计算机科学
COMPUTER SCIENCE

基于空频联合卷积神经网络的GAN生成人脸检测

王金伟, 曾可慧, 张家伟, 罗向阳, 马宾

引用本文

王金伟, 曾可慧, 张家伟, 罗向阳, 马宾基于空频联合卷积神经网络的GAN生成人脸检测[J]. 计算机科学, 2023, 50(6): 216-224.

WANG Jinwei, ZENG Kehui, ZHANG Jiawei, LUO Xiangyang, MA Bin. GAN-generated Face Detection Based on Space-Frequency Convolutional Neural Network [J]. Computer Science, 2023, 50(6): 216-224.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.220400268

基于空频联合卷积神经网络的GAN生成人脸检测

王金伟[1,2,3]　曾可慧[1]　张家伟[1]　罗向阳[3]　马　宾[4]
1 南京信息工程大学计算机学院、软件学院、网络空间安全学院　南京 210044
2 南京信息工程大学江苏省大气环境与装备技术协同创新中心　南京 210044
3 数学工程与高级计算国家重点实验室　郑州 450001
4 齐鲁工业大学山东省计算机网络重点实验室　济南 250353
(wjwei_2004@163.com)

摘　要　生成式对抗网络(GAN)的快速发展使其在图像生成领域取得了前所未有的成功。StyleGAN等新型GAN的出现使得生成的图像更真实且具有欺骗性,对国家安全、社会稳定和个人隐私都构成了较大威胁。文中提出了一种基于空频联合的双流卷积神经网络的检测模型。鉴于GAN图像在生成过程中因上采样操作在频谱上留下了清晰可辨的伪影,设计了可学习的频率域滤波核以及频率域网络来充分学习并提取频率域特征。为了减弱图像变换至频域过程中丢弃部分信息而带来的影响,同样设计了空间域网络来学习图像内容本身具有差异化的空间域特征,最终将两种特征融合来实现对GAN生成人脸图像的检测。在多个数据集上的实验结果表明,所提模型在高质量生成数据集上的检测精度及在跨数据集的泛化性上都优于现有算法,且对于JPEG压缩、随机剪裁、高斯模糊等图像变换具有更强的鲁棒性。不仅如此,所提方案在GAN生成的局部人脸数据集上也有不错表现,进一步证明了所提模型有着更好的通用性以及更加广泛的应用前景。

关键词:数字图像取证;人脸伪造检测;卷积神经网络;生成式对抗网络;频率域
中图法分类号　TP391.41;TP183

GAN-generated Face Detection Based on Space-Frequency Convolutional Neural Network

WANG Jinwei[1,2,3],ZENG Kehui[1],ZHANG Jiawei[1],LUO Xiangyang[3] and MA Bin[4]
1 College of Computer,College of Software,College of Cyberspace Security,Nanjing University of Information Science and Technology,Nanjing 210044,China
2 Jiangsu Collaborative Innovation Center of Atmospheric Environment and Equipment Technology,Nanjing University of Information Science and Technology,Nanjing 210044,China
3 State Key Laboratory of Mathematical Engineering and Advanced Computing,Zhenzhou 450001,China
4 Shandong Provincial Key Laboratory of Computer Networks,Qilu University of Technology,Jinan 250353,China

Abstract　The rapid development of generative adversarial networks(GANs) has led to unprecedented success in the field of image generation. The emergence of new GANs such as StyleGAN makes the generated images more realistic and deceptive,posing a greater threat to national security,social stability,and personal privacy. In this paper,a detection algorithm based on a space-frequency joint two-stream convolutional neural network is proposed. Since GAN images will leave clearly discernible artifacts on the spectrum due to the up-sampling operation during the generation process,a learnable frequency-domain filter kernel and frequency domain network are designed to fully learn and extract frequency-domain features. In order to reduce the influence of the information discarded from the image transformation to the frequency domain,a spatial domain network is also designed to

到稿日期:2022-04-26　返修日期:2022-10-09
基金项目:国家自然科学基金(62072250,62172435,U1804263,U20B2065,61872203,71802110,61802212);中原科技创新领军人才项目(214200510019);江苏省自然科学基金(BK20200750);河南省网络空间态势感知重点实验室开放基金(HNTS2022002);江苏省研究生研究与实践创新项目(KYCX200974));广东省信息安全技术重点实验室开放项目(2020B1212060078);人文社会科学教育部项目(19YJA630061);江苏高校优势学科建设工程项目
This work was supported by the National Natural Science Foundation of China(62072250,62172435,U1804263,U20B2065,61872203,71802110,61802212),Zhongyuan Science and Technology Innovation Leading Talent Project of China(214200510019),Natural Science Foundation of Jiangsu Province,China(BK20200750),Open Foundation of Henan Key Laboratory of Cyberspace Situation Awareness(HNTS2022002),Post Graduate Research & Practice Innvoation Program of Jiangsu Province(KYCX200974),Opening Project of Guangdong Province Key Laboratory of Information Security Technology(2020B1212060078),Ministry of Education of Humanities and Social Science Project(19YJA630061) and Priority Academic Program Development of Jiangsu Higher Education Institutions(PAPD).
通信作者:罗向阳(xiangyangluo@126.com)

learn that the image content itself has differentiated spatial domain features. Finally, the two features are fused to detect the face image generated by GAN. Experimental results on multiple datasets show that the proposed model outperforms existing algorithms in detection accuracy on high-quality generated datasets and generalization across datasets. And for JPEG compression, random cropping, Gaussian blur, and other operations, this method has stronger robustness. In addition, the proposed method also performs well on the local face dataset generated by GAN, which further proves that this model has better generality and wider application prospects.

Keywords Digital image forensics, Face forgery detection, Convolutional neural network, Generative adversarial networks, Frequency domain

1 引言

近年来，随着互联网技术的快速发展以及人工智能的不断成熟，信息传递已不再耗时耗力。其中，数字图像凭借其内容丰富和存储便利等优势，已成为信息传递的主要方式，并在金融、医疗、新闻媒体等领域得到了广泛应用。然而，随着图像编辑软件的不断发展，普通用户都可以借助这些工具创建逼真的合成图片，而无需具备摄影编辑方面的专业知识。例如 FaceApp 等“一键式”合成 APP 的出现，使得用户只需在移动设备上轻触一下，就可以变换面部表情、合成虚拟语音[1]，甚至是与明星换脸。但是，这项技术不仅带来了这些新奇有趣的体验，还埋藏着一些潜在的威胁。例如，一些不法分子利用该技术恶意篡改图像，并通过互联网广泛传播，这不仅侵犯了个人隐私，打破了社会稳定，更有甚者对国家的安全和利益都造成了非常恶劣的影响。

GAN 在多媒体领域取得的巨大成功同样也引起了安全领域专家学者的高度关注。其中，最具影响力的事件就是 Karras 等[2]于 2018 年在国际知名会议 ICLR 上提出的新型 GAN 结构——PGGAN。与已有模型不同的是，PGGAN 利用逐层训练的方式首次生成了分辨率高达 1024×1024 的清晰图像，不仅改善了 GAN 生成图像分辨率低的问题，而且生成的图像不存在明显异常痕迹，平滑逼真，质量惊人。图 1 给出了多种 GAN 生成的伪造人脸图像和多个真实人脸图像数据集采集的样本。如今，眼见为实已不再是真理，这给维护数字图像的真实性和完整性也带来了前所未有的威胁和挑战。近年来，有许多研究者剖析了 GAN 结构，并基于此分析了 GAN 生成图像过程中所留下的异常痕迹，鉴于这些痕迹也提出了多种性能良好的伪造人脸的检测方法[3-7]。

(a)GAN 生成图像

(b)真实图像

图 1 GAN 生成图像和真实图像

Fig. 1 Images generated by GAN and real images

然而，经研究发现，这些算法在简单场景下均具有较好的检测效果[5-8]，但是当面对未知 GAN 模型生成的图像或是一些图像编辑操作，如 JPEG 压缩、下采样、高斯噪声等，算法的检测性能便会有明显下降[9-10]。除此之外，现有的算法往往将空间域特征和频率域特征视为两个独立的特征域，并且单独地使用空间域图像或频率域图像作为检测依据，未能将空频结合起来共同探究 GAN 生成图像的分布特点。

为了应对上述挑战，本文提出了一种基于空频结合的双流卷积神经网络的检测模型。其核心是将频率域与空间域联合起来，共同挖掘 GAN 生成图像与真实图像间的显著差异，高效精准地完成对待测样本的真伪鉴别。由于 GAN 图像在生成过程中因上采样操作在频谱上会留下棋盘伪影，因此设计了可学习的频率域滤波核以及频率域网络，用于充分学习并提取频率域特征。为了减弱图像变换至频域过程中丢弃部分信息而带来的影响，同样设计了空间域网络来学习图像内容本身具有差异化的空间域特征，最终将两种特征融合来实现对 GAN 生成人脸图像的检测。实验结果表明，所提模型与现有方法相比无论是在精度、泛化性和鲁棒性上都有一定提高。

2 相关工作

2.1 GAN 生成人脸检测算法

现有的代表性算法主要分为两类：基于空间域信息的算法[3-6,8-13]和基于频率域信息[7,14-16]的算法。基于空间域信息的算法具体又可分为基于手工特征和基于卷积神经网络的检测算法。

Marra 等[3]直接使用隐写分析领域的富模型特征来进行伪造图像检测。McCloskey 等[4]通过分析 GAN 生成图像颜色分量的生成过程，提出了一个基于颜色特征的检测系统和一个用于最终分类的线性支持向量机。Li 等[8]研究了真伪图像间的颜色分量差异，提取空间域中色度分量的高频残差，计算基于共生矩阵的检测特征，最终结合分类器进行分类。

Mo 等[9]首次使用 CNN 来检测 GAN 生成的人脸图像。他们通过修改 CNN 架构，并以监督学习的方式实现了对伪造图像的检测，但是这种方法极易受到对抗样本的攻击。Nataraj 等[10]通过提取图像 RGB 通道上的共生矩阵，将其按通道维度进行堆叠并输入 CNN 进行分类，此算法忽略了颜色通道间的相关性。基于上述研究，Barni 等[5]提出了一种新的模型，即 Cross-Net。此模型运用 RB，RG 和 GB 这 3 种跨颜色通道组合的形式，并在此基础上计算跨颜色通道共生矩阵。实验结果表明，相比仅利用单一颜色通道信息，Cross-Net 能进一步提升检测算法的鲁棒性。Fu 等[11]提出了一种双流 CNN 检测算法。其中，一流使用高斯低通滤波器计算低频分量，而另一流则使用高通滤波器计算高频残差，然后将

两流融合输入至最后的全连接层进行最终的分类。实验结果表明，此方法有效提高了模型的鲁棒性，但仍难以抵抗 JPEG 压缩对检测精度的干扰。

在实际检测场景中，检测算法通常会受到两方面的挑战，即图像编辑操作的影响和未知 GAN 生成模型的威胁。针对上述情况，研究者们还从计算机视觉领域引入了一些先进思想进行算法优化。Wang 等[12]引入了数据增强算法，Marra 等[13]采用了增量学习技术，都使得模型的性能得到一定程度的提高。

Frank 等[7]指出，由于 GAN 在生成过程中反复使用上采样，使得图像在高频分量上存在异常。一旦转换到频率域，这种异常便清晰可见，表现为高频分量上的尖峰。基于此，提出了一种基于 DCT 频域变换的检测方式。类似地，Agarwal 等[14]设计了一种将图像颜色通道频谱作为输入，将胶囊网络作为核心架构的检测算法，同样取得了不错的效果。通过分析图像像素间的回归关系，Bonettini 等[15]发现，可将 GAN 图像的生成过程看作一组有限脉冲响应滤波器进行信号处理的过程，通过 DCT 系数的首位数字分布表征 GAN 图像的异常痕迹也同样可以达到较好的检测精度。

综上分析，这些算法在简单场景下均具有较好的检测效果，但是当面对未知 GAN 模型生成的图像或是一些图像编辑操作时，算法的检测性能便会明显下降。因此，如何提高算法的泛化性和鲁棒性，设计出一种对复杂场景、大多数数据都适用的通用性检测模型是本文研究的重点。

2.2 GAN 图像中的上采样伪影

GAN 在生成图像的过程中，其生成器会进行上采样操作以逐步提高图像的分辨率，这是图像生成必不可少的环节，且对于不同的 GAN 模型，上采样操作也是不同的[17]，如 DCGAN[18]采用转置卷积操作来完成上采样，而 PGGAN[2]则采用最近邻插值。图 2 给出了这两种操作的细节，其中 $*$ 代表卷积。给定一个低分辨率特征张量作为输入，上采样器将水平和垂直分辨率扩大 m。为了便于说明，本文假设 $m=2$，这是最常见的设置。上采样器在低分辨率特征张量中的每一行/列之后插入一个零行/列，并应用卷积运算来为补零的位置分配适当的值。转置卷积和最近邻插值的区别在于转置卷积中的卷积核是可学习的，而最近邻插值是固定的。

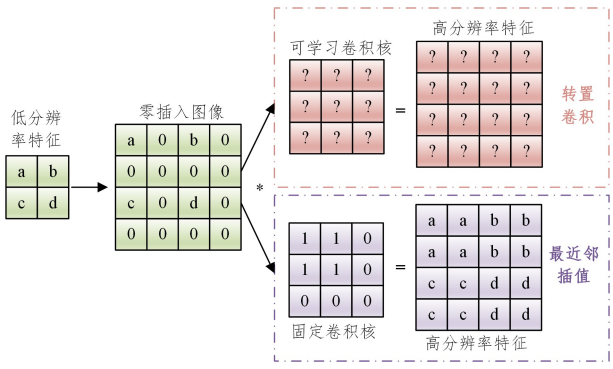


图 2 上采样操作(转置卷积和最近邻插值)

Fig. 2 Up-sampling operation(transpose convolution and nearest neighbor interpolation)

转置卷积的上采样伪影被称为“棋盘伪影”，Odena 等[19]在空间域中对此进行了研究，Zhang 等[20]在频域中也给出了深入的分析。根据离散傅里叶变换(Discrete Fourier Transform，DFT)的性质，在低分辨率图像中补零，相当于在图像的高频部分复制原始低分辨率图像的频谱的多个副本，以构成最终高分辨率图像的幅度频谱，如图 3 所示。为了方便展示，将频谱中的低频分量移至中间位置。

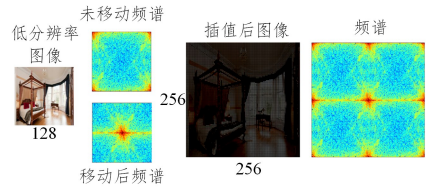
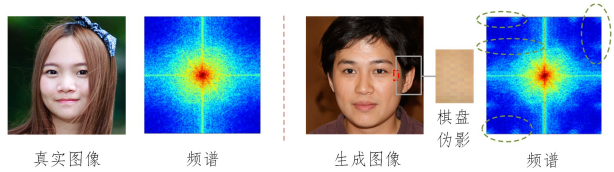


图 3 低分辨率图像和零插入图像的幅度频谱

Fig. 3 Spectral of low resolution images and zero insertion images

可以清楚地看到，经过插值后，图像频谱的高频分量存在异常。为了避免最终生成图像中存在这些伪影，需要移除或减少高频分量。因此，图 3 中的后续卷积核通常为低通滤波器。由于转置卷积中的卷积核是可学习的，不能保证卷积核一定为低通，因此仍然可以在许多图像中观察到棋盘伪影。如图 4 所示，其中最左边的两个图像是一个真实的人脸图像及其幅度频谱；右边的图像是 GAN 生成人脸及其对应的频谱。生成图像中红色高亮框中显示的是空间域中的棋盘效应，在其对应的幅度频谱中，本文用绿色椭圆形框出的明亮的斑点，对应于生成器中上采样模块产生的伪影，即频域中的棋盘效应。最近邻上采样器使用的卷积核是固定的低通滤波器，它确实可以更好地消除伪影。然而，伪造痕迹仍然没有被完全移除。并且，如果低通滤波器去除了太多的高频内容，最终图像则可能会变得模糊，影响视觉质量，从而很容易与真实图像区分开来。



注：放大细节为棋盘效应

图 4 真实图像及生成图像所对应的频谱(电子版为彩图)

Fig. 4 Spectrum of real image and generated image

基于上述理论分析，本文拟在频率域中分析真实图像和 GAN 生成图像，且将图像幅度频谱作为输入来让网络学习特征。除此之外，本文将频率域学习到的特征与空间域特征相结合，提出了基于空频联合的神经网络的 GAN 人脸检测模型，并将在第 3 节中对其进行详细介绍。

3 基于空频联合的 GAN 生成图像检测模型

基于上述对自然图像和 GAN 生成图像在频率域伪影的分析，本文构建了频率域网络来提取和学习图像的频谱特征。由于空域到频域的转换是一把双刃剑，可以将人脸图像能量集中，但也因此丢弃了许多代表细节信息的分量。为了使模型效果达到最大化，同时本文也构建了空间域网络来学习图像内容间的差异。最终，本文将各自学习到的特征进行融合，

构成了基于空频联合的 GAN 图像检测模型，具体如图 5 所示。

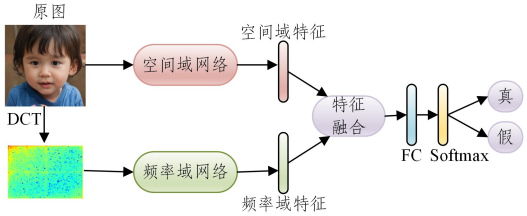


图 5　基于空频联合的 GAN 生成图像检测模型

Fig. 5　GAN-generated image detection model based on space-frequency combination

3.1　频率域网络的构建

由 2.2 节的分析可知，GAN 在生成图像的过程中，生成器会进行上采样操作以逐步提高图像的分辨率。而上采样会使得 GAN 生成的图像在频率域留下与自然图像不同的痕迹。如图 3 和图 4 所示，上采样操作不仅会在频谱的高频部分留下异常，而且会使得整张频谱出现分布不均的棋盘效应。基于这些特征，本文拟构建网络来充分学习频率域出现的伪影，以提高真伪图像的识别效率。

3.1.1　图像从空间域到频率域的转换

为了在频率域中对人脸图像进行特征学习，本文利用离散余弦变换（DCT）将图像从空间域转换到频率域。最初，DCT 常被用于处理一维数据（如文本、语音等），后来基于其能量集中的特性，在图像处理中也得到了广泛的应用。目前，有 8 种离散余弦变换的形式，本文使用最常用的形式 DCT-II，其计算式如下：

$$F(u,v)=4\sum_{x=0}^{M-1}\sum_{y=0}^{N-1}S(x,y)\cos\frac{(2x+1)u\pi}{2M}\cos\frac{(2y+1)v\pi}{2N}$$

$$1\leqslant u\ll M,1\leqslant v\ll N,S(x,y)\in[0,255] \tag{1}$$

其中，S 和 F 分别代表图像在空间域和频率域的形式，图像尺寸大小为 $M\times N$，$S(x,y)$ 表示图像空间域中坐标为 (x,y) 的像素值，$F(u,,v)$ 代表图像频率域中坐标为 (x,y) 的频率分量。

图 6 给出了真实图像及不同 GAN 模型生成图像的平均幅度频谱。虽然每种 GAN 的结构都有所不同，但是它们在频率域中都会留下相似的伪影，即每张图中清晰可见的棋盘伪影。接下来，本文通过构建网络来充分学习频率域出现的伪影，以提高真伪图像的识别效率。

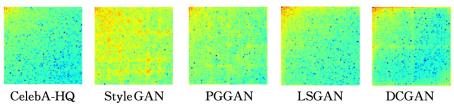


图 6　真实图像及不同 GAN 模型生成图像的平均幅度频谱

Fig. 6　Average spectra of real images and images generated by different GAN models

3.1.2　可学习的滤波核

当图像从空间域转换至频率域后，一些常用的特性（如噪声、空间信息冗余等）变得更为显著且更易操作。因此，本文利用频率域的这些优势来对人脸图像进行进一步的分析。由于图像滤波是通过矩阵之间的元素点乘实现的，在频率域中的计算非常高效，因此本文拟利用矩阵的元素点乘充当图像的频域滤波，来实现有效的特征提取。例如常见的高通滤波和低通滤波，前者可以保留高频信息，突出图像的细节，生成一幅锐化的图像，而后者保留低频信息，以此来去除图像的噪声，从而使图像变得平滑。然而，若要过滤掉无用信息而提取所需的图像特征，则需要针对性地设计滤波。为了高效地生成合适的滤波，本文使用可学习的频率域滤波核（Learnable Filter Kernel，LFK），它能够充当一个可学习的滤波进行特征提取。由于卷积神经网络具有反向传播的特性，这使得滤波核能够在网络中反复学习，自动调整自身的多个权值。与手工设定参数值相比，这样的方式更为简洁高效，且能提取出最有效的人脸信息。单个滤波核的定义如下：

$$\boldsymbol{K}=\begin{bmatrix} w_{11} & \cdots & w_{1n} \\ \vdots & \ddots & \vdots \\ w_{m1} & \cdots & w_{mn} \end{bmatrix} \tag{2}$$

在滤波核 $\boldsymbol{K}$ 中，$w(x,y)$ 表示可训练的权重，其数值根据截断正态分布随机初始化，其中截断正态分布的均值为 0，标准差为 0.1。滤波核的尺寸大小为 $m\times n$，与输入图像相等。由于滤波核可通过在网络中进行训练来自动调整权值，因此非常省时高效，而现有的工作大多是根据先验知识通过手工设计的方法来设定滤波核的权值，该方法未必能得到最有效的人脸信息，且会浪费大量时间。

本文所指的频率域特征图在形式上与卷积神经网络中的特征图相同，都以矩阵的形式存在，差异在于一个包含图像频域信息，而另一个代表图像的空间域信息。当图像转换至频率域输入网络后，滤波核就会对输入进行频域滤波的操作，而此操作是通过点乘的方式进行的，最终输出的特征图尺寸与原图相同。可学习滤波核的具体过程如图 7 所示。

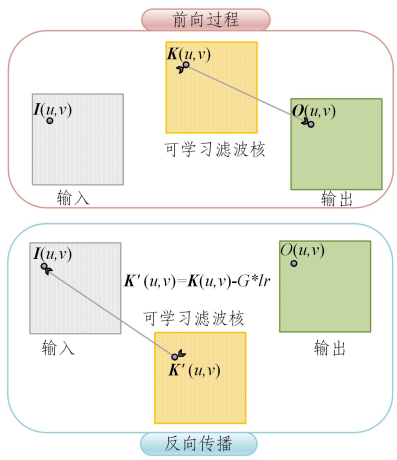


图 7　可学习的滤波核

Fig. 7　Learnable filter kernel

图 7 中，$\boldsymbol{I}$ 为图像的输入频谱，滤波核 $\boldsymbol{K}$ 与输入尺寸相同，$\boldsymbol{O}$ 为输出的特征图。该滤波前向过程的定义如下：

$$\boldsymbol{O}(u,v)=\boldsymbol{I}(u,v)\cdot\boldsymbol{K}(u,v) \tag{3}$$

其中，$1\leqslant u\leqslant m,1\leqslant v\leqslant n$，$m$ 和 n 分别为输入矩阵 $\boldsymbol{I}$ 的行数和列数。就输出特征图中的单个元素点而言，该元素点由一次乘法操作计算而来。相比卷积滤波中单个元素由多个点加权

求和得来的方式，这种计算更加高效，这也正是传统频域处理图像的优势之一。

该滤波反向传播过程的定义如下：

$$\boldsymbol{K}'(u,v)=\boldsymbol{K}(u,v)-G*lr \tag{4}$$

其中，G 为梯度，lr 为学习率。通过不断地迭代来调整自身权值，以提取最有效的人脸信息。

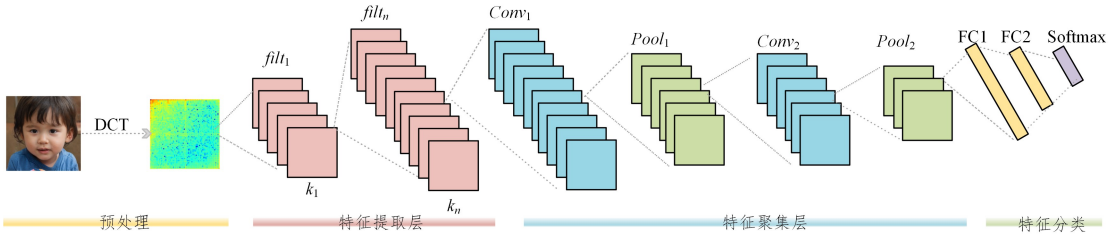


图 8 频率域网络

Fig. 8 Frequency domain network

在特征提取层中，本文使用可学习的滤波核作为图像特征提取的工具，即对于每个输入的特征频谱，都有 K 个频率域滤波核。其中，本文定义 K 为该层特征提取层的特征图扩充系数，具体设置如表 1 所列，该系数在网络中的作用就是增加特征图的数量，以便更全面地提取到频谱中所蕴含的信息。具体地，n 为特征提取层的总数，$filt_i$ 表示第 i 层特征提取层。为了生成充足的特征图，设置 K_i 为第 i 层的扩充系数。n 的数值会影响网络的深度和模型的性能，太小可能会导致特征未得到充分提取，太大可能会导致模型参数过多以致于不收敛，耗费大量时间。经过后续的实验验证，本文将 n 设置为 3。

特征聚集层主要包括卷积和池化。由于本文使用的可学习滤波核输出特征图后，特征维度并没有降低，因此特征聚集层最主要目的就是增加输出特征图之间的关联信息，并实现特征的降维。因此，本文首先采用卷积来实现频率信息的局部聚合，接着使用池化操作在尽可能保留图像信息的前提下降低特征维度，加快模型收敛。由于在图像的频率域中，低频信息包含图像绝大部分信息，且绝对值较大，因此本文采用最大池化来实现特征的降维。频率域网络的详细配置如表 1 所列。

表 1 频率域网络结构及参数配置

Table 1 Frequency domain network structure and parameter configuration

网络结构	变量	数据维度	参数设置
预处理	—	256×256	None
特征提取	$filt_1$	256×256×32	$K_1=32$
	$filt_2$	256×256×32	$K_2=1$
	$filt_3$	256×256×32	$K_3=1$
特征聚集	$conv_1$	256×256×32	conv(3×3×32)
	$pool_1$	128×128×32	max pool(2×2)
	$conv_1$	128×128×48	conv(3×3×48)
	$pool_1$	64×64×48	max pool(2×2)

3.2 空间域网络的构建

3.1 节构建了频率域网络来提取图像频谱中的伪影，以实现最终的分类任务。虽然该过程利用了离散余弦变换能量集中的性质，减小了计算的难度，但是丢弃了代表图像细节信息的分量，在一定程度上影响了检测的效果。且对于不同的 GAN 模型而言，上采样使用的频次不同，存在伪影的区域大小也不尽相同。尤其是对于清晰度不佳的生成图像而言，光靠频率域的信息来鉴别真伪不够准确。因此，除了频率域，本文进一步从空间域的角度出发，挖掘图像内容本身的差异，构建空间域网络来提取空间域特征。

通过构建空频联合模型，来实现更好的 GAN 人脸图像检测效果，该模型主要由 3 个层构成。不同的是，受到 VGG 网络结构的启发，本文采用双层卷积结构来进一步扩充卷积核的感受野，且在每个卷积层后都配备非饱和激活函数 LReLU[21]。空域模型的结构及参数配置如表 2 所列。

表 2 空间域网络结构及参数配置

Table 2 Spatial domain network structure and parameter configuration

数据变量	数据维度	参数设置
输入	256×256	None
$conv_{11}$	256×256×24	conv(3×3×24)
$conv_{12}$	256×256×24	conv(3×3×24)
$pool_1$	128×128×24	max pool(2×2)
$conv_{21}$	128×128×36	conv(5×5×36)
$conv_{22}$	128×128×36	conv(5×5×36)
$pool_2$	64×64×36	max pool(2×2)
$conv_3$	64×64×48	conv(7×7×48)
$pool_3$	32×32×48	max pool(2×2)

3.3 基于空频联合的双流 CNN 网络

3.1 节和 3.2 节分别构建了频率域和空间域网络，以分别提取各自的特征信息。为了更好地实现真伪人脸检测的任务，本文将两种特征进行融合，构成最终基于空频联合的双流 CNN 模型。

经过卷积和池化操作后，频率域特征和空间域特征的长度不同，无法直接进行融合。因此，本文首先进行维度转换，以确保这两种特征在特征融合时权重相等。具体而言，通过全连接的方式，将两种特征都转化为长度为 1024 的特征，通过逐点相加的方式对两个特征进行融合。最后，本文将融合的特征送入由 512 个神经元单位组成的全连接层中，并使用非饱和函数 LReLu 进行激活。除此之外，本文还在全连通层中进行 L2 正则化，其中参数 λ 为 0.0005。由于本文研究的

3.1.3 频率域网络的构建

频率域网络的构建是为了提取和学习自然图像与真实图像在频率域中具有可分辨性的特征。网络的具体结构如图 8 所示，由预处理阶段、特征提取层、特征聚集层和最终的特征分类 4 部分组成。其中预处理阶段是将图像从空间域转换到频率域的过程，如式(1)所示。

是针对真伪人脸图像检测的二分类问题，因此最后利用 Softmax 层产生分类概率并进行最终的分类，分类过程如图 5 所示。

4 实验结果与分析

本节首先介绍了用于评估所提算法有效性的实验设置，包括数据集以及评价指标。其次，通过大量实验证明了本文提出的基于空频联合的检测模型的有效性，不仅在 GAN 生成的全局人脸中，对于 GAN 生成的局部人脸，所提方法依然有较好的检测性能。

4.1 数据集和评价指标

对于伪造人脸数据集，本文选用由 6 种不同结构 GAN 网络生成的图像，包括生成质量极高的先进 GAN 模型 StyleGAN II[22]和 StyleGAN I[23]、中期的生成模型 PGGAN[2]和 WGAN[24]，以及生成质量不够清晰的早期模型 BEGAN[25]和 LSGAN[26]。对于真实人脸图像，本文选用的是 CelebA-HQ 数据集[27]，为了与 GAN 生成的人脸图像所匹配，本文也同样选择了 3 种分辨率的图像。所有数据集的分辨率及数量如表 3 所列。

表 3 实验中所使用的伪造人脸及真实人脸数据集

Table 3 Fake face and real face datasets used in experiment

真假性	名称	分辨率	数量
伪造人脸数据集	StyleGAN II	1024×1024	8000
	StyleGAN I	1024×1024	8000
	PGGAN	256×256	10000
	WGAN	256×256	10000
	BEGAN	128×128	10000
	LSGAN	128×128	10000
真实人脸数据集	CelebA-HQ	1024×1024	12000
	CelebA-HQ	256×256	12000
	CelebA-HQ	128×128	12000

数据如此多样化的原因是，为了使训练后的模型能够适应更实际且更广泛的需求，该模型不仅需要对多种结构 GAN 生成的图像有效，而且对不同结构 GAN 生成的不同分辨率的图像同样可以达到理想的检测效果。由于此双流模型的输入图像尺寸为 256×256，因此本文将所有图像的尺寸均调整为 256×256。在所有的实验中，本文只在 StyleGAN I 和 CelebA-HQ 上训练双流 CNN，其中真实图像选用的分辨率与 StyleGAN I 保持一致，即 1024×1024。其余 5 个 GAN 生成的人脸数据集仅用于测试训练模型的泛化及鲁棒能力。最终，模型的性能通过精度(ACC)来进行衡量，计算式如下：

$$Accuracy=\frac{T_p+T_N}{P+N} \tag{5}$$

其中，P 和 N 分别为数据集中生成样本和真实样本的个数，T_p 和 T_N 分别为正确检测到的生成样本和正确检测到的真实样本的个数。

4.2 实验结果与分析

4.2.1 消融实验

本节首先就单独的频率域网络、空间域网络以及基于空频联合的整体模型进行了消融研究。训练中使用的真实数据集来自 CelebA-HQ(包含 8000 张)，伪造人脸数据集来自 StyleGAN I，PGGAN 以及 WGAN(各包含 8000 张)，对应的分辨率为 1024×1024，256×256 以及 128×128。训练结束后，模型在包含 3000 张真实图像和 3000 张伪造图像的测试集中进行评估，实验结果如表 4 所列。表中的第二行为单独使用频率域网络(FreNet)所获得的真伪人脸分类的准确率；第三行为单独使用空间域网络(SpaNet)所获得的准确率；第四行是将频率域特征与空间域特征融合后(Spa-Fre Net)所获得的人脸鉴别准确率。可以清楚地看到，在所有数据集的测试中，所提出的基于空频联合的网络模型都获了最高的分类精度，证明了所提方法对 GAN 人脸检测的有效性。除此之外，如表 4 所列，单独的 FreNet 对于 StyleGAN I 这种生成质量较高的、非常逼真的人脸有着惊人的准确率，而对于早期的 GAN 模型，如 WGAN 生成的图像准确率只有 95.20%，这也归因于 StyleGAN 模型在生成图像的过程中经历了多次上采样操作，最终使得生成的图像有较高的分辨率。且由于经历了多次上采样，也使得图像在高频部分留下了清晰可辨的伪影，这些伪影单独利用频率域网络也可以很好地进行检测。而对于像 WGAN 这种清晰度欠佳的数据集来说，仅使用频率域网络来实现分类，效果并不理想，会出现很多误判。这也是设计空间域网络的重要原因。本文通过弥补图像在频率域被丢弃的细节信息，在空间域进行同步的提取和学习，以此来弥补频率域的不足，构成最终的基于空频联合的达到理想效果的检测模型。

表 4 消融实验的结果

Table 4 Ablation results

(单位：%)

精度	StyleGAN I	PGGAN	WGAN
FreNet	99.68	98.87	95.20
SpaNet	98.59	98.68	99.68
Spa-FreNet	**99.86**	**99.49**	**99.77**

4.2.2 频率域特征提取层数量分析

频率域特征提取层的数量 n 影响着频率域网络的深度和训练的收敛程度，这也直接影响到了模型的检测性能。在本次实验中，本文设置不同的 n(1,2,3,4,5)进行对比实验，实验结果如表 5 所列。可以清楚地看到，当 n 设置为 1 时，网络并未充分学习到有用的特征，分类精度只有 65%左右。而从提取层设置为 2 开始，模型渐渐发挥出了它的优势，分类精度显著上升。当 n 设置为 3 或 4 时，在数据集上已经有非常好的表现，准确率接近 100%。而当 $n=5$ 时，由于层数太深以致于网络无法收敛。经过综合考量，本文最终设置 n 为 3，这样既可以达到理想的精度，且由于比 $n=4$ 时的训练参数少，因此模型收敛更快，还可以缩短训练时间。

表 5 关于特征提取层层数的分析

Table 5 Analysis on the number of feature extraction layers

(单位：%)

提取层层数	精度
1	65.00
2	97.55
3	99.86
4	**99.88**
5	—

4.2.3 泛化性实验

随着研究的推进，GAN 网络的发展不会止步于此，未来

将会生成更多且质量更高的新型模型。想要使得算法真正应用于实际，泛化能力是非常重要的检验标准。因此，本文测试了基于空频联合的 GAN 人脸检测模型的泛化能力，验证其在面对未知 GAN 生成的图像时，是否也能获得令人满意的效果。

本文使用 6 种不同结构 GAN 生成的图像进行泛化实验，具体如表 6 和表 7 所列。表 6 中，用于训练的 3 组伪造图像分别来自 StyleGAN I，BeGAN 和 LsGAN。对于每种 GAN，都随机选择了 4000 张图像进行训练，总共有 12000 张伪造人脸图像。相应地，本文在 Celeba-HQ 数据集中选择了包含 3 种分辨率的 12000 张真实图像，使得训练图像总数达到 24000 张。对于剩余的 3 种 GAN(StyleGAN II，PGGAN 以及 WGAN)，本文使用上述训练好的模型分别对其进行测试，选取的真实图像也会与其分辨率相匹配。

表 6 不同算法在 StyleGAN I，BeGAN 和 LsGAN 生成的数据集上的泛化性实验对比

Table 6 Generalization comparison of different algorithms on datasets generated by StyleGAN I, BeGAN, and LsGAN

训练集	StyleGAN I+BeGAN+LsGAN		
测试集	StyleGAN II	PGGAN	WGAN
Nataraj	92.97	98.04	**99.53**
Mo	94.37	97.64	99.26
Zhang	98.05	97.45	95.25
FreNet	98.56	97.37	97.06
Fre-Spa Net	**99.24**	**98.67**	98.90

表 7 不同算法在 StyleGAN II，PGGAN 和 WGAN 生成的数据集上的泛化性实验对比

Table 7 Generalization comparison of different algorithms on datasets generated by StyleGAN II, PGGAN, and WGAN

训练集	StyleGAN I+BeGAN+LsGAN		
测试集	StyleGAN II	PGGAN	WGAN
Nataraj	94.41	98.45	**99.89**
Mo	93.28	98.77	99.75
Zhang	98.78	98.27	95.63
FreNet	98.91	98.22	96.02
Fre-Spa Net	**99.36**	**98.90**	98.83

交换用于训练和测试的第一组 GAN 生成数据集的位置并进行泛化性实验，其效果如表 7 所列，即使用 StyleGAN II，PGGAN 和 WGAN 生成的图像来训练模型，并使用 StyleGAN I，BeGAN 和 LsGAN 来测试模型性能。

为了更直观地展现所提方案的优势，本文将模型的性能与一些 GAN 检测算法[9-10,20]进行了对比。其中，Zhang 等[20]提的为基于频率域的检测方法。总体来看，所提模型的泛化性可圈可点，这也证实了尽管每一种 GAN 模型的结构都不尽相同，但是只要使用了上采样操作，频率域就会留下异常痕迹。一旦模型学习到了这种伪影，就可以提升在鉴别其他 GAN 生成图像时的表现。最引人注意的是，本文提出的空频联合结构大大提升了 StyleGAN 的检测率，这是目前生成效果最好的、质量最高的伪造图像，分类结果达到了 99%以上。

与高分辨率的图像相比，类似 LsGAN 这种清晰度较低的数据集的高频特征不容易被发现，但是有了空间域特征的辅助，分类精度也维持在较好的水平。

4.2.4 鲁棒性实验

为了判断当图像上传到互联网或被下载到手机上时，这些用于分辨图像的伪影在通过后期处理操作后是否还存在并持续发挥作用，本文评估了所提的空频联合检测模型对常见图像干扰的鲁棒性，即压缩、剪切、模糊、随机噪声以及所有这些干扰的组合，具体设置如下：

(1)图像采用 JPEG 压缩，压缩所选择的质量因子从[10，90]中随机取样；

(2) 图像被随机裁剪，裁剪的百分比从[5，20]中随机取样，裁剪后的图像将上采样到其原始分辨率；

(3)图像被添加高斯模糊，核尺寸从[0.5，2.5]中随机取样；

(4)图像被加入高斯噪声，方差从[5，20]中随机取样；

(5)将上述所有扰动组合。

在本实验中，训练及测试的数据集均来自 Celeba-HQ(真实)和 StyleGAN I(伪造)，测试模型是表 4 中第一列第四行所保存的模型。实验结果如表 8 所列。

表 8 鲁棒性测试结果

Table 8 Robustness experiments results

(单位：%)

模型	压缩	剪裁	模糊	噪声	组合
Nataraj	75.76	90.78	78.67	53.26	70.22
Mo	95.24	94.08	93.45	**95.26**	95.01
Zhang	94.83	98.83	93.61	89.56	92.27
FreNet	95.04	98.24	95.77	91.23	94.96
Fre-Spa Net	**97.82**	**99.68**	**97.88**	93.76	**96.56**

可以清楚地看到，本文提出的基于空频联合的网络除了在应对图像剪裁上获得了 99.68%的高精度外，在其余的图像干扰下，其精度都有明显的下降。表 8 中的第五行列出了单独的频率域网络的检测效果。由于频率域网络的核心是抓住高分辨率图像频谱中的棋盘伪影来进行检测，这就导致上采样操作使用得越多，越高清的图像反而更有利于网络的学习和提取特征。然而，除了剪裁外，无论是压缩、模糊或是噪声都对图像的清晰程度造成了非常严重的影响，这也就间接破坏了本身频谱中清晰可辨的伪影，使得最终的模型效果不够理想。值得注意的是，在频域的基础上添加了空域的辅助，使得每一项分类精度都有了明显的提升，这也说明将空域特征和频域特征联合起来鉴别图像真伪的思路是正确的，但是如何进一步提高模型的鲁棒性也是未来值得思考和研究的方向。

4.2.5 GAN 生成的局部人脸分类测试

目前，对于 GAN 图像检测来说，几乎所有的目光都聚焦在全局人脸上，很少有人关注 GAN 生成的局部人脸。但实际上，用 GAN 进行人脸修复的情况很多。有些是出于好意，有些却是故意为之，改变了原有人脸图像的眼神、表情等信息，故意歪曲事实，造成了不好的影响。因此，本次实验尝试探索所提模型是否具有更广泛的应用场景。所用的真实数据集来自 FFHQ，局部伪造人脸数据集来自 LGGF[28]。LGGF 数据集共有 952000 张图像，生成的区域具有不同的形状和大小，具体如图 9 所示。

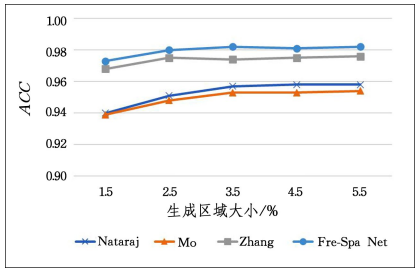
(a)规则掩码

(b)不规则掩码

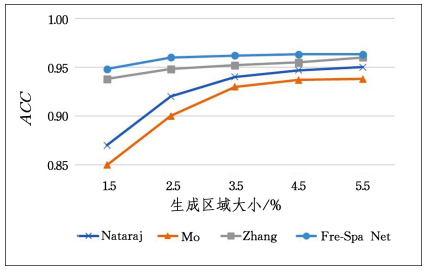
图 9 LGGF 中的部分数据集样本

Fig. 9 Some datasets samples in LGGF

本文将对不同生成区域大小进行评估，整个测试集包含 6 种尺寸，生成图像的比例为 0.5%～5.0%。图 10(a)给出了规则生成区域的检测精度，图 10(b)给出了不规则生成区域的检测精度。可以清楚地看到，所有模型的检测精度都随着生成区域大小的增大而增加，且本文提出的基于空频联合的模型结构在两种区域都取得了最佳的效果。这表明，在 GAN 生成的局部人脸中，空间域特征中纹理的差异以及频率域信息中高频的痕迹照样起到了鉴别真伪的作用，这也引发了更多对模型更广泛的适用性以及未来优化的思考。



(a)



(b)

图 10 不同生成区域的检测精度

Fig. 10 Detection accuracy in different generated areas

结束语 本文提出了一种基于空频联合的卷积神经网络的检测模型，可学习滤波核和频率域网络的构建来学习频率域中的棋盘伪影，同时设计了空间域网络来学习图像内容本身的差异化特征，最终将空频特征融合来鉴别人脸图像的真伪。实验结果证明，所提算法在高质量生成数据集的检测精度及泛化性上均优于现有算法，且有着更加广泛的应用前景。当然，本文方案还存在着一些不足，当图像经过模糊、压缩、噪声等操作后，会严重影响频率域的伪影，从而影响了模型的鲁棒性。如何弱化这些影响，进一步提升模型鲁棒性是未来可以优化的方向。

参 考 文 献

[1] STUPP C. Fraudsters used AI to mimic CEO's voice in unusual cybercrime case[J]. The Wall Street Journal, 2019, 30(8).

[2] KARRAS T, AILA T, LAINE S, et al. Progressive growing of gans for improved quality, stability, and variation[J]. arXiv: 1710.10196, 2017.

[3] MARRA F, GRAGNANIELLO D, COZZOLINO D, et al. Detection of GAN-generated fake images over social networks[C]// 2018 IEEE Conference on Multimedia Information Processing and Retrieval(MIPR). IEEE, 2018: 384-389.

[4] MCCLOSKEY S, ALBRIGHT M. Detecting GAN-generated imagery using saturation cues[C]// 2019 IEEE International Conference on Image Processing(ICIP). IEEE, 2019: 4584-4588.

[5] BARNI M, KALLAS K, NOWROOZI E, et al. CNN detection of GAN-generated face images based on cross-band co-occurrences analysis[C]// 2020 IEEE International Workshop on Information Forensics and Security(WIFS). IEEE, 2020: 1-6.

[6] GUO Z, YANG G, CHEN J, et al. Fake face detection via adaptive manipulation traces extraction network[J]. Computer Vision and Image Understanding, 2021, 204: 103170.

[7] FRANK J, EISENHOFER T, SCHÖNHERR L, et al. Leveraging frequency analysis for deep fake image recognition[C]// International Conference on Machine Learning. PMLR, 2020: 3247-3258.

[8] LI H, LI B, TAN S, et al. Detection of deep network generated images using disparities in color components[J]. arXiv: 1808.07276.

[9] MO H, CHEN B, LUO W. Fake faces identification via convolutional neural network[C]// Proceedings of the 6th ACM Workshop on Information Hiding and Multimedia Security. 2018: 43-47.

[10] NATARAJ L, MOHAMMED T M, MANJUNATH B S, et al. Detecting GAN generated fake images using co-occurrence matrices[J]. Electronic Imaging, 2019(5): 532-1-532-7.

[11] FU Y, SUN T, JIANG X, et al. Robust gan-face detection based on dual-channel cnn network[C]// 2019 12th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics(CISP-BMEI). IEEE, 2019: 1-5.

[12] WANG S Y, WANG O, ZHANG R, et al. Cnn-generated images are surprisingly easy to spot…for now[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 8695-8704.

[13] WANG S Y, WANG O, ZHANG R, et al. Cnn-generated images are surprisingly easy to spot for now[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 8695-8704.

[14] AGARWAL S, GIRDHAR N, RAGHAV H. A Novel Neural Model based Framework for Detection of GAN Generated Fake Images[C]// 2021 11th International Conference on Cloud Computing, Data Science & Engineering(Confluence). IEEE, 2021: 46-51.

[15] BONETTINI N, BESTAGINI P, MILANI S, et al. On the use of Benford's law to detect GAN-generated images[C]// 2020 25th

International Conference on Pattern Recognition(ICPR). IEEE, 2020:5495-5502.
[16] DURALL R, KEUPER M, KEUPER J. Watch your up-convolution:Cnn based generative deep neural networks are failing to reproduce spectral distributions[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020:7890-7899.
[17] HE P S, LI W C, ZHANG J Y, et al. Overview of passive forensics and anti-forensics techniques for GAN-generated image[J]. Journal of Image and Graphics, 2022, 27(1):0088-0110.
[18] RADFORD A, METZ L, CHINTALA S. Unsupervised representation learning with deep convolutional generative adversarial networks[J]. arXiv:1511.06434, 2015.
[19] ODENA A, DUMOULIN V, OLAH C. Deconvolution and checkerboard artifacts[J]. Distill, 2016, 1(10):e3.
[20] ZHANG X, KARAMAN S, CHANG S F. Detecting and simulating artifacts in gan fake images[C]// 2019 IEEE International Workshop on Information Forensics and Security(WIFS). IEEE, 2019:1-6.
[21] LI C L, RAVANBAKHSH S, POCZOS B. Annealing Gaussian into ReLU:a new sampling strategy for leaky-ReLU RBM[J]. arXiv:1611.03879, 2016.
[22] KARRAS T, LAINE S, AITTALA M, et al. Analyzing and improving the image quality of stylegan[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020:8110-8119.
[23] KARRAS T, LAINE S, AILA T. A style-based generator architecture for generative adversarial networks[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019:4401-4410.
[24] ARJOVSKY M, CHINTALA S, BOTTOU L. Wasserstein generative adversarial networks[C]// International Conference on Machine Learning. PMLR, 2017:214-223.
[25] BERTHELOT D, SCHUMM T, METZ L. Began:Boundary equilibrium generative adversarial networks[J]. arXiv: 1703.10717, 2017.
[26] MAO X, LI Q, XIE H, et al. Least squares generative adversarial networks[C]// Proceedings of the IEEE International Conference on Computer Vision. 2017:2794-2802.
[27] SUN Y, CHEN Y, WANG X, et al. Deeplearning face representation by joint identification-verification[J/OL]. Advances in Neural Information Processing Systems, 2014, 27. https://proceedings. neurips. cc/paper/2014/hash/e5e63da79fcd2bebbd7cb8bf1c1d0274-Abstract. html.
[28] CHEN B, JU X, XIAO B, et al. Locally GAN-generated face detection based on an improved Xception[J]. Information Sciences, 2021, 572:16-28.

WANG Jinwei, born in 1978, Ph.D, professor, Ph.D supervisor, is a member of China Computer Federation. His main research interests include artificial intelligence security, color image forensics, color image reversible watermark, robust watermark and image encryption.

LUO Xiangyang, born in 1978, Ph.D, professor, Ph.D supervisor, is a member of China Computer Federation. His main research interests include network and information security.

(责任编辑:喻藜)