•网络空间安全•

DOI:10.15961/j.jsuese.201900902

本刊网刊

基于时空特征一致性的Deepfake视频检测模型

赵　磊[1,2], 葛万峰[1,2], 毛钰竹[2], 韩　萌[2], 李文欣[2], 李　学[2]

(1.武汉大学 空天信息安全与可信计算教育部重点实验室, 湖北 武汉 430000; 2.武汉大学 国家网络安全学院, 湖北 武汉 430000)

摘　要:针对目前大部分研究仅关注Deepfake单幅图像的空间域特征而设计检测模型的问题, 以Deepfake视频中人物面部表情变化存在细微的不一致、不连续等现象为出发点, 提出一种基于时空特征一致性的检测模型。该模型使用卷积神经网络对待检测图像提取空域特征, 利用光流法在待检测图像的连续帧间进行时域特征的捕获, 同时利用卷积神经网络对时域特征进行深层次特征提取, 在时域特征和空域特征经过多重的特征变换后, 使用全连接神经网络对空域特征和时域特征的组合空间进行分类实现检测目标。将本文提出的模型在Faceforensics++开源Deepfake数据集上开展模型的训练, 并对模型的检测效果进行实验验证。实验结果表明, 本文模型的检测准确率可达98.1%, AUC值可达0.998 1。通过与现有的Deepfake检测模型进行对比, 本文模型在检测准确率和AUC取值方面均优于现有模型, 验证了本文模型的有效性。

关键词:虚假图像; Deepfake检测; 时域特征; 空域特征

中图分类号:TP309　　**文献标志码**:A　　**文章编号**:2096-3246(2020)04-0243-08

Deepfake Video Detection Model Based on Consistency of Spatial–Temporal Features

ZHAO Lei[1,2], *GE Wanfeng*[1,2], *MAO Yuzhu*[2], *HAN Meng*[2], *LI Wenxin*[2], *LI Xue*[2]

(1.Key Lab. of Aerospace Info. Security and Trusted Computing, Ministry of Education, Wuhan Univ., Wuhan 430000, China;
2.School of Cyber Sci. and Eng., Wuhan Univ., Wuhan 430000, China)

Abstract: In order to improve the feature utilization rate of the image to be detected, a Deepfake video detection model based on consistency of spatial-temporal features was proposed, inspired by the observation that there are slight inconsistency and discontinuity in the facial expression changes of the characters in Deepfake videos. In the model, the convolutional neural network (CNN) was used to extract the spatial features from the video to be detected, and an optical flow method was used to perform temporal features between consecutive frames of the video. Then another CNN was used to extract the abstract and in-depth features from the optical flow map. After the temporal features and spatial features were transformed from original representation space to a new feature space by neural networks, a fully connected neural network was used to classify the combined spatial and temporal features space to achieve the detection target. The model proposed in the paper was trained on the Faceforensics++, an open source Deepfake dataset. The experimental results indicated that the detection accuracy of the proposed model reaches 98.1%, and the AUC value reaches 0.998 1. By comparing with the existing Deepfake detection models, the proposed model is superior to the existing models in terms of detection accuracy and AUC value, which verifies the effectiveness of the proposed model.

Key words: fake images; Deepfake detection; temporal features; spatial features

虚假图片常常伴随着恶意的虚假讯息而大量传播, 一直是危害网络空间信息安全的一个重要因素。近年来, 随着计算机技术尤其是图像处理技术的发展, 虚假图片不断充斥着网络空间, 严重危害着网络空间的舆论形势。已有研究表明, 人工智能已经被应用于黑产/灰产, 攻击者利用先进的人工智能技术,

收稿日期:2019 – 09 – 17
基金项目:国家自然科学基金项目(61672394); 中央高校基本科研业务费专项资金(2042019kf0017)
作者简介:赵　磊(1985—), 男, 副教授, 博士. 研究方向: 信息系统安全. E-mail: leizhao@whu.edu.cn
网络出版时间:2020 – 07 – 11 13 : 44 : 56　　**网络出版地址**:https://kns.cnki.net/kcms/detail/51.1773.TB.20200709.2127.004.html

生成更逼真的虚假图片或虚假视频，混淆视听，给网络安全带来新的挑战。更重要的是，相比于早先的虚假图像处理依赖于很强的专业技术，需要人力物力支撑等，AI技术的飞速发展使得“AI赋能”的虚假图像生成技术门槛大大降低，利用AI在计算机视觉方面的巨大突破，AI换脸成为现实，其中最著名的就是Deepfake技术。

基于深度学习的Deepfake换脸技术所制作的视频最早出现在国外社区论坛Reddit，从此掀开了使用Deepfake技术的浪潮。自2018年以来，有多种开源实现的Deepfake软件或者代码在网络上公布，例如，FaceSwap[1]、Deepfacelab[2]、FakeApp[3]等，这些开源的软件由于简易的操作性使得Deepfake技术被大范围滥用，不仅严重地侵害了换脸对象的肖像权[4]，而且有大部分人员使用该项技术制作淫秽色情图像，挑战法律的底线[5]；更有甚者利用Deepfake技术制作与政治相关人物的图像视频，严重危害了社会的稳定和国家的安全。针对Deepfake视频的检测是近期网络空间安全和计算机视觉的热点和难点问题。

由于Deepfake独特的生成机理，传统虚假图像检测技术在Deepfake检测上面临着诸多挑战。目前，Deepfake图像的检测依据其检测手段可分为两类，基于位置特征的检测和基于内容特征的检测。基于内容特征的Deepfake检测是当前主要的技术手段。Güera等[6]以人脸合成时造成的视频抖动问题入手，采用卷积神经网络（convolutional neural networks，CNN）模型和长短期记忆网络（long short-term memory，LSTM）模型相结合的方法进行特征提取和分类，该方法以完整视频帧作为特征提取对象，存在粒度过大的问题，同时也进一步导致了模型计算量过大。Li等[7]提出在Deepfake制作的视频中人物的眨眼频率与真实的频率存在差异，因此设计了一种名为LRCN（long-term recurrent convolutional neural networks）的深度学习模型用于检测视频中人物眨眼频率，该模型使用CNN对视频中人物眼部状态进行特征提取，结合LSTM对特征进行判别分类，但其仅仅使用眼部状态特征信息进行检测，在一定程度上降低了模型的检测精度。Li等[8]使用几种经典的巨大深层神经网络，希望能对Deepfake视频提取出更加细致的特征并用于分类，但该方法会随着Deepfake技术的迭代训练而降低特征提取的效果。Yang等[9]以Deepfake人物头部姿势为出发点，以人脸各器官部位的位置信息为特征点，设计一种基于支持向量机（support vector machine，SVM）算法作为分类器的模型，但该模型存在特征提取不充分的缺陷，严重影响了分类器的检测效果。Afchar等[10]提出一种基于Inception模块构成的轻量型卷积神经网络，Nguyen等[11]设计一种将VGG（visual geometry group）网络和Capsule Network相结合的网络模型用以检测换脸图像，但是这些模型都易受到Deepfake技术网络迭代训练的影响而降低模型的检测效果。

针对目前模型大都是以单张视频帧为待检测对象从中提取内容特征并以此建立检测模型所存在特征利用率不足的问题，本文提出一种基于细微表情时空特征一致性的Deepfake视频检测模型。在特征构建方面，帧与帧之间的特征变化可以反映更加细粒度的关联信息，因此，本文以视频中连续的帧图像作为输入数据对象，通过计算两帧之间的光流[12]表征人脸表情的变化，得到所提取的时域特征；利用卷积神经网络可以提取更深层次特征的巨大优势，提取空域特征，并对时域特征进行再提取，同时，结合原有图像的空间信息特征，使用深度学习模型对Deepfake视频进行检测，以此提高待检测图像的特征利用率，从而提升模型的检测性能。为了验证该模型的有效性，在FaceForensics++数据集[13]上进行实验，实验结果表明，本文模型在检测准确率和AUC取值方面有着很好的表现效果。

1 Deepfake生成机理

Deepfake技术的核心原理是基于AI实现的一种编解码模型，整个技术核心的实现分为两个阶段。第1个阶段为训练阶段，该阶段主要训练两个神经网络，每个神经网络都由编码器网络（encoder）和解码器网络（decoder）构成，对于两个不同的输入人脸图片A、B，神经网络先通过编码器将人脸数据编码压缩成一个低维向量；之后，通过解码器对低维向量进行解码，并得到解码后的图片，通过最小化解码后的图片与输入图片的差异对网络进行优化训练。在训练阶段，A与B的编码器网络权重保持一致，目的在于使编码器网络对脸部特征的编码具有一致性。该阶段的主要过程如式（1）所示：

$$\begin{aligned}\min L_{\mathrm{A}} &= \frac{1}{N}\sum_{i=1}\|F_i - D_{\mathrm{A}}(E(F_i;\boldsymbol{\theta});\boldsymbol{\phi}_{\mathrm{A}})\|^2,\\ \min L_{\mathrm{B}} &= \frac{1}{N}\sum_{i=1}\|F_i - D_{\mathrm{B}}(E(F_i;\boldsymbol{\theta});\boldsymbol{\phi}_{\mathrm{B}})\|^2\end{aligned} \tag{1}$$

式中：L_{A}、L_{B}分别为编解码网络A、B的损失值；N为网络输入的数据量大小；F_i为网络的输入人脸图像；E为编码网络；$\boldsymbol{\theta}$为编码网络的权值，其中网络A、B的编码器网络E与权重$\boldsymbol{\theta}$一致；D_{A}、D_{B}分别为网络A、B的解码器网络；$\boldsymbol{\phi}_{\mathrm{A}}$、$\boldsymbol{\phi}_{\mathrm{B}}$分别为其网络权重。

第2个阶段是生成阶段，即换脸阶段。在该阶段，

使用已经训练完毕的解码器B（decoder B）去解码经由编码网络编码后的人脸A的低维特征向量，便可得到A的换脸图像，同样地，使用训练完毕的A的解码器解码由编码网络编码后的人脸B的低维特征向量，可得到B的换脸图像。该阶段的主要过程如式（2）所示：

$$F'_{\mathrm{A}} = D_{\mathrm{B}}(E(F_{\mathrm{A}};\boldsymbol{\theta});\boldsymbol{\phi}_{\mathrm{B}}),$$
$$F'_{\mathrm{B}} = D_{\mathrm{A}}(E(F_{\mathrm{B}};\boldsymbol{\theta});\boldsymbol{\phi}_{\mathrm{A}}) \quad (2)$$

式中，F_{A}、F_{B}为原始的人脸图像A、B，F'_{A}与F'_{B}分别为仅由编解码网络替换后的虚假人脸图像，编码器网络权值$\boldsymbol{\theta}$与解码器D_{A}、D_{B}的网络权值$\boldsymbol{\phi}_{\mathrm{A}}$、$\boldsymbol{\phi}_{\mathrm{B}}$为固定值。

2　检测模型设计与实现

本文检测模型的总体架构与流程如图1所示，检测模型的架构主要由数据预处理模块、特征提取模块和深度学习模型模块3个主要部分构成。首先，在数据预处理模块，将视频数据集处理为帧图片，由于视频帧中只有人物脸部是本文的检测目标，所以对抽取的视频帧进行人脸截取。然后，在得到视频中连续的人脸图片之后，利用光流法计算连续变化的两张人脸图片之间的光流，捕获人脸表情随时间变化的特征来表征时域特征。最后，利用CNN对人脸图像提取空间上的特征表征空域特征，同时，借由CNN强大的特征自提取能力对光流图提取更加深层次的特征以更加精确地表征时域特征。深度学习模型分为训练和验证两个阶段，训练集和验证集为不同数据集。

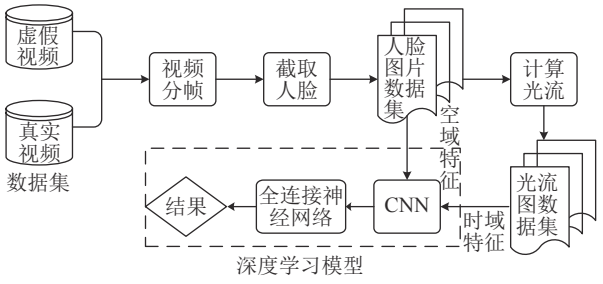

图 1　检测模型的总体架构与流程

Fig. 1　Architecture and pipeline of the detection model

2.1　数据预处理

数据预处理阶段主要包含3个步骤，分别是从视频中抽取图像帧、对图像帧进行脸部截取、图像的顺序存储。

本文使用FFmpeg[14]对原始的视频数据集进行视频帧提取操作，FFmpeg是当前领先的多媒体框架，可以用于编码、解码、转换数字音频和流式传输的开源计算机程序。本文使用Python语言调用FFmpeg程序进行视频帧提取操作。

在帧图像人脸截取部分，本文主要使用基于CNN人脸识别模型的Dlib[15]工具实现，Dlib是一个包含多种机器学习算法的工具包，其在人脸检测方面有HOG–SVM人脸检测和基于CNN的人脸检测等多个检测算法，在实验中发现，Dlib的基于CNN的人脸检测效果更好。通过人脸检测得到的脸部位置信息，使用OpenCV对图像裁剪，并保存为像素256×256的3通道png格式图片，在保存人脸图像时，按照视频帧顺序编号、顺序存储。

2.2　空域特征和时域特征提取

本文主要关注Deepfake的两个特征，分别为空域特征和时域特征。对于这两部分的特征分别采用不同的方法进行提取。

对于空域特征，卷积神经网络（CNN）是目前图像处理领域非常强大的一种自提取特征的网络，具有良好的特征表达能力。本文在模型设计中采用CNN对人脸图像进行特征提取，利用CNN强大的特征自提取的优势来表征人脸图像的空域特征。

对于时域特征，主要表征人物脸部表情随着时间的变化差异，例如，张口、微笑、眨眼等细微动作，由于Deepfake视频按帧进行制作，所以这些面部表情的细微变化存在着不连贯、不一致的现象，因此提取时域特征极为重要。本文采用光流法，通过计算连续的两帧脸部图像的光流场来表征时域特征。

光流场是一个2维矢量场，反映了被处理图像上每一点灰度的变化趋势。计算光流场有两个必要的前提条件[12]：1）亮度恒定，2）时间连续或者运动是“小运动”。实验场景完全符合这两项前提条件，这也是本文选择光流法提取时域特征的重要原因之一。计算光流场的基本原理如下所述。定义函数$I(x,y,t)$为图像中位置(x,y)处在t时刻的亮度，则在Δt时间上，(x,y)移动了$(\Delta x,\Delta y)$的距离，此时根据假设条件1）亮度不变，则有：

$$I(x,y,t) = I(x+\Delta x, y+\Delta y, t+\Delta t) \quad (3)$$

对式（3）进行1阶泰勒展开有：

$$I(x+\Delta x, y+\Delta y, t+\Delta t) = I(x,y,t) + \frac{\partial I}{\partial x}\mathrm{d}x + \frac{\partial I}{\partial y}\mathrm{d}y + \frac{\partial I}{\partial t}\mathrm{d}t + \xi \quad (4)$$

式中，ξ为泰勒展开式中2阶无穷小，可忽略不计。将式（3）代入式（4），有：

$$\frac{\partial I}{\partial x}\mathrm{d}x + \frac{\partial I}{\partial y}\mathrm{d}y + \frac{\partial I}{\partial t}\mathrm{d}t = 0 \quad (5)$$

对式(5)两边同时除以dt,则有:

$$\frac{\partial I}{\partial x}\frac{\mathrm{d}x}{\mathrm{d}t}+\frac{\partial I}{\partial y}\frac{\mathrm{d}y}{\mathrm{d}t}+\frac{\partial I}{\partial t}=0 \tag{6}$$

不难看出,在式(6)中$\frac{\mathrm{d}x}{\mathrm{d}t}$和$\frac{\mathrm{d}y}{\mathrm{d}t}$分别表示所追踪的像素点在$x$轴方向和$y$轴方向的运动矢量,令$\boldsymbol{u}=\frac{\mathrm{d}x}{\mathrm{d}t}$、$\boldsymbol{v}=\frac{\mathrm{d}y}{\mathrm{d}t}$,则$(\boldsymbol{u},\boldsymbol{v})$即为所求光流。从光流的计算过程可以看出,其很好地描述了所计算对象的变化趋势和运动幅度,这一特点非常适合用于本文中的时域特征提取。计算稠密光流有多种不同的方案,本文采用文献[16]方案,并使用OpenCV[12]计算稠密光流,用于提取时域特征。

2.3 检测模型

在经过数据预处理和时空特征提取环节得到特征之后,采用深度学习模型,对提取到的特征进行进一步的学习,具体的操作是使用CNN对原始图像提取空域特征,同时,对计算得到的光流图利用CNN以提取更深层次的时域特征。模型的细节如图2所示。

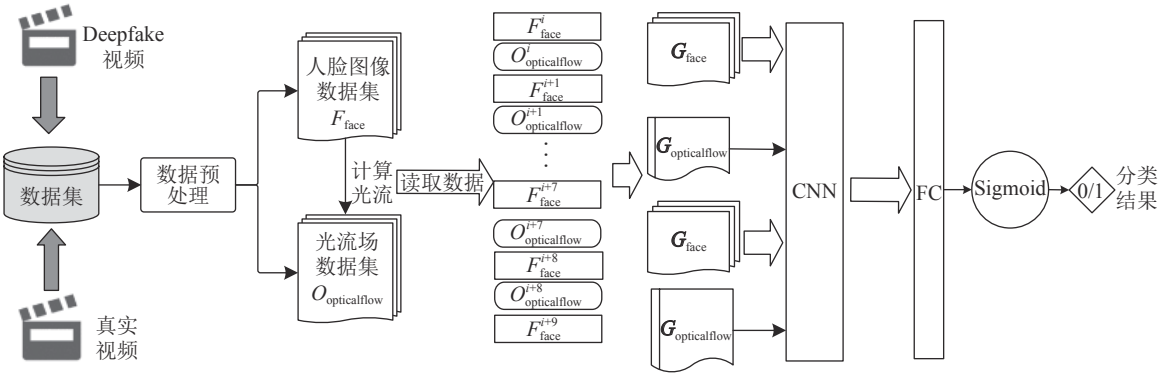

图 2 训练模型

Fig. 2 Training model

在模型的数据预处理阶段,首先将人脸图片数据和计算后的光流图按照帧顺序存储,综合考虑计算量和信息损耗,实验中对视频每秒钟抽样10帧RGB图片。深度学习模型在读取训练集时,同时读取10帧人脸图片和与这些人脸图片相对应的光流图。在实验中发现,采用CNN对多幅光流进行高层次特征提取效果明显优于单幅光流图,所以本文对9张光流图采取拼接操作,为了保证数据的一致性,取前8张光流图对每4张光流图进行拼接,并舍弃最后1张光流图。由于光流图为二通道,所以拼接后的光流图为256×256×8的形状。对于原始人脸图像和拼接后的光流图分别采用CNN提取空域特征和高层次时域特征,在实验中共设置4层卷积层和最大池化层,这样设置成4层的理由是层数太深的话极易造成过拟合现象,层数过少又难以提取有效特征,经过反复实验,本文将卷积池化层设置为4层。在经过CNN提取特征后,采用全连接层(fully connected layers, FC)进行二分类,得出判定结果。为了使模型能够快速收敛,本文在卷积层与池化层间加入BN层[17],并使用Relu非线性激活函数使模型快速收敛。算法1展示了模型的学习与优化过程。

算法1 本文模型学习算法

输入:真实视频集D^{R}、Deepfake视频集D^{F}、模型迭代次数e、模型读取训练集批量数据大小$B_{batchsize}$;

输出:权重w确定的深度学习模型。

```
D=D^R∪D^F;
For V in D
    对V抽取帧F_frame;
    For P in F_frame
        从P截取人脸F_face;
    End For
    对F_face计算光流O_opticalflow;
End For
while e−− ≥ 0 do
    G_opticalflow=∅, G_face=∅;
    For i=1 to B_batchsize do
        读取连续的F_face和O_opticalflow;
        拼接O_opticalflow形成C_opticalflow;
            G_opticalflow[i]←C_opticalflow;
            G_face[i]←F_face;
    End For
    读取 G_opticalflow、G_face 并输入训练模型;
    w←反向传播优化算法Adam;
End while
```

算法1主要分为数据预处理和训练模型两个过程,在数据预处理阶段,模型的时间复杂度为$O(n^2)$。在深度学习模型训练阶段,以该模型的浮点运算次

数（floating point operations，FLOPs）表示时间复杂度，本文模型中浮点运算次数主要由卷积层构成，因此本文模型FLOPs的计算公式如下：

$$F=\sum_{i=1}^{N}\sum_{l=1}^{D}M_l^2\times K_l^2\times C_{l-1}\times C_l \tag{7}$$

式中：F为FLOPs；N为输出深度学习模型中的人脸图片与光流图的个数；D为卷积层的个数；M为每个特征图的边长；K为每个卷积核的边长，本文实验中特征图和卷积核的长宽都是相等的；C_{l-1}为第$l-1$层的卷积核输出通道数；C_{l-1}为第l层的卷积核输出通道数。

3　实　验

3.1　实验数据与环境

为了验证本文基于时域特征一致性检测模型的性能与有效性，从FaceForensics++[13]Deepfake开源数据集中选取视频进行实验。实验中选取Deepfake视频段和真实视频段各950个（850个视频段作为训练集，100个视频段作为验证集）。由于本文深度学习模型是按连续帧序列读取训练集，所以经过预处理后的实验数据集如表1所示。

表 1　实验数据集

Tab. 1　Datasets in test

类别	训练集个数	验证集个数
Deepfake图像序列	17 116	1 924
真实图像序列	17 191	1 962
总计	343 07	3 886

本文实验均在CPU为Intel(R)Core(TM)i7-8700CPU@3.20 GHz、内存为16 GB、GPU为Nvidia GTX1080ti的Windows10台式机上完成。深度学习模型使用Keras2.2.4[18]构建，并使用Tensorflow1.8.0[19]作为后端引擎，开发环境为Pycharm 2018.2.4。

3.2　评判标准

为了评价本文基于时空特征一致性检测模型的性能，选取多种度量指标来评价模型。首先对于分类模型，准确率常常用于评估一个模型的全局准确度，准确率值愈高则模型的准确度越好。准确率A_{accuracy}的计算公式如下：

$$A_{\text{accuracy}}=\frac{T_{\text{TP}}+T_{\text{TN}}}{T_{\text{TP}}+F_{\text{FP}}+T_{\text{TN}}+F_{\text{FN}}} \tag{8}$$

式中：T_{TP}为真正例（true positive，TP），即被正确分类的Deepfake图像的个数；T_{TN}为真反例（true negative，TN），即被正确分类的真实图像的个数；F_{FP}为假正例（false positive，FP），即被错误分类的Deepfake图像个数；F_{FN}为假反例（false negative，FN），即被错误分类的真实图像个数。

为了更加全面地评估模型，本文除选用准确率以外，还选用受试者工作特征曲线（receiver operating characteristic curve，ROC）下面积（area under roc curve，AUC）作为评价指标。ROC曲线是根据模型的预测结果大小对样例进行排序，按此顺序逐一把每个样例的预测概率作为阈值，计算出假正例率（false positive rate，FPR）和真正例率即召回率（true positive rate，TPR），并以FPR作为横轴，TPR作为纵轴的曲线，其中，FPR与TPR的计算公式如下：

$$F_{\text{FPR}}=\frac{F_{\text{FP}}}{F_{\text{FP}}+T_{\text{TN}}} \tag{9}$$

$$T_{\text{TPR}}=\frac{T_{\text{TP}}}{T_{\text{TP}}+F_{\text{FN}}} \tag{10}$$

ROC曲线可以很好地描述模型的泛化性能。AUC是ROC曲线下面积，AUC值越大，则可说明模型的性能越好，AUC的计算公式如下：

$$A_{\text{AUC}}=\frac{1}{2}\sum_{i=1}^{m-1}(F_{\text{FPR}}^{(i+1)}-F_{\text{FPR}}^{(i)})\times(T_{\text{TPR}}^{(i)}+T_{\text{TPR}}^{(i+1)}) \tag{11}$$

式中，m为样例个数。

除使用准确率、AUC等关键度量指标，真正例率TPR、真负例率（true negative rate，TNR）、正例预测精度（positive predicyive value，PPV）、负例预测精度（negative predicyive value，NPV）等指标也被本文所采用，以评价模型的优劣，相同情况下这些指标的值越高则模型的效果越好。

3.3　实验过程及结果分析

为了使模型能够充分学习数据集的时域特征及空域特征信息，本文将模型的迭代次数e设置为80次，并将深度学习模型的损失函数设置为均方误差（mean-square error，MSE）损失函数，计算方法如下：

$$L_{\text{loss}}=\frac{1}{2m}\sum_{i=1}^{m}(\hat{y}_i-y_i)^2 \tag{12}$$

式中，m为样例个数，y_i为该样例的标签，$\hat{y}_i$为模型的预测值。损失函数可以很好地表示模型的预测结果和真实标签的拟合程度，数值越小代表拟合程度越好。深度学习中采用优化算法对模型进行优化，使得L_{loss}的值趋于最小值，增加模型的拟合能力，本文采用的优化算法为Adam[20]，并将学习率设置为5×10^{-5}。实验数据集分为训练集和验证集，在训练集上训练模型，在验证模型上验证模型。在训练集上本文模型的准确率和损失函数值如图3、4所示。

从图3可以看出，随着训练次数的增加，本文模型的Loss函数值逐渐降低，说明本文模型的拟合能力越来越强，这充分说明了本文模型的有效性。

图4显示了随模型迭代次数的增加，本文模型的分类预测率也逐渐增加，说明模型的效果越来越好。从图4中可以看出，本文模型的准确率在第60轮时趋于平稳，最终模型的训练准确率可以达到99.47%左右，说明本文模型具有良好的分类和检测效果。

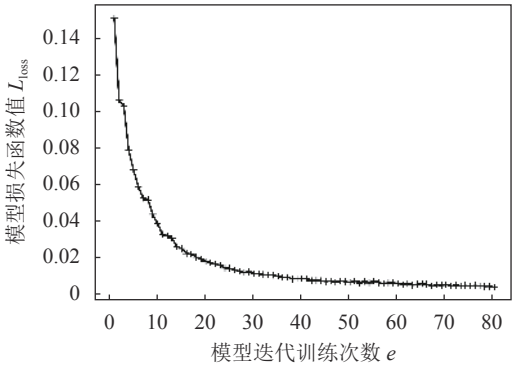

图 3 损失值随训练次数变化曲线

Fig. 3 Changing curve of loss values with training times

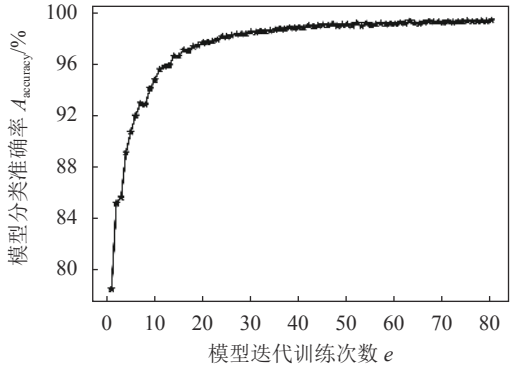

图 4 准确率随训练次数变化曲线

Fig. 4 Changing curve of accuracy rates with training times

随着模型迭代次数的增加，本文模型的拟合能力越来越强，但是由于深度学习模型强大的拟合能力，极其容易出现过拟合训练数据的情况。为了避免过拟合，在实验中采取提前停止策略。提前停止策略经常被用在深度学习模型训练中，即当模型的损失函数值在一段时间不再改善则终止模型的训练。

表2为本文模型在验证集上检测结果的混淆矩阵，混淆矩阵可以清晰地展示模型的预测结果和各项指标。从表2中可以看出，本文模型在召回率TPR、真负例率TNR以及预测精度PPV和NPV都表现良好。具体而言，根据模型检测结果的混淆矩阵可以计算出模型的检测准确率为98.1%，召回率为97.14%，真负例率为99.03%，正例预测精度和负例预测精度分别为98.99%、97.24%。

图5为本文模型的预测结果的ROC曲线和AUC取值，从图5中可以看出，模型的AUC取值较高，模型的预测效果和泛化性能表现良好。

表 2 验证集模型预测结果混淆矩阵

Tab. 2 Confusion matrix of the prediction results on valid dataset

真实标签	模型预测	
	positive个数	negative个数
positive个数	1 869	55
negative个数	19	1 943

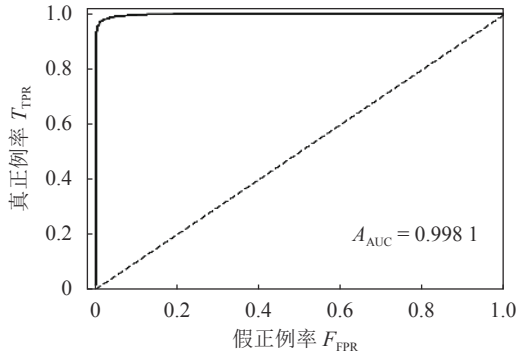

图 5 模型预测结果ROC曲线和AUC

Fig. 5 ROC curve and AUC value of the model's prediction results

为了验证模型的有效性，以FaceForensics++[13]中提到的模型开展对比实验，对比结果如表3所示。从表3可看出，本文模型的准确率明显优于特征+SVM[21]、Cozzolino[22]、Bayar [23]、Rahmouni[24]、MesoNet[10]等模型。

表 3 不同模型的实验准确率结果对比

Tab. 3 Comparison of experimental accuracy results by different models

模型	准确率/%
特征+SVM[21]	73.85
Cozzolino[22]	85.23
Bayar[23]	89.50
Rahmouni[24]	93.81
MesoNet[10]	89.31
XceptionNet[25]	95.03
CNN+LSTM[7]	97.20
本文模型	98.10

需要说明的是，本文对相关模型的实现与FaceForensics++[13]中提到的参数及配置相同。由于FaceForensics++[13]没有给出具体的训练集和验证集的分配方法，本文均按照相同的配置（850个训练样本和100个验证样本）开展实验。如图6所示，在对比实验中发现，XceptionNet模型在训练集上的效果最高只达95.03%，且该模型在验证集上的检测效果存在严重的过拟合现象，即该模型在验证集上分类准确率严重降低。XceptionNet在本文验证集上的ROC曲线和AUC值如图7所示。从图7可知，由于XceptionNet过拟合的原因导致其在验证集中的AUC值明显偏低。

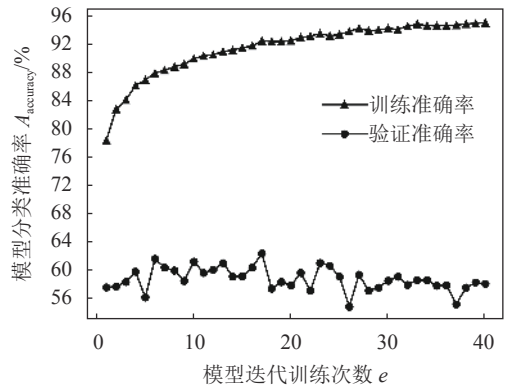

图 6　XceptionNet在本文数据集上的训练和验证准确率

Fig. 6　Accuracy of training and verification of XceptionNet on the dataset in this paper

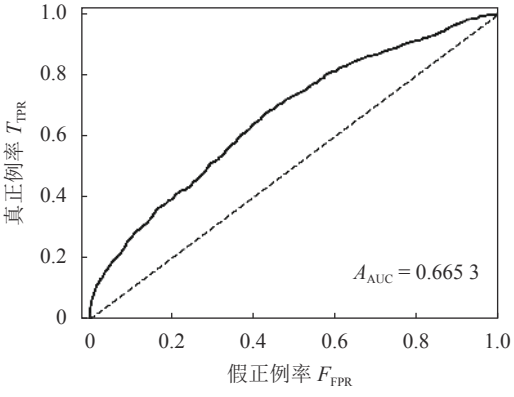

图 7　XceptionNet在验证集上的ROC曲线和AUC取值

Fig. 7　ROC curve and AUC value of XceptionNet on the valid dataset

为了多方面比较本文模型的性能，另外选取了AUC作为比较指标，由于一些现有的模型难以在实验中复现，且相关文献没有公布模型各项指标，所以本文仅选取可复现或公布了AUC指标的模型进行对比。对比结果如表4所示。

表 4　实验AUC值对比

Tab. 4　AUC values comparison of experimental results

模型	AUC值
XceptionNet[25]	0.665 3
VGG16[8]	0.845 0
ResNet50[8]	0.987 0
ResNet101[8]	0.991 0
特征点+SVM[9]	0.890 0
MesoNet[10]	0.984 0
LRCN[7]	0.990 0
ResNet152[8]	0.978 0
本文模型	0.998 1

AUC值不仅可以反映出模型的检测效果，还可代表模型的泛化能力。从表4可以看出，本文模型相比于对比模型具有明显的优势，说明本文模型的检测效果和泛化性能明显优于对比模型，具有优异的性能。

4　结　论

针对目前的Deepfake检测模型大都基于内容特征，忽视了时域特征的重要性的问题，本文从换脸图像人脸表情变化存在的细微的不连贯、不一致等现象为出发点，提出了一种基于时空特征一致性的检测模型。使用卷积神经网络对Deepfake图像进行空域特征提取，使用光流法对Deepfake时间域上的特征进行描述；为了对光流场描述的时域特征进行组织利用，使用卷积神经网络对时域特征进行特征再提取；最后，将Deepfake图像的时域特征与空域特征相融合进行检测。实验表明该模型具有良好的检测效果，通过实验对比显示本文模型的在检测准确率、AUC取值方面较现有模型有较大优势。

由于换脸技术不仅有Deepfake还有Face2Face等多种多样的技术手段，因此下一步的研究重点在于改进本文的检测模型，使得模型具有更加广泛的通用性。

参考文献：

[1] Github deepfakes.faceswap[CP/DK].(2019–02–23)[2019–04–20].https://github.com/deepfakes/faceswap.

[2] Github iperov.Deepfacelab[CP/DK].(2019–02–22)[2019–04–22].https://www.deepfacelabs.com/.

[3] Todd Ditchendorf.Fakeapp[CP/DK].[2019–04–22].http://www.fakeapp.com/.

[4] 朱昌俊.规定恶搞换脸侵犯肖像权为AI应用划界[N].中国青年报,2019–04–23(002).

[5] 吕俊延.人工智能视频换脸技术的法律困境[N].中国社会科学报,2019–03–26(006).

[6] Güera D,Delp E J.Deepfake video detection using recurrent neural networks[C]//Proceedings of the 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS).New Zealand:IEEE,2018:1–6.

[7] Li Y,Chang M C,Lyu S.In ictu oculi:Exposing ai created fake videos by detecting eye blinking[C]//Proceedings of the 2018 IEEE International Workshop on Information Forensics and Security (WIFS).Hong Kong:IEEE,2018:1–7.

[8] Li Y,Lyu S.Exposing deepfake videos by detecting face warping artifacts[EB/OL].(2018–11–01)[2019–04–22].https://arxiv.org/abs/1811.00656.

[9] Yang X,Li Y,Lyu S.Exposing deep fakes using inconsistent head poses[C]//Proceedings of 2019 IEEE International Conference on Acoustics,Speech and Signal Processing (ICASSP 2019).Brighton:IEEE,2019:8261–8265.

[10] Afchar D,Nozick V,Yamagishi J,et al.Mesonet:A compact facial video forgery detection network[C]//Proceedings of the 2018 IEEE International Workshop on Information Forensics and Security (WIFS).Hong Kong:IEEE,2018:1–7.

[11] Nguyen H H,Yamagishi J,Echizen I.Capsule-forensics:Using capsule networks to detect forged images and videos[C]//Proceedings of 2019 IEEE International Conference on Acoustics,Speech and Signal Processing (ICASSP 2019).Brighton:IEEE,2019:2307–2311.

[12] Bradski G,Kaebler A.学习OpenCV[M].于仕琪,刘瑞祯,译.北京:清华大学出版社,2009.

[13] Rössler A,Cozzolino D,Verdoliva L,et al.Faceforensics:A large-scale video dataset for forgery detection in human faces[EB/OL].(2018–03–24)[2019–04–22].http://arxiv.org/abs/1803.09179.

[14] Fabrice Bellard.Ffmpeg[CP/DK].(2018–11–12)[2019–03–20].http://ffmpeg.org/.

[15] Davis E.King.Dlib[CP/DK].(2009–07–10)[2019–03–20].http://dlib.net/.

[16] Farnebäck G.Polynomial expansion for orientation and motion estimation[D].Linköping:Linköping University Electronic Press,2002.

[17] Ioffe S,Szegedy C.Batch normalization:Accelerating deep network training by reducing internal covariate shift[C]//Proceedings of the 32nd International Conference on Machine Learning.Lille:JMLR,2015:448–456.

[18] Keras[CP/DK].(2015–06–13)[2018–07–10].https://keras.io/.

[19] Google.Tensorflow[CP/DK].(2015–11–19)[2018–07–10].https://tensorflow.google.cn/.

[20] Kingma D P,Ba J.Adam:A method for stochastic optimization[EB/OL].(2014–12–22)[2019–04–22].https://arxiv.org/abs/1412.6980.

[21] Fridrich J,Kodovsky J.Rich models for steganalysis of digital images[J].IEEE Transactions on Information Forensics and Security,2012,7(3):868–882.

[22] Cozzolino D,Poggi G,Verdoliva L.Recasting residual-based local descriptors as convolutional neural networks:An application to image forgery detection[C]//Proceedings of the 5th ACM Workshop on Information Hiding and Multimedia Security (IHMMSec'17).New York:ACM,2017:159–164.

[23] Bayar B,Stamm M C.A deep learning approach to universal image manipulation detection using a new convolutional layer[C]//Proceedings of the 4th ACM Workshop on Information Hiding and Multimedia Security (IH & MMSec'16).New York:ACM,2016:5–10.

[24] Rahmouni N,Nozick V,Yamagishi J,et al.Distinguishing computer graphics from natural images using convolution neural networks[C]//Proceedings of the 2017 IEEE Workshop on Information Forensics and Security (WIFS).Rennes:IEEE,2017:1–6.

[25] Chollet F.Xception:Deep learning with depthwise separable convolutions[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.Honolulu:IEEE,2017:1251–1258.

（编辑　赵　婧）

引用格式: Zhao Lei,Ge Wanfeng,Mao Yuzhu,et al.Deepfake video detection model based on consistency of spatial–temporal features[J].Advanced Engineering Sciences,2020,52(4):243–250.[赵磊,葛万峰,毛钰竹,等.基于时空特征一致性的Deepfake视频检测模型[J].工程科学与技术,2020,52(4):243–250.]