杂的数据分布(基本结构如图1所示). 其中,生成网络G通过生成与训练样本高度一致的数据分布,欺骗判别网络D将其分类为真实的数据样本. 这一结果表明生成网络G能够在与判别网络的对抗中获取训练数据的语义特征信息. 然而,尽管生成对抗网络可以实现从伪随机噪声到任意数据分布学习,但该框架并不包含从训练样本数据到潜在特征表示的映射. 双向生成对抗网络(BiGAN)在基于随机噪声生成与训练样本高度一致数据分布的同时,通过加入编码网络E对抗学习样本数据的隐空间特征信息,实现生成网络的逆映射,从而能够提取训练数据的内在特征信息.

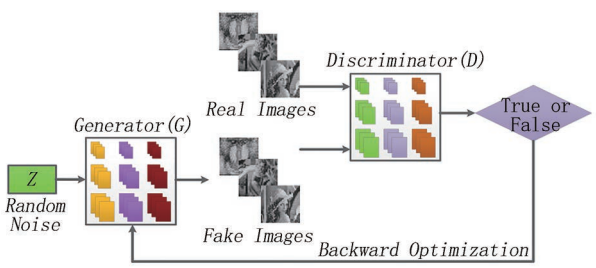

图1 生成对抗网络GAN原理示意图

双向生成对抗网络包含生成网络G、编码网络E和判别网络D三部分组成(如图2所示). 其中,编码网络E将真实数据x映射成为隐空间特征表示$E(x)$,形成分布$P_E(E(x)|x)$;生成网络G将预设的噪声变量z映射到与训练数据一致的数据分布$G(z)$,形成分布$P_G(G(z)|z)$. 判别网络不仅在数据

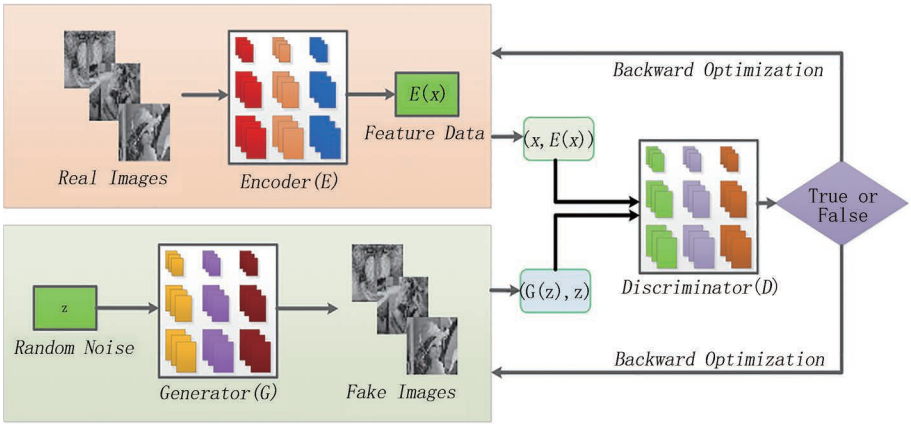

图2　双向生成对抗网络BiGAN原理示意图

空间鉴别x和$G(z)$，同时也在数据隐特征空间区分z和$E(x)$，判断联合数据元组$(x,E(x))$与$(G(z),z)$是来自真实数据分布还是来自生成数据分布，输出判别概率$P_D\big(R|(x,E(x)),(G(z),z)\big)$. 当数据元组来自真实数据分布$(x,E(x))$时，判别概率趋近于1；当数据元组自生成数据分布$(G(z),z)$时，判别概率趋近于0.

双向生成对抗网络是一个高度智能的无监督特征学习模型，其不需要预设数据的结构和类型，即可以主动的学习训练样本的隐含特征. 编码网络E通过提取样本数据的语义属性，产生训练样本隐空间特征表示$E(x)$，生成网络G通过学习训练样本不同维度的特征，输出数据分布$G(z)$. 通过判别网络D与编码网络E、生成网络G间的对抗训练，同步提高生成网络G输出数据分布的质量，以及编码网络E生成训练样本隐空间特征表示的能力.

双向生成对抗网络的目标优化函数为：

$$\min_{G,E}\max_{D}V(D,E,G)=$$
$$E_{x\sim p_x}[\log D(x,E(x))]+$$
$$E_{z\sim p_z}[\log(1-D(G(z),z))] \tag{1}$$

由式(1)可知，双向生成对抗网络的训练过程也是一个极小和极大博弈的过程，通过判别网络D、生成网络G和编码网络E间的多重迭代对抗，激励生成网络G生成与训练样本更近似的数据分布，同时促使编码网络E输出更具代表性的数据特征编码.

3.1　基于BiGAN的感知哈希图像内容取证网络

双向生成对抗网络采用对抗元组的策略，不但能够生成与训练样本一致的数据分布，而且能够输出训练样本的隐空间特征表示. 通过网络间的多重迭代对抗，双向生成对抗网络取得捕捉训练样本数据隐空间特征表示的能力. 因而，充分利用双向生成网络的图像语义特征抽象能力，生成具有图像本质特征描述能力的特征序列，可构建具有更强图像内容表示能力的感知哈希码，提高图像内容取证性能. 本研究提出基于BiGAN的感知哈希图像内容取证算法，通过对双向生成对抗网络结构进行优化和增强处理，提高双向生成对抗网络学习复杂数据分布的能力，生成与复杂训练样本(如：自然图像)一致的数据分布；同时，增强网络对样本数据潜在特征的学习性能，输出样本图像的隐空间特征表示并量化生成感知哈希码，实现基于感知哈希的图像内容认证与版权保护.

基于BiGAN的图像感知哈希生成网络的基本模型由编码网络E、生成网络G、联合判别网络D和跳接层网络S四个子网络组成(如图3所示). 其中，编码网络E实现从原始图像数据x到潜在特征表示$E(x)$的映射$E:x\to E(x)$其输入为归一化后的训练图像，输出为图像隐空间特征编码. 生成网络G将预设的噪声分布z映射为与目标图像样本一致的数据分布$G(z)$，$G:z\to G(z)$. 联合判别网络D区分输入的数据元组是来自编码网络还是生成网络D：$\big((x,E(x)),(G(z),z)\big)\to\{0,1\}$. 针对双向生成对抗网络所存在的生成图像质量不高、输出特征码表示能力不足的问题，通过在编码网络E和生成网络G之间添加跳接层网络S，实现编码网络E和生成网络G之间不同维度的特征信息传递，并在网络优化损失中添加均方误差(MSE)损失，增强生成网络G所生成图像的内容表示能力，使得生成网络G能够输出具有复杂纹理分布的高质量图像. 同时，基于联合判别网络D的对抗损失，反向激励编码网络E输出具有更强代表性的图像隐空间特征编码，提高生成图像感知哈希码的质量.

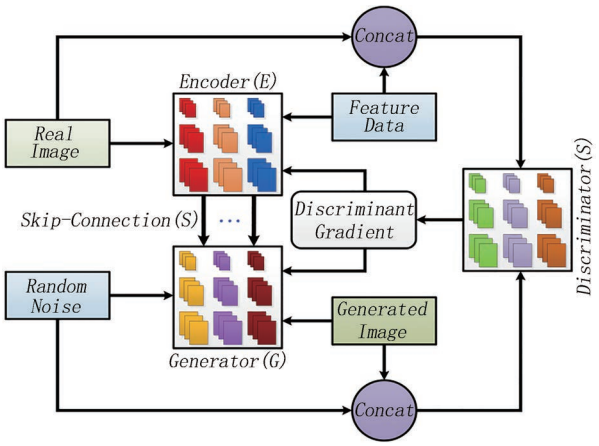

图3 基于BiGAN的感知哈希图像内容取证算法结构图

其中，E代表编码网络，G代表生成网络，D代表联合判别网络，S是加入的跳接层网络，Real Image为训练样本图像，Generated Image代表生成图像.

基于BiGAN的感知哈希生成算法充分利用双向生成对抗网络的自学习能力，将编码网络E输出的隐空间特征编码量化生成图像感知哈希码. 通过生成网络G、编码网络E和联合判别网络D之间的迭代对抗，以及跳接层网络S和均方误差损失MSE的优化增强，不断提升图像感知哈希码的隐空间特征表示能力，实现感知哈希码认证鲁棒性与区分敏感性之间的有效平衡.

3.1.1 编码网络设计

基于BiGAN的感知哈希生成网络中，编码网络E的作用是从原始图像提取隐空间特征信息生成图像感知哈希序列. 本研究中设计编码网络包含10层卷积神经网络，每层卷积神经网络包括图像卷积(*Conv2d*)、批归一化(*BatchNorm*)和激活(*LeakRelu*)三种数据操作处理. 编码网络的初始输入为归一化处理后的训练样本图像；同时，为了提高输出特征编码的隐空间特征表示能力，编码网络E的第五、六、七层卷积层输出通过跳接层网络S传递到生成网络G的同维度网络层中作为输入，将编码网络E提取的不同维度图像特征信息链接到生成网络G中，提高生成图像的视觉质量与网络收敛速度. 编码网络E最后一层卷积输出作为图像隐空间特征表示序列生成图像感知哈希码. 基于BiGAN的感知哈希生成网络中编码网络E的详细设置参数信息如表1所示. 其中，配置栏第一列参数代表滤波器数量，可知编码网络E采用倍乘方法增加滤波器数量，最终形成具有良好隐空间特征表示能力的图像感知哈希码. 第二列参数是卷积操作的感受野大小、卷积步长以及添加到输入侧的行/列数；第三列参数为所采用激活函数*LeakRelu*的斜率. 除最后一层卷积层外，实验中将批归一化操作*BatchNorm*应用到每一次卷积运算之后，以保障训练样本数据分布在激活函数比较敏感的区域，从而避免梯度消失，加快模型收敛速度. 实验中，综合考虑卷积层数量对网络运行效率与生成感知哈希码特征表示能力的影响，选择具有10层网络结构的自学习编码网络结构模型，其中包含10个卷积层、9个批归一化层，9个激活层，以保障网络输出具有较强表示能力的隐空间特征编码.

表1 编码网络详细参数设计

	层数	函数	配置
E_X	1	Conv2d,BatchNorm2d,LeakyRelu	32,[3,1,1],0.01
	2	同上	64,[4,2,1],0.01
	3	同上	128,[4,2,1],0.01
	4	同上	256,[5,1,0],0.01
	5	同上	512,[4,2,0],0.01
	6	同上	512,[4,1,0],0.01
	7	同上	512,[4,2,0],0.01
	8	同上	1024,[4,1,0],0.01
	9	同上	2048,[1,1,0],0.01
	10	Conv2d	1024,[1,1,0]

3.1.2 生成网络设计

生成网络G的目标是将预设的随机噪声分布映射成高质量图像分布，并基于联合判别网络D提供的损失梯度，不断优化生成图像质量. 生成网络G同样采用卷积神经网络架构，共包含8个反卷积(*ConvTranspose2d*)层，7个批归一化(*BatchNorm*)层和7个激活(*LeakRelu*)层，初始输入为符合高斯分布的预设随机噪声. 第二、三、四反卷积层的输入为前一层的输出与编码网络E对应层输出的叠加，通过跳接层将编码网络E中不同维度的图像特征信息传递到生成网络G中，从而增强生成网络G对样本图像特征的学习能力，提升生成图像质量和网络收敛速度；实验中在生成网络G的最后一层使用*Tanh*激活函数增强生成图像输出细节信息的能力. 生成网络G的详细参数如表2所示，分别为反卷积层数，每层反卷积所采用的二维卷积(*ConvTranspose2d*)、批归一化(*BatchNorm*)和激活(*LeakRelu*)三种数据处理操作，以及每层卷积的参数配置(具体参数含义与表1一致). 生成网络通过

表2　哈希生成网络详细参数设计

	层数	函数	配置
G_Z	1	ConvTranspose2d,BatchNorm2d,LeakyRelu	2048,[4,1,0],0.01
	2	同上	1024,[4,2,0],0.01
	3	同上	512,[4,1,0],0.01
	4	同上	256,[4,2,0],0.01
	5	同上	128,[5,1,0],0.01
	6	同上	64,[4,2,1],0.01
	7	同上	32,[4,2,1],0.01
	8	ConvTranspose2d , Tanh	1,[1,1,0],0.01

逐层反卷积操作，不断增大矩阵维数，最终生成和训练样本图像维数大小一致的图像，同时，通过跳接层S为生成网络G传递样本图像不同维度的特征信息，加快生成网络G的学习能力与收敛速度，增强网络性能.

3.1.3　联合判别网络设计

联合判别网络D判断数据元组$(x,E(x))$和$(G(z),z)$是来自编码网络E还是来自生成网络G，并输出二者的差异作为反向传播误差损失优化生成网络G和编码网络E. 联合判别网络D的输出为介于0～1间的数据分布，其目标是对来自编码网络E的数据元组尽量赋高值(靠近1)，而对来自生成网络G的数据元组赋低值(靠近0)，以优化生成网络G生成与训练样本一致的数据分布，同时激励编码网络E输出更具代表性的样本图像隐空间特征表示. 联合判别网络采用梯度上升策略，通过多轮迭代优化提升对输入数据元组的鉴别能力，为生成网络G和编码网络E提供性能优化的损失梯度.

联合判别网络D采用多层卷积神经网络提取输入的数据元组特征，并采用*Sigmoid*激活函数输出数据元组的真实性评价. 联合判别网络D的详细参数如表3所示，分别为卷积层数，每层卷积所包含的二维卷积(*Conv2d*)、激活(*LeakRelu*)和消除(*Dropout2d*)三种数据操作处理，以及每层卷积的参数配置. 配置栏中的参数分别为输出通道数、卷积核大小、步长、填充大小、*LeakyRelu*激活函数斜率和*Dropout2d*丢失率. 考虑图像本身数据分布的复杂性，判别网络采用13层卷积神经网络，其中8层用于图像真实性判别，2层用于感知哈希码的真实性判别，3层用于联合编码的真实性判别，并在输出层叠加采用*Sigmoid*激活函数输出对数据元组的真实性判定结果.

3.1.4　跳接层网络设计

本研究通过在编码网络E和生成网络G之间添加跳接层网络S，将编码网络E中间层的特征信息链接到生成网络G相应的网络层中(编码网络第五、六、七层的输出特征链接到生成网络的四、三、二层的输入). 编码网络通过多层网络卷积，基于不同的卷积核形成不同的特征通道，每个特征通道所提取的图像特征以特征向量的形式进行表征. 同时，采用跳接层网络将编码网络与生成网络同维度特征向量进行concatenate链接处理，增加生成网络的卷积层通道数量，从而提升网络特征信息描述能力. 如图4所示，编码网络将第5层的输出(13×13维度)512通道特征信息与生成网络第4层输入(13×13维度)256通道特征信息进行concatenate链接处理；编码网络将第6层的输出(10×10维度)512通道特征信息与生成网络第3层输入(10×10维度)512通道特征信息进行concatenate链接处理；编码网络将第7层的输出(4×4维度)512通道特征信息与生成网络第2层输入(4×4维度)1024通道特征信息进行concatenate链接处理. 基于跳接层网络特征信息

表3　联合判别网络详细参数设计

	层数	函数	配置
X_M	1	Conv2d,LeakyRelu,Dropout2d	32,[3,1,1],0.01,0.2
	2	同上	64,[4,2,1],0.01,0.2
	3	同上	128,[4,2,1],0.01,0.2
	4	同上	256,[5,1,0],0.01,0.2
	5	同上	512,[4,2,0],0.01,0.2
	6	同上	512,[4,1,0],0.01,0.2
	7	同上	512,[4,2,0],0.01,0.2
	8	同上	1024,[4,1,0],0.01,0.2
Z_M	1	同上	2048,[1,1,0],0.01,0.2
	2	同上	1024,[1,1,0],0.01,0.2
$Joint_M$	1	同上	2048,[1,1,0],0.01,0.2
	2	同上	1024,[1,1,0],0.01,0.2
	3	Conv2d, LeakyRelu, Sigmoid	1,[1,1,0],0.01

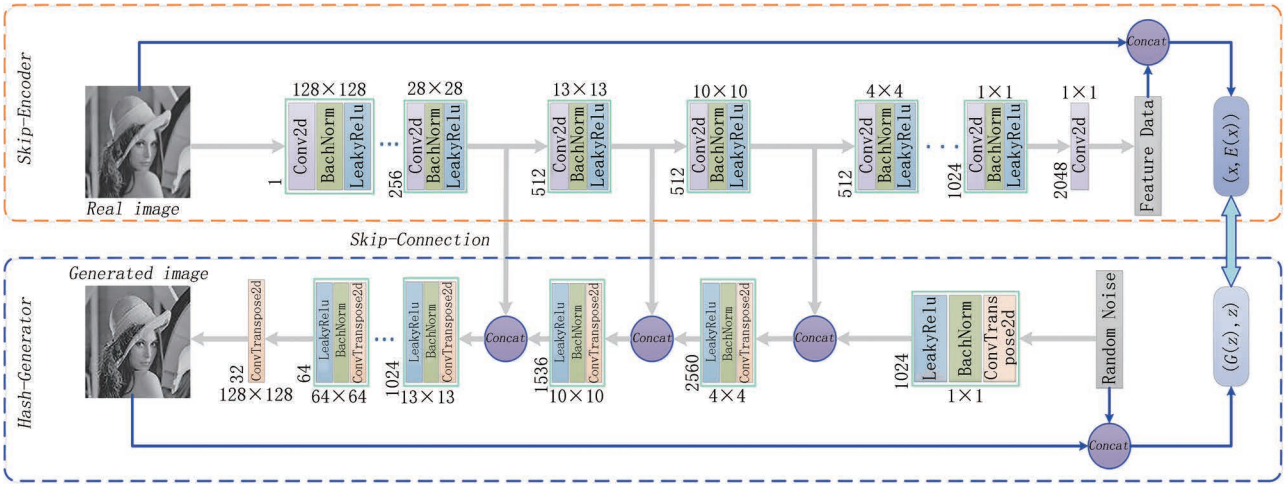

图4 跳接层增强网络结构示意图

增强处理，在生成网络反卷积过程中，通过添加来自编码网络所采集的图像隐空间特征信息，提升生成网络特征学习能力，从而输出与训练样本数据分布更加一致的生成图像. 跳接层网络结构通过将编码网络E与生成网络G的同维度特征信息进行链接，将原始图像不同维度的编码特征传递给生成网络G，拼接来自编码网络E的不同卷积层特征信息与生成网络G的对应层特征信息，从而融合不同维度下的样本图像特征信息，生成高质量的图像与更具代表性的感知哈希编码，并加速网络收敛.

跳接层增强网络中，卷积运算模块右侧数字表示卷积核个数，上/下方数字表示该层的输入特征图大小，跨网络跳接箭头表示通过跳接层网络将编码网络E所提取的图像特征信息传递并拼接到生成网络G对应维度的卷积层.

采用同层跳接结构将编码网络E在特定网络层所提取的特征信息传递到生成网络G，通过拼接来自编码网络的特征信息和生成网络的反卷积信息，可充分利用样本图像不同维度的特征，增强生成网络学习效率并生成高质量图像. 同时，基于网络判别损失同步优化编码网络E与生成网络G，激励编码网络产生更具代表性的隐空间特征信息，增强输出感知哈希码的语义特征表示能力，提高感知哈希码图像识别的鲁棒性能.

3.2 损失函数设计

基于BiGAN的图像感知哈希算法核心是通过训练编码网络E、生成网络G、联合判别网络D以及跳接层网络S，生成具有较强代表性的图像隐空间特征表示，并构造图像感知哈希码，实现相同内容图像鲁棒性认证和不同内容图像敏感性区分. 一方面，联合判别网络D通过梯度上升策略增强对来自编码网络E和生成网络G数据元组的区分能力，并输出反向传播误差损失优化编码网络E和生成网络G. 另一方面，编码网络E和生成网络G采用梯度下降算法不断缩小输出元组$(x, E(x))$和$(G(z), z)$间的差异，通过双向生成对抗网络BiGAN的多重迭代，最终形成与训练样本高度一致的输出图像分布和具有较强代表性的隐空间特征编码. 实验中采用基于Adam的误差迭代算法，同时考虑误差损失的一阶和二阶矩估计计算网络更新步长，通过自适应网络步长优化和系统多轮迭代，增强算法收敛的稳定性.

进一步地，本研究在编码网络E和生成网络G的损失函数中增加了均方误差(Mean-Square Error，MSE)损失，提高编码网络输出数据元组$(x, E(x))$与生成网络输出数据元组$(G(z), z)$的一致性程度，并加速网络收敛. MSE通过计算预测数据和原始数据对应点误差平方和的均值，评价数据的变化程度. MSE的值越小，说明预测模型描述原始数据的精确度越高. 因而，在双向生成对抗网络中添加MSE损失，一方面，通过计算生成图像与原始图像对应像素间的误差，表征生成图像与原始图像间的数据一致性程度，增强生成图像的细节表示能力；另一方面，通过缩小编码网络输出特征序列与原始序列特征值之间的差异，不断提升输出特征编码的图像语义特征表示能力. 实验中，通过在基于BiGAN的感知哈希生成网络中添加MSE损失，可以为编码网络E和生成网络G提供更好的性能优化梯度，增强生成网络G的细节特征学习能力，从而提高生成图像细节描述性能；同时，激励编码网络输出更具图

像内容特征表示准确度的感知哈希序列，增强感知哈希码对不同内容图像的区分敏感性.

为保障BiGAN双向生成对抗网络快速稳定收敛，并提高生成图像和感知哈希码质量，设计基于BiGAN的感知哈希生成网络损失函数为：

$$Loss_D=-\frac{1}{N}\sum_{i=1}^{N}[\log(D(x_i,E(x_i)))+\log(1-D(G(z_i),z_i))] \tag{2}$$

$$Loss_{E,G}=-\frac{1}{N}\sum_{i=1}^{N}[\log(D(G(z_i),z_i))+\log(1-D(x_i,E(x_i)))]+\lambda Loss_{MSE} \tag{3}$$

$$Loss_{MSE}=-\frac{1}{N}\sum_{i=1}^{N}(x_i-G(z_i))^2 \tag{4}$$

式中，$x\in X$，$z\in Z$；x_i表示第i个训练图像；z_i表示第i个随机噪声向量；λ为$Loss_{MSE}$的权值系数；N表示样本数量. 实验中采用公式(2)优化联合判别网络D，公式(2)取值越大，$D\big(x,E(x)\big)$取值越大(靠近1)，$D\big(G(z),z\big)$取值越小(靠近0)，也即判别能力越强；采用公式(3)优化编码网络E与生成网络G，公式(3)取值越小，$D\big(G(z),z\big)$取值越大(靠近1)，$D\big(x,E(x)\big)$取值越小(靠近0)，生成网络G输出元组$\big(G(z),z\big)$与编码网络E输出元组$(x,E(x))$的相似度越高，也即生成的数据分布具有更多的样本图像细节信息，同时，生成的隐空间特征表示包含更多的图像语义特征.

3.3　基于BiGAN的感知哈希网络训练方法

作为一种典型的无监督深度学习网络框架，生成对抗网络可以将预设的随机噪声映射为任意有意义数据分布，当预设随机分布发生改变时，所生成的数据分布也随之发生改变，这一现象表明特定的随机分布中包含生成数据分布的隐特征信息. 相应的，基于任意有意义的数据分布也应该可以通过一个无监督学习网络映射出其对应的隐空间特征表示，因而，采用双向生成对抗网络学习图像有意义的特征表示在理论上是可行的.

由前述分析可知，每一个双向生成对抗网络BiGAN的组成模块都是一个可以通过参数优化的深度卷积神经网络. 设编码网络E、生成网络G和联合判别网络D的参数分别为用φ、ζ、θ表示. 基于BiGAN的感知哈希图像内容取证网络可通过如下交替优化策略进行模型训练；一方面，在每一轮的迭代过程中，联合判别网络D的参数采用梯度上升策略进行优化(沿着梯度变化的正方向$\nabla_\theta V(D,E,G)$)，由于联合判别网络D的变化速度较慢，每一轮交替训练中，判别网络都进行多次迭代优化，以最大化联合判别网络D的识别性能. 另一方面，编码网络E和生成网络G同时采用梯度下降法进行优化(沿着梯度变化的负方向$-\nabla_{\varphi,\zeta}V(D,E,G)$)，每一轮交替训练中，编码网络$E$和生成网络$G$都只进行一次迭代优化. 双向生成对抗网络价值函数$V(D,E,G)$基于输入样本$\{x_i\sim p_X\}_{i=1}^{N}$和$\{z_i\sim p_Z\}_{i=1}^{n}$进行优化更新. 其中，编码网络$E$和生成网络$G$是一对“互逆”的优化过程，二者通过同步优化降低双向生成对抗网络损失，加速网络收敛.

基于BiGAN的感知哈希网络采用Jensen－Shannon散度计算数据元组$(x,E(x))$和$(G(z),z)$的概率分布差异. 编码器E和生成器G采用梯度下降法迭代降低误差损失，实现网络参数的优化调整. 理想生成网络G应满足$G(E(x))=x$，当生成网络输出图像与样本图像分布完全一致时，存在$E(x)=z$. 此时，判别网络无法区分数据源组来自编码网络E还是生成网络G(联合判别器的输出为0.5)，编码器输出元组$(x,E(x))$和生成器输出元组$(G(z),z)$完全一致，编码器E和生成器G满足$E=G^{-1}$. 编码网络输出目标图像最优隐空间特征编码，基于此编码量化生成感知哈希码，可以最大限度实现图像内容认证与版权保护.

实验过程中，首先从训练集X中随机抽取N个真实数据样本. 然后，随机初始化编码网络E、生成网络G以及联合判别网络D的参数φ、ζ、θ，将经过预处理的图像输入到编码网络并输出特征编码：

$$E(x_i)=z_i',\ (i=1,2,\cdots,N) \tag{5}$$

在生成网络G中，采用预设的随机高斯分布噪声序列$z_i(i=1,2,\cdots,N)$作为输入，对抗生成与样本图像近似的数据分布：

$$G(z_i)=x_i',\ (i=1,2,\cdots,N) \tag{6}$$

考虑到基础双向生成对抗网络生成图像质量不高，实验中在编码网络E和生成网络G的卷积层之间添加跳接层网络结构，将原始图像不同维度的特征信息传递到生成网络中，增强生成网络G的学习能力，提高生成图像的视觉质量，并加速哈希生成网络的学习速度.

在基于BiGAN的感知哈希网络中，编码网络E的输入x和特征编码输出$E(x)$组成的数据元组$(x,E(x))$，以及生成网络G的输入z和生成图像输出$G(z)$组成的数据元组$(G(z),z)$共同作为联合判

别网络D的输入变量，联合判别网络基于梯度上升迭代算法，通过多重迭代训练将来自编码网络E的输入判别为真（靠近1的分布），而将来自生成网络G的分布判别为假（靠近0的分布）. 实验中利用$Adam$优化算法更新联合判别网络参数θ，联合判别网络参数采用梯度上升迭代方法：

$$\text{即：}\theta \leftarrow \theta + \gamma_D \nabla V_\theta \tag{7}$$

$$\nabla V_\theta = \frac{\partial}{\partial \theta}\left[\log(D(x_i, E(x_i))) + \log(1 - D(G(z_i), z_i))\right] \tag{8}$$

其中，γ_D为学习速率. 训练过程中，联合判别网络D通过梯度上升迭代算法，不断增强对输入元组的判别能力.

实际训练过程中，当G的性能较弱时，联合判别网络很容易区分出生成图像和原始图像的差异，此时$\log(1-D(G(z)))$处于饱和的位置，不能为生成网络G提供足够的梯度. 因而，实验中采用“逆”对抗损失函数Λ对生成网络G进行优化：

$$\Lambda = \log(D(G(z_i), z_i)) + \log(1 - D(x_i, E(x_i))) \tag{9}$$

如公式(9)所示，在“逆”对抗损失函数Λ中，通过引入分量 $\log(1-D(x_i, E(x_i)))$获取较大的初始优化梯度，可有效提高网络收敛速度.

此外，基于BiGAN的感知哈希网络还通过添加MSE损失，增强编码网络E和生成网络G的优化能力，加速哈希生成网络收敛进程并提升生成图像细节表示能力，激励编码网络输出更具区分敏感性的感知哈希码. 基于BiGAN的感知哈希生成网络总的优化函数为：

$$\min_{E,G}\max_{D}\Lambda = \log(D(G(z_i), z_i)) + \log(1 - D(x_i, E(x_i))) + \lambda(x_i - G(z_i))^2 \tag{10}$$

编码网络E与生成网络G的网络参数ψ，ζ也采用$Adam$优化算法进行更新，其参数更新迭代方法如下：

$$\psi \leftarrow \psi - \gamma_E \Lambda, \zeta \leftarrow \zeta - \gamma_G \Lambda \tag{11}$$

其中，γ_E，γ_G为学习速率. 通过编码网络E、生成网络G与联合判别网络D之间的多轮迭代和对抗，生成网络G产生与样本图像一致的数据分布，编码网络输出具有更强代表性的输入图像隐空间特征表示. 基于BiGAN的感知哈希算法迭代优化流程如算法1所示：

算法1. 基于BiGAN的感知哈希优化算法

输入：D，E，G，训练数据x_i

输出：更新后的D，E，G

1. In Each Iteration：
2. 从数据集X中取出N个样本$x_1, x_2, \cdots, x_N$
3. $z_i' = E(x_i)$
4. 从正态随机分布Z中生成N个样本$z_1, z_2, \cdots, z_N$
5. $x_i' = G(z_i)$
6. 根据公式(2)优化D的参数
7. 根据公式(3)优化E，G的参数
8. Return 优化后的D，E，G网络

3.4 感知哈希码量化生成与置乱加密

编码网络E所输出的图像特征编码序列由一组实数构成，其数值为不同卷积核与通道特征矩阵的内积. 因而，如果直接采用编码网络输出的特征序列作为感知哈希码，会大大增加感知哈希码的复杂度，降低感知哈希码的使用效率. 为了提升感知图像哈希的编码效率，降低编码复杂性和空间占用，实验中采用均值量化策略对特征编码序列进行处理并生成二进制感知哈希码. 首先计算特征编码序列中所有元素的平均值. 然后，根据公式(13)比较每个元素与平均值的大小，当元素的值大于等于平均值时，取感知哈希码的元素值为“1”，当元素值小于平均值时，感知哈希码的元素值取“0”. 最终，根据均值量化策略得到一个二进制特征向量，作为原始图像的感知哈希码. 特征序列量化过程如式(12-13)所示：

$$\bar{A} = \frac{1}{m}\sum_{k=0}^{m} A(k), \quad A(k) \in R \tag{12}$$

$$H_m(k) = \begin{cases} 1, & A(k) > \bar{A} \\ 0, & \text{otherwise} \end{cases} \tag{13}$$

其中，m为特征编码序列长度，$A(k)$代表编码网络输出的第k位实数特征编码，$H_m(k)$表示哈希序列中的第k位二进制数.

同时，考虑到感知哈希码的唯一性是实现图像版权认证的重要特征，实验中采用随机置乱的方式对所产生的二进制哈希序列进行置乱处理，不同参与者采用不同的随机置乱密钥对感知哈希进行置乱加密，从而保障不同用户所产生的图像感知哈希码的唯一性，避免针对图像感知哈希码的伪造和碰撞攻击. 随机置乱过程如式(14)所示：

$$H(k) = H_m(S(k)) \tag{14}$$

其中，$S(k)$表示随机置乱序列中的第k位随机数，$H(k)$是置乱后的第k位二进制数.

4　实验结果与分析

在这一部分中，我们基于双向生成对抗网络的无监督自学习特征，首次选择采用CelebA Mask-HQ大型数据集验证基于BiGAN的感知哈希图像内容取证算法性能，将编码网络E输出的隐空间特征信息进行量化后，作为感知哈希编码实现图像内容取证，并将试验结果与当前主流的高性能图像感知哈希算法进行比较，评价本算法的先进性.

4.1　实验设计

实验中从数据库中随机选取12 000幅图像作为训练集，8000幅作为测试集. 原始图像均为1024×1024的彩色人脸图像，考虑面部图像特征的复杂性，为提高感知哈希运算效率，降低服务器运算代价，实验中将所有的图像进行灰度化并采用双线性插值运算统一将原始图像变换为128×128的灰度图像. 实验结果表明尽管将原始图像缩放为128×128的灰度图像会产生一些图像细节内容的损失，基于BiGAN的感知哈希生成算法仍然可以实现高精度图像内容取证，有效提高算法的图像内容认证与版权保护性能.

由于本研究首次采用大规模的CelebA Mask-HQ数据库开展基于感知哈希的图像内容取证研究，并采用无监督的BiGAN深度学习网络自动提取样本图像的特征信息，为避免出现特征信息碰撞，准确描述每幅图片的隐空间特征信息，实验中采用1024位感知哈希编码，以提高感知哈希码对相同内容图像的认证鲁棒性和对不同内容图像的区分敏感性. 实验选取pytorch框架开展基于BiGAN的感知哈希图像内容取证算法研究，并随机初始化网络参数. 所有实验都采用配置Intel 酷睿i9CPU，32GB内存，2TB硬盘，英伟达V100 32GB显存的DeLL R740图形工作站实现.

同时，为保持算法性能评价度量的一致性，方便与其他优秀的感知哈希算法进行比较，实验中分别采用互相关值、受试者操作特征(Receiver Operating Characteristic Curve，ROC)曲线、查准率-召回率(Precision-Recall，PR)曲线以及样本识别准确率[48]作为评价指标对算法性能进行验证.

经典的感知哈希算法基于有限的数据样本检验哈希算法的性能，采用汉明距离统计所生成的感知哈希码中不同元素的个数作为感知哈希相似性的评价手段. 而当样本图像数量加大时，仅仅通过计算感知哈希码间不同元素的个数，显然忽略了元素所在位置的权值对实验结果的影响，难以准确描述不同感知哈希码间的相似性程度. 因而，实验中采用互相关值计算不同感知哈希码的数值距离，表征其相似性程度，互相关值能够同时反映感知哈希码中不同元素的数值及其所在位置权重的影响，因而更适合基于大型数据库的样本图像感知哈希码相似性评价与图像内容取证. 实验中选择ROC曲线评价基于BiGAN的感知哈希图像内容取证算法的性能，采用真阳性率(TPR)计算相同内容的图像被正确识别的概率，采用假阳性率(FPR)检验不同内容图像被识别为相同内容图像的概率. 同时，实验中还进一步选择PR曲线验证基于BiGAN的感知哈希图像内容取证算法在样本不平衡条件下的性能表现，并采用样本识别准确率描述分类器对正样本和负样本都能准确识别的能力.

4.2　算法性能验证

基于BiGAN的感知哈希图像内容取证算法可以实现相同内容图像的认证，即使原始图像受到一定程度的攻击，仍能产生相似度非常高的感知哈希序列，实现相同内容图像的鲁棒性识别；另一方面，针对不同内容的图像，应能产生差别较大的感知哈希序列，实现不同内容图像的敏感性区分. 由于识别鲁棒性与区分敏感性之间是相互制约的，一个性能优良的感知哈希算法应能够实现二者之间的良好平衡，从而成为图像内容认证和版权保护的有效手段.

4.2.1　相同内容图像识别鲁棒性验证

首先从训练集中随机选取500幅图像，验证基于BiGAN的感知哈希图像内容取证网络对相同内容图像认证的鲁棒性能(如图5所示). 采用表4所

图5　随机选取的CelebAMask-HQ部分图像测试样本

示的内容保持攻击方法为每幅测试图像生成85幅具有相同内容,且视觉感知差异的副本图像,共生成42 500幅副本图像.然后,分别采用(*a*)基于基础BiGAN网络、(*b*)基于跳接层增强的BiGAN网络、(*c*)基于MSE损失和跳接层增强的BiGAN网络三种不同结构网络生成图像感知哈希码,计算原始图像与受攻击图像感知哈希码的互相关值并求取平均数进行对比,验证感知哈希算法的鲁棒性能.互相关值越靠近1,说明感知哈希码对图像隐含特征的表示能力越强;互相关值越接近0,说明感知哈希码对图像隐含特征的表示能力比较弱.

表4 不同类型的内容保持攻击方法

攻击方式	攻击方法	攻击强度
亮度	Photoshop scale	−20,−10,10,20
对比度	Photoshop scale	−20,−10,10,20
高斯低通滤波	标准差	0.3,0.4,…,0.9,1.0
伽马滤波	gamma	0.7,0.9,1.1,1.3
缩放	比率	0.5,0.7,0.9,1.1,1.3,1.5
椒盐噪声	密度	0.001,0.002,…,0.009,0.01
高斯噪声	方差	0.001,0.002,…,0.009,0.01
旋转+裁剪+缩放	旋转角度	1,2,3,4,5
JPEG压缩	质量因子	30,40,…,90,100
水印嵌入	量化步长	0.3,0.4,…,0.9,1.0
随机擦除	比率	0.005, 0.01, 0.015, 0.02
中心裁剪	比率	0.8, 0.85, 0.9, 0.95

实验结果表明,图像在受到不同类型攻击后,基于BiGAN网络所生成的感知哈希码互相关值仍然具有较强的相似度(如表5所示).由实验结果还可以看出,本算法在受到“亮度”、“伽马滤波”、“旋转+裁剪+缩放”和“中心裁剪攻击”等攻击时感知哈希码的互相关值较低.而在增加跳接层网络和MSE优化损失后,“亮度”、“伽马滤波”攻击图像的互相关值迅速增加,表现出较好的抗攻击能力,而“旋转、裁剪、缩放”和“中心裁剪攻击”等攻击所产生的互相关值仍然未出现明显改善.这是因为“旋转、裁剪、缩放”和“中心裁剪攻击”都在较大程度上影响了原始图像的内容信息,使得原始图像的语义出现缺失,而图像“亮度”和“伽马滤波”对图像的语义内容影响较小,而基于BiGAN的感知哈希网络在增加跳接层网络与MSE优化损失后,对原始图像的语义学习能力进一步增加,对于图像内容变化的语义表达更加准确,因而互相关值明显增强.同时,实验结果也表明,“椒盐噪声”、“高斯噪声”等类型的攻击在添加跳接层网络与MSE优化损失后,互相关值略有下降,这是因为随着感知哈希图像内容取证网络对图像细节的学习能力增强,增强了感知哈希码的语义表示能力,使得图像相关性回归到更准确的数值区间.

表5 相同内容图像的感知哈希互相关值

攻击	均值		
	BiGAN	Skip-Connection+BiGAN	Skip-Connection+MSE+BiGAN
亮度	0.6729	0.9119	0.8243
对比度	0.9735	0.9496	0.9316
高斯低通滤波	0.9892	0.9856	0.963
伽马滤波	0.5227	0.8592	0.7392
缩放	0.985	0.9838	0.9578
椒盐噪声	0.994	0.968	0.9635
高斯噪声	0.9839	0.914	0.8892
旋转+裁剪+缩放	0.6448	0.7807	0.7086
Jpeg压缩	0.7245	0.9254	0.8348
水印嵌入	0.9988	0.9981	0.9975
随机擦除	0.9025	0.9077	0.8893
中心裁剪	0.7458	0.6892	0.6555
不同攻击平均互相关	0.8448	0.9061	0.8629

另一方面,由表5还可以看出,采用基础BiGAN网络生成的感知哈希码平均互相关值为0.8448;而当添加跳接层网络结构后,图像感知哈希码平均互相关值进一步改善,达到0.9061,添加跳接层结构明显增强了生成感知哈希码对相同内容图像识别的鲁棒性.尽管在使用MSE损失增强图像细节描述能力后,受攻击图像生成感知哈希码的互相关值略有下降,但仍然保持在0.8629.

实验中取得相同图像互相关值分布如图6所示(条件(*c*)情况下).其中,纵轴表示互相关值的数量,横轴为互相关值.结果表明,本算法所生成的感知哈希序列具有很强的互相关性.实验中发现在所有42 500张受攻击图像中,仅有1457张图片的相关系数低于0.5.其中,297张是由“旋转+裁剪+缩放”组合攻击产生,896张是由中心裁剪攻击产生,264张是由伽马滤波攻击产生.而这几种攻击类型都在一定程度上改变了原始图像的结构与内容信息,因而降低了感知哈希码的特征表示能力.由以上分析可知,基于BiGAN的感知哈希网络具有很好的图像隐空间特征学习能力,可生成具有较强图像语义特征表示能力的感知哈希码,从而有效提升感知哈希码对相同内容图像认证的鲁棒性.

4.2.2　不同内容图像区分敏感性验证

为了检验基于BiGAN的感知哈希图像内容取证模型对不同内容图像的区分能力，实验中重新从测试集中随机选取500幅图像，并采用表4中所示12种攻击手段分别对图像进行攻击和篡改，分别采用三种不同结构的BiGAN网络生成图像感知哈希码．如前述分析，考虑到互相关值相较于汉明距离更能表征感知哈希码相似性，此处仍然采用互相关值评价不同内容图像的感知哈希码．从而，最终能够通过不同内容图像与相同内容图像感知哈希码互相关值变化曲线，寻求最优的互相关判别阈值作为图像内容取证的评价标准．实验中计算不同图像感知哈希码的互相关值，每次产生 $C_{500}^{2}=500\times(500-1)/2=124\,750$ 个互相关值，检验不同内容图像的感知哈希码分布特征．

实验结果如表6所示，在每种类型的内容保持攻击情况下，不同内容图像的感知哈希码互相关值的均值都不大于0.2630，相比于相同内容的图像，互相关值的取值范围明显下降．由表6还可以看出，基础BiGAN网络所产生的感知哈希码互相关值的总体均值为0.2506，基于跳接层增强的BiGAN网络所产生的感知哈希互相关总体均值下降到0.1329，而在增加MSE网络损失后，互相关总体均值进一步下降为0.0836．这是因为添加MSE损失增强了基于BiGAN的感知哈希生成网络对图像细节特征的学习能力，有利于增强感知哈希码对不同内容图像的区分性能．

表6　不同内容图像的感知哈希互相关值

操作	均值		
	BiGAN	Skip-Connection+BiGAN	Skip-Connection+MSE+BiGAN
亮度	0.2545	0.1198	0.0828
对比度	0.2551	0.1213	0.0858
高斯低通滤波	0.2552	0.1196	0.0858
伽马滤波	0.2361	0.1122	0.0728
缩放	0.2210	0.1224	0.0832
椒盐噪声	0.2551	0.1214	0.0840
高斯噪声	0.2553	0.1230	0.0836
旋转+裁剪+缩放	0.2579	0.1232	0.0863
Jpeg压缩	0.2630	0.1242	0.0881
水印嵌入	0.2550	0.1205	0.0825
随机擦除	0.2549	0.2549	0.0839
中心裁剪	0.2450	0.1333	0.0844
不同攻击平均互相关	0.2506	0.1329	0.0836

不同内容图像间的互相关值分布如图6所示（条件（c）情况下）．实验结果表明本算法所产生的不同内容图像感知哈希码互相关值均分布在靠近0的位置，也即不同内容图像感知哈希码的互相关性很低．因此，基于BiGAN的感知哈希图像内容取证算法可以有效增强感知哈希码对不同内容图像的区分敏感性．

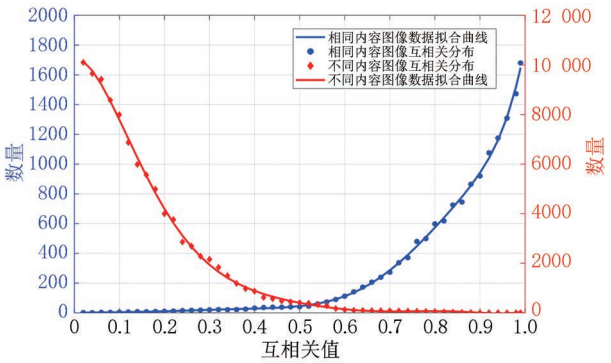

图6　相同/不同内容图像互相关值分布

4.3　跳接层对感知哈希生成网络性能影响

在双向生成对抗网络中，编码网络E通过多重卷积编码提取原始图像的隐空间特征信息，生成网络G通过反卷积运算生成内容丰富的图像分布，并通过联合判别网络D为生成网络G和编码网络E提供优化损失，对网络参数进行迭代优化．直接采用双向生成网络生成感知哈希码存在收敛速度慢、训练周期长、甚至出现网络振荡难以收敛的情况，并且感知哈希码对原始图像特征信息表示能力弱、生成图像质量不高．因而，实验中在生成网络G和编码网络E之间添加跳接层网络S，提取原始图像不同维度的特征信息传递到生成网络G，以提升网络学习能力，产生与原始图像更近似的生成图像；同时，反向激励编码网络E产生更具图像隐空间特征表示能力的感知哈希序列．

跳接层结构最早应用于U-net图像语义分割网络[49]，其在网络的降采样和升采样过程中通过跳接层将编码阶段获得的特征信息同解码阶段获得的特征信息拼接在一起，更好地融合图像不同层次的特征信息，提升图像语义分割能力．实验中采用跳接的方式将编码网络E的第五、六、七层网络输出与生成网络G的第四、三、二层网络输入信息进行链接（对应编码层的向量维数大小一致），将编码网络E提取的图像特征与生成网络G中特征信息进行融合，通过卷积运算传递到下一层网络中，从而增强生成网络G的特征学习能力和收敛性能．实验中分别

选择编码网络E第五、六、七卷积层的输出将样本图像低维、中维和高维信息传递到生成网络G中，从而激励生成网络输出与样本图像在全局和细节部分都具有较高近似度的数据分布，并加速哈希生成网络的收敛进程．同时，实验中还从CelebA Mask-HQ数据库中随机选取2000幅图像，采用表7所示的12种攻击方案生成$2000\times12\times2=48\,000$幅相似图像(为方便不同感知哈希算法间的性能比较，实验中选择最具代表性的感知哈希攻击类型进行性能验证)，每次计算$C_{2000}^{2}=2000\times(2000-1)/2=1\,999\,000$个互相关值，验证基于跳接层增强的BiGAN感知哈希图像内容取证算法性能．

表7 对比测试采用的不同攻击类型与攻击方法

攻击类型	攻击方法	攻击强度
亮度	Photoshop scale	−20, ,20
对比度	Photoshop scale	−20, ,20
高斯低通滤波	标准差	0.6,1.0
伽马滤波	gamma	0.7 ,1.3
缩放	比率	0.5,1.5
椒盐噪声	密度	0.002,0.006
高斯噪声	方差	0.002,0.006
旋转、裁剪、缩放	旋转角度	3,5
JPEG压缩	质量因子	40,80
水印嵌入	量化步长	0.6,1.0
随机擦除	比率	0.01, 0.02
中心裁剪	比率	0.85, 0.95

为检验跳接层网络结构对本算法性能的增强效果，实验中采用ROC曲线对比评价双向生成对抗网络在添加跳接层前后所生成感知哈希码的图像内容取证性能．由图7可知，基于跳接层增强的双向生成对抗网络所产生的感知哈希码对图像内容取证能力明显提升，其ROC曲线更靠近图像的左上角，而且具有更大的线下面积．当FPR为0.05时，基于BiGAN网络的感知哈希识别能力为87.23%，而添加三重跳接层网络后BiGAN感知哈希网络的识别能力上升为97.52%以上．这是因为基于BiGAN的感知哈希网络在添加跳接网络后，生成网络G在反卷积过程中接收到来自样本图像不同维度的特征信息，增强了网络生成图像的语义表达能力，提升了图像生成质量．同时，通过联合判别网络的梯度反向传播，激励编码网络E生成更具代表性的图像隐空间特征编码，促使BiGAN输出具有更强语义表示能力的感知哈希码．实验结果表明：采用跳接层网络结构可以有效增强BiGAN的感知哈希码生成能力，所生成感知哈希码的图像内容取证能力明显提升．

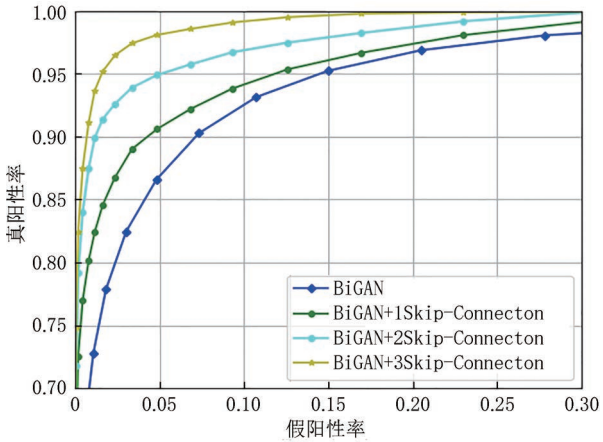

图7 跳接层网络对感知哈希码性能影响

4.4 均方误差损失对感知哈希生成网络性能影响

均方误差(MSE)通过计算生成图像与原始图像像素间的差异，评价生成图像与原始图像分布的一致性程度．增加MSE作为网络优化损失，可以激励生成网络G生成与原始网络更趋于内容分布一致的图像，更好地恢复原始图像的细节信息，进一步提升生成图像质量，增强生成图像与原始自然图像的数据分布相似度，同时反向激励和提升编码网络E输出感知哈希码对图像内容表示的精确度．然而，增强感知哈希码对图像内容表示的精确度在提高其对不同内容图像的区分敏感性的同时，也会降低其对相同内容图像的识别鲁棒性．因而，实验中需选择合适的MSE损失强度优化提升感知哈希码的生成质量，平衡感知哈希码对相同内容图像的识别鲁棒性和对不同内容图像的区分敏感性．

实验中分别采用不同强度系数(0.1、0.3、0.5、0.8、1.0)的MSE损失验证本算法性能(采用与4.3节一致的网络参数)．实验结果如图8所示，当MSE权值系数为0.5时，ROC曲线最靠近坐标左上角的位置，具有最大的线下面积，也即感知哈希生成网络取得最好的图像内容识别性能．由实验结果分析可知：当MSE强度小于0.5时，MSE损失对生成图像细节的优化能力不足，生成图像质量不高，对增强生成感知哈希码的语义特征表示能力作用不明显．而当MSE强度大于0.5时，网络输出感知哈希码的识别能力开始下降，这是因为过大的MSE损失权值降低了网络对图像深层特征的学习能力，所形成的感知哈希码的语义表示能力降低；另一方面，较大的MSE强度增强了BiGAN学习原始图像细节信息的能力，生成的图像与原始图像的细节相似度迅速提

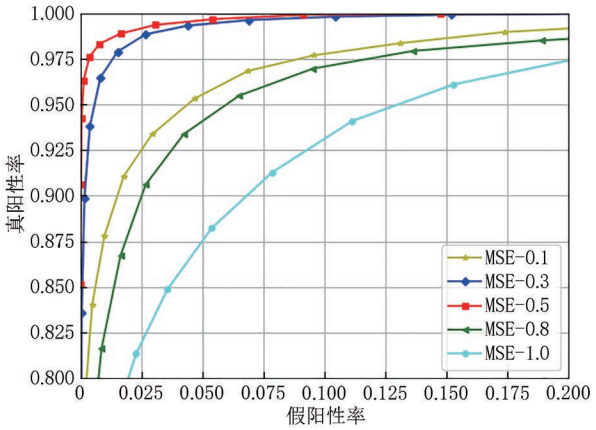

图8　MSE强度对感知哈希码性能影响

高，网络优化梯度快速达到饱和，而此时生成的感知哈希码对图像深层特征的表示能力不足，而对图像的细节变化过于敏感，导致其对相同内容图像识别的鲁棒性下降．由以上分析和实验结果可知，当MSE损失的权值系数为0.5时，基于BiGAN感知哈希码生成网络取得最优性能，所输出的图像感知哈希码具有最佳的图像内容取证能力．

4.5　不同方法间性能的比较

经典的感知哈希方案多采取人工设计的算法提取图像的特定特征信息，并基于小样本数据集进行训练生成图像感知哈希，导致所生成的哈希值只能针对某一种或几种特定类型的攻击具有较强的取证能力．本方案首次提出基于大型数据库的无监督深度学习感知哈希图像内容取证算法，充分利用双向生成对抗网络(BiGAN)的深度特征提取能力，生成更具代表性的图像隐空间特征编码序列，提高输出感知哈希码的图像本质特征表示能力．为全面评价本算法所生成感知哈希码的性能，本试验方案分别选取了基于人工设计特征生成图像感知哈希的算法以及基于深度学习的感知哈希生成算法进行对比验证．由于基于深度学习算法生成图像感知哈希，并采用大型数据库实现感知哈希算法性能验证的研究刚刚起步，相关研究成果较少．为公平起见，实验中选取了目前最新的几种代表性感知哈希生成算法，并首次采用CelebA Mask-HQ大型训练集对不同感知哈希算法进行对比验证，充分探究所提出的基于BiGAN的无监督感知哈希图像内容取证算法性能．

实验中分别选取Tang等人提出的基于色彩相位角统计特征的方案[13]，Huang等人提出的基于纹理与不变特征距离的方案[23]、Tang等人构造的基于DCT变换与矩阵压缩的算法[32]、Li等人提出的基于分层神经网络的方法[40]、Qin等人提出的基于卷积神经网络的方案[41]，以及Sun等人提出的基于哈希中心的深度学习感知哈希方案[42]与本文所提出的方法进行对比验证，评价基于BiGAN的感知哈希图像内容取证算法性能．实验中分别采用ROC曲线、PR曲线和识别准确率曲线对比分析本文所提出算法与其他几种算法的图像内容取证能力．

首先从CelebAMask-HQ测试集中随机选取5000幅图像，并采用表7中所示12种不同的攻击算法对原始图像进行内容保持攻击，选取不同感知哈希生成算法提取图像感知哈希码，并计算感知哈希码互相关性，以对比检验不同感知哈希生成算法实现图像内容取证的性能．

实验中首先采用真正率(*TPR*)和假正率(*FPR*)的联合分布绘制ROC曲线，对比不同算法间的分类性能，实验结果如图9所示，从图中可以看出本算法所得到的ROC曲线取得更靠近坐标图左上角的分布，具有更大的线下面积(AUC)，也即本算法具有最佳的图像内容取证能力．实验结果表明：当阈值为0.52时，本算法取得最佳分类效果．此时，基于本算法生成的感知哈希码取得图像识别正确率 $TPR=98.8\%$，识别错误率 $FPR=1.65\%$，具有最大的线下面积．方法[42]由于采用了基于哈希中心的多重卷积感知哈希生成网络，取得较好的图像内容取证能力，其最优图像识别正确率 $TPR=97.4\%$，识别错误率 $FPR=2.85\%$；方法[41]和方法[40]都采用了基于深度学习的感知哈希生成网络，并取得近似的图像分类效果，具有相似的线下面积；其中，方法[41]最优图像识别正确率 $TPR=96.4\%$，识别错误率 $FPR=3.15\%$；方法[40]最优图像识别正确率 $TPR=95.3\%$，识别错误率 $FPR=3.23\%$．方法[23]采用了基于图像纹理与不变特征向量距离的感知哈希生成算法，兼顾了图像的全局和局部特征信息，取得了较好的感知哈希图像内容取证性能．最优图像识别正确率达到 $TPR=94.9\%$，识别错误率 $FPR=3.35\%$．方法[32]采用DCT变化与频域矩阵压缩算法，图像内容取证性能出现一定程度下降，其最优图像识别正确率 $TPR=91.6\%$，识别错误率 $FPR=3.39\%$．而方法[13]的分类能力明显降低．图像识别正确率和识别错误率分别为：$TPR=79.9\%$，$FPR=5.16\%$．由以上对比分析可知，基于BiGAN的感知哈希图像内容取证算法可以有效结合目标图像的深层与浅层特征信息，生成更具图像语义特征表示能力的感知

哈希码，从而取得更好的图像内容认证与版权保护能力.

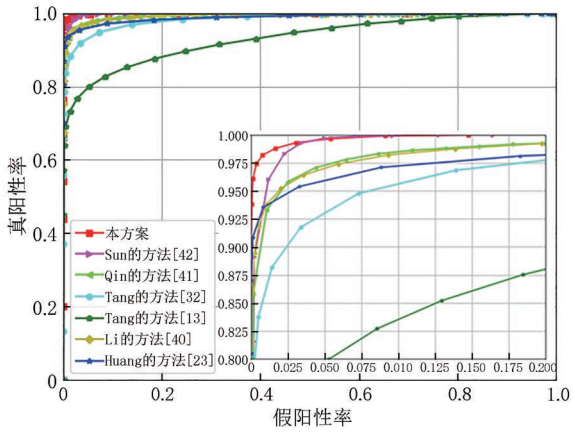

图9 不同感知哈希算法ROC变化曲线

此外，考虑到生成的数据样本在进行混淆处理后，同一类型的样本占比相对较少(经过12种攻击类型扩展后，每种类型的样本数据约占1/12)，实验中进一步采用PR关系图，验证不平衡样本条件下本算法的性能，计算任意两幅图像间的互相关值，并通过查准率、召回率构建PR曲线，比较不同感知哈希算法基于样本内容的认证能力. 实验结果如图10所示，当阈值$\eta=0.52$时，分类器的查全率$Recall=98.3\%$，查准率为$Precision=97.8\%$，明显高于其他感知哈希分类方案. 实验结果再次表明了基于BiGAN的感知哈希生成算法具有更强的内容取证能力，可以用于图像内容认证与版权保护.

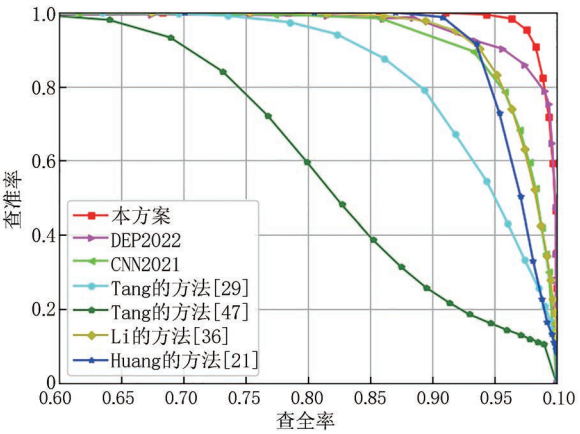

图10 不同感知哈希算法PR变化曲线

进一步地，为了检验本算法实现图像内容识别的性能，实验中采用样本识别准确率评价其对于相同内容图像和不同内容图像的正确分类能力，并基于分类准确率曲线对比不同感知哈希算法的最优分类阈值. 实验结果如图11所示.

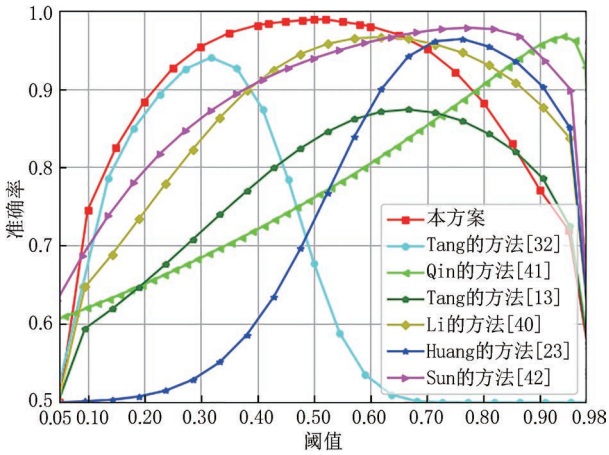

图11 不同感知哈希算法的Accuracy变化曲线

由图11可知，本文所提出的感知哈希图像内容取证算法在所有感知哈希算法中取得了最好的图像识别能力. 当阈值为$\eta=0.52$时，本文所提出的感知哈希图像内容取证算法可以对99.2%的图像实现正确分类，算法取得最好的图像分类性能. 方法[42]，[41]和[40]分别在阈值为0.79，0.93和0.62时取得最优的图像内容取证能力，可以实现98.3%，97.6%和96.8%的图像正确分类比率. 而方法[23]在0.76时，实现96.3%的图像正确分类，方法[32]和[13]则分别在阈值为0.32和0.67时取得各自最优分类效果93.9%和87.8%. 实验结果表明，基于深度学习的感知哈希生成算法相比于人工设计特征的感知哈希算法，取得了更好的图像内容取证能力. 这是因为基于深度学习网络的感知哈希算法基于大量训练样本的特征学习和提取训练，能够实现原始图像本质特征信息的准确描述能力. 而基于人工设计特征的感知哈希算法只能基于图像特定类别的特征生成图像感知哈希，因而在一定程度上限制了图像感知哈希码的特征描述能力. 实验结果还表明，方法[32]在阈值为0.32时就已达到最强的分类性能，且随着阈值升高，分类能力快速下降，当阈值大于0.7时，所有图像都被识别为不同内容的图像，显示该算法对相同内容图像的识别鲁棒性能有待提升. 这是因为方法[32]采用离散余弦变换(DCT)和局部线性嵌入(LLE)技术生成图像感知哈希码. 其利用DCT变换提取图像特征矩阵，并将LLE数据约简应用于特征矩阵，采用LLE结果的方差进行量化构造图像感知哈希码. 这种基于原始图像特征约简所生成的感知哈希码对原始图像的语义

特征表示能力不足，鲁棒性能不强．而且，该算法仅采用图像DCT频域分量进行局部线性嵌入，也一定程度上损失了原始图像的细节特征信息，降低了感知哈希码对原始图像的细节描述能力．因而，基于该算法所计算出的图像感知哈希互相关系数普遍较小，在阈值较小时就达到了图像内容取证的峰值，所生成感知哈希码的图像内容取证性能也存在提升空间．由以上分析可知，基于BiGAN的感知哈希图像内容取证算法与当前其他最优的感知哈希算法相比，可以实现最佳的图像内容分类性能，算法对相同内容的图像具有识别鲁棒性，而对不同内容的图像具有区分敏感性．即使在图像受到各种不同类型的内容保持攻击后，仍能取得优秀的图像内容取证能力，实现基于内容的图像认证与版权保护．

4.6　密钥依赖性

为保障图像感知哈希码的唯一性，避免感知哈希码的伪造和碰撞攻击，实验中对生成感知哈希码的密钥依赖性进行了验证．

密钥依赖性是指使用不同密钥生成的感知哈希码应存在极大不相关，也即相关系数应非常小．考虑到不同内容图像的感知哈希码相关性极低，采用相同密钥加密不同内容图像的感知哈希值，会进一步降低感知哈希之间互相关特性．因而，此处仅采用不同密钥对相同内容图像的感知哈希码进行加密，以验证采用不同的密钥可产生完全不同的感知哈希码．从而保障即使图像感知哈希生成算法公开，不掌握密钥的攻击者也不能伪造出与目标感知哈希一致的散列码．实验中采用随机置乱的方式对生成的二进制哈希码进行了加密处理，从CelebA Mask-HQ数据库的测试集中随机抽选出1000张图作为测试图像，并采用不同密钥生成图像感知哈希码，计算不同感知哈希码之间的互相关值．首先使用正确的密钥生成图像感知哈希码，然后保持其他所有参数不变，选择100个不同的密钥生成图像感知哈希码，计算初始密钥与其他不同密钥所产生的图像感知哈希码之间的互相关值并求取平均数如图12所示．其中，X轴是密钥索引，Y轴是不同密钥产生哈希值之间的互相关值．试验结果表明：基于不同密钥产生的感知哈希码最大相关系数小于0.07．也即任何一个不掌握生成密钥的攻击者在理论上都不可能产生与目标值相同的感知哈希码，基于密钥顺序置乱的感知哈希码具有很好的密钥依赖性，可以满足图像内容认证与版权保护的要求．

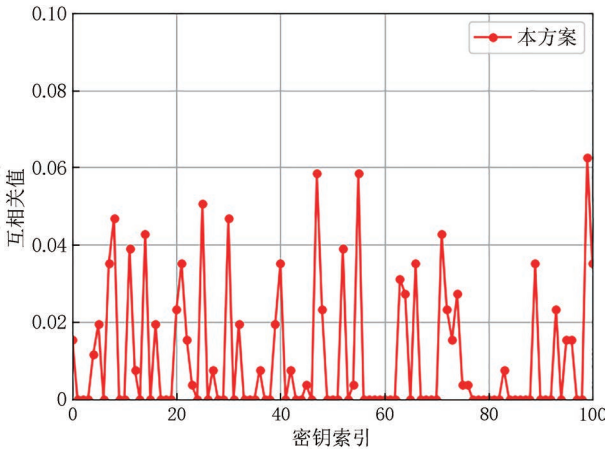

图12　不同加密密钥所产生的哈希码互相关值分布

4.7　不同算法运行效率比较

本研究还对不同算法生成哈希码的生成效率和码长进行了对比分析．具体做法是：随机选取1000张图像，分别采用不同算法生成图像感知哈希码，并计算生成每个感知哈希码所需平均时间．实验结果如表8所示，本算法只采用基础BiGAN网络时需要0.72s，添加跳接层网络后，生成感知哈希码的时间为0.83s，当采用基础网络＋跳接层网络结构，并增加MSE损失后，生成感知哈希码所需要的平均时间为0.88s，其生成感知哈希码所需要的时间明显低于其他大部分方法．虽然文献[36]算法生成图像感知哈希仅需0.04s，但是该算法仅适用于尺寸较小的图像，且面对高强度攻击时鲁棒性能较差．

同时，表8还显示本研究所采用的感知哈希码相比于其他感知哈希序列具有略长的位数，这是因为本研究是首次提出基于大规模图像数据库的无监督感知哈希生成算法，为避免针对所生成感知哈希码的伪造和碰撞攻击，提高感知哈希码对图像内在特征的表示能力，实验中选取1024 bits长度的感知哈希码，进一步增强基于感知哈希的图像内容取证

表8　感知哈希生成时间与感知哈希长度对比表

算法	文献[32]	文献[13]	文献[40]	文献[23]	文献[41]	文献[42]	基础GAN	基础GAN＋跳接层	基础GAN＋跳接层＋MSE
时间(s)	7.22	31.55	0.04	2.56	0.98	0.96	0.72	0.83	0.88
长度(bits)	64	400	50	720	400	64	1024	1024	1024

能力. 试验结果表明,基于BiGAN的感知哈希图像内容取证算法可以实现图像内容取证性能与算法运行效率间的优化平衡.

5 结 语

本文提出了一种基于双向生成对抗网络的感知哈希图像内容取证算法实现图像内容认证与版权保护. 本算法充分利用双向生成对抗网络对图像语义特征的学习能力,通过判别网络和编码网络、生成网络间的双向迭代对抗,生成具有较强图像语义特征表示能力的感知哈希码. 同时,算法通过在双向生成对抗网络中添加跳接层网络结构传递样本图像不同维度的特征信息,增强生成网络的特征学习能力并生成高质量图像,从而提升图像感知哈希码语义表示能力与网络收敛速度. 更进一步的,通过在网络中增加MSE误差损失,激励生成网络产生具有更强细节描述能力的图像,增强编码网络生成感知哈希码对图像内容的表示精确度. 本研究中首次采用大型图像数据库CelebA Mask-HQ检验感知哈希生成算法的性能,充分利用双向生成对抗网络的无监督学习能力,自适应的学习图像隐空间特征信息,提高图像内容认证精度. 实验中分别探讨了跳接层网络、MSE误差损失以及加密密钥等参量对感知哈希生成网络性能的影响,并将实验结果与当前最优秀的感知哈希算法进行比较. 实验结果表明,基于双向生成对抗网络的感知哈希图像内容取证算法具有优异的图像内容认证与版权保护能力. 当分类阈值为0.52时,本算法可以取得99.2%的图像内容识别正确率,性能强于当前其他优秀的感知哈希生成算法. 基于BiGAN的感知哈希生成算法可以产生具有更强图像语义特征和细节表示能力的感知哈希码,实现相同内容图像识别鲁棒性与不同内容图像区分性敏感性间的良好平衡,从而取得更好的图像内容认证与版权保护能力.

下一步研究将继续针对不同类型的数据库,探索海量图像数据环境中基于无监督学习的感知哈希图像内容取证能力提升和优化方法.

参 考 文 献

[1] Ma B, Shi Y Q. A reversible data hiding scheme based on code division multiplexing. IEEE Transactions on Information Forensics and Security, 2016, 11(9): 1914-1927.

[2] Ma B, Chang L, Wang C, et al. Robust image watermarking using invariant accurate polar harmonic Fourier moments and chaotic mapping. Signal Processing, 2020, 172: 107544.

[3] Srivastava M, Siddiqui J, Ali M A. Local binary pattern based technique for content based image copy detection//Proceedings of the International Conference on Power Electronics & IoT Applications in Renewable Energy and its Control (PARC). Mathura, India, 2020: 374-377.

[4] Qin C, Liu E, Feng G, et al. Perceptual image hashing for content authentication based on convolutional neural network with multiple constraints. IEEE Transactions on Circuits and Systems for Video Technology, 2020, 31(11): 4523-4537.

[5] Donahue J, Krähenbühl P, Darrell T. Adversarial feature learning. arXiv preprint arXiv:1605.09782, 2016.

[6] Schneider M, Chang S F. A robust content based digital signature for image authentication//Proceedings of the 3rd IEEE International Conference on Image Processing. Lausanne, Switzerland, 1996: 227-230.

[7] Tang Z, Dai Y, Zhang X, et al. Perceptual image hashing with histogram of color vector angles//Proceedings of the International Conference on Active Media Technology. Macau, China, 2012: 237-246.

[8] Zhao Y, Wang S, Zhang X, et al. Robust hashing for image authentication using Zernike moments and local features. IEEE Transactions on Information Forensics and Security, 2012, 8(1): 55-63.

[9] Chen Y, Yu W, Feng J. Robust image hashing using invariants of Tchebichef moments. International Journal for Light and Electron Optics, 2014, 125(19): 5582-5587.

[10] Tang Z, Zhang X, Li X, et al. Robust image hashing with ring partition and invariant vector distance. IEEE Transactions on Information Forensics and Security, 2015, 11(1): 200-214.

[11] Hosny K M, Khedr Y M, Khedr W I, et al. Robust image hashing using exact Gaussian-Hermite moments. IET Image Processing, 2018, 12(12): 2178-2185.

[12] Zhao Y, Yuan X. Perceptual image hashing based on color structure and intensity gradient. IEEE Access, 2020, 8(1): 26041-26053.

[13] Tang Z, Huang L, Zhang X, et al. Robust image hashing based on color vector angle and Canny operator. AEU-International Journal of Electronics and Communications, 2016, 70(6): 833-841.

[14] Abbas, S Q, Shirazi, S J, Chen Y P P. Aggregated bidirectional local binary pattern for robust perceptual image hashing//Proceedings of the IEEE International Symposium on Multimedia (ISM), Naples, Italy. 2022: 50-57

[15] Wang X, Xue J, Zheng Z, et al. Image forensic signature for content authenticity analysis. Journal of Visual Communication and Image Representation, 2012, 23(5): 782-797.

[16] Tang Z, Huang L, Zhang X, et al. Robust image hashing based on color vector angle and Canny operator. AEU-International Journal of Electronics and Communications, 2016, 70(6): 833-841.

[17] Ouyang J, Liu Y, Shu H. Robust hashing for image

authentication using SIFT feature and quaternion Zernike moments. Multimedia Tools and Applications，2017，76(2)：2609-2626.

[18] Vadlamudi L N，Vaddella R P V，Devara V. Robust image hashing using SIFT feature points and DWT approximation coefficients. Information & Communications Technology，2018，4(3)：154-159.

[19] Yuan，X，Zhao Y. Perceptual image hashing based on three-dimensional global features and image energy. IEEE Access，2021.9(3)：49325-49337.

[20] Lin C Y，Chang S F. A robust image authentication method distinguishing JPEG compression from malicious manipulation. IEEE Transactions on Circuits and Systems for Video Technology，2001，11(2)：153-168.

[21] Tang Z，Yang F，Huang L，et al. Robust image hashing with dominant DCT coefficients. International Journal for Light and Electron Optics，2014，125(18)：5102-5107.

[22] Qin C，Chen X，Dong J，et al. Perceptual image hashing with selective sampling for salient structure features. Displays，2016，45(12)：26-37.

[23] Huang Z，Liu S. Perceptual image hashing with texture and invariant vector distance for copy detection. IEEE Transactions on Multimedia，2020，23(6)：1516-1529.

[24] Venkatesan R，Koon S M，Jakubowski M H，et al. Robust image hashing//Proceedings of the International Conference on Image Processing. Vancouver，Canada，2000：664-666.

[25] Tang Z，Ling M，Yao H，et al. Robust image hashing via random Gabor filtering and DWT. Computers，Materials & Continua，2018，55(2)：331-344.

[26] Swaminathan A，Mao Y，Wu M. Robust and secure image hashing. IEEE Transactions on Information Forensics and Security，2006，1(2)：215-230.

[27] Qin C，Chang C C，Tsou P L. Robust image hashing using non-uniform sampling in discrete Fourier domain. Digital Signal Processing，2013，23(2)：578-585.

[28] Yu，M，Tang，Z，Zhang，X，et al. Perceptual hashing with complementary color wavelet transform and compressed sensing for reduced-reference image quality assessment. IEEE Transactions on Circuits and Systems for Video Technology，2022，32(11)：7559-7574.

[29] Kozat S S，Venkatesan R，Mihçak M K. Robust perceptual image hashing via matrix invariants//Proceedings of the International Conference on Image Processing，Singapore，2004，5：3443-3446.

[30] Tang Z，Zhang X，Zhang S. Robust perceptual image hashing based on ring partition and NMF. IEEE Transactions on Knowledge and Data Engineering，2013，26(3)：711-724. .

[31] Tang Z，Ruan L，Qin C，et al. Robust image hashing with embedding vector variance of LLE. Digital Signal Processing，2015，43(8)：17-27.

[32] Tang Z，Lao H，Zhang X，et al. Robust image hashing via DCT and LLE. Computers & Security，2016，62(9)：133-148.

[33] Zhu X，Li X，Zhang S，et al. Graph PCA hashing for similarity search. IEEE Transactions on Multimedia，2017，19(9)：2033-2044.

[34] Tang Z，Huang Z，Zhang X，et al. Robust image hashing with multidimensional scaling. Signal processing，2017，137(8)：240-250.

[35] Huang Z，Tang Z，Zhang X，et al. Perceptual image hashing with locality preserving projection for copy detection. IEEE Transactions on Dependable and Secure Computing，2023. 20(1)：463-477.

[36] Erin L V，Lu J，Wang G，et al. Deep hashing for compact binary codes learning//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston，USA，2015：2475-2483.

[37] Li W J，Wang S，Kang W C. Feature learning based deep supervised hashing with pairwise labels. arXiv：1511.03855，2015.

[38] Deng C，Chen Z，Liu X，et al. Triplet-based deep hashing network for cross-modal retrieval. IEEE Transactions on Image Processing，2018，27(8)：3893-3903.

[39] Liu X，Liang J，Wang Z Y，et al. Content-based image copy detection using convolutional neural network. Electronics，2020，9(12)：2029.

[40] Li Y，Wang D，Tang L. Robust and secure image fingerprinting learned by neural network. IEEE Transactions on Circuits and Systems for Video Technology，2019，30(2)：362-375.

[41] Qin C，Liu E，Feng G，et al. Perceptual image hashing for content authentication based on convolutional neural network with multiple constraints. IEEE Transactions on Circuits and Systems for Video Technology，2020，31(11)：4523-4537.

[42] Sun X，Zhou J. Deep Perceptual Hash Based on Hash Center for Image Copyright Protection. IEEE Access，2022，10(11)：120551-120562.

[43] Song J，He T，Gao L，et al. Binary generative adversarial networks for image retrieval//Proceedings of the AAAI Conference on Artificial Intelligence. New Orleans，USA，2018：394-401.

[44] Lin K，Lu J，Chen C S，et al. Learning compact binary descriptors with unsupervised deep neural networks//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas，USA，2016：1183-1192.

[45] Goodfellow I，Pouget A J，Mirza M，et al. Generative adversarial nets//Proceedings of the 27th International Conference on Neural Information Processing Systems. Montreal，Canada，2014：2672-2680.

[46] Liu D，Long C，Zhang H，et al. Arshadowgan：Shadow generative adversarial network for augmented reality in single light scenes//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle，USA，2020：8139-8148.

[47] Islam A，Long C，Basharat A，et al. DOA-GAN：Dual-order attentive generative adversarial network for image copy-move forgery detection and localization//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle，USA，2020：4676-4685.

[48] Davis J，Goadrich M. The relationship between Precision-Recall and ROC curves//Proceedings of the 23rd International

Conference on Machine Learning. Pennsylvania, USA, 2006: 233-240.

[49] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation//Proceedings of International Conference on Medical Image Computing and Computer-Assisted Intervention. Munich, Germany, 2015: 234-241.

MA Bin, Ph. D., professor. His research interests include reversible data hiding, image forensics and multimedia security.

WANG Yi-Li, M. S. candidate. His research interests focus on generative adversarial networks and perceptual image hashing.

XU Jian, Ph. D., associate professor. Her research interests focus on information hiding and image processing.

WANG Chun-Peng, Ph. D., associate Professor. His research interests focus on image processing and multimedia information security.

LI Jian, Ph. D., associate professor. His research interests focus on digital image and video forensics.

ZHOU Lin-Na, Ph. D., professor. Her research interests focus on information hiding and multimedia content forensics.

SHI Yun-Qing, Ph. D., professor. His research interests focus on multimedia forensics and multimedia security.

Background

With rapid developments of intelligent terminal and digital image processing technology, the cost of obtaining a high-precision image drops continually, and images stored on the internet are growing exponentially. At the same time, image modification is getting easier and easier with the emergence of image editing software. Seeing is not necessarily true. In recent years, digital image content forensics has increasingly become a research hotspot. Although the digital watermark algorithm has been proposed for image copyright authentication, the watermark embedding inevitably destroys the internal structure and degrades the image's visual quality. As a new multimedia security technology, the image perceptual hash algorithm is a method that can produce a fixed-length sequence based on image visual content features. Due to its high content representation capability, the image perceptual hash algorithm is becoming an ideal choice for achieving image content authentication and copyright protection. As identification robustness and discrimination sensitivity are mutually constrained, an excellent image perceptual hash algorithm should maintain a good balance between the two ends.

On the one hand, traditional image perceptual hash algorithms produce image perceptual hash codes depending on the pre-designed feature extractors and quantizers, requiring much expert knowledge and experience. As it is difficult to capture the intrinsic features of an image, the representation capability of the perceptual hash code is limited to a great extent. Therefore, a perceptual hash algorithm that can make full use of image latent semantic property is highly desired to enable it to achieve superior image content forensics.

On the other hand, with the tremendous growth of computing power and deep learning technology, more and more researchers have begun to learn the deep features of an image by employing deep learning networks so as to improve the representativity of the image perceptual hash code. And thus, the identification robustness and the discrimination sensitivity are both enhanced. According to the powerful learning ability of the Generative Adversarial Network, this paper proposed an image perceptual hash algorithm for image content forensics based on the Bi-directional Generative Adversarial Network (BiGAN). By making full use of BiGAN's image latent semantic property extraction ability, the intrinsic features of an image are captured comprehensively via the mutual competition among the generative network, the discriminative network, and the decoding network. As a result, a high-performance perceptual hash code robust to image identification and sensitive to image discrimination is obtained.

In addition, to evaluate the performance of the proposed image perceptual hash algorithm, a large image database (CelebA Mask-HQ) is chosen for the experiment. Extensive experimental results show that the BiGAN-based image perceptual hash algorithm can achieve higher identification robustness for images with identical content and discrimination sensitivity for images with different contents compared with most advanced perceptual hash algorithms.