第 60 卷 第 10 期/2023 年 5 月/激光与光电子学进展 研究论文

激光与光电子学进展 先进成像

面向浅层特征高频分量的深度伪造检测算法

彭舒凡，蔡满春*，马瑞，刘晓文

中国人民公安大学信息网络安全学院，北京 100038

摘要 近年来，深度伪造技术大幅提升了合成人脸的真实感，且相较于传统伪造方法，其生成的虚假视频更加难以分辨。基于深度伪造图像视觉伪影常常存在于特征提取网络浅层特征高频分量中这一特性，设计了一种面向浅层特征高频分量的深度伪造图像检测算法。针对高通滤波器的缺陷，本实验在拉普拉斯金字塔的基础上设计了一种具有更好的过滤性能的高频残差提取模块。在增强模块中，使用Convolutional Block Attention Module（CBAM）增加特征图关键区域以及关键特征通道的权重，提升特征图的空间以及通道相关性。针对深层网络中高频分量学习优先级低的问题，设计了一种图像梯度损失算法，防止高频信息随着网络的加深而丢失。将梯度中心化引入AdamW优化器，解决了深度伪造检测模型训练时间长、泛化性差的问题。所提两种模型在FaceForensics++和Celeb-DF数据集上的准确率均优于主流算法，证明了算法的有效性以及泛化性。

关键词 机器视觉；深度伪造；深度伪造检测；高频分量；图像梯度损失；梯度中心化

中图分类号 O436 **文献标志码** A **DOI**：10.3788/LOP213318

Deepfake Detection Algorithm for High-Frequency Components of Shallow Features

Peng Shufan, Cai Manchun*, Ma Rui, Liu Xiaowen

College of Information and Cyber Security, People's Public Security University of China, Beijing 100038, *China*

Abstract Deepfake techniques have dramatically improved the realism of synthetic faces in recent years. And the fake videos it generates are more difficult to distinguish than traditional forgery methods. Based on the characteristic that visual artifacts of depth forgery images often exist in the high frequency components of shallow features in feature extraction network, a detection algorithm for depth forgery images oriented to the high frequency components of shallow features is designed. First, a high-frequency residual extraction module based on Laplace's pyramid with better filtering performance is designed to address high-pass filters' shortcomings. Second, the Convolutional Block Attention Module (CBAM) is used to increase the weights of key regions of the feature map and key feature channels to improve the spatial and channel correlation of the feature map in the enhancement module. Then, an image gradient loss is designed to prevent the loss of high-frequency information as the network deepens to address the problem of low learning priority of high-frequency components in deep networks. Finally, gradient-centralization is introduced into the AdamW optimizer to solve the problems of long training time and poor generalization of deep forgery detection models. Two models proposed outperform mainstream algorithms in terms of accuracy when validated on the FaceForensics++ and Celeb-DF datasets, demonstrating the algorithms' effectiveness and generalization.

Key words machine vision; deepfake; deepfake detection; high-frequency component; image gradient loss; gradient-centralization

1 引 言

随着数字化社会的发展，“人脸”在网络中常常起到多种虚拟身份之间关联以及认证的作用，例如，人脸识别常常用于访问控制以及移动支付[1]。然而，这些进步也诱使了“深度伪造”技术的诞生。深度伪造技术

收稿日期：2021-12-23；**修回日期**：2022-01-17；**录用日期**：2022-02-14；**网络首发日期**：2022-02-24

基金项目：“十三五”国家密码发展基金密码理论研究重点课题（MMJJ20180108）、中国人民公安大学2020年基本科研业务费重大项目（2020JKF101）

通信作者：*caimanchun@ppsuc.edu.cn

大幅提升了合成人脸的真实感：相较于传统伪造方法，其生成的虚假视频更加难以分辨；其生成的图片能够通过社交媒体之类的互联网媒介[2-4]迅速传播，给个人、社会、国家带来潜在威胁。因此，为了减轻深度伪造技术带来的负面影响、保护公共安全以及个人隐私，开发有效的深度伪造检测技术已经成为当前的热门研究方向。

根据数据驱动的不同，常用的深度伪造检测方法可以分为基于视频级学习的检测方法和基于图片级学习的检测方法。由于生成深度伪造视频的过程是逐帧进行的，帧与帧之间的伪造操作相互独立，因此帧间常常会存在差异。基于视频级的检测方法可以利用循环神经网络（RNN）等与序列数据有关的算法学习到伪造视频与真实视频的不一致性，如生理信号的差异、被篡改区域的不稳定、前后帧的不一致等缺陷。例如：Amerini 等[5]利用 VGG-16 检测视频连续帧之间光流矢量的差异；Agarwal 等[6]首先对帧间的人脸以及头部生理信号进行编码建模，然后利用支持向量机（SVM）进行检测；Güera 等[7]针对伪造视频与背景融合性差的缺陷，利用 RNN 进行检测；Sabir 等[8]将视频中人脸对齐后利用循环卷积模型进行检测。基于视频级的学习方法对于视频预处理的要求很高，无法判断单帧的真伪，并且由于所用方法往往针对特定数据集，泛化能力差。

基于图片级学习的检测方法常常利用卷积神经网络（CNN）作为主干网络从视频的帧内图像中提取特征信息，然后利用特征信息进行分类。例如：直接使用经典的 ResNet[9]、XceptionNet[10]和 EfficientNet[11]等网络进行检测；Zhou 等[12]提出的 Region-CNN 双流网络结构包含 RGB 流以及噪声流，最终经特征融合达到检测定位的效果；Nguyen 等[13]利用胶囊网络在特征提取、抵御噪声方面优于 CNN 的特性，对多种伪造手段进行检测；Afchar 等[14]提出 MesoNet，利用图像的中层语义信息进行检测；耿鹏志等[15]提出针对深度伪造样本的数据增强方法，在降低 XceptionNet 参数量的同时提升了模型精度。当前图片级检测方法一般包括特征提取网络和分类网络，其中特征提取网络常常为深层结构，存在对于高频分量学习优先级低的固有缺陷[16]，因此在最终的分类决策时，高频信息对于最终分类结果的贡献低。而深度伪造生成时会在伪造区域留下很多高频痕迹[17-18]，因此需要在检测时对于图像的高频信息进行增强。在以往的相关工作中：Mo 等[19]利用图像的高频分量进行深度伪造检测，将原始图像经过高通滤波器后直接对高频分量进行检测；Masi 等[20]使用双流网络，通过原始图像的高频流以及 RGB 流实现检测以及伪造区域的定位。然而，现有的基于高频特征的检测方法通常是直接基于原始图像的高频特征的，但在压缩的数据集中，由于原始图像伪影不明显，该类方法在压缩数据集中表现不佳。而特征提取网络的浅层则更能提取到图像的一些细粒度信息，伪造图像的视觉伪影等特征在这些信息的高频分量中是显著的。之前基于高频特征的检测方法并未注意到随着网络层数的上升会导致高频分量丢失，因此，本实验提出一种面向浅层特征高频分量的检测算法。

本文的主要工作与贡献如下：针对以往算法直接使用压缩图像的高频分量出现精度下降的缺陷，使用主流检测模型浅层特征的高频分量更好地捕捉图像的细粒度信息；针对一般高通滤波器提取高频信息不充分的缺陷，提出一种具有更好的高频信息提取能力的高频信息残差提取方法；在高频信息增强模块中，将 Convolutional Block Attention Module（CBAM）[21]嵌入残差块（ResBlock）结构，增加特征图关键区域以及关键特征通道的权重，提升特征图的空间以及通道相关性；针对深层网络中高频分量学习优先级低的问题，设计了一种能够使模型更加有效地学习高频分量图像的梯度损失算法；将梯度中心化（GC）引入 AdamW 优化器，解决深度伪造检测模型训练时间长、泛化性差的问题。

2 面向浅层特征高频分量的深度伪造检测算法

2.1 算法整体架构

本实验为基于图片级学习的深度伪造检测算法，算法整体架构如图 1 所示。

所提算法分为 3 个模块：浅层特征提取模块、高频信息残差提取模块以及高频信息增强模块。浅层特征提取模块采用主流检测网络的浅层提取输入图像 $\boldsymbol{i}$ 的第 t 层的浅层特征图 $\mathrm{FL}_t(\boldsymbol{i})$，选取主流网络的原因是主流模型能更好地证明算法的泛化性；高频信息残差提取模块改进了一般高通滤波器的不足，借鉴拉普拉斯残差金字塔的思想更充分地提取浅层特征图的高频信息，输出高频特征图 $\mathrm{HL}_t(\boldsymbol{i})$；高频信息增强模块借鉴了文献[19]中的设计，采用 3 层 ResBlock 对高频特征进行增强，在此基础上将空间与通道注意力机制引入原有的 ResBlock，使其能够有针对性地增强重要特征，输出增强后的高频特征图 $\mathrm{En}\left[\mathrm{HL}_t(\boldsymbol{i})\right]$。同时设计了一种图像梯度损失算法，防止随着网络的加深，高频信息丢失；由于深度伪造模型通常存在训练时间长等问题，本实验将 GC 引入 AdamW 优化器，解决了深度伪造检测模型训练时间长、泛化性差的问题。

2.2 浅层特征提取模块

为了证明所提算法的通用性，本实验选取了深度伪造检测中使用最广泛的模型 XceptionNet 和 EfficientNet-B4，采用其网络的浅层作为特征提取模块。

XceptionNet 在基于图片级学习的深度伪造检测和图像分类任务中应用十分广泛，网络架构如表 1 所

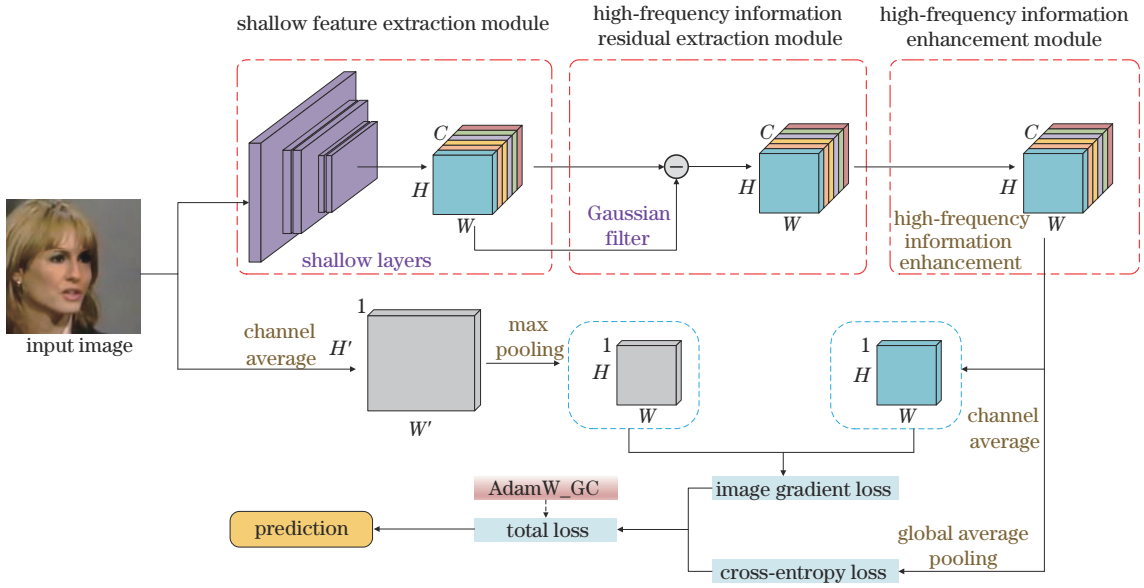

图 1　面向浅层特征高频分量实现深度伪造检测的算法整体架构

Fig. 1　Overall architecture of deepfake detection algorithm for high-frequency components of shallow features

表 1　XceptionNet 架构

Table 1　Architecture of XceptionNet

Layer	Operator		Output, Size	Channels
L_0	Entry flow	Conv1		32
		Conv2		64
L_1		Block1		128
L_2		Block2	$FL_2(\boldsymbol{i})$, (40×40)	256
L_3		Block3	$FL_3(\boldsymbol{i})$, (20×20)	728
L_4	Middle flow	Block4	$FL_4(\boldsymbol{i})$, (20×20)	728
L_5–L_{11}		Block5-11		728
L_{12}		Block12		1024
Final	Exit flow	SeparableConv2d		1536
		SeparableConv2d		2048
Logits		Linear		

示。其由 Incption v3 发展而来，核心思想是认为通道相关性和空间相关性是可分离的。网络主体模块是深度可分离卷积，先进行 Depthwise 卷积即各个通道独立对空间进行卷积，然后进行 Pointwise 卷积即 1×1 卷积。深度可分离卷积结构在不增加模型复杂度的前提下提升了模型的性能。

EfficientNet 是指使用复合系数和 AutoML[22]技术从深度、宽度和输入图像分辨率这 3 个维度对 CNN 进行放缩的一组特征提取网络（EfficientNet-B0～B7），网络架构如表 2 所示。相较于其他网络模型，EfficientNet 的参数量较小、分类准确率高。本实验选取对深度伪造检测表现最好且参数量适中的 EfficientNet-B4。EfficientNet 的网络主体模块是 Mobile Inverted Bottleneck Convolution（MBConv）模块，由深度可分离卷积和 Squeeze-and-Excitation Net（SENet）[23]即通道注意力模块组成。

表 2　EfficientNet-B4 架构

Table 2　Architecture of EfficientNet-B4

Layer	Operator	Numbers	Output, Size	Channels
L_0	Conv 3×3	1		48
L_1	MBConv1	2		24
L_2	MBConv6	4	$FL_2(\boldsymbol{i})$, (80×80)	32
L_3	MBConv6	4	$FL_3(\boldsymbol{i})$, (40×40)	56
L_4	MBConv6	6	$FL_4(\boldsymbol{i})$, (20×20)	112
L_5	MBConv6	6		160
L_6	MBConv6	8		272
L_7	MBConv6	2		448
Logits	Linear	1		

将输入的图片表示为 $\boldsymbol{i}$，选取的浅层特征指 XceptionNet 和 EfficientNet-B4 的 t 层（L_t）卷积层输出的特征图，记为 $FL_t(\boldsymbol{i})$。

2.3　高频信息残差提取模块

以往利用高频信息的检测算法[19]一般直接使用高通滤波器进行高频信息的提取，但是由于浅层特征图通常较小，一般的高通滤波器存在提取高频信息不充分的缺陷。本实验借鉴拉普拉斯残差金字塔的思想，提出一种利用残差结构进行高频信息提取的方案。实验证明，相较于文献[19]中效果最好的内核大小为(3, 3)的拉普拉斯滤波器，所提算法具有更充分的高频信息提取性能。该模块使用的高斯滤波器(GLPF)[24]是一种低通滤波器，传递函数为

$$H(x,y)=\exp\left[-D^2(x,y)/2\sigma^2\right], \tag{1}$$

式中：$D(x,y)$是滤波器中心到滤波器上任意一点(x,y)的距离；$H(x,y)$为各个元素的计算值；σ为中心分离度的测度。定义高频特征图$HL_t(\boldsymbol{i})$为$F_gL_t(\boldsymbol{i})$和$FL_t(\boldsymbol{i})$的差值：

$$HL_t(\boldsymbol{i})=FL_t(\boldsymbol{i})-F_gL_t(\boldsymbol{i}), \tag{2}$$

式中：$F_gL_t(\boldsymbol{i})$是$FL_t(\boldsymbol{i})$经过高斯滤波器得到的低频特征图。由于该模块是为了提取尽可能多的视觉伪影特征，而视觉伪影所占整张图像的比例较小，为了获取尽可能多的包含视觉伪影的高频信息而不引入更多的冗余信息，同时尽量减少区域信息扰动，因此选取核的大小为(3, 3)。σ与图像的平滑程度正相关：如果σ过大，则$F_gL_t(\boldsymbol{i})$会过于平滑，$HL_t(\boldsymbol{i})$在得到特征图纹理特征的同时也会引入特征图的噪声；如果σ过小，则$HL_t(\boldsymbol{i})$有可能无法充分获取特征图的纹理特征。因此，本实验选取内核的标准差为(1.5, 1.5)。高频信息残差提取模块结构如图2所示。

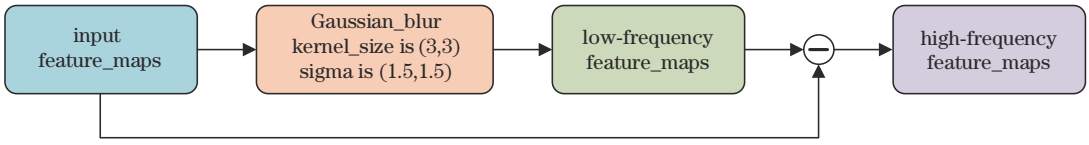

图2　高频信息残差提取模块结构

Fig. 2　Module structure of high-frequency information residual extraction

图3为原始特征图、高频信息残差提取模块所得到的高频特征图、内核大小为(3, 3)的拉普拉斯滤波器所得到的高频特征图的3D傅里叶频谱图。

由图3可以明显看出，相较于一般的高通滤波器，原始特征图经过高频信息残差模块之后所得到的高频特征图在3D傅里叶频谱图中的高频部分(边缘区域)颜色更深，即高频信息残差模块的高频信息提取能力更强。

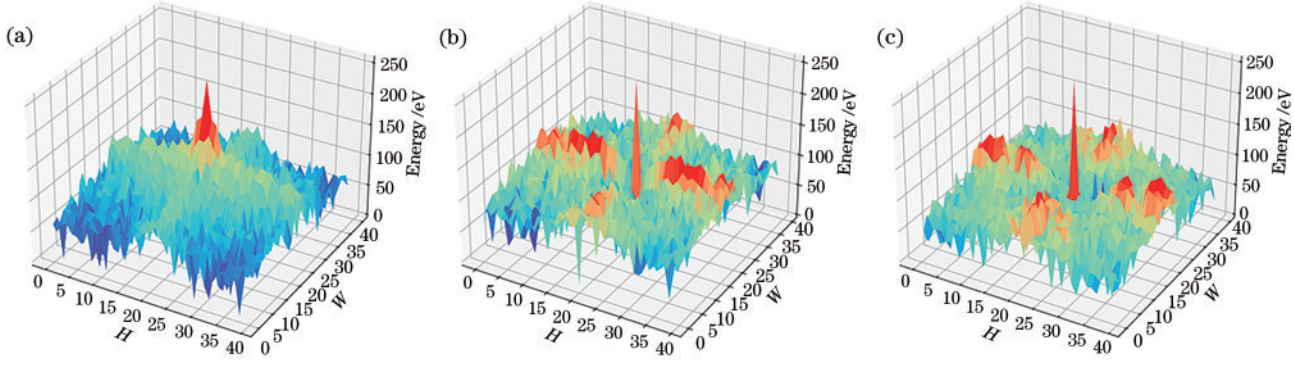

图3　频谱图对比。(a)原始图像；(b)残差提取模块得到的图像；(c)拉普拉斯滤波器得到的图像

Fig. 3　Spectrogram comparison. (a) Original image; (b) image from residual extraction module; (c) image from Laplace filter

2.4　高频信息增强模块

文献[20]中使用ResBlock进行高频信息的增强，然而由于$HL_t(\boldsymbol{i})$不同通道以及同一特征图不同空间所承载信息的重要性不同，在处理时对重要性不同的区域不加以区分是缺乏合理性的。因此，本实验在ResBlock的基础上引入空间与通道注意力机制即CBAM，使增强模块能够更好地抽取并增强重要的特征。CBAM的核心思想是增加特征图关键区域以及关键特征通道的权重，提升特征图的空间以及通道相关性，减少无关信息的干扰，最终提高检测结果的准确性。CBAM的结构如图4所示。

CBAM由通道注意力模块(CAM)和空间注意力模块(SAM)模块构成，属于融合通道与空间注意力机制的混合注意力机制模块。设输入的特征图为$\boldsymbol{F}$，大小为$C\times H\times W$。在CAM中，进行空间维度的压缩：使$\boldsymbol{F}$经过并行的最大池化操作以及平均池化操作变为2个大小为$C\times 1\times 1$的特征向量；然后经过一个参数共享的多层感知机，先将其压缩为大小为$\frac{C}{r}\times 1\times 1$的向量然后进行还原；最后将得到的2个特征向量相加，将结果经过激活函数后与$\boldsymbol{F}$相乘，得到CAM的输出$\boldsymbol{F}'$。计算过程为

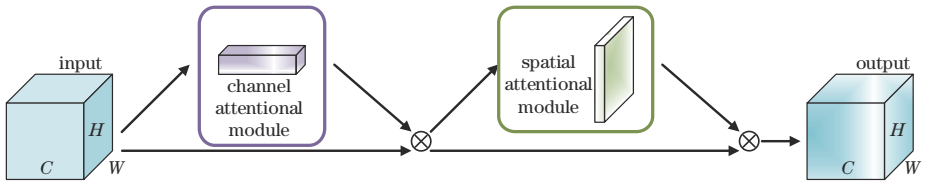

图 4　CBAM 结构

Fig. 4　Structure of CBAM

$$M_c(\boldsymbol{F})=\sigma\left\{\mathrm{MLP}\left[\mathrm{AvgPool}(\boldsymbol{F})\right]+\mathrm{MLP}\left[\mathrm{MaxPool}(\boldsymbol{F})\right]\right\},\tag{3}$$

$$\boldsymbol{F}'=M_c(\boldsymbol{F})\cdot\boldsymbol{F},\tag{4}$$

式中：$\boldsymbol{F}$ 为输入 CAM 的特征图；AvgPool 表示平均池化；MaxPool 表示最大池化；MLP 表示参数共享的多层感知机；σ 是 Sigmoid 激活函数；$M_c(\boldsymbol{F})$ 代表 channel_out；$\boldsymbol{F}'$ 是 CAM 的输出。

在 SAM 模块中，进行通道维度的压缩：将新特征图 $\boldsymbol{F}'$ 经过最大池化与平均池化变为 2 个大小为 $H\times W\times 1$ 的张量；然后将 2 个张量在通道维度上进行拼接成为大小为 $H\times W\times 2$ 的张量，经过卷积操作将张量大小变回 $H\times W\times 1$；最后将结果经过激活函数后与 $\boldsymbol{F}'$ 相乘得到最终的输出 $\boldsymbol{F}_{\mathrm{out}}$。计算过程为

$$M_s(\boldsymbol{F}')=\sigma\left\{\mathrm{Conv}_{7\times7}\left\{\left[\mathrm{AvgPool}(F');\mathrm{MaxPool}(F')\right]\right\}\right\},\tag{5}$$

$$\boldsymbol{F}_{\mathrm{out}}=M_s(\boldsymbol{F}')\cdot\boldsymbol{F}',\tag{6}$$

式中：$\boldsymbol{F}'$ 为输入 SAM 的特征图；$\mathrm{Conv}_{7\times7}$ 为卷积核大小为 7×7 的卷积层；$M_s(\boldsymbol{F}')$ 代表 spatial_out；$\boldsymbol{F}_{\mathrm{out}}$ 是 CABM 的输出。

本实验利用在每个 ResBlock 中加入 CBAM 的方式抽取出重要的高频信息，然后与原始输入通过通道拼接的方式进入下一层 ResBlock。改进后的高频信息增强模块结构如图 5 所示，定义该模块的输出为增强后的高频特征图 $\mathrm{En}\left[\mathrm{HL}_t(\boldsymbol{i})\right]$。

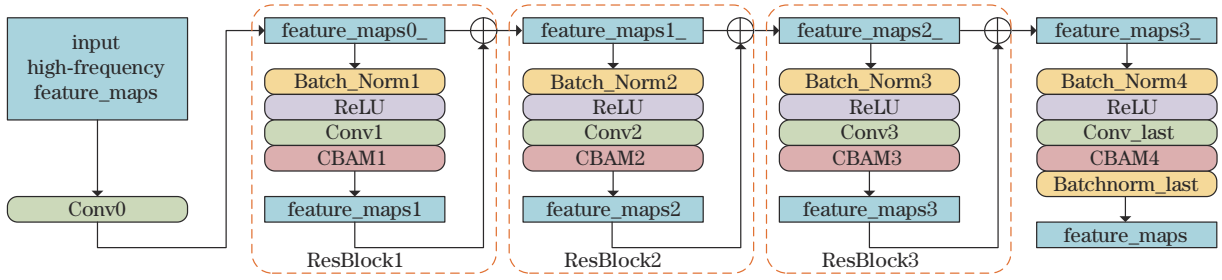

图 5　高频信息增强模块结构

Fig. 5　Structure of high-frequency information enhancement module

DeepFakes 数据集中对应真假样本的高频特征图增强前后效果如图 6 所示。

由图 6 可以看出，$\mathrm{HL}_t(\boldsymbol{i})$ 保留了原始图像的纹理信息 $\boldsymbol{i}$，例如假样本图 6(b)中的视觉伪影也很好地体现在了图 6(d)中。而增强后的高频特征图图 6(e)、(f)则在增强图 6(c)、(d)中高频信息的同时增加了对于重要区域(人脸五官区域)的关注。

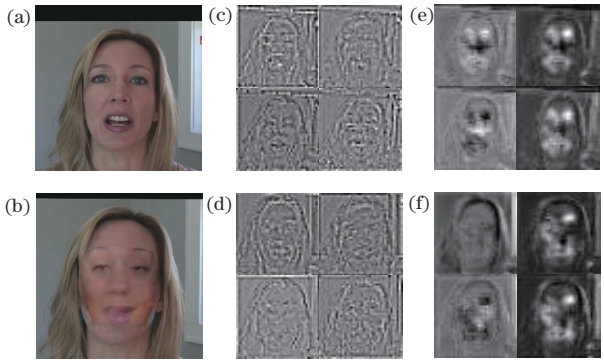

图 6　对应真假样本的高频特征图增强前后效果图。(a) (b)原始图像 $\boldsymbol{i}$；(c) (d)增强前的高频特征图 $\mathrm{HL}_t(\boldsymbol{i})$；(e) (f)增强后的高频特征图 $\mathrm{En}\left[\mathrm{HL}_t(\boldsymbol{i})\right]$

Fig. 6　Effect before and after enhancement of high-frequency feature map corresponding to real and fake samples. (a) (b) Original images $\boldsymbol{i}$; (c) (d) high-frequency feature maps $\mathrm{HL}_t(\boldsymbol{i})$ before enhancement; (e) (f) enhanced high-frequency feature map $\mathrm{En}\left[\mathrm{HL}_t(\boldsymbol{i})\right]$

2.5　图像梯度损失

由于深层网络对于高频信息的学习优先级低，因此，随着网络的加深，原始图像的高频信息会出现丢失。而以往利用高频信息检测的算法都忽视了此现象。针对这一问题，本实验设计了一种图像梯度损失算法，可以起到高频信息增强的作用。对于原始图像 $\boldsymbol{i}$，先在通道维度上进行平均，然后采用自适应最大池化操作尽可能保留其高频分量，并使其长宽与 $\mathrm{En}\left[\mathrm{HL}_t(\boldsymbol{i})\right]$ 相同，记为 $\mathrm{AMP}\left[\mathrm{CAvg}(\boldsymbol{i})\right]$，将 $\mathrm{En}\left[\mathrm{HL}_t(\boldsymbol{i})\right]$ 在通道维度上进行平均记为 $\mathrm{CAvg}\left\{\mathrm{En}\left[\mathrm{HL}_t(\boldsymbol{i})\right]\right\}$。拉普拉斯算子作为应用最广泛的边缘检测算子，本实验将其作为梯度算子，设图像为 $f(x,y)$，则拉普拉斯算子[25]定义为

$$\nabla^2 f=\frac{\partial^2 f}{\partial x^2}+\frac{\partial^2 f}{\partial y^2}, \tag{7}$$

式中：$\frac{\partial^2 f}{\partial x^2}$、$\frac{\partial^2 f}{\partial y^2}$ 为 $f(x,y)$ 沿 x、y 方向的二阶微分。由此可以推导出2个变量的离散拉普拉斯算子为

$$\nabla^2 f(x,y)=f(x+1,y)+f(x-1,y)+f(x,y+1)+f(x,y-1)-4f(x,y)。 \tag{8}$$

由式(8)可以得到该拉普拉斯算子模板，如图7所示。

0	1	0
1	−4	1
0	1	0

图7　拉普拉斯算子模板

Fig. 7　Template for Laplace operator

在本实验中，图像梯度损失表示为

$$L_{\mathrm{Gd}}=\frac{1}{WH}\sum_{x=1}^{W}\sum_{y=1}^{H}\left\{\nabla^2 f\left\{\mathrm{CAvg}\left\{\mathrm{En}\left[\mathrm{HL}_t(\boldsymbol{i})\right]\right\}_{(x,y)}\right\}-\nabla^2 f\left\{\mathrm{AMP}\left[\mathrm{CAvg}(\boldsymbol{i})\right]_{(x,y)}\right\}\right\}^2, \tag{9}$$

式中：∇^2 表示梯度算子。图像梯度损失 L_{Gd} 用 $\nabla^2 f\left\{\mathrm{CAvg}\left\{\mathrm{En}\left[\mathrm{HL}_t(\boldsymbol{i})\right]\right\}_{(x,y)}\right\}$ 和 $\nabla^2 f\left\{\mathrm{CAvg}\left[\mathrm{AMP}(\boldsymbol{i})\right]_{(x,y)}\right\}$ 的均方根误差来定义。其中，浅层特征图 $\mathrm{FL}_t(\boldsymbol{i})$ 到高频特征图 $\mathrm{HL}_t(\boldsymbol{i})$，再到增强后的高频特征图 $\mathrm{En}\left[\mathrm{HL}_t(\boldsymbol{i})\right]$ 的变化如图8所示。

此变化用3D傅里叶频谱图表示，如图9所示。从图9可以直观地看出，相较于原始特征图 $\mathrm{FL}_t(\boldsymbol{i})$，$\mathrm{En}\left[\mathrm{HL}_t(\boldsymbol{i})\right]$ 的高频部分(边缘区域)被明显增强。

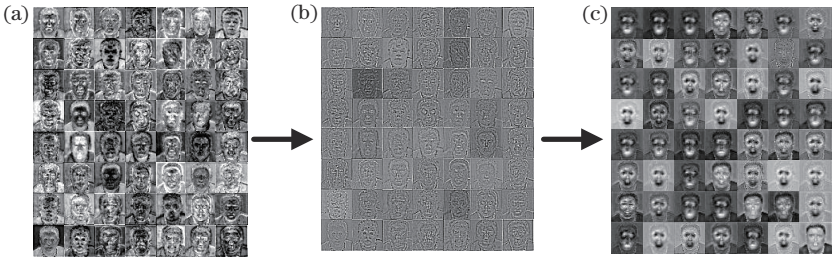

图8　特征图变化。(a)浅层特征图 $\mathrm{FL}_t(\boldsymbol{i})$；(b)高频特征图 $\mathrm{HL}_t(\boldsymbol{i})$；(c)增强后的高频特征图 $\mathrm{En}\left[\mathrm{HL}_t(\boldsymbol{i})\right]$

Fig. 8　Feature map variations. (a) Shallow feature maps $\mathrm{FL}_t(\boldsymbol{i})$; (b) high-frequency feature maps $\mathrm{HL}_t(\boldsymbol{i})$; (c) enhanced high-frequency feature maps $\mathrm{En}\left[\mathrm{HL}_t(\boldsymbol{i})\right]$

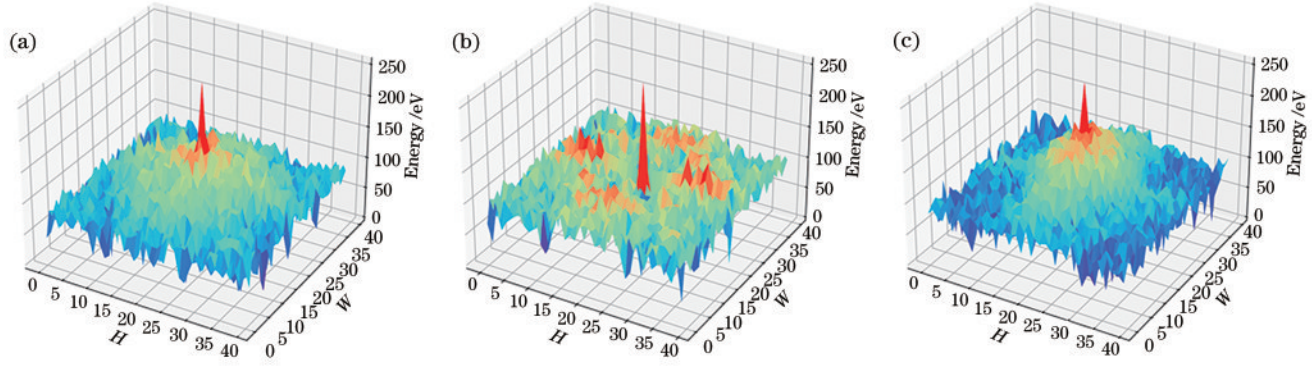

图9　频谱图变化。(a) 浅层特征图 $\mathrm{FL}_t(\boldsymbol{i})$；(b) 高频特征图 $\mathrm{HL}_t(\boldsymbol{i})$ (c) 增强后的高频特征图 $\mathrm{En}\left[\mathrm{HL}_t(\boldsymbol{i})\right]$

Fig. 9　Spectrogram variations. (a) Shallow feature maps $\mathrm{FL}_t(\boldsymbol{i})$; (b) high-frequency feature maps $\mathrm{HL}_t(\boldsymbol{i})$; (c) enhanced high-frequency feature maps $\mathrm{En}\left[\mathrm{HL}_t(\boldsymbol{i})\right]$

3 仿真实验与结果分析

3.1 实验环境

本实验在NVIDIA GeForce RTX 3090上使用版本为1.10.0的PyTorch深度学习框架实现。实验平台为版本号为Ubuntu 16.04.6 LTS的64位Linux操作系统，显卡内存为24 GB。CPU版本为Intel(R) Core (TM) i9-10920X@3.50 GHz，内存为32 GB。Anaconda版本号为2020.11。

3.2 数据集

实验在FaceForensics++(FF++)[2]和Celeb-DF[26]这2个主流数据集上进行验证。FF++是目前规模较大、种类最丰富的数据集，采集自YouTube上1000个含有无遮挡人脸的短视频，并确保连续帧内都含有人脸。根据压缩质量高低，FF++分为无损压缩(Raw)、高质量压缩(c23)和低质量压缩(c50)；根据伪造方式的不同，FF++分为DeepFakes、Face2Face、FaceSwap和Neural-Textures。其中，DeepFakes与FaceSwap属于换脸伪造，Face2Face与Neural-Textures属于换表情伪造。DeepFakes使用自编码器一对一生成模型生成，FaceSwap则基于3D图像的方法生成。本实验中，为了验证所提算法对于检测压缩数据集的有效性，选择c23中的FaceSwap与DeepFakes数据集。

Celeb-DF解决了以往数据集的一些缺陷，如分辨率低、合成质量差、篡改痕迹明显、人脸闪烁等问题，是目前公认的高质量深度伪造数据集。真实视频由YouTube上59个性别、年龄、种族差别各异的名人采访视频组成，共有590个。伪造视频共有5639个，采用改进的DeepFakes生成方法：使用颜色迁移算法减小篡改区域的不一致性；使用更精准的人脸关键点定位方法减轻人脸闪烁问题；使用更平滑的人脸面部覆盖掩膜去除大部分视觉伪影。

对于FF++和Celeb-DF数据集，本实验对每个视频等间隔选取30帧作为实验样本，之后使用RetinaFace通过检测人脸五官关键点来确定人脸面部矩形，并对人脸面部矩形放大至原始图像的1.3倍，调整图片大小为(320, 320)。按照7∶3的比例将其分为训练集和测试集，如表3所示。

表3 所用数据集

Table 3 Datasets used

Dataset	DeepFakes	FaceSwap	Celeb-DF
Train	41936	41945	143192
Test	17089	15600	41618

3.3 实验设置

针对深度伪造模型常常存在的模型参数量大、泛化性差等问题，本实验将GC[27]与AdamW优化器相结合为AdamW_GC。GC是一种对于梯度的预处理，首先获取网络模型反向传播梯度，再将对每个列向量中心化之后的梯度矩阵传递给下游的优化器，使得模型的权重规范化，起到加快训练速度、提高模型泛化能力、增强损失函数的抗扰动能力的作用。用公式可以表示为

$$\Phi_{GC}(\nabla w_i L)=\nabla w_i L-\frac{1}{n}\sum_{j=1}^{n}\nabla w_{i,j}L, \tag{10}$$

式中：$\nabla w_i L$表示梯度；w_i表示权重；i表示梯度矩阵中的第i列；j表示第i列中的第j个元素。超参数设定如下：学习率为1×10^{-4}，每经过5个epoch调整为原来的0.5；weight decay为1×10^{-6}；batch size为16；epoch为20。每次实验均设置相同的随机种子以确保实验结果的稳定性。

本实验将深度伪造检测抽象为二分类真假问题，分类器为Softmax，判断阈值$\theta=0.5$。使用准确率(Acc)、接受者操作特征曲线下的面积(AUC)这2个评价指标。测试时，每个视频中截取部分帧进行测试，每个视频内取预测结果的平均值作为总体的预测结果，最终再求分类准确率。Acc与AUC的计算公式如下：

$$A_{Acc}=\frac{N_{TP}+N_{TN}}{N_{TP}+N_{FP}+N_{TN}+N_{FN}}, \tag{11}$$

$$A_{AUC}=\frac{1}{2}\sum_{i=1}^{n-1}\left[\left(\frac{N_{FP}}{N_{FP}+N_{TN}}\right)^{(i+1)}-\left(\frac{N_{FP}}{N_{FP}+N_{TN}}\right)^{(i)}\right]\times\left[\left(\frac{N_{TP}}{N_{TP}+FN}\right)^{(i+1)}+\left(\frac{N_{TP}}{N_{TP}+N_{FN}}\right)^{(i)}\right], \tag{12}$$

式中：N_{TP}为真正例的数量；N_{TN}为真反例的数量；N_{FP}为假正例的数量；N_{FN}为假反例的数量；n为样例个数。网络模型的整体损失函数为

$$L=(1-\lambda)L_{Log}+\lambda L_{Gd}, \tag{13}$$

式中：L_{Log}表示分类损失；L_{Gd}表示图像梯度损失；λ是用来平衡2个损失的权重参数。

3.4 仿真实验与结果分析

本实验在DeepFakes、FaceSwap和Celeb-DF这3个数据集上进行训练和测试，使用Acc、AUC这2个评价指标。其中，记以EfficientNet-B4为特征提取模块的模型为En_model，以XceptionNet为特征提取模块的模型为X_model。

实验1：比较将EfficientNet-B4和XceptionNet模型的L_2、L_3和L_4作为浅层特征提取模块的优劣，本实验中设置$\lambda=0.5$，如图10、11所示。

由图10中的数据可知：特征提取模块选取EfficientNet-B4第3层时在3个数据集上的平均Acc为99.04%，与第2层和第4层相比分别高了0.89个百分点和0.30个百分点，虽然在FaceSwap数据集上第4层的精度比第3层高，但这可能与数据集本身的差异性有关；平均AUC指数为0.9784，与第2层和第4层相比分别高了0.0155和0.0094。由图11中的数据可

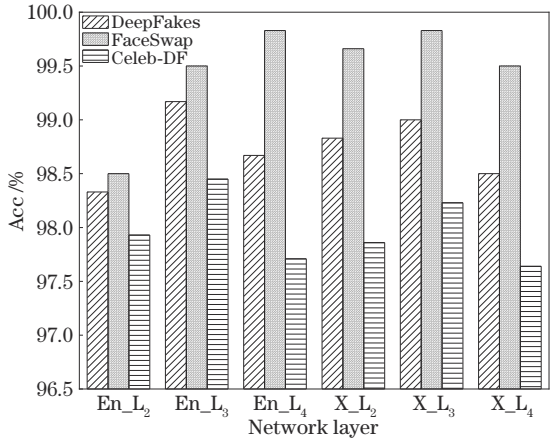

图 10　3种数据集下模型 Acc 指数随网络层数变化

Fig. 10　Variation of model Acc index with the number of network layers on 3 datasets

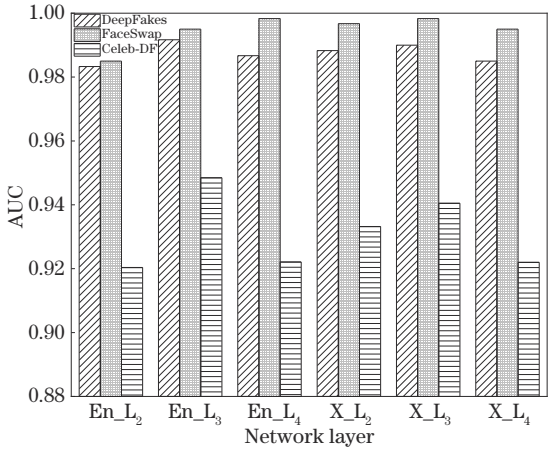

图 11　3种数据集下模型 AUC 指数随网络层数变化

Fig. 11　Variation of model AUC index with the number of network layers on 3 datasets

知：选取 XceptionNet 第 3 层时在 3 个数据集上的平均 Acc 为 99.13%，与第 2 层和第 4 层相比分别高了 0.35 个百分点和 0.58 个百分点；平均 AUC 指数为 0.9774，与第 2 层和第 4 层相比分别高了 0.0047 和 0.01。综上所述，选取 EfficientNet-B4 和 XceptionNet 的第 3 层作为本实验模型的特征提取模块最为合理。因此，浅层特征提取模块在 2 个主干网络上选取的 L_f 为 L_3。

实验 2：比较权重参数 λ 取不同参数对于模型性能的影响。为了在图表中直观地体现权重参数，Acc、AUC 分别取模型在 3 个数据集上测试结果的平均 Acc、AUC 进行比较，其结果如图 12、13 所示。

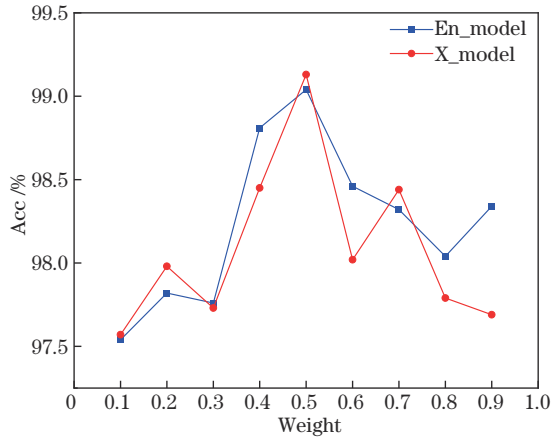

图 12　平均 Acc 指数随 λ 的变化折线图

Fig. 12　Line graph of change in average Acc index with λ

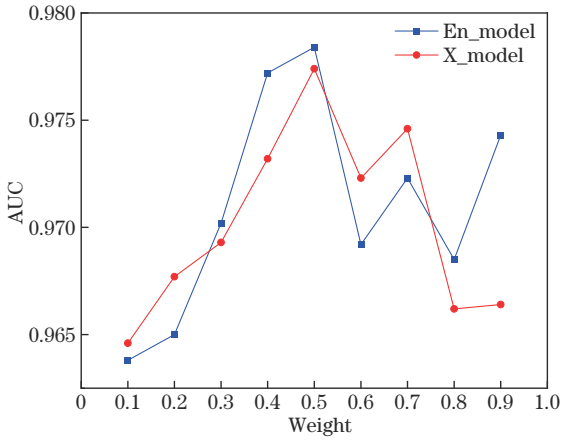

图 13　平均 AUC 指数随 λ 的变化折线图

Fig. 13　Line graph of change in average AUC index with λ

由图 12、13 可以看出：当 $\lambda=0.5$ 时，模型的 Acc 和 AUC 取值最高。所以，权重参数最终取值为 0.5，此时网络模型的整体损失函数为

$$L=(L_{\mathrm{Log}}+L_{\mathrm{Gd}})/2。\tag{14}$$

实验 3：为了研究所提改进方法对算法产生性能增益的大小，在原始的基线网络上，分别添加 CBAM 以及梯度损失(GD)，然后计算所提算法在 3 个数据集上的 Acc 值与 AUC 值，比较结果如表 4、5 所示。

由表 4 可知：在以 XceptionNet 为基础的 Baseline 上，引入 CBAM 之后，模型的平均 Acc 提升了 0.36 个百分点，平均 AUC 提升了 0.0118；引入图像梯度损失后，模

表 4　比较不同模块在以 XceptionNet 为基础的 Baseline 上产生的增益

Table 4　Comparison of gains produced by different modules on XceptionNet-based Baseline

X_model			DeepFakes		FaceSwap		Celeb-DF	
Baseline	CBAM	GD	Acc /%	AUC	Acc /%	AUC	Acc /%	AUC
√			98.67	0.9867	99.33	0.9933	97.49	0.9037
√	√		98.83	0.9883	99.67	0.9967	98.07	0.9342
√		√	98.83	0.9883	99.50	0.9950	98.01	0.9384
√	√	√	99.33	0.9933	99.83	0.9983	98.23	0.9405

表 5　比较不同模块在以 EfficientNet-B4 为基础的 Baseline 上产生的增益

Table 5　Comparison of gains produced by different modules on EfficientNet-B4-based Baseline

En_model			DeepFakes		FaceSwap		Celeb-DF	
Baseline	Cbam	GD	Acc /%	AUC	Acc /%	AUC	Acc /%	AUC
√			98.50	0.9850	99.00	0.9900	97.56	0.9221
√	√		98.83	0.9883	99.50	0.9950	97.86	0.9263
√		√	98.67	0.9867	99.17	0.9917	97.63	0.9220
√	√	√	99.17	0.9917	99.50	0.9950	98.45	0.9485

型的平均 Acc 提升了 0.28 个百分点，平均 AUC 提升了 0.0127；同时引入 CBAM 和图像梯度损失后，模型的平均 Acc 提升了 0.63 个百分点，平均 AUC 提升了 0.0161。

由表 5 可知：在以 EfficientNet-B4 为基础的 Baseline 上，引入 CBAM 之后，模型的平均 Acc 提升了 0.38 个百分点，平均 AUC 提升了 0.0042；引入图像梯度损失后，模型的平均 Acc 提升了 0.14 个百分点，平均 AUC 提升了 0.0011；同时引入 CBAM 和图像梯度损失后，模型的平均 Acc 提升了 0.69 个百分点，平均 AUC 提升了 0.0127。这证明了 CBAM 以及图像梯度损失对于提升模型性能的有效性。

由图 8(b)、(c)的对比可以看出，CBAM 是通过在通道以及空间维度上关注重要的特征、抑制不必要的特征的方式提升模型性能的。而图像梯度损失可以有效防止由于模型深度的加深，特征图中高频信息丢失的问题，如图 8(c)很好地保留并增强了图 8(b)中的纹理特征，而深度伪造图像的视觉伪影常常存在于这些高频信息中，最终能够提升模型的性能。同时，以 XceptionNet 和 EfficientNet-B4 为主干网络进行实验都取得了理想的结果，证明所提算法对主流检测模型而言具有一定的泛化性。

实验 4：验证引入 GC 后对于模型训练速度的提升。本次实验中的 Acc、损失值取 X_model、En_model、X_model_GC、En_model_GC 在 3 个数据集上的测试结果数值的平均，其结果如图 14、15 所示。

从图 14、15 可以看出：在引入 GC 之后，前几个

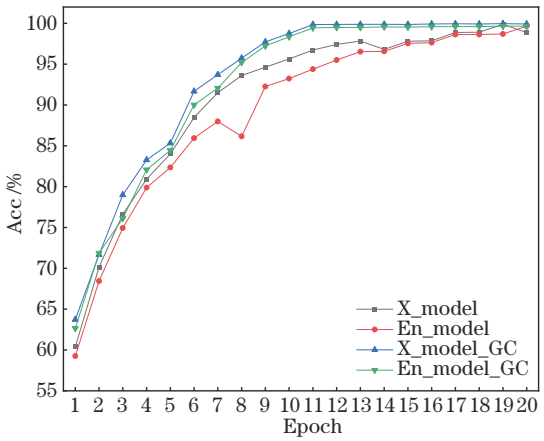

图 14　引入 GC 后在 3 个数据集上的平均 Acc 对比图

Fig. 14　Comparison of average Acc on three datasets after introduction of GC

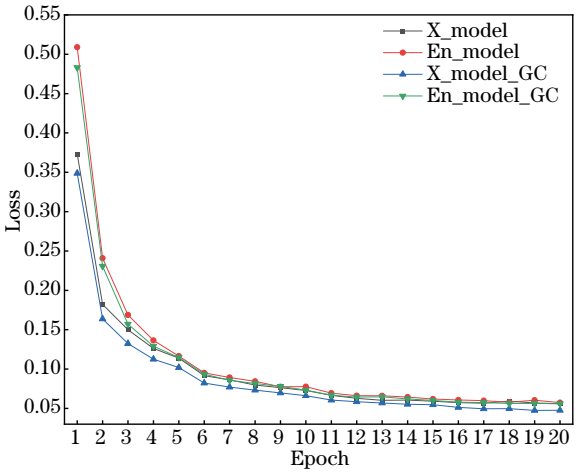

图 15　引入 GC 后在 3 个数据集上的平均损失对比图

Fig. 15　Comparison of average loss on three datasets after introduction of GC

Epoch 的损失值明显降低，加快了模型的收敛速度；在最后的几个 epoch 中，Acc 曲线趋于稳定，提高了模型的泛化能力。这证明了 GC 可以通过加快模型收敛速度的方式提升模型的训练速度。

实验 5：将所提算法与主流模型进行对比，实验结果如表 6 所示。

表 6　各个算法在 3 个数据集上的 Acc 指数比较

Table 6　Comparison of Acc index of each algorithm on three datasets

Acc /%	DeepFakes	FaceSwap	Celeb-DF
EfficientNet-B4[11]	98.33	98.83	97.42
XceptionNet[10]	97.83	98.17	96.97
MesoNet[14]	95.50	93.33	91.72
Mo et al[18]	96.67	97.00	95.79
Sabir et al[8]	96.50	96.33	95.42
ResNet34[9]	93.83	94.17	93.50
En_model	**99.17**	**99.50**	**98.45**
X_model	**99.33**	**99.83**	**98.23**

由表 6 可以看出，面向浅层特征高频分量的深度伪造检测算法在准确率方面具有一定的优势。

在仿真实验 1～5 中，选取了 2 个主流的深度伪造检测模型为主干网络。通过对照实验，比较了 2 个主流检测网络的不同分层作为特征提取网络的优劣，验

证了引入GC后对于模型训练速度、泛化性的提升，确定了损失函数中权重参数λ的选取。在此基础上，通过消融实验验证了CBAM以及图像梯度损失对于提升模型准确率的有效性，并与主流算法比较证明了该算法的可行性以及准确率优势。

4 结 论

所提算法面向深度伪造图像视觉伪影常常存在于浅层特征高频分量中这一特性，解决了高频信息提取不充分、无法自适应地增强特征图关键区域以及关键通道、网络对于高频分量学习优先级低以及深度伪造模型存在的训练时间长、泛化性差等问题。与主流检测算法相比，模型的准确率优于其他主流检测模型且网络层数大大减小，并且该检测方法对于目前主流模型以及数据集具有一定的泛化性。未来的工作：1)将浅层特征的高频分量与图像的高层语义信息进行特征融合，在此基础之上对融合后的特征进行检测；2)由于基于空域的算法对于数据集有很大的依赖性，因此需要结合图像的频域信息提升算法的跨库测试能力。

参 考 文 献

[1] Tran L, Yin X, Liu X M. Representation learning by rotating your faces[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 41(12): 3007-3021.

[2] Rössler A, Cozzolino D, Verdoliva L, et al. FaceForensics: learning to detect manipulated facial images [C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV), October 27-November 2, 2019, Seoul, Korea (South). New York: IEEE Press, 2019.

[3] Shu K, Sliva A, Wang S H, et al. Fake news detection on social media: a data mining perspective[J]. ACM SIGKDD Explorations Newsletter, 2017, 19(1): 22-36.

[4] Tolosana R, Vera-Rodriguez R, Fierrez J, et al. Deepfakes and beyond: a survey of face manipulation and fake detection[J]. Information Fusion, 2020, 64: 131-148.

[5] Amerini I, Galteri L, Caldelli R, et al. Deepfake video detection through optical flow based CNN[C]//2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), October 27-28, 2019, Seoul, Korea (South). New York: IEEE Press, 2019: 1205-1207.

[6] Agarwal S, Farid H, Gu Y, et al. Protecting world leaders against deep fakes[C]//CVPR workshops, June 16-20, 2019, Long Beach, CA, USA. New York: IEEE Press, 2019: 38-45.

[7] Güera D, Delp E J. Deepfake video detection using recurrent neural networks[C]//2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance, November 27-30, 2018, Auckland, New Zealand. New York: IEEE Press, 2018.

[8] Sabir E, Cheng J X, Jaiswal A, et al. Recurrent convolutional strategies for face manipulation detection in videos[EB/OL]. (2019-05-02) [2021-04-05]. https://arxiv.org/abs/1905.00582.

[9] He K M, Zhang X Y, Ren S Q, et al. Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition, June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE Press, 2016: 770-778.

[10] Chollet F. Xception: deep learning with depthwise separable convolutions[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition, July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 1800-1807.

[11] Tan M, Le Q. Efficientnet: rethinking model scaling for convolutional neural networks[C]//International Conference on Machine Learning, June 9-15, 2019, Long Beach, California, USA. New York: ACM Press, 2019: 6105-6114.

[12] Zhou P, Han X T, Morariu V I, et al. Learning rich features for image manipulation detection[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-22, 2018, Salt Lake City, UT, USA. New York: IEEE Press, 2018: 1053-1061.

[13] Nguyen H H, Yamagishi J, Echizen I. Capsule-forensics: using capsule networks to detect forged images and videos[C]//ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing, May 12-17, 2019, Brighton, UK. New York: IEEE Press, 2019: 2307-2311.

[14] Afchar D, Nozick V, Yamagishi J, et al. MesoNet: a compact facial video forgery detection network[C]//2018 IEEE International Workshop on Information Forensics and Security, December 11-13, 2018, Hong Kong, China. New York: IEEE Press, 2018.

[15] 耿鹏志, 唐云祁, 樊红兴, 等. 基于CutMix算法和改进Xception网络的深度伪造检测研究[J]. 激光与光电子学进展, 2022, 59(16): 1615007.
Geng P Z, Tang Y Q, Fan H X, et al. Research on deep forgery detection based on CutMix algorithm and improved xception network[J]. Laser & Optoelectronics Progress, 2022, 59(16): 1615007.

[16] Rahaman N, Baratin A, Arpit D, et al. On the spectral bias of neural networks[C]//International Conference on Machine Learning, June 9-15, 2019, Long Beach, California, USA. New York: ACM Press, 2019: 5301-5310.

[17] 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, December 9-12, 2019, Delft, Netherlands. New York: IEEE Press, 2019.

[18] Zhao H Q, Wei T Y, Zhou W B, et al. Multi-attentional deepfake detection[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 19-25, 2021, Nashville, TN, USA. New York: IEEE Press, 2021: 2185-2194.

[19] Mo H X, Chen B L, Luo W Q. Fake faces identification via convolutional neural network[C]//IH&MMSec '18: Proceedings of the 6th ACM Workshop on Information

Hiding and Multimedia Security, June 20-22, 2018, Innsbruck, Austria. New York: ACM Press, 2018: 43-47.

[20] Masi I, Killekar A, Mascarenhas R M, et al. Two-branch recurrent network for isolating deepfakes in videos [M]// Vedaldi A, Bischof H, Brox T, et al. Computer vision-ECCV 2020. Lecture notes in computer science. Cham: Springer, 2020, 12352: 667-684.

[21] Woo S, Park J, Lee J Y, et al. CBAM: convolutional block attention module[M]// Ferrari V, Hebert M, Sminchisescu C, et al. Computer vision-ECCV 2018. Lecture notes in computer science. Cham: Springer, 2018, 11211: 3-19.

[22] He Y H, Lin J, Liu Z J, et al. AMC: AutoML for model compression and acceleration on mobile devices[M]// Ferrari V, Hebert M, Sminchisescu C, et al. Computer vision-ECCV 2018. Lecture notes in computer science. Cham: Springer, 2018, 11211: 815-832.

[23] Hu J, Shen L, Sun G. Squeeze-and-excitation networks [C]// Proceedings of the IEEE conference on computer vision and pattern recognition, June 18-22, 2018, Salt Lake City, UT, USA. New York: IEEE Press, 2018: 7132-7141.

[24] Young I T, van Vliet L J. Recursive implementation of the Gaussian filter[J]. Signal Processing, 1995, 44(2): 139-151.

[25] Belkin M, Sun J, Wang Y S. Discrete Laplace operator on meshed surfaces[C]// SCG '08: Proceedings of the twenty-fourth annual symposium on Computational geometry, June 9-11, 2008, College Park, MD, USA. New York: ACM Press, 2008: 278-287.

[26] Li Y Z, Yang X, Sun P, et al. Celeb-DF: a large-scale challenging dataset for DeepFake forensics[C]// 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 13-19, 2020, Seattle, WA, USA. New York: IEEE Press, 2020: 3204-3213.

[27] Yong H W, Huang J Q, Hua X S, et al. Gradient centralization: a new optimization technique for deep neural networks[M]// Vedaldi A, Bischof H, Brox T, et al. Computer vision-ECCV 2020. Lecture notes in computer science. Cham: Springer, 2020, 12346: 635-652.