基于图像分割网络的深度假脸视频篡改检测

胡永健*①② 高逸飞① 刘琲贝① 廖广军③

①(华南理工大学电子与信息学院 广州 510641)

②(中新国际联合研究院 广州 511356)

③(广东警官学院 广州 510230)

摘 要：随着深度学习技术的快速发展，利用深度神经网络模型伪造出的深度假脸(deepfake)视频越来越逼真，假脸视频造成的威胁也越来越大。文献中已出现一些基于卷积神经网络的换脸视频检测算法，他们在库内获得较好的检测效果，但跨库检测性能急剧下降，存在泛化能力不足的问题。该文从假脸篡改的机制出发，将视频换脸视为特殊的拼接篡改问题，利用流行的神经分割网络首先预测篡改区域，得到预测掩膜概率图，去噪并二值化，然后根据换脸主要发生在人脸区域的前提，提出一种计算人脸交并比的新方法，并进一步根据换脸处理的先验知识改进人脸交并比的计算，将其作为篡改检测的分类准则。所提出方法分别在3个不同的基础分割网络上实现，并在TIMIT, FaceForensics++, FFW数据库上进行了实验，与文献中流行的同类方法相比，在保持库内检测的高准确率同时，跨库检测的平均错误率显著下降。在近期发布的合成质量较高的DFD数据库上也获得了很好的检测性能，充分证明了所提出方法的有效性和通用性。

关键词：假脸视频；图像分割网络；人脸交并比；信任机制；泛化能力

中图分类号：TN911.73 **文献标识码**：A **文章编号**：1009-5896(2021)01-0162-09

DOI: 10.11999/JEIT200077

Deepfake Videos Detection Based on Image Segmentation with Deep Neural Networks

HU Yongjian①② GAO Yifei① LIU Beibei① LIAO Guangjun③

①(*School of Electronic and Information Engineering, South China University of Technology, Guangzhou* 510641, *China*)

②(*Sino-Singapore International Joint Research Institute, Guangzhou* 511356, *China*)

③(*Guangdong Police College, Guangzhou* 510230, *China*)

Abstract: With the rapid development of deep learning technology, videos with changed faces generated by deep neural networks (i.e., Deepfake videos) become more and more indistinguishable. As a result, the threat raised by Deepfake videos becomes greater and greater. In literature, there are some convolutional neural networks-based detection algorithms for fake face videos. Although those algorithms perform well when the training set and the testing set are from the same dataset, their performance could deteriorate dramatically in cross-dataset scenario where the training and the testing sets are from different sources. Motivated by the fabrication course of fake face videos, this article attempts to solve the problem of fake faces detection with the way of image splicing detection. A neural network borrowed from image segmentation is adopted for predicting

收稿日期：2020-01-17；改回日期：2020-07-10；网络出版：2020-07-22

*通信作者： 胡永健 eeyjhu@scut.edu.cn

基金项目：国家重点研发计划项目(2019QY2202)，广州市开发区国际合作项目(201902010028)，中新国际联合研究院项目(206-A017023, 206-A018001)，广东省自然科学基金博士科研启动项目(2017A030310320)，中央高校基本科研业务费专项资金(2019MS025)，广东省教育厅特色创新类项目(2017KTSCX132)

Foundation Items: The National Key R & D Program (2019QY2202), The International Cooperation Project of Guangzhou Development Zone (201902010028), The Sino Singapore International Joint Research Institute Project (206-A017023, 206-A018001), The Doctoral Research Project of Natural Science Foundation of Guangdong Province (2017A030310320), The Special Fund for Basic Scientific Research of Central University (2019MS025), The Department of Education of Guangdong Province Characteristic Innovation Project (2017KTSCX132)

the tampered face area from which a tampering mask is obtained through denoising and thresholding the probability map. Using the prior knowledge of face tampering that the changing of face mainly happens in face region, a new way is proposed to determine the Face-Intersection over Union (Face-IoU) and to further improve the ratio calculation method. The Face-Intersection over Union with Penalty (Face-IoUP) is used as the classification criterion for deepfake video detection. The proposed method is impletmented using three basic image segmentation neural networks separately and is tested them on datasets of TIMIT, FaceForensics++, Fake Face in the Wild(FFW). Compared with current methods in literature, the HTER (Half Total Error Rate) in cross-dataset test decreases significantly while the detection accuracy in intra-dataset test keeps high. For the Deep Fake Detection(DFD) dataset with higher synthesis quality, the proposed method still performs very well. Experimental results validate the proposed method and demonstrate its good generality.

Key words: Deepfake videos; Image segmentation networks; Face-Intersection over Union(Face-IoU); Confidence mechanism; Generalization

1　引言

在众多的生物特征中，人脸是最具有代表性的特征之一，可辨识度较高。因此，随着人脸识别技术的突飞猛进，人脸篡改所带来的安全威胁越来越大，特别是在手机高度流行和社交网络日益成熟的当代。虚假新闻、丑闻、名人色情视频以及报复性的色情视频在各种网络社区上涌出，困扰着从政治家、名人到普通民众，甚至威胁到国家安全。目前，有3种常见的假脸视频篡改技术[1]：Face2Face, Faceswap和Deepfakes[2]。Face2Face通过重建3维模型只对面部表情进行控制，而Faceswap和Deepfakes则将源视频中的整张人脸替换到目标视频中，基于传统计算机图形学方法的技术统称为Faceswap，而基于深度学习的技术统称为Deepfakes。Deepfakes主要利用深度学习中的深度卷积神经网络、自编码器(auto-encoder)和对抗生成网络(Generative Adversarial Networks, GANs)等技术，使网络可以学习到目标视频中更深层次的人脸特征，从而实现精准替换，并且能够匹配目标视频中人脸的动作和变化，达到较为理想的换脸效果。Deepfakes有时直接指利用深度网络换脸所得到的视频。

假脸视频对社会的威胁日益增加，引起了学术界和工业界的广泛关注，已经出现了一些相关研究，甚至出现针对换脸视频检测的国际大赛[1)]。根据所使用的特征，现有的假脸视频检测技术大致分为3大类：基于传统手工特征、基于生物特征和基于神经网络提取特征。第1类方法主要借鉴了图像取证思想，对单帧图像进行分析，典型做法包括使用图像质量测度和主成分分析[2]以及使用局部二值模式(Local Binary Pattern, LBP)特征[3]。第2类方法主要利用了人脸特有的生物信息，文献[4]将脸部标志点根据篡改过程时的受影响程度大小分为两组，使用不同标志点估计出头部姿态方向后比较差异作为判别依据；文献[5]发现假脸中两只眼睛的瞳孔部位呈现出的漫反射信息存在不一致的现象；文献[6]同时利用视频图像和音频的信息，通过比较真假视频中唇部动作和声音匹配的差异甄别有无篡改；文献[7]指出每一个人都有独特的运动动作模式，而换脸导致目标对象与源对象动作模式的不匹配，可从额头、脸颊、鼻子等区域的运动变化中提取特征进行分类判决。第3类方法主要通过构建卷积神经网络对人脸进行学习，提取较高维度的语义特征进行分类。一些研究者将其视为常规的分类问题，文献[3]使用了AlextNet,VGG-19,ResNet[8],Inception和Xception[9]等用于图像识别的经典分类模型进行检测；文献[10]搭建了Meso-4和MesoInception-4，文献[11]搭建了ShallowNet对单帧图像进行检测；文献[12]指出在篡改检测问题中篡改痕迹比图像内容信息更加重要，据此构建了带有受约束卷积层的MISLnet，在提取浅层特征时抑制图像内容；文献[13]考虑视频中的时域信息，将卷积神经网络与序列神经网络结合，寻找假脸视频中连续帧特征的不一致性；文献[14]使用ResNet-50[8]网络模型对不同GAN合成图像和Deepfake假脸图像进行检测。

从以上研究给出的实验结果看，基于神经网络提取特征的算法往往能在库内检测中获得较高的准确率，但其主要缺陷在于跨库检测时性能均急剧下降，存在泛化性能不足的问题[3]。

本文针对换脸视频检测网络泛化性能不足的问题提出一种解决方案。与上述基于特征检测的方法不同，本文直接从图像像素出发，认为假脸篡改是

1) DeepFake Detection Challenge: < https://www.kaggle.com/c/deepfake-detection-challenge>

一种特殊的拼接篡改问题，根据换脸主要对部分人脸区域进行操作而未修改其他图像内容的事实，提出了利用预测篡改区域和人脸框的交并比作为是否发生换脸的依据进行检测的方法。主要贡献包括：(1)利用图像分割网络逐像素地估计篡改区域；(2)解决盲检测时没有真实篡改区域作为参考信号的不足，提出一种计算人脸交并比(Face-Intersection of Union, Face-IoU)的新方法，作为是否发生换脸篡改的判断依据；(3)根据换脸视频的先验知识，构建信任机制，进一步改进Face-IoU的计算，得到Face-IoUP(Face-IoU with Penalty)。本文分别基于FCN-8s, FCN-32s[15]和DeepLabv3[16]3个分割网络实现检测算法，实验结果表明，本文检测模型不仅在库内具有很高的准确率，在跨库检测中，平均错误率与现有流行的方法相比也有显著下降。

2 算法介绍

本文算法整体流程如表1和表2的伪代码所示，包括网络训练和样本测试。网络训练部分利用训练集训练分割模型，利用验证集计算最小等错误率时的二值化阈值和判决阈值；样本测试部分对分帧预处理后的图像分割得到掩膜，平滑滤波去除掩膜噪点，二值化处理得到预测篡改区域的位置和面积信息，以待测视频的人脸区域为参考信号计算面部交并比，并根据换脸视频相关的先验知识建立信任机制改进面部交并比的计算，作为最终的判决依据，完成检测。下面对关键环节进行描述。

表 1 网络训练

输入：训练集数据X，验证集数据Z
输出：训练好的分割网络模型Model、二值化阈值T_1^*和判决阈值T_2^*
(1) Begin(算法开始)
(2) 初始化Model的权重
(3) 将X输入Model中进行训练，更新得到训练好的权重模型
(4) 将Z输入Model中，计算最小等错误率下的T_1^*和T_2^*
(5) End(算法结束)

表 2 样本测试

输入：测试样本视频V，训练好的分割网络模型Model、二值化阈值T_1^*和判决阈值T_2^*
输出：检测结果Y
(1) Begin(算法开始)
(2) 对V分帧并定位裁剪出人脸区域，得到$I=\{I_1,I_2,\cdots,I_Q\}$。
(3) For q=1 to Q do:
(4) 将I_q输入Model，得到预测篡改区域掩膜M_q
(5) 对M_q滤波得到MF_q
(6) 根据T_1^*对MF_q二值化，得到MB_q
(7) 对MB_q计算$\mathrm{Face}-\mathrm{IoUP}_q$
(8) 根据T_2^*对$\mathrm{Face}-\mathrm{IoUP}_q$进行二分类判决，得到$y_q$
(9) end For
(10) End(算法结束)

2.1 利用深度分割网络得到篡改区域的预测掩膜概率图

传统的图像拼接篡改通常是指将非同源图像不作任何修饰地粘贴到目标图像上的篡改技术[17]。一般而言，视频换脸和传统拼接篡改的共同点在于两者都使用非同源的图像对目标图像的部分区域进行替换，篡改方式类似；其不同点在于后者使用的人脸大都来自于真实图像，而前者所使用的人脸可能是通过计算机图形学技术或深度网络生成。虽然人眼难以区分这两类图像，但这两类图像的内部结构和纹理特征存在差异，其在颜色空间特征[18]和共生矩阵[19]上有所反映。

鉴于此，本文将换脸篡改认为是一种特殊的拼接篡改问题，借助像素级的精确标签让神经网络按纹理差异区分真假脸像素点，从而将来源不同的两种图像分割成两类不同的图像区域。以语义分割FCN网络[15,20]为例，分割网络首先借助诸如VGG-16的分类网络作为骨架网络提取特征，然后利用跨层架构将来自浅且精细网络层的表象信息与来自深且粗糙网络层的语义信息相结合，即表征图像内部结构差异的低水平特征和表征边缘不连续的深层特征相结合，逐像素地对图像进行来源判断，最终输出值在0–1之间的篡改区域预测掩膜概率图M，其尺寸与输入图像大小一致。

2.2 去噪与二值化处理获得预测的篡改区域

尽管分割网络能够预测出像素级的掩膜概率图，但由于提取特征过程中的卷积和池化处理使深层语义特征的感受野越来越大，在转化为特征表达图像时，跨层结构的上采样处理不可避免地引入噪声，造成预测掩膜概率图出现细小孤立的噪声点。为了获得准确的预测掩膜概率图，便于机器自动计算篡改面积，本文利用传统的图像去噪方法对概率图去噪，利用邻域信息抑制孤立噪声点，得到去噪后的预测掩膜概率图MF。

去噪后的预测掩膜概率图需要进行二值化处理后才能逐像素计算面积。一种直接取二值化门限的方法是使其等于0.5，然而这种方法过于简单，无法获得最佳的分类效果。本文将二值化处理和篡改判断相结合，提出一种等错误率最小时获取最优二值化门限的方法，详见2.5节。设所得到的二值化阈值为T_1，当预测概率大于T_1时置为1(即篡改)，否则置为0(即未篡改)。具体公式为

$$\mathrm{MB}(x,y)=\begin{cases}0, & \mathrm{MF}(x,y)<T_1\\ 1, & \mathrm{MF}(x,y)\geq T_1\end{cases} \tag{1}$$

式中$\mathrm{MF}(x,y)$表示去噪后预测掩膜概率图中(x,y)点处的值，$\mathrm{MB}(x,y)$表示对应点二值化后的0/1值，二值化掩膜MB即是预测的篡改区域。

2.3 人脸交并比Face-IoU的提出

图1(a)、图1(b)和图1(c)分别展示包含人脸框的待检测区域示例图、对应的实际篡改区域示例图和对应的预测篡改区域示例图，图1(d)是一张广义的示意图。其中黑色方框为整体待检测区域；绿色方框代表人脸框，记为S_1；蓝色线条包围的区域为预测篡改区域，记为S_2；红色线条包围的区域为实际或称真实的篡改区域，记为S_3。

在传统的分割问题中，交并比是以真实篡改区域为基准(即参考信号)，衡量预测区域的准确性。以图1(d)为例，交并比的计算式为$(S_2\cap S_3)/(S_2\cup S_3)$。然而，在本文的应用场景中，此式并不能直接作为盲分类判决的依据，主要是因为：第一，若输入图像是换脸图像，盲检测器可以通过分割网络预测得到篡改区域S_2,但无法得到真实的参考信号S_3，因此，无法计算$(S_2\cap S_3)/(S_2\cup S_3)$；第二，若输入图像是未篡改图像，即$S_3=0$，则不论预测篡改区域$S_2$多大，交并比$(S_2\cap S_3)/(S_2\cup S_3)$的计算始终为0。

考虑到换脸篡改的前提是图像中存在人脸区域S_1，若不存在S_1就不会存在S_3及其预测。因此，将S_1作为参考信号无论对篡改图像还是未篡改图像均合理，在盲检测中能够提供有效的计算。据此，本文直接将人脸区域S_1替代S_3得到Face-IoU计算式为

$$\text{Face}-\text{IoU}=(S_1\cap S_2)/(S_1\cup S_2) \tag{2}$$

2.4 信任机制下改进的人脸交并比Face-IoUP

式(2)尽管解决了交并比的计算问题，但由常识可知，换脸篡改不可能总是一个规则的正方形，事实上，图1(d)的示意图已经显示，无论是S_3还是S_2都有可能落在人脸区域之外。考虑到S_3无法得到，为了修正式(2)的偏差，只能从S_1和S_2的关系着手。考虑到大部分实际的篡改区域都出现在人脸区域这个事实，本文以此作为重要的先验知识，构建了信任机制。本文认为预测篡改区域落在$S_1\cap S_2$是合理的，而S_2落在人脸区域之外的部分信任程度低一些，即对预测篡改区域$S_1\cup S_2-S_1$产生一定的不信任，认为分割网络的预测结果中出现了一定程度的虚警，据此，提出信任机制，改进Face-IoU的计算。具体而言，在分母项中增加以p为惩罚因子的惩罚项，得到Face-IoUP为

$$\begin{aligned}\text{Face}-\text{IoUP}=&(S_1\cap S_2)/((S_1\cup S_2)\\&+p\times(S_1\cup S_2-S_1))\end{aligned} \tag{3}$$

式(3)显示，落在人脸区域外的预测篡改像素越多，$S_1\cup S_2-S_1$越大，惩罚项越大，则Face-IoUP越小，这张图像被判决成假脸图像的概率越小，从而有效地降低了虚警。

Face-IoUP能够在训练和测试阶段有效描述预测的篡改区域，本文以此作为依据，设置判决阈值T_2，当Face-IoUP大于T_2时，判定为假脸图像，否则为真脸图像。判决公式为

$$y=\begin{cases}0, & \text{Face}-\text{IoUP}<T_2\\ 1, & \text{Face}-\text{IoUP}\geq T_2\end{cases} \tag{4}$$

2.5 最优二值化阈值T_1^*和判决阈值T_2^*的获取

为了获得最优二值化阈值T_1^*和判决阈值T_2^*，依据式(1)和式(4)，在验证集上采取网格联合搜索，以0.001为步长在0～1的范围内搜索T_1和T_2，记录验证集上满足虚警率和漏检率相同的阈值对，最终选取等错误率最小时的阈值对作为T_1^*和T_2^*。

3 实验场景设置

本文在4个常见的深度假脸视频数据库上进行了实验，分别为TIMIT[2], FaceForensics++[1], FFW[3]和DFD[1]。对于TIMIT, FaceForensics++以及DFD数据库，类似文献[21]，以按人划分的准则依7:2:1的比例将它们分为训练集、验证集和测试集。而对FFW数据库，由于只有假脸视频，正负样本不平衡而无法进行全面的评估，本文从Face-

(a) 待检测区域示例图　(b) 实际篡改区域示例图　(c) 预测篡改区域示例图　(d) 广义示意图

图 1 待检测区域、实际篡改区域和预测篡改区域示例及其广义示意图

Forensics++数据库的测试集中随机选取了50段真实视频进行补充。

为了有效评估算法的学习能力和泛化能力，实验分库内检测和跨库检测两个内容。由于FFW数据库中补充了FaceForensics++数据库的视频，为了不影响性能评测，在跨库检测中不使用FaceForensics++数据库训练的模型测试FFW数据库。

为了获得统一尺寸的输入图像，将各数据库中的视频分帧，逐帧使用Dlib库中卷积神经网络检测器定位人脸，以人脸框为中心取$k(k>1)$倍于人脸大小的图像区域，采样至256×256的尺寸，作为输入图像。可以看到，$k-1$倍的图像区域(即人脸框周边区域)是背景。需要指出的是，k值太小，惩罚项的作用就小；k值太大，会包含一些远离人脸的像素，不符合换脸的事实。本文统一取经验值k=1.3，得到的输入图像即为图1(a)。

4 去噪滤波器的选取和惩罚因子p的确定

检测模型的性能与预测掩膜概率图的去噪程度以及落在人脸框外的预测像素的可信程度有直接关系。本小节在TIMIT，FaceForensics++和FFW这3个数据库中以FCN-8s和FCN-32s这两个分割网络为例，分别讨论去噪滤波器和惩罚因子的作用。

4.1 去噪滤波器的选取

二值化预测掩膜上每一点处的{0,1}值都会影响交并比的计算，因此，必须对预测掩膜概率图进行去噪预处理，排除噪声点引起的交并比计算误差。表3列出未滤波和分别采用均值、中值以及高斯滤波这4种情形下检测模型的平均错误率。滤波器核的大小分别为3×3和5×5。

表3显示，总的来看，去噪处理有助于改善检测模型的性能。而在均值、中值和高斯这3种滤波器中，以核大小为3×3的高斯滤波器对降低检测模型的平均误差率效果最好。本文在不作特殊声明的场合均采用核大小为3×3的高斯滤波器。

4.2 信任机制中惩罚因子p的选取

本文信任机制是由式(3)分母中的惩罚项来体现，惩罚强度会直接影响分类的性能。本小节将惩罚因子p分别设置为0(无惩罚)，0.5, 1.0和1.5进行实验，结果如表4所示。

由表4可知，与不使用惩罚项相比，有惩罚时两个分割网络的库内和跨库检测错误率总的来说是有所降低，且在p=1.0时效果最好。本文在不作特殊声明的场合均采用p=1.0。

5 实验结果

换脸视频检测目前并无统一比较基准。为展现所提出检测模型的良好性能，本文以同类且较新的4个检测网络MesoInception-4[10]，ShallowNetV1[11]，MISLnet[12]，ResNet-50[8,14]和Xception[9]作为对象进行比较。为了展示本文算法良好的可扩展性，除用前述FCN-8s和FCN-32s[15]，还以DeepLabv3[16]分割网络作为分割手段，实现本文算法。在TIMIT，FaceForensices++和FFW数据库上的实验结果见表5和表6，可视化结果见图2，其中热力图红色部分表示篡改区域，蓝色部分表示未篡改区域。

表 3 检测模型在不同滤波器下的平均错误率(%) p=1

网络	训练数据库 测试数据库 滤波器类型	核大小	TIMIT			FaceForensics++	
			TIMIT(库内)	FaceForensics++(跨库)	FFW(跨库)	FaceForensics++(库内)	TIMIT(跨库)
FCN-8s	无	无	2.5	23.2	24.4	2.2	24.1
	均值	3×3	2.6	23.4	23.0	2.1	24.8
		5×5	2.5	23.2	22.9	2.3	25.2
	中值	3×3	2.6	23.1	22.9	2.1	25.3
		5×5	2.6	22.9	23.4	2.2	24.0
	高斯	3×3	2.4	22.7	22.9	1.8	22.6
		5×5	2.5	23.2	22.9	1.9	24.8
FCN-32s	无	无	5.8	27.2	20.8	1.9	29.2
	均值	3×3	5.8	26.0	20.1	1.9	29.6
		5×5	5.2	26.3	20.4	1.9	29.7
	中值	3×3	5.9	27.4	20.7	1.9	30.4
		5×5	5.6	27.0	20.3	1.8	29.7
	高斯	3×3	5.7	26.8	20.5	1.8	27.5
		5×5	6.0	27.0	20.7	1.7	30.4

表 4　检测模型在不同惩罚因子下的平均错误率(%)

训练数据库		TIMIT			FaceForensics++	
测试数据库		TIMIT(库内)	FaceForensics++(跨库)	FFW(跨库)	FaceForensics++(库内)	TIMIT(跨库)
网络	惩罚因子					
FCN-8s	0	2.5	23.6	23.3	1.9	24.3
	0.5	2.5	23.1	23.5	1.8	24.5
	1.0	2.4	22.7	22.9	1.8	22.6
	1.5	2.5	22.7	23.1	1.9	23.7
FCN-32s	0	6.0	27.2	20.5	2.2	29.8
	0.5	5.8	27.0	20.8	1.9	30.6
	1.0	5.7	26.8	20.5	1.8	27.5
	1.5	5.9	27.2	20.6	1.8	29.6

表 5　以TIMIT数据库训练模型所得到的测试结果(%)

测试数据库	TIMIT(库内)			FaceForensics++(跨库)	FFW(跨库)
网络	等错误率	平均错误率	准确率	平均错误率	平均错误率
MesoInception-4[10]	11.2	14.4	86.1	37.7	40.1
ShallowNetV1[11]	1.4	4.3	95.8	38.2	42.3
MISLnet[12]	5.4	5.2	94.8	30.3	41.0
ResNet-50[8,14]	0.8	2.5	97.6	44.9	45.7
Xception[9]	1.6	2.4	97.8	35.4	35.7
FCN-8s(本文算法)	4.0	2.4	97.7	22.7	22.9
FCN-32s(本文算法)	6.2	5.7	94.4	26.8	20.5
DeepLabv3(本文算法)	1.1	3.7	96.4	30.0	25.0

表 6　以FaceForensics++数据库训练模型所得到的测试结果(%)

测试数据库	FaceForensics++(库内)			TIMIT(跨库)
网络	等错误率	平均错误率	准确率	平均错误率
MesoInception-4[10]	4.6	5.6	94.4	28.2
ShallowNetV1[11]	0.8	2.1	96.4	35.1
MISLnet[12]	3.0	3.5	96.4	19.3
ResNet-50[8,14]	2.8	3.5	96.4	38.3
FCN-8s(本文算法)	2.1	1.8	98.2	22.6
FCN-32s(本文算法)	1.0	1.8	98.3	27.5
DeepLabv3(本文算法)	0.8	1.0	99.0	22.5

表5显示，基于FCN-8s和FCN-32s分割网络的检测模型在TIMIT的库内检测均有较好准确率，尤其是基于FCN-8s的模型，库内准确率位于次高，平均错误率位于并列最低；其在FaceForensics++中的平均错误率比目前文献算法降低超过12%，位于最低；其在FFW库中的平均错误率比目前文献算法降低超过12%，仅高于本文基于FCN-32s的模型，综合表现最佳，充分证明本文基于分割网络进行换脸视频检测的有效性和优良的泛化性能。基于DeepLabv3检测模型的良好表现也证明所提出方法具有很好的可扩展性，其在库内的等错误率只有1.1%。表6的结果趋势大致与表5的类似。

图2进一步给出基于FCN-8s分割网络的检测模型在FaceForensics++数据库上检测结果示例图。第1行为针对换脸视频的检测结果，从左到右依次为：假脸视频帧，输入图像的正确热力图，通过FaceForensics++库训练模型后得到的预测热力图(即库内检测热力图)，通过TIMIT训练模型后得到的预测热力图(即跨库检测热力图)。可以看到，本文算法无论在库内还是库外都能够较精确地检测出

输入图像的篡改区域。第2行为针对真实为篡改视频的检测结果，从左到右分别为：真脸视频帧，输入图像的正确热力图，通过FaceForensics++库训练模型后得到的库内检测热力图，通过TIMIT训练模型后得到的跨库检测热力图。

图2结果显示，本文算法无论在库内还是库外都能正确判断输入图像中的篡改区域，证明该算法的有效性。尤其在假脸图像中，能够有效区分篡改区域和真实区域，通过像素级的精确标签让网络学习到纹理像素点之间的差异而降低对内容信息的依赖，进而降低了在单一数据库上训练造成的过拟合现象，提升了跨库测试的性能。

针对由演员表演为素材，假脸合成质量普遍较高的DFD数据库，表7给出了本文基于FCN-8s和FCN-32s两个分割网络的部分测试结果。DFD数据库中包括无压缩库(C0)、压缩比为23(C23)的高质量视频库，以及压缩比为40(C40)的低质量视频库。检测模型在C23视频库上训练，得到的库内检测结果以及在TIMIT, FaceForensics++(C0和C23)和FFW这3个数据库上的跨库检测结果。可以看到本文算法有相当突出的表现，库内检测的平均错误率低于2%，且跨库检测的平均错误率比上述跨库检测的结果更低。

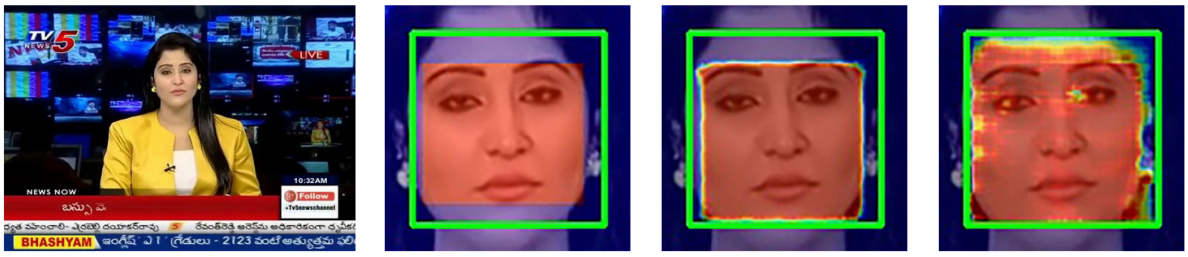
在实际检测中存在多人脸情况，由于本文算法逐像素判断当前像素是落在真脸区域还是假脸区域，因此检测过程并不受人脸数目的影响，只与标签图像有关。限于文章篇幅，此处仅给出FFW测试库中(4hMa-gKljhw_0.000_6.773.avi)的检测结果，该段视频中的右边人脸为假脸。图3显示，可以看到本文算法同样可以有效检测出假脸。

5.1 算法时间对比

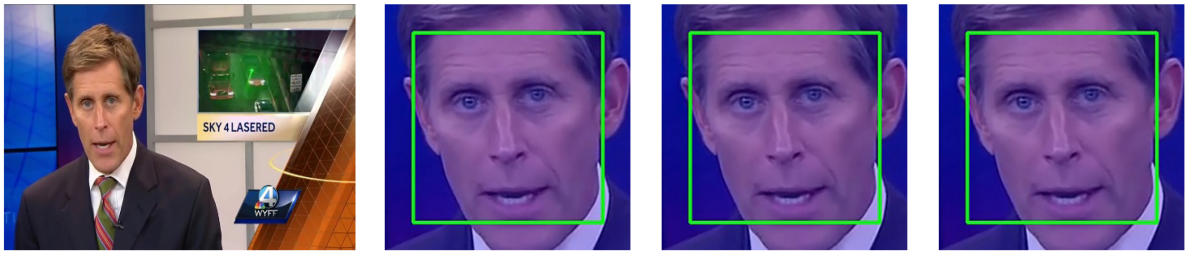
本小节讨论各算法计算复杂度，按浮点运算次数(FLOPs)和检测时长(Time)进行对比。检测时长包括了对100段10 s视频的逐帧检测，每一视频检测过程均包括“分帧—人脸定位—人脸检测—结果判决”4个步骤。结果如表8所示。由于分割网络是对每个像素进行判断，其网络结构与其他分类器相比增加了上采样部分，因此FLOPs和运行总时长均有所增加。以FCN-8s实现的本算法为例，测试一个10 s视频的平均时长为37.8 s，与最快的MesoInception-4相比，时长增加了约0.5倍，应在可接受的范围之内。通过牺牲少量时长来提高检测精度在很多应用场景下都是有意义的。

6 结束语

目前流行的假脸视频检测算法大多利用深度网络提取特征进行，这类方法跨库性能欠佳的主



(a) 假脸视频帧示例图 (b) 假脸正确热力图 (c) 假脸库内测检测热力图 (d) 假脸跨库检测热力图



(e) 真脸视频帧示例图 (f) 真脸正确热力图 (g) 真脸库内检测热力图 (h) 真脸跨库检测热力图

图 2 FaceForensics++数据库视频检测结果示例图

表 7 通过DFD的C23数据库训练模型所得到的平均错误率(%)

测试数据库	DFD(C23)(库内)	TIMIT(跨库)	FaceForensics++(C0)(跨库)	FaceForensics++(C23)(跨库)	FFW(跨库)
FCN-8s(本文算法)	1.7	15.9	14.2	16.9	21.5
FCN-32s(本文算法)	1.9	17.9	7.9	11.4	20.2

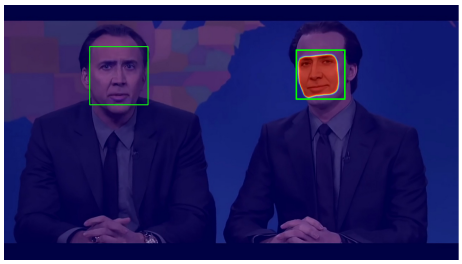


图 3 同时含有真脸和假脸的检测热力图示例

表 8 算法复杂度与时间对比

网络	FLOPs(M)	时长(s/100段视频)
MesoInception-4[10]	0.5	2402.6
ShallowNetV1[11]	65.1	2535.1
MISLnet[12]	31.5	2351.8
ResNet-50[8,14]	47.3	2680.5
Xception[9]	41.9	2564.3
FCN-8s(本文算法)	268.5	3780.0
FCN-32s(本文算法)	268.5	3782.9
DeepLabv3(本文算法)	37.7	2510.1

要原因是深度网络容易学到过多的库内特征，导致泛化性能不好。与文献中的方法不同，本文将假脸视频检测视为一个特殊的拼接篡改检测问题，利用图像分割网络逐像素对篡改区域进行预测，降低不同训练数据库的影响，提高检测算法的泛化性能。此外，利用去噪、优化的二值化门限和根据换脸先验知识改进的人脸交并比等措施提高检测的准确性。在多个流行换脸视频测试库上的实验结果表明，与其他同类算法相比，本文方法在库内检测保持高准确率的同时大幅降低了跨库检测平均错误率，算法具有很好的通用性。本文方法在不同分割网络的实现均能获得优良的假脸视频检测性能，说明本文提高泛化性能的思想具有一般性。将来的改进方向包括解决侧脸人脸框的确定、不同尺寸人脸的精确检测以及优化分割网络模型等方面。

参 考 文 献

[1] RÖSSLER A, COZZOLINO D, VERDOLIVA L, *et al*. FaceForensics++: Learning to detect manipulated facial images[C]. 2019 IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 2019: 1–11. doi: 10.1109/iccv.2019.00009.

[2] KORSHUNOV P and MARCEL S. DeepFakes: A new threat to face recognition? Assessment and detection[EB/OL]. https://arxiv.org/abs/1812.08685, 2018.

[3] KHODABAKHSH A, RAMACHANDRA R, RAJA K, *et al*. Fake face detection methods: Can they be generalized?[C]. 2018 International Conference of the Biometrics Special Interest Group, Darmstadt, Germany, 2018: 1–6. doi: 10.23919/BIOSIG.2018.8553251.

[4] YANG Xin, LI Yuezun, and LÜ Siwei. Exposing deep fakes using inconsistent head poses[C]. ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing, Brighton, England, 2019: 8261–8265. doi: 10.1109/icassp.2019.8683164.

[5] MATERN F, RIESS C, and STAMMINGER M. Exploiting visual artifacts to expose deepfakes and face manipulations[C]. 2019 IEEE Winter Applications of Computer Vision Workshops, Waikoloa Village, USA, 2019: 83–92. doi: 10.1109/WACVW.2019.00020.

[6] KORSHUNOV P and MARCEL S. Speaker inconsistency detection in tampered video[C]. The 26th European Signal Processing Conference, Rome, Italy, 2018: 2375–2379. doi: 10.23919/eusipco.2018.8553270.

[7] AGARWAL S, FARID H, GU Yuming, *et al*. Protecting world leaders against deep fakes[C]. The IEEE/CVF Conference on Computer Vision and Pattern Recognition, California, USA, 2019: 38–45.

[8] HE Kaiming, ZHANG Xiangyu, REN Shaoqing, *et al*. Deep residual learning for image recognition[C]. 2016 IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, USA, 2016: 770–778. doi: 10.1109/CVPR.2016.90.

[9] CHOLLET F. Xception: Deep learning with depthwise separable convolutions[C]. 2017 IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, USA, 2017: 1251–1258. doi: 10.1109/CVPR.2017.195.

[10] 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, Hong Kong, China, 2018: 1–7. doi: 10.1109/WIFS.2018.8630761.

[11] TARIQ S, LEE S, KIM H, *et al*. Detecting both machine and human created fake face images in the wild[C]. The 2nd International Workshop on Multimedia Privacy and Security, Toronto, Canada, 2018: 81–87. doi: 10.1145/3267357.3267367.

[12] BAYAR B and STAMM M C. Constrained convolutional neural networks: A new approach towards general purpose image manipulation detection[J]. *IEEE Transactions on Information Forensics and Security*, 2018, 13(11): 2691–2706. doi: 10.1109/TIFS.2018.2825953.

[13] GÜERA D and DELP E J. Deepfake video detection using recurrent neural networks[C]. The 15th IEEE International Conference on Advanced Video and Signal Based Surveillance, Auckland, New Zealand, 2018: 1–6. doi:

10.1109/AVSS.2018.8639163.

[14] WANG Shengyu, WANG O, ZHANG R, *et al*. CNN-generated images are surprisingly easy to spot... for now[EB/OL]. https://arxiv.org/abs/1912.11035, 2019.

[15] SHELHAMER E, LONG J, and DARRELL T. Fully convolutional networks for semantic segmentation[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017, 39(4): 640–651. doi: 10.1109/TPAMI.2016.2572683.

[16] CHEN L C, PAPANDREOU G, SCHROFF F, *et al*. Rethinking atrous convolution for semantic image segmentation[EB/OL]. https://arxiv.org/abs/1706.05587, 2017.

[17] 毕秀丽, 魏杨, 肖斌, 等. 基于级联卷积神经网络的图像篡改检测算法[J]. 电子与信息学报, 2019, 41(12): 2987–2994. doi: 10.11999/JEIT190043.

BI Xiuli, WEI Yang, XIAO Bin, *et al*. Image forgery detection algorithm based on cascaded convolutional neural network[J]. *Journal of Electronics & Information Technology*, 2019, 41(12): 2987–2994. doi: 10.11999/JEIT190043.

[18] LI Haodong, LI Bin, TAN Shunquan, *et al*. Detection of deep network generated images using disparities in color components[EB/OL]. https://arxiv.org/abs/1808.07276, 2018.

[19] 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. doi: 10.2352/ISSN.2470-1173.2019.5.MWSF-532.

[20] 杨宏宇, 王峰岩. 基于深度卷积神经网络的气象雷达噪声图像语义分割方法[J]. 电子与信息学报, 2019, 41(10): 2373–2381. doi: 10.11999/JEIT190098.

YANG Hongyu and WANG Fengyan. Meteorological radar noise image semantic segmentation method based on deep convolutional neural network[J]. *Journal of Electronics & Information Technology*, 2019, 41(10): 2373–2381. doi: 10.11999/JEIT190098.

[21] 高逸飞, 胡永健, 余泽琼, 等. 5种流行假脸视频检测网络性能分析和比较[J]. 应用科学学报, 2019, 37(5): 590–608. doi: 10.3969/j.issn.0255-8297.2019.05.002.

GAO Yifei, HU Yongjian, YU Zeqiong, *et al*. Evaluation and comparison of five popular fake face detection networks[J]. *Journal of Applied Sciences*, 2019, 37(5): 590–608. doi: 10.3969/j.issn.0255-8297.2019.05.002.

胡永健：男，1962年生，教授，博士生导师，研究方向为多媒体信息安全、图像处理、人工智能及其应用.

高逸飞：男，1996年生，硕士生，研究方向为多媒体信息安全、图像处理和机器学习.

刘琲贝：女，1980年生，讲师，研究方向为多媒体信息安全、图像处理和机器学习.

廖广军：男，1981年生，副教授，研究方向为多媒体信息安全、图像处理和机器学习.

责任编辑：余 蓉