计算机科学
COMPUTER SCIENCE

基于关键帧与时空特征融合的人脸伪造检测

程燕

引用本文

程燕. 基于关键帧与时空特征融合的人脸伪造检测[J]. 计算机科学, 2024, 51(11): 191-197.

CHENG Yan. Facial Forgery Detection Based on Key Frames and Fused Spatial-Temporal Features [J]. Computer Science, 2024, 51(11): 191-197.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.240100063

基于关键帧与时空特征融合的人脸伪造检测

程　燕
华东政法大学信息科学与技术系　上海 201620

摘　要　基于深度学习的人脸真伪检测是一个典型的二分类问题,模型训练结果的精度不仅受到训练数据质和量的影响,还与训练策略、网络架构设计等有关。以光流法为基础,提出了一种基于关键帧与时空特征融合的人脸伪造检测方法。首先,采用加权光流能量阈值分析法筛选出视频中能量较大的关键帧,将关键帧的光流和LBP纹理特征进行融合,构成具有时间和空间特性的融合特征图,经过增强处理后输入CNN模型进行学习。在FaceForensics++和Celeb-df数据集上的测试表明,所提算法的检测率较传统算法均有明显提升。跨库实验中,所提算法采用Efficientnet-V2结构在FaceForensics++数据集上表现出最优的跨库检测性能,准确率达到90.1%,XceptionNet结构的整体性能优于其他方法,准确率均达到80%以上,具有优越的泛化性能。

关键词:光流;关键帧;LBP纹理;CNN模型
中图分类号　TP391

Facial Forgery Detection Based on Key Frames and Fused Spatial-Temporal Features

CHENG Yan
Department of Information Science and Technology,East China University of Political Science and Law,Shanghai 201620,China

Abstract　The deep learning-based facial forgery detection is commonly approached as a binary classification problem. The accuracy of model training results is not only affected by the quality and quantity of training data,but also related to training strategy and network architecture design.. In this paper,we propose a new method based on key frames and spatial-temporal features. Firstly,the weighted optical flow energy analysis is used to detect the key frames in a video. Then,the optical flow and LBP features of the key frames are fused to form feature maps with spatial and temporal characteristics. After data augmentation,the feature maps are fed into the CNN model for training. Evaluations conducted on the FaceForensics++ and Celeb-df datasets demonstrate that the proposed method achieves superior or comparable detection accuracy. Experimental results on cross-datasets show that the proposed method,utilizing the Efficientnet-V2 structure,achieves the best performance on the FaceForensics++ database with the accuracy of 90.1%. Furthermore,the overall performance of the XceptionNet structure surpasses that of other methods,achieving the accuracy over 80%,thus demonstrating superior generalization performance of the proposed method.

Keywords　Optical flow,Key frames,LBP texture,CNN model

1　引言

人脸伪造是深度伪造技术应用领域的热点。网络上海量人脸数据的出现,以及生成对抗网络(Generative Adversarial Network,GAN)和自编码器技术的蓬勃发展,使得现有的伪造技术能够生成极为逼真的人脸图像和视频。利用目前网络上广泛流传的Face2Face,FaceSwap,FakeApp,DeepFake,DeepNude以及Github共享的伪造开发程序,可以轻松实现全脸合成、换脸、唇形同步、面部复现等操作。深度伪造技术的门槛不断降低,滥用该技术对图像和视频等多媒体内容进行篡改的现象也层出不穷,甚至一些非专业技术人员都可以快速制作出以假乱真的多媒体内容,给现有法律体系尤其是司法举证和鉴定带来很大的挑战。

目前主流的人脸伪造检测方法大多基于深度学习来进行判别。按照检测的原理和思路,现有的检测算法可分为图像级和视频级检测两大类[1]。图像级检测将视频处理成帧,通过设计不同的网络结构,对图像帧内特征进行识别,最终形成对视频真伪性的二分类判别。然而该方法由于缺少视频帧间的时间和运动信息,检测准确率较低,跨数据集的泛化性能也较差。视频级检测虽然充分考虑了视频帧间的时序信息,但其检测对视频的预处理(如视频压缩、背景光线变化等)很敏感。特别是在未知篡改算法的前提下,现有算法大多是将新生成的算法数据集加入训练集,以此来提高算法的检测率。然而,选取帧数过少会导致样本取样不充分,模型无法充分

到稿日期:2024-01-04　返修日期:2024-05-11
基金项目:教育部人文社科一般项目(23YJA820015)
This work was supported by the Humanities and Social Sciences of the Ministry of Education(23YJA820015).
通信作者:程燕(chengyan@ecupl.edu.cn)

学习视频帧的空间特征，而选取帧数过多会导致样本冗余，使模型过拟合。以往的算法在实现过程中，普遍是从每个视频中随机选取 30～60 帧作为输入进行模型训练，选取的帧或为某一段视频的连续帧，或是对该视频进行均匀抽帧，这将导致输入数据不具有代表性，检测结果可信度不高或训练性能有限。

因此，针对现有工作的局限性，本文提出了一种基于关键帧和时空域特征融合的深度伪造检测方法。本文的主要贡献如下：

1)以帧间光流信息为基础，利用加权光流能量阈值分析法计算视频中包含主要人脸变化信息的关键帧，以提高模型训练输入数据的有效性。

2)本文将空间域的 LBP(Local Binary Pattern)纹理信息作为时间域光流特征的补充，能够充分体现视频在空域和时域上的特性，反映细微的伪造痕迹或者压缩误差，有利于对低质量的人脸深度伪造图像的检测。

3)算法直接采用在常规图像分类任务中性能表现优异的神经网络进行特征提取及模型训练，实验结果表明该算法具有易实现性和扩展性的特点，且在跨库检测时也可以达到较理想的效果。

2 相关工作

无论是基于图像级还是视频级的深度学习检测算法，都需要构建适当的网络结构来提取视频特征并进行分类。模型训练结果的精度不仅受到训练数据质和量的影响，还与训练策略、网络架构设计等有关。

伪造人脸是对局部区域进行操作，可能导致人脸五官变形、边缘区域模糊、人脸区域与非人脸区域存在差异等现象；且在视频生成过程中，若没有考虑视频帧的关联性，则生成的视频帧间的一致性会遭到破坏，前后两帧生成的细节也可能存在差异，例如文献[2]提出基于局部二值模式、方向梯度直方图特征和孪生神经网络提取帧间差异特征的检测框架。Han 等利用频域变换及滤波等方式提取高频噪声作为输入进行学习[3]。由于人脸伪造过程很可能会丢失人物的生理特性信息，因此，挖掘心跳[4]、眨眼[5]、口型[6]等生物学特征或头部的 3D 姿态建模[7]而获得的面部表情以及头部运动的位移和旋转等矢量特征，也常被作为人脸伪造特征检测的重要依据。

此外，优化神经网络模型也可以增强模型学习能力，从而提升检测效果。这类方法不依赖于特定的人脸伪造图像特征，而是将大量真实与伪造的人脸图像作为训练数据，通过训练神经网络模型为二分类器，来判断人脸伪造与否。最直接的方法是采用性能表现优异的现有神经网络进行训练。Rossler 等[8]在 FaceForensics++数据集上对不同压缩率的 4 种伪造人脸图像进行检测评估，证明了 XceptionNet 架构是性能最好的检测器。Amerini 等将相邻帧间计算的光流图序列输入 VGG16 网络进行训练，给出整个视频被伪造的概率[9]。文献[10]进一步利用 canny 边缘检测方法获取每一帧的边缘信息，与 RGB 图像以不同方式融合，并根据融合方式的不同选择不同的 Xception[11]结构，接入 Bi-LSTM 网络以挖掘帧间的关联性与相关性，最后使用全连接与 Sigmoid 进行二分类。Sabir 等[12]结合 DenseNet 和循环神经网络(Recurrent Neural Network，RNN)挖掘帧间的不一致性，设计了一种基于时域信息的检测方法。类似地，Guera 等[13]提出结合 InceptionNet-V3 和 RNN 检测伪造视频，其中 InceptionNet-V3 用于提取单帧图像中的视觉特征，RNN 则采用长短时记忆网络(LSTM)捕捉视频的时序特征。在此基础上，许多学者进一步开展了对神经网络结构的研究。Afchar 等[14]提出了两种神经网络模型 Meso-4 和 MesoInception-4，其中 Meso-4 为小规模网络，包含 4 个卷积层和两个全连接层；MesoInception-4 为大规模网络，其将 Meso-4 中前两层普通卷积层替换为改进的 Inception[15]模块，将单个视频作为输入，利用卷积神经网络捕捉帧内信息，进而检测伪造视频。Nguyen 等[16]设计了胶囊网络来判别造假的图像或视频，通过抽取人脸，用 VGG19[17]提取特征编码，再输入胶囊网络进行分类。文献[18]结合 SegCaps 和卷积神经网络(CNN)方法的优点用于改进图像特征提取，再进行胶囊网络训练以增强泛化能力。Yu 等[19]利用 U-net 结构为给定的伪造方法分别训练特定的伪造特征提取器，并考虑了三元组损失、位置损失、分类损失和自动加权损失等，以确保特征提取器对相应伪造方法的检测能力，最终取得了泛化能力的提升。此外，相比二维神经网络，三维卷积神经网络具有同时捕获空间和时间特征的能力，因此运用该网络进行真伪检测也能取得显著效果[20]。然而，RNN 和 3D 卷积神经网络虽然可以保留视频中的时间信息，但是网络复杂度较高，权值参数数量庞大，且需要大量的训练样本进行训练，对计算资源要求较高。

一些针对现有神经网络结构的修改也被用来进一步提升伪造视频的检测能力，如 Wang 等[21]考虑到 I，B，P 帧在视频压缩编码中信息丢失的影响，选择 I 帧作为关键帧进行模型训练，再联合利用卷积池化和再注意力机制(re-Attention Mechanism，re-AM)方法丰富全局特征的学习；Hsu 等[22]采用对比损失寻找不同生成器生成的图像的特征；Dang 等[23]在主干网络增加注意力机制以聚焦篡改区域；Rahmouni 等[24]在网络中增加了计算统计数据的全局池化层。这些方法均达到了提升视频真伪检测性能的效果，但网络复杂性的增加意味着计算成本也会增加，甚至可能导致模型出现过拟合。

3 理论基础

3.1 光流模型

光流表征了三维空间中运动物体表面的点在成像平面上投影的瞬时速度，利用图像序列中像素在时间域上的变化以及相邻帧之间的相关性来确定相邻两帧之间像素点的运动信息[25]。设像素点(x,y)在 t 时刻的亮度为 $I(x,y,t)$，根据光流法中亮度一致的基本假设，像素点从 t 时刻的位置(x,y)运动到 $t+\Delta t$ 时刻的位置$(x+\Delta x,y+\Delta y)$，其亮度保持不变，则有：

$$I(x,y,t)=I(x+\Delta x,y+\Delta y,t+\Delta t) \tag{1}$$

将式(1)进行一阶泰勒级数展开，忽略高阶项，得到光流的约束方程为：

$$I_x u_x + I_y u_y + I_t = 0 \tag{2}$$

其中,$I_x = \frac{\partial I}{\partial x}$,$I_y = \frac{\partial I}{\partial y}$表示同一张图像上相邻像素点相对 X 和 Y 轴方向的光强变化,$I_t = \frac{\partial I}{\partial t}$ 表示同一个像素点在相邻时刻的强度变化,$\vec{u} = (u_x, u_y)^{\mathrm{T}}$ 即光流,u_x 和 u_y 分别是像素点在 x 与 y 方向的偏移量。

虽然一个视频序列包含庞大的图像信息量,但其中很多图像信息对于光流特征分析是没有任何价值的,主要体现在视频中可能存在大量运动幅度微小或几乎不动的对象。以帧率 30FPS 的视频为例,若人脸的头部或面部表情有 1s 状态保持不动或动态幅度很小,则对应有 30 个连续帧的画面几乎相同。在网络模型训练中,若将此 30 帧图像或光流图像作为模型训练的随机输入数据,则从这些数据中提取的特征量的价值就非常小。若简单地通过增加输入网络模型的图像帧数来提升特征提取的性能,则会加大模型训练的复杂度。为了提高真伪检测算法的准确率,降低计算复杂度,有必要在模型训练前去掉视频序列中无价值的图像帧,保留有价值的图像帧,即确定视频中有效的关键帧以供模型训练。

3.2　LBP 纹理特征

真假人脸之间的差异通常是细微的,并且往往存在于低层的纹理特征中,这不容易被单一的结构网络捕获。本文增加纹理信息并通过神经网络提取特征可有效提升人脸真伪检测性能。

LBP,也被称作局部二值模式,是一种用来描述图像局部纹理特征的算子,其定义为在 3×3 的窗口内,以窗口中心像素为阈值,将相邻的 8 个像素的灰度值与其进行比较,若周围像素值大于中心像素值,则该像素点的位置被标记为 1,否则为 0。如图 1 所示,该 3×3 范围内的 8 个像素点经过与中心像素比较可产生 8 位二进制数 00010011,将该二进制转换为十进制数即得到对应窗口中心像素点的 LBP 码值,该值可反映区域的纹理信息。

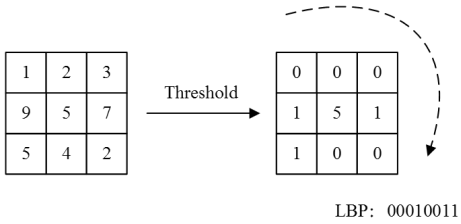


图 1　LBP 算子计算过程

Fig. 1　Computing process of LBP operator

LBP 具有旋转不变性和灰度不变性的特点,在网络模型训练中,对于数据增强处理的操作,该特征不会因为旋转或者光照变化而发生明显变化,因此提取的人脸图像的局部纹理特征对光照和微小平移具有较强的顽健性,将其作为网络训练输入特征更有助于网络对图像特征分布的理解,进一步降低网络学习到不利的特征描述的可能性。此外,LBP 特征的算法计算比较简单,处理视频人脸表情识别更有效,实时性更好。

4　所提算法

4.1　关键帧检测

光流特征蕴含丰富的运动信息。光流的长短反映了物体运动速度的快慢,方向则反映了物体运动的方向。将这些特征结合起来可以有效表征目标运动时的形态变化特征和运动变化规律。如图 2 所示,以眼睛、嘴唇为例,对于变化较小的帧,其速度方向比较单一,光流图变化平稳(如图中方框所示区域);而变化较大的帧,其速度方向信息则较为丰富,光流图变化剧烈(如图中圆框所示区域)。通常,真实视频的光流变化会有一定的规律性和连续性,但是经过篡改的视频则无法兼具到时序特征,且与真实视频相比,伪造视频的光流变化很小,因此,光流特征可作为人脸真伪检测的一个重要依据[9]。

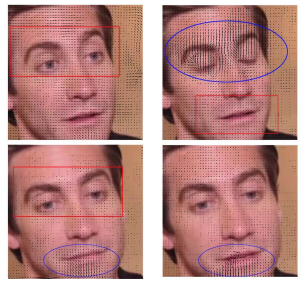


图 2　人脸光流示意图

Fig. 2　Diagram of facial optical flow

设第 n 帧中像素点 (i,j) 处的光流矢量值为 $\vec{u} = (u_x, u_y)^{\mathrm{T}}$,则该像素点的光流速度为:

$$v_{i,j}(n) = \sqrt{u_x^2 + u_y^2} \tag{3}$$

运动方向与水平方向的夹角为:

$$\angle Angle_{i,j} = \arctan \frac{u_x}{u_y} \tag{4}$$

该帧中第 k 个运动区域的光流能量为:

$$E_k(n) = \sum_{i=1}^{W} \sum_{j=1}^{H} v_{i,j}^2(n) \tag{5}$$

其中,W,H 分别表示目标区域的宽度和高度。式(5)的局限性在于,仅利用像素点运动速度而得到的光流能量不能很好地表征运动的方向特征。人脸中存在运动状态的对象主要是人脸轮廓、眉毛、眼睛和嘴等,如头部的转动、眉毛耸动、眨眼、嘴唇开闭等。在同一帧图像中,五官部分的运动状态相对独立,没有统一的幅度和方向,反映在光流特征上即为像素速度值的差异和运动方向各不相同。若综合考虑运动角度对光流能量的影响[26],可定义加权光流能量为:

$$E_k(n) = \sum_{i=1}^{W} \sum_{j=1}^{H} w_{i,j}(n) v_{i,j}^2(n) \tag{6}$$

其中,$w_{i,j}(n)$为光流能量的权值,定义为:

$$w_{i,j}(n) = \left(\frac{|\angle Angle_{i,j} - \angle AngleAvg|}{\pi} \times \lambda \right)^2 + \left(\frac{|\angle Angle_{i,j} - \angle AngleMax|}{\pi} \times \lambda \right)^2 \tag{7}$$

其中,$\angle Angle_{i,j}$为当前像素与水平方向的夹角,$\angle AngleMax = \max_{(i,j)\in\Omega} (\angle Angle_{i,j})$为运动目标区域中速度最大值在像素的角度,$\angle AngleAvg = \underset{(i,j)\in\Omega}{mean}(\angle Angle_{i,j})$为运动目标与水平方向

夹角的平均值;λ 为调节光流能量权值的大小。

图 3 给出了某段视频中对人脸区域按加权和非加权两种方法计算的光流能量曲线对比图。如图所示,非加权的光流能量曲线较为平缓,很难区分视频帧间存在明显变化的运动行为;而加权的光流能量曲线变化显著,图中峰值的高低可以清晰地反映出视频中存在明显运动变化的情况。根据图 3 中峰值的变化情况,从该视频的前 60 帧中任意选择连续帧、等间隔帧,以及峰值处图像帧进行对比。如图 4 所示,连续帧和等间隔帧的画面几乎静止或存在着微小变化,而对应峰值处的人脸则变化明显。因此,设置阈值 T,可定义视频的关键帧为:

$$key(n)=\begin{cases}1, & E(n)\geqslant T\\ 0, & \text{others}\end{cases} \tag{8}$$

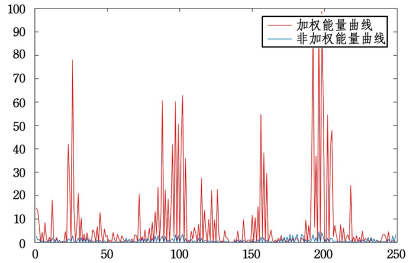


图 3 加权和非加权的光流能量曲线比较图

Fig. 3 Comparison of weighted and unweighted optical flow energy curves

(a)连续图像帧:第 25,26,27,28 帧 (b)等间隔图像帧:第 20,24,28,32 帧

(c)关键图像帧:第 20,24,37,49 帧

图 4 关键帧与非关键帧对比结果

Fig. 4 Comparison of keyframes and non-keyframes

4.2 算法模型结构

本文基于传统的 CNN 检测算法展开研究,提出了基于关键帧的人脸伪造检测模型,将图像的 LBP 纹理特征与光流特征相融合,形成具有时空特征的“纹理-光流特征图”。图 5 给出了本文所提检测算法的总体流程,其由数据预处理、关键帧检测、数据增强以及模型训练 4 个模块组成。

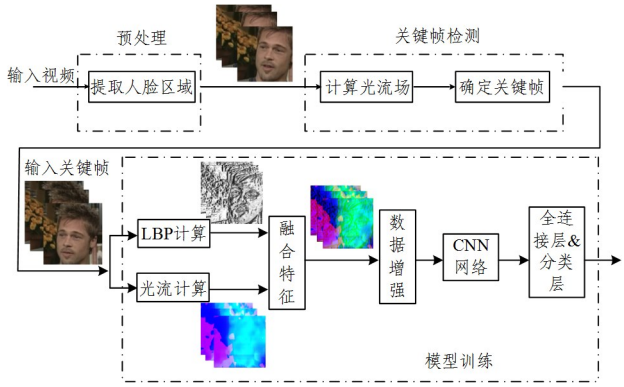


图 5 算法总体流程图

Fig. 5 Overall algorithm process

4.2.1 数据预处理模块

由于伪造人脸只能发生在人的面部区域,因此在视频真伪检测过程中,只需要对人脸面部区域进行标记、计算,即可避免光流计算复杂度高的问题。本文使用 MTCNN(Multi-task Convolutional Neural Network)[27]算法裁剪出人脸矩形区域,并将其大小调整为 224×224。

4.2.2 关键帧检测模块

选择实时性和精度均较高的 Lucas-Kanade 金字塔算法[28]对视频的人脸区域进行光流计算。运用式(6)—式(8)加权光流能量阈值分析法筛选出关键帧。为了分析阈值 T 对关键帧检测结果的影响,我们从 Celeb-df[29]数据集中随机选取真假视频各 100 个,每个视频任意选择连续的 60 帧进行检测。假设按 $E_k(n)$最大值的不同比例进行阈值筛选,即:

$$T=i\times\max(E_k(n)) \tag{9}$$

其中 $i\in[0,1]$,则实际检测出的关键帧数如图 6 所示。可以看出,T 的取值影响到关键帧数目的检测结果。T 值越大,检测出的关键帧数越少,当 T 取值在 $0.5\max(E_k(n))$左右时,每个视频可确定约 3~20 个关键帧。

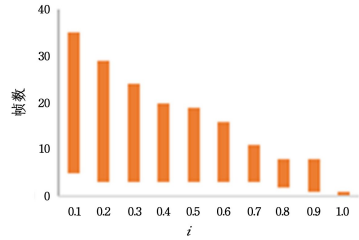


图 6 关键帧检测数目示例

Fig. 6 Example of detected number of keyframe

4.2.3 数据增强模块

经过关键帧检测模块提取的关键帧数约为 3~20 帧。为了增加模型训练的样本数量,减轻模型训练时的过拟合,在训练时采用了高斯噪声、翻转、旋转、色彩调整及标准化等处理以增强训练样本的多样性,使神经网络不会过分依赖人脸图像的某个特征来区分图像的真假,从而提高模型的泛化能力和鲁棒性。

4.2.4 模型训练模块

计算关键帧图像的光流和 LBP 值,每帧的光流场分解为幅度与方向两个矩阵,与 LBP 的结果形成 224×224×3 的数

据阵列,定义为融合特征图,经数据增强处理后,输入 CNN 网络进行模型训练。算法采用基于图像训练学习的思想,即采用每轮 epoch 对正确分类的图像样本进行标记,若预测类别与视频真实类别一致的样本数超过 95%,则将此视频视为真视频,反之则为假视频。

5　实验结果及分析

5.1　实验环境参数

实验选取 FaceForensics++(FF++)[8] 和 Celeb-df[29] 数据集进行测试。实验平台使用 Win10 操作系统,显卡 NVIDIA GeForce RTX3090,CUDA 版本 11.6,运行环境为 Python3.7,Pytorch1.12.1。实验中,分别将 FF++和 Celeb-df 数据集的视频按照 8:2 的比例划分为训练集和验证集。从训练集中每个视频连续选取 60 帧进行关键帧检测,设置阈值 $T=mean(E(n))$,$\lambda=5$,则从每个视频中提取约 3～15 个关键帧。LBP 计算采用图 1 所示 3×3 正方形窗口的 8 邻域来描述图像的局部纹理特征。对于融合特征图,实验采用的数据增强操作如表 1 所列。模型训练过程中使用 Adam 优化器和交叉熵损失函数,初始学习率为 0.0001,Dropout 设置为0.5,batchsize 为 16,训练轮次 epochs 为 100。

表 1　数据增强处理说明

Table 1　Description of data augmentation

编号	处理方法	说明
1	剪切	随机裁剪
2	旋转	以一定概率对图像进行随机旋转,旋转角度在[-15°,15°]之间
3	翻转	以 30%的概率对图像进行水平翻转
4	标准化	进行标准化处理

5.2　实验结果及分析

5.2.1　CNN 模型选择的对比实验

为了验证算法的准确性和易扩展性,本文采用 XceptionNet[8],VGG16[9],ResNet50[30],MobileNet-V2[31] 以及 Efficientnet-V2[32] 这 5 种常见的主流 CNN 模型进行对比实验。实验中使用准确率(Accuracy)作为评价指标,其定义为分类正确的样本数与总样本数的比值,计算式为:

$$Acc=\frac{TP+TN}{TP+FP+FN+TN} \tag{10}$$

其中,真正 TP(True Positive)指真视频被预测为真的数量,真负 TN(True Negative)指假视频被预测为假的数量,假正 FP(False Positive)指假视频被预测为真的数量,假负 FN(False Negative)指真视频被预测为假的数量。

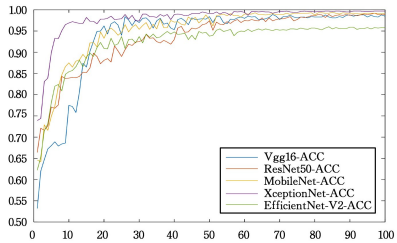
各模型检测结果如表 2 所列。在选取的 5 种主流 CNN 模型中,采用本文算法进行真伪检验的准确率明显优于传统检测算法,尤其在 FF++数据库中的检测率几乎达到 100%,主要原因在于本文算法集中对关键帧图像进行训练,可以提高模型训练输入数据的有效性,并且增加对纹理特征的考虑,反映了细微的伪造痕迹或者压缩误差,有利于对低质量的伪造人脸进行检测。而 Celeb-df 数据集的多样性和复杂性使得对抗样本的检测更具有挑战,故而 Celeb-df 数据集的检测率整体低于 FF++数据集的检测率。图 7、图 8 所示为各模型在不同数据库中训练的检测率和损失函数曲线对比图。在 FF++数据集中,5 种 CNN 网络结构的检测性能差异不大,但在 Celeb-df 数据集中,XceptionNet,Efficientnet-V2 作为骨干网络的性能优于其他网络结构,在 FF++和 Celeb-df 数据集中均能取得较高的准确率,且作为轻量级 CNN 模型的代表,Efficientnet-V2 具有比 XceptionNet 更少的参数,训练时间也更短。因此,综合考虑上述 5 种 CNN 结构进行真伪检测的性能,本文选取 XceptionNet,Efficientnet-V2 作为骨干网络进行后续的对比实验。

表 2　各方法在不同数据集上的检测结果

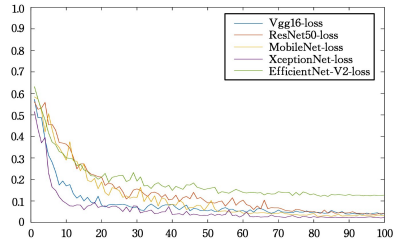
Table 2　Detection results of each model on different datasets

(%)

网络模型	FaceForensics++		Celeb-df	
	本文算法	传统算法	本文算法	传统算法
VGG16	99.9	81.6	88.7	73.4
ResNet50	100.0	92.3	90.7	76.9
MobileNet-V2	99.9	87.3	92.3	75.5
XceptionNet	100.0	91.3	96.9	89.1
Efficientnet-V2	95.7	91.9	98.8	82.5



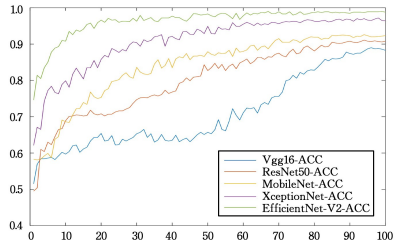
(a)检测率对比结果



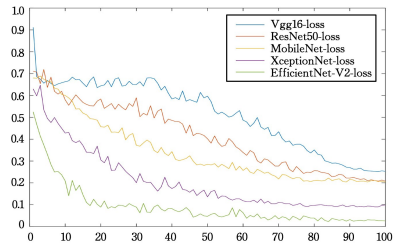
(b)损失函数对比结果

图 7　模型在 FF++数据集上的检测性能结果

Fig. 7　Detection performance of different models on FF++ dataset



(a)检测率对比结果



(b)损失函数对比结果

图 8　模型在 Celeb-df 数据集上的检测性能结果

Fig. 8　Detection performance of different models on Celeb-df dataset

5.2.2 与经典算法的对比实验

本文算法基于图像帧进行训练学习，如表3所列，与传统的图像级或视频级检测算法相比，本文算法的检测率具有明显的优势。其原因在于，传统的视频级检测方法虽然可以同时捕捉伪造视频时间和空间上的特征，但在CNN特征提取阶段丢失了一些低级特征，因此对效果有一定的影响，且计算复杂度较高。本文算法有针对性地增加了纹理特征的信息及关键帧处理，使得算法检测率表现出较高的优越性。

表3 与传统检测算法的结果对比

Table 3 Comparison of detection results between the proposed algorithm and traditional detection algorithms

(%)

	算法模型	FaceForensics++	Celeb-df
图像级检测方法	VGG16[9]	81.6	73.4
	ResNet50[30]	92.3	76.9
	Xception[8]	91.3	89.1
	胶囊网络[16]	93.6	61.7
	Efficientnet-V2[32]	97.9	82.5
	Keyframes+re-AM[21]	92.1	—
	SegCaps+胶囊网络[18]	97.3	98.9
	Cross Efficient ViT[33]	94.6	98.7
视频级检测方法	3D CNN[34]	93.7	81.6
	DenseNet+RCN[12]	93.2	81.3
	Inception-v3+LSTM[13]	95.9	77.3
本文算法	XceptionNet	100.0	96.9
	Efficientnet-V2	95.7	98.8

注："—"表示文献中未给出该项，且本文未进行复现。

5.2.3 泛化性实验

为了评估本文算法的泛化性，我们分别将在FF++和Celeb-df数据集上训练好的模型进行跨数据集检验，如表4所列，跨库检测性能比同库的检验率低。跨库性能普遍不理想的原因之一在于缺乏有效手段来保证网络学习仅受到伪造痕迹而非其他不相关信息的影响；此外，各数据集中所使用的伪造算法的差异也是导致跨库检测性能下降的原因。本文算法采用的Efficientnet-V2结构由于在浅层网络中引入了Fused-MBConv模块，可以更好地训练低质量人脸的特征信息，故在FF++数据库上表现出最优的跨库检测率，在Celeb-df数据集上的检测率与DenseNet+RCN和CFFs相当，比XceptionNet的检验率降低了12.6%。综合两个数据库上的跨库检测性能来看，基于XceptionNet结构的算法的整体性能优于其他方法，准确率均可达到80%以上，具有优越的泛化性能。

表4 各方法的跨库检测结果对比

Table 4 Detection results comparison of each method on cross-datasets

(%)

网络模型		FaceForensics++	Celeb-df
胶囊网络[16]		—	57.5
Keyframes+re-AM[21]		—	63.3
CFFs[19]		—	72.1
3D CNN[34]		—	70.1
DenseNet+RCN[12]		—	73.4
Inception-v3+LSTM[13]		—	52.7
本文算法	XceptionNet	85.9	84.9
	EfficientnetV2	90.1	72.3

注："—"表示文献中未给出该项，且本文未进行复现。

结束语 针对现有检测算法中模型训练输入数据不具有代表性、时空域信息提取特征不充分、检测结果可信度不高的问题，本文提出基于关键帧和时空融合特征进行模型训练的思想。在传统光流真伪检验算法的基础上，利用加权光流能量阈值分析对视频进行预处理，去除视频中不重要的图像帧，保留关键帧，使网络集中学习有意义的部分。之后通过改进的融合时空特征的"纹理-光流特征图"进行CNN模型训练，提高了算法检测的效率和准确性。在FF++数据集和Celeb-df两个数据集上进行了实验论证，该模型可以有效检测到深度伪造视频在时间和空间特征上的人脸差异，在FF++数据库中的检测准确率几乎达到100%，在Celeb-df数据库中的检测准确率最高可达98.8%以上。在跨数据集测试中，基于XceptionNet结构的算法检测率可达80%以上。实验结果表明，本文提出的优化模型具有较高的准确性和泛化性。后续工作将关注完善关键帧阈值的选择机制、优化网络结构和训练策略等，以进一步提升检测性能。

参 考 文 献

[1] LI X R, JI S L, WU C M, et al. Survey on Deepfakes and Detection Techniques[J]. Journal of Software, 2021, 32(2): 496-518.

[2] ZHANG Y X, LI G, CAO Y, et al. A Method for Detecting Human-face-tampered Videos based on Interframe Difference[J]. Journal of Cyber Security, 2020(2): 49-72.

[3] HAN B, HAN X G, ZHANG H, et al. Fighting Fake News: Two Stream Network for Deepfake Detection via Learnable SRM[J]. IEEE Transactions on Biometrics, Behavior, and Identity Science, 2021, 3(3): 320-331.

[4] QI H, GUO Q, XU J F, et al. DeepRhythm: Exposing DeepFakes with Attentional Visual Heartbeat Rhythms [C]// Proceedings of the 28th ACM International Conference on Multimedia. 2020: 1318-1327.

[5] JUNG T, KIM S, KIM K. DeepVision: Deepfakes Detection Using Human Eye Blinking Pattern [J]. IEEE Access, 2020, 8: 83144-83154.

[6] AGARWAL S, FARID H, GU Y M, et al. Protecting World Leaders Against Deep Fakes [C]// Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. 2019: 38-45.

[7] YANG X, LI Y Z, LYU S. Exposing Deep Fakes using Inconsistent Head Poses[C]// Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing. 2019: 8261-8265.

[8] ROSSLER A, COZZOLINO D, VERDOLIVA L, et al. FaceForensics++: Learning to Detect Manipulated Facial Images [C]// Proceedings of IEEE International Conference on Computer Vision. 2019: 1-11.

[9] AMERINI I, GALTERI L, CALDELLI R, et al. Deepfake Video Detection through Optical Flow based CNN[C]// Proceedings of International Conference on Computer Vision Workshop. 2019: 1205-1207.

[10] AKASH C, AISHWARYA R, SANIAT S, et al. Leveraging Edges and Optical Flow on Faces for Deepfake Detection[C]// Pro-

ceedings of IEEE/IAPR International Joint Conference on Biometrics. 2020.

[11] CHOLLET F. Xception:Deep Learning with Depthwise Separable Convolutions[C]// Proceedings of the 30th IEEE Conference on Computer Vision and Pattern Recognition. 2017:1800-1807.

[12] SABIR E,CHENG J X,JAISWAL A,et al. Recurrent Convolutional Strategies for Face Manipulation Detection in Videos [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. 2019:80-87.

[13] GUERA D,DELP E J. DeepfakeVideo Detection using Recurrent Neural Network[C]// Proceedings of the 15th IEEE International Conference on Advanced Video and Signal Based Surveillance. 2018:1-6.

[14] AFCHAR D,NOZICK V,YAMAGISHI J,et al. Mesonet:A Compact Facial Video Forgery Detection Network[C]// Proceedings of IEEE International Workshop on Information Forensics and Security. 2018:1-7.

[15] SZEGEDY C,VANHOUCKE V,LOFFE S,et al. Rethinking theInception Architecture for Computer Vision [C]// Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2016:2818-2826.

[16] NGUYEN H H,YAMAGISHI J,ECHIZEN I. Capsule-forensics:Using Capsule Networks to Detect Forged Images and Videos[C]// Proceedings of IEEE International Conference on Acoustics,Speech and Signal Processing. 2019:2307-2311.

[17] SIMONYAN K,ZISSERMAN A. Very Deep Convolutional Networks for Large-scale Image Recognition[C]// Proceedings of the 3rd International Conference on Learning Representations. 2015.

[18] ARASH H,NIMAJAFARI N,HASAN D,et al. A Novel Blockchain-based Deepfake Detection Method using Federated and Deep Learning Models[J]. Cognitive Computation,2024,16(3):1073-1091.

[19] YU P P,FEI J W,XIA Z H,et al. Improving Generalization by Commonality Learning in Face Forgery Detection[J]. IEEE Transactions on Information Forensics and Security,2022(17):547-558.

[20] XING H,LI M. Deepfake Video Detection based on 3D Convolutional Neural Networks[J]. Computer Science, 2021,48(7):86-92.

[21] WANG T Y,CHENG H,CHOW K P,et al. Deep Convolutional Pooling Transformer for Deepfake Detection[J]. ACM Transactions on Multimedia Computing,Communications,and Applications,2023,19(6):1-20.

[22] HSU C C,ZHUANG Y X,LEE C Y. Deep Fake Image Detection based on Pairwise Learning[J]. Applied Sciences,2020,10(1):370.

[23] DANG H,LIU F,STEHOUWER J,et al. On the Detection of Digital Face Manipulation[C]// Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2020:5780-5789.

[24] RAHMOUNI N,NOZICK V,YAMAGISHI J,et al. Distinguishing Computer Graphics from Natural Images using Convolution Neural Networks[C]// Proceedings of the IEEE Workshop on Information Forensics and Security. 2017:1-6.

[25] ZHU S H,HU J J,SHI Z. LocalAbnormal Behavior Detection based on Optical Flow and Spatio-temporal Gradient[J]. Multimedia Tools and Applications,2016,75(15):9445-9459.

[26] FU B,LI W H,CHEN B,et al. Abnormal Behavior Detection based on Weighted Energy of Optical Flow[J]. Journal of Jilin University(Engineering and Technology Edition),2013,43(6):1644-1649.

[27] ZHANG K P,ZHANG Z P,LI Z F,et al. Joint Face Detection and Alignment using Multitask Cascaded Convolutional Networks[J]. IEEE Signal Processing Letters,2016,23(10):1499-1503.

[28] BATTITI R,AMALDI E,KOCH C. Computing Optical Flow Across Multiple Scales:An adaptive coarse-to-fine strategy[J]. International Journal of Computer Vision,1991,6(2):133-145.

[29] LI Y Z,YANG X,SUN P,et al. Celeb-df:ALarge-scale Challenging Dataset for Deepfake Forensics[C]// Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2020:3204-3213.

[30] LI Y Z,LYU S W. Exposing Deepfake Videos by Detecting Face Warping Artifacts[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition Workshops. 2019:46-52.

[31] SANDLER M,HOWARD A,ZHU M L,et al. Mobilenetv2:Inverted Residuals and Linear Bottlenecks[C]// Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2018:4510-4520.

[32] DENG L W,SUO H F,LI D J. Deepfake Video Detection based on EfficientNet-V2 Network[J]. Computational Intelligence and Neuroscience, 2022: 1-13. https://doi. org/10. 1155/2022/3441549.

[33] COCCOMINI D A,MESSINA N,GENNARO C,et al. Combining EfficientNet and Vision Transformers for Video Deepfake Detection[C]// Proceedings of the 21st International Conference on Image Analysis and Processing. 2022:219-229.

[34] WANG Y H,DANTCHEVA A. AVideo is Worth More than 1000 Lies. Comparing 3DCNN Approaches for Detecting Deepfake [C]// Proceedings of the 15th IEEE International Conference on Automatic Face and Gesture Recognition. 2020:515-519.

CHENG Yan, born in 1978, Ph.D, associate professor. Her main research interests include image/video forensic and artificial intelligence.

(责任编辑:杨雪敏)