基于帧间差异的人脸篡改视频检测方法

张怡暄[1,2], 李 根[1,2], 曹 纭[1,2], 赵险峰[1,2]

1. 中国科学院信息工程研究所 信息安全国家重点实验室 北京 中国 100093
2. 中国科学院大学 网络空间安全学院 北京 中国 100093

摘要 近几年, 随着计算机硬件设备的不断更新换代和深度学习技术的不断发展, 新出现的多媒体篡改工具可以让人们更容易地对视频中的人脸进行篡改。使用这些新工具制作出的人脸篡改视频几乎无法被肉眼所察觉, 因此我们急需有效的手段来对这些人脸篡改视频进行检测。目前流行的视频人脸篡改技术主要包括以自编码器为基础的 Deepfake 技术和以计算机图形学为基础的 Face2face 技术。我们注意到人脸篡改视频里人脸区域的帧间差异要明显大于未被篡改的视频中人脸区域的帧间差异, 因此视频相邻帧中人脸图像的差异可以作为篡改检测的重要线索。在本文中, 我们提出一种新的基于帧间差异的人脸篡改视频检测框架。我们首先使用一种基于传统手工设计特征的检测方法, 即基于局部二值模式(Local binary pattern, LBP)/方向梯度直方图(Histogram of oriented gradient, HOG)特征的检测方法来验证该框架的有效性。然后, 我们结合一种基于深度学习的检测方法, 即基于孪生网络的检测方法进一步增强人脸图像特征表示来提升检测效果。在 FaceForensics++数据集上, 基于 LBP/HOG 特征的检测方法有较高的检测准确率, 而基于孪生网络的方法可以达到更高的检测准确率, 且该方法有较强的鲁棒性; 在这里, 鲁棒性指一种检测方法可以在三种不同情况下达到较高的检测准确率, 这三种情况分别是: 对视频相邻帧中人脸图像差异用两种不同方式进行表示、提取三种不同间隔的帧对来计算帧间差异以及训练集与测试集压缩率不同。

关键词 视频篡改; 篡改检测; 帧间差异; 孪生网络; Deepfake; Face2face
中图法分类号 TN915.08 **DOI 号** 10.19363/J.cnki.cn10-1380/tn.2020.02.05

A Method for Detecting Human-face-tampered Videos based on Interframe Difference

ZHANG Yixuan[1,2], LI Gen[1,2], CAO Yun[1,2], ZHAO Xianfeng[1,2]

1. State Key Laboratory of Information Security, Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093, China
2. School of Cyber Security, University of Chinese Academy of Sciences, Beijing 100093, China

Abstract With the continuous upgrade of computer hardware and the continuous development of deep learning techniques in recent years, new multimedia tampering tools make it easier for people to tamper human faces in videos. Human-face-tampered videos created with these new tools can hardly be noticed by naked eyes, thus we urgently need effective methods to detect these human-face-tampered videos. At present, popular techniques used to tamper human faces in videos mainly include the autoencoder-based Deepfake and the computer-graphics-based Face2face. We have noticed that interframe differences between human face regions in human-face-tampered videos are significantly greater than those of untampered videos, so the differences between human face images in adjacent frames of videos can be utilized as an important clue for tampering detection. In this paper, we propose a new detection framework for human-face-tampered videos based on interframe differences. We first use a detection method based on artificially designed features which is traditional, namely Local Binary Pattern(LBP)/Histogram of Oriented Gradient(HOG)-feature-based detection method to verify the effectiveness of the proposed detection framework. Then, with a deep-learning-based detection method, namely Siamese-network-based detection method, we further strengthen feature representation of human face images to improve detection performance. In FaceForensics++ dataset, the LBP/HOG-feature-based detection method can have relatively high detection accuracy; while the Siamese-network-based detection method can reach higher detection accuracy, and the method has relatively strong robustness; here, the robustness refers to that a detection method can reach relatively high detection accuracy in three different situations, they are expressing differences of human face images in adjacent frames of videos in two different ways, extracting pairs of frames in three different intervals for calculating interframe differences, the training dataset and the testing dataset have different compression rates.

Key words video tampering; tampering detection; interframe difference; siamese network; Deepfake; Face2face

通讯作者: 赵险峰, 博士, 研究员, zhaoxianfeng@iie.ac.cn

本课题得到国家重点研发计划课题(No. 19QY2202, No.19QY(Y)0207); 中国科学院信息工程研究所攀登计划项目
收稿日期: 2019-12-20; 修改日期: 2020-03-09; 定稿日期: 2020-03-10

1 引言

近几年, 随着互联网通信技术的快速发展和广泛应用以及功能不断完善的摄影摄像设备的不断普及, 同时随着图像处理技术和人工智能技术的不断实用化, 人们可以轻松拍摄出效果精良且清晰度高超的图像与视频, 并能够迅速将这些多媒体素材发布到互联网上。这些多媒体素材成为自媒体信息的一个主要来源, 对信息的实时传播起到很大的促进作用。同时, 便捷的拍摄技术也为社会活动的记录带来了方便, 例如交通事故处理、保险理赔以及司法取证等。但是另一方面, 目前图像视频制作和传播的便捷性也带来了一些负面影响, 例如一些人对拍摄的图像视频进行一些不恰当的篡改后将它们传播到互联网上。如果我们不对这些篡改后的图像和视频进行甄别和限制, 它们会迅速在网上各大平台广泛传播, 并可能对社会造成非常严重的影响。因此如何对这些被篡改的图像和视频进行检测变得越来越重要, 也越来越受到人们的关注。

在多年以前, Photoshop 和 ACDSee 等图像编辑软件通常是人们进行多媒体篡改的主要工具。以 Photoshop 为例, 该软件能够对图像进行旋转、缩放、镜像等各种几何操作, 能够任意调整图像的亮度、对比度、饱和度等颜色属性, 还可以对图像进行修补、去噪和风格转换等操作, 再加上 Photoshop 软件中存在的多种滤镜功能和该工具使用的多通道多图层的构架, 可以说, Photoshop 几乎能满足一个用户对于图像修改的任何要求。然而, Photoshop 等图像处理工具很难对视频进行加工与篡改。这是因为用这些工具去处理视频, 则必须要先将视频分解成帧序列, 然后再逐帧进行大量细致的修改工作, 而这样的工作量是及其巨大的。

随着近些年以来多媒体处理硬件设施的不断升级和新技术的不断进步, 尤其是大数据处理技术和人工智能技术爆发式地增长, 出现了新的多媒体伪造篡改技术。其中最为流行的技术当属对抗生成网络 (Generative adversarial networks, GAN)[1] 技术, Deepfake[2-3]和 Face2face[4]篡改技术。GAN 是一种新型的网络结构, 该结构中的判别网络和生成网络互相促进, 训练好的 GAN 可以生成大量难以用肉眼进行区分的伪造样本; Deepfake 借助自编码器等技术手段实现对视频中人脸的替换; 而 Face2face 利用计算机图形学的有关方法来实现视频中人物脸部表情的迁移。

Deepfake 和 Face2face 等新技术的出现使得人们能够更加容易地对整段视频进行篡改。以 Deepfake 为例, 目前在 Github 网站上有封装完好的软件和详细的安装和使用教程。这些教程非常简单, 一个人只需要懂得一些基本的计算机常识就可以读懂该教程, 并能够在短时间内完成对人物脸部转换模型的训练和篡改视频的制作, 所需要的硬件设备仅仅是一台带有 GPU 的电脑而已。

除此以外, 这些视频人脸篡改工具自身也在不断升级进步, 使用这些工具制作出的篡改样本质量越来越高。以 Deepfake 为例, 在该工具的早期版本中, 训练的假视频有很大瑕疵, 比如视频在被篡改后, 里面的人物不会眨眼等。然而随着 Deepfake 的开发者对该工具的不断升级改造, 使用最新版本的 Deepfake 软件所制作出的视频里, 篡改人脸几乎和真实人脸一模一样。

与图像不同的是, 视频所含信息极为丰富, 人们往往会对视频中传达的信息深信不疑, 所以篡改视频所带来的危害将远远大于篡改图像。例如, 如果一段伪造视频中播出的是一个国家或组织的重要领导人在不合时宜的场合发表了一些令人匪夷所思的言论, 势必将会在国际上引起轩然大波。综上所述, 目前视频人脸篡改新技术将会对国际社会的各个方面造成非常严重的威胁, 我们急需有效的篡改检测手段来对这些篡改视频和真实视频加以区分, 以维护国际社会的安定与和谐。

目前互联网上存在的视频人脸篡改方法, 主要有以下两类: (1)以 Deepfake 技术为代表的人脸替换方法: 这类方法主要将视频中的人脸替换为另一个人的人脸, 并保持原人脸的面部神态和表情; (2)以 Face2face 为代表的表情迁移方法: 这类方法不会对原视频中人脸的面部特征做出改变, 只会让原视频中人物的面部的表情动作完全按照另一个人面部的表情动作做出改变, 此类技术主要利用的是计算机图形学的有关方法。

针对 Deepfake 和 Face2face 等视频人脸篡改技术的检测方法已成为篡改取证领域新的关注点。目前有许多检测效果良好的人脸识别方法, 例如 DeepFace[5]、FaceNet[6]和较新的 LightCNN[7]等, 但对于人脸篡改检测, 尤其是视频中的人脸篡改检测, 现有方法的检测效果仍然需要进一步加强。

本文专注于对 Deepfake 和 Face2face 等人脸篡改视频检测方法的研究。我们发现, 由于 Deepfake、Face2face 等人脸篡改视频技术都是逐帧进行篡改的, 所以用这些方法制作出的篡改视频中相邻帧的人脸区域很难保持原有的帧间连续性, 因此这些篡改视

频中人脸区域的帧间差异要明显大于未被篡改的视频中人脸区域的帧间差异。由此, 我们提出了一种基于帧间差异的人脸篡改视频检测框架, 并在该框架的基础上使用基于孪生网络[8-9]的检测方法来进行检测。使用我们所提出的方法对 FaceForensics++数据集[10]中未压缩、压缩量化系数为 23 和压缩量化系数为 40 的 Face2face 篡改视频进行检测, 分别可以达到 99.83%、99.33%和 91.83%的视频级检测准确率。

本文的工作和贡献, 主要在于如下三个方面:

(1) 我们描述了 Deepfake、Face2face 等视频人脸篡改技术对原视频的帧间差异的改变背后的原因和具体体现, 进而指出利用这些帧间的差异性可以作为对人脸篡改视频进行检测的重要线索。在此基础之上, 我们提出了一种基于视频帧间差异的人脸篡改视频检测框架。

(2) 我们使用基于局部二值模式(Local binary pattern, LBP)[11]/方向梯度直方图(Histogram of oriented gradient, HOG)[12]特征的检测方法来初步验证所提出的基于帧间差异的人脸篡改视频检测框架的有效性。LBP 特征和 HOG 特征可以反应出人脸区域的局部纹理特性。实验结果表明, 这种基于 LBP/HOG 特征的检测方法可以在 FaceForensics++数据集上达到较高的检测准确率, 初步验证了我们所提出的检测框架的有效性。

(3) 在所提出的基于帧间差异的人脸篡改视频检测框架的基础上, 使用基于孪生网络的检测方法可以通过衡量两分支输入的相似程度来自动对两分支提取的特征进行优化。实验结果表明, 基于孪生网络的方法在 FaceForensicss++数据集上可以达到比基于 LBP/HOG 检测方法更高的准确率, 且基于孪生网络的方法对 FaceForensics++数据集中 Face2face 篡改视频的检测准确率可以超过现存的许多其他方法。除此以外, 基于孪生网络的方法也有较强的鲁棒性; 这里的鲁棒性指的是一种检测方法可以在三种不同情况下拥有较高的检测准确率, 这三种情况是: 对视频相邻帧中人脸图像差异用两种不同方式进行表示、提取三种不同间隔的帧对来计算帧间差异以及训练集与测试集压缩率不同。

值得说明的是, 考虑到实际应用的泛化性能, 对于人脸篡改视频的检测, 我们采用由粗到精的策略, 即首先使用基于 LBP/HOG 特征的粗识别方法, 再使用基于孪生网络的精确识别方法, 实现人工特征和深度学习特征的互补。

本文的组织结构如下: 在第 2 节中对人脸篡改视频制作和检测的相关工作进行描述; 在第 3 节中讲述所提出的人脸篡改视频检测方法的技术基础; 在第 4 节具体讲述所提出的基于帧间差异的人脸篡改视频检测方法; 第 5 节对所进行的实验进行描述; 第 6 节是对全文的总结和对未来工作的展望。

2 相关工作

本节简要介绍现存视频人脸篡改及其检测技术的一些重要的概念和近期发展概况。本节通过基于 GAN 的伪造图像合成技术、基于 Deepfake 的篡改视频生成技术、基于 Face2face 的篡改视频生成技术和视频中人脸篡改的检测技术四个方面来对相关工作进行论述。

2.1 基于 GAN 的伪造图像合成技术

GAN[1]是在 2014 年由 Goodfellow 等人提出的一种神经网络框架。GAN 主要由两个部分组成: 判别网络和生成网络。GAN 中的判别网络和生成网络在训练的过程中互相促进, 最终训练好的 GAN 能够生成非常逼真的伪造样本。

GAN 的训练过程主要分为以下两个阶段: (1)生成网络训练阶段: 在这个阶段中, 需要固定判别网络的参数, 只训练生成网络。随机噪声经过生成网络后, 将生成的伪造样本进入判别网络, 并利用损失函数来对生成网络进行优化。(2)判别网络训练阶段: 在这个阶段中, 需要固定生成网络参数, 只训练判别网络。真实样本和由生成网络生成的伪造样本分别作为判别网络的输入来优化判别网络的参数。

GAN 网络训练阶段的目标函数如公式(1)所示, 其中 D 与 G 分别代表生成网络和判别网络, 而 z 代表生成网络中输入的随机噪声。

$$\min_G \max_D V(D,G) = = \mathrm{E}_{x\sim P_{data}(x)}\left[\log D(x)\right] + \mathrm{E}_{z\sim P_z(z)}\left[\log\left(1-D\left(G(z)\right)\right)\right] \tag{1}$$

原始 GAN 网络的缺点是生成的伪造样本的类别是完全随机的, 实验者不能将输出限定在指定的类别范围内。在这之后, 出现了条件 GAN(condition-GAN, C-GAN)[13], 与原始 GAN 所不同的是, 即使生成网络生成的伪造样本质量足够高, 但若判别网络认为该样本与生成网络输入的类别标签不符, 这类样本依然会被判别网络给予与低质量样本一样的评分。C-GAN 网络用这种方式将输出的伪造图像限制在一定类别之中。但是, C-GAN 只能将网络的输出限制在一定类别范围中, 而无法根据特定输入样本的特点来进行输出。随后出现的 Pix2pix[14]、Cycle-GAN[15]以及文献[16]解决了这个问题, 以上几

种网络均可以按照输入样本的特点来生成对应的输出样本, 并让输入样本按照指定的风格来进行转换。

在以上工作的基础之上, 新出现的 GAN 网络可以完成越来越复杂和越来越精细的伪造人脸生成和人脸篡改任务。TP-GAN[17]可以在单一侧脸的基础上生成正脸, Star-GAN[18]可以对人脸图像的风格及面部表情进行转换, SC-FEGAN[19]甚至可以根据用户对人脸图像的涂鸦来改动该人脸的面部细节, 这些新的 GAN 技术都可能在将来被用于制作新的视频人脸篡改工具。目前 Deepfake 的一些新的实现方法中也采取了与 GAN 结合的方式[3], 用这些与 GAN 结合的新方法制作出的人脸篡改视频效果更为逼真。

2.2 基于 Deepfake 的篡改视频生成技术

Deepfake 篡改技术的一般流程如图 1 所示, 具体步骤描述如下: (1)首先要将视频拆分为帧, 然后对原始视频逐帧进行篡改。(2)识别出视频帧中人物脸部区域, 如图 1(a1)所示, 并标记出人脸区域中的特征点, 如图 1(a2)所示。(3)接着, 根据找到特征点来计算出围绕人脸部区域的最小矩形框, 如图 1(a3)所示, 并根据该矩形框截取出人的脸部图像。(4)接下来我们要将脸部图像通过仿射变换进行校正并缩放到固定的大小以方便输入到自编码器中进行换脸, 如图 1(a4)(a5)所示。再将校正后的脸部区域送入自编码器来将其转化成篡改人脸, 如图 1(a5)(b5)所示。训练好的自编码器可以对人脸图像进行换脸操作,并保持原脸的表情和神态。在将脸部图像校正和缩放的过程中, 我们还需要记录下仿射变换的参数, 以方便将自编码器生成的篡改人脸图像(图 1(b5))的旋转角度和尺寸变换到和校正前的真实人脸图像相一致, 以方便将篡改后的人脸安放回原视频帧中。整个脸部替换过程如图 1(a4)(a5)(b4)(b5)所示。(5)当我们将篡改后的人脸图像安放回原视频帧(图 1(b3))后, 使用泊松克隆或遮罩覆盖等技术, 可以消除篡改人脸在嵌入到原视频帧背景区域的过程中造成的不自然的边界, 如图 1(b2)所示, 这样可以使篡改后的视频帧显得更加真实。(7)对每一个视频帧进行篡改后, 再将所有被篡改的帧连接在一起, 就得到了 Deepfake 篡改视频。

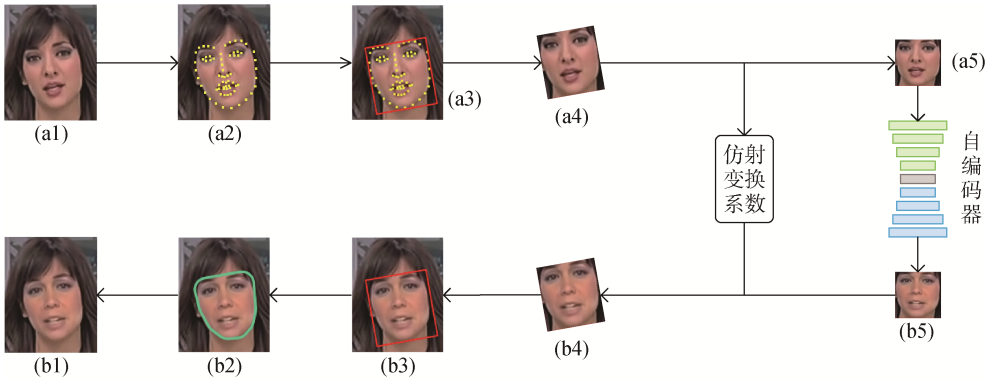

图 1 Deepfake 篡改视频制作流程图[20]

Figure 1 Flow diagram of creating Deepfake tampered videos

需要特别指出的是, 目前的 Deepfake 篡改技术在人脸合成方面还存在一些缺陷, 例如如果目标人脸和源人脸一个戴眼镜一个不戴, 或者一个有浓密的胡须而另一个没有的话, 生成的伪造视频会在人脸篡改边缘处出现“断层”现象, 导致该篡改视频出现明显的篡改痕迹。另外, 目前基于 Deepfake 篡改技术开发的软件有很多参数需要调节, 一旦一些关键参数调节不好, 生成的篡改视频会有非常大的瑕疵。

2.3 基于 Face2face 的篡改视频生成技术

Face2face 是另一种能够对视频中人脸进行篡改的技术。不同于 Deepfake 可以将视频中的人脸换成另一个人脸, Face2face 技术可以让源脸的表情神态迁移到目标脸上, 而篡改后的新目标脸依然保持原来的面部特征。

Face2face 篡改的原理大致如下: 首先对视频中的目标脸和源脸进行跟踪, 获得目标脸与源脸的脸部表情遮罩。再对表情遮罩中的表情进行转移, 以使目标脸表情遮罩中的表情和源脸表情遮罩一致, 最终使用新的目标脸遮罩来完成对新目标脸图像的生成。Face2face 的篡改效果图如图 2 所示, 生成的新目标脸继承了原目标脸的面部特征和源脸的表情神态以及嘴部形态。值得注意的是, Face2face 对人物嘴形的迁移方法进行了改变, 不同于其他论文中使用的直接将源脸的嘴部区域粘贴到目标脸中的方法, Face2face 从目标视频里寻找与源脸嘴形最为一致的

目标脸嘴形来合成新目标脸。在这种方式的作用下, 新合成的目标脸的嘴部区域会非常逼真。Face2face 篡改技术利用的是基于计算机图形学的相关方法, 目前已经能够实现实时的人物表情迁移。

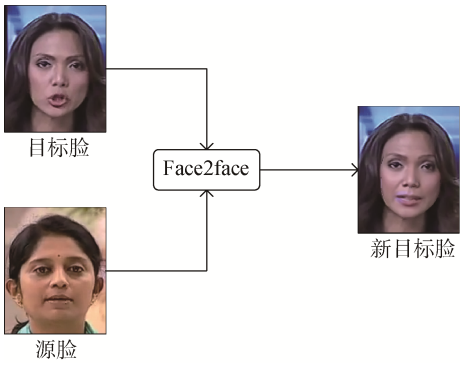

图 2 Face2face 的效果图

Figure 2 Result of Face2face

2.4 视频中人脸篡改的检测技术

传统的对篡改图像进行检测的方法有基于彩色滤波阵列(Color filter array, CFA)的检测方法[21-25]、基于相机响应非一致性(Photo response non-uniformity, PRNU)噪声的检测方法[26-28]、基于模糊不一致性的检测方法[29-30]、基于成像色差的检测方法[31-33]和基于几何变换及插值痕迹[34-35]的检测方法等等。但是, 绝大部分视频都是经过高度压缩的, 在这种情况下这些方法所依赖的篡改痕迹很难再被检测到。因此这些检测方法并不适用于检测 Deepfake 和 Face2face 等人脸篡改视频。

文献[36]提出了一种专门针对 Deepfake 篡改视频的检测方法, 该文指出, 由于在早期人们训练 Deepfake 中 GAN 网络的过程中很少会选择人眼闭合状态的人脸图像样本, 因此用这种方式生成的假视频中的人物将会一直处于睁眼状态。由此, 该文提出可以通过检测视频中人物眨眼的状况来进行篡改检测。但是如果实验者在 GAN 网络的训练过程中加入了处于闭眼状态的人脸图像的样本, 则文献[36]所中所提出的检测方法将会失效。文献[37]指出由于 Deepfake 脸部转换网络输出的人脸图像特征点的位置会与输入的人脸图像特征点的位置存在偏差, 且由于 Deepfake 只对人内脸区域进行篡改, 因此 Deepfake 篡改视频中人物内外脸夹角的差异将明显大于真实视频中人物内外脸的夹角, 文献[37]提出可以利用视频中人物内外脸夹角的大小来进行篡改检测。但随着 Deepfake 所使用的脸部转换网络不断的进化, 输入输出脸特征点的不一致性会逐渐减弱, 此时文献[37]中所用方法的检测效果也会大打折扣[38]。

文献[39]利用在 HSV 通道和 YCbCr 通道中颜色分量分布的不一致性来区分真实拍摄的人脸图像和人工智能工具生成的虚假人脸图像。文献[40]指出真实人脸图像和 GAN 生成的伪造人脸图像特征点分布的不一致, 因此可以利用这种不一致性来对真假人脸加以区分。类似地, 文献[41]利用在人脸图像上提取的各种面部细节信息来对篡改人脸进行检测。文献 [42-43] 利用精心设计的卷积神经网络(Convolutional neural network, CNN)对在视频帧中提取的人脸图像进行检测。然而, 以上这些文献所提及的方法都没有考虑到视频帧间的时序性关系。文献[44]利用循环神经网络(Recurrent neural network, RNN)对视频序列中提取的人脸图像进行检测。相比于直接使用 CNN 进行检测的方法, 利用 RNN 来进行检测能够对视频帧间的时序关系加以利用, 但由于 RNN 直接从视频帧序列中提取内容信息, 因此使用 RNN 来进行检测的方式会在一定程度上受到视频内容的干扰。

另外, 鉴于人脸识别和人脸篡改检测的相通性, 可以借鉴一些典型的人脸识别方法来用于处理人脸篡改检测问题, 例如 FaceNet、LightCNN 等。FaceNet 能够直接获取从输入的人脸图像到欧氏空间的映射关系, 并能利用三元组(Triplet)损失来提升网络的性能。LightCNN 利用提出的 MFM(Max-Feature-Map)激活层对网络中的信号进行筛选, 以获得对信息更为精简的表示并以此降低计算代价。实验结果表明虽然 LightCNN 的参数数量较少, 但依然能够达到很高的检测准确率。

在本文中, 我们所提出的人脸篡改视频检测框架利用视频相邻帧中人脸图像的差异作为特征来进行篡改检测, 这种方式能充分地利用视频的时序关系, 并不太容易受到视频内容信息的干扰。除此以外, 在我们所提出的篡改检测框架下基于孪生网络的检测方法中, 孪生网络的两个分支可以对视频相邻帧中的人脸图像提取深度特征, 之后该网络的对比损失函数依据相似度标签来不断优化在这两个分支中所提取的特征。实验证明, 这种基于孪生网络的检测方法可以在 FaceForensics++数据集上达到较高的检测准确率且有较强的鲁棒性; 此处鲁棒性特指一种检测方法可以在三种不同情况下拥有较高的检测准确率, 这三种情况分别是: 对视频相邻帧中人脸图像差异用两种不同方式进行表示、提取三种不同间隔的帧对来计算帧间差异以及训练集与测试集压缩率不同。

3 所提出的人脸篡改视频检测方法的技术基础

本节叙述本文中人脸篡改视频检测方法涉及的有关的基础概念和知识, 包括 LBP 特征的计算、HOG 特征的计算、LightCNN 的结构、Inception 网络[45-48]的结构、残差网络[49]的结构和孪生网络的结构。

3.1 LBP 特征的计算

局部二值模式(Local binary pattern, LBP)[11]特征, 用于描述图像局部纹理特征, 其优点在于具有旋转不变性和灰度不变性。该特征原理简单, 检测效果优异, 广泛应用于人脸检测等任务中。

最经典的LBP特征采取的是半径为1, 采样点为8 的提取模式, 该模式下特征的具体提取步骤如下: (1)计算图像中每个像素的二进制模式, 如图 3 所示。首先将图像转化为灰度图, 让待计算像素作为中心像素, 并与周围的 8 个相邻像素做比较。如果相邻的像素值大于等于中心像素, 则将该相邻像素的值置为 1, 否则置为 0。选定一个相邻像素作为初始像素, 并沿着一定方向提取特征。以图 3 为例, 我们选取中心像素左上角的相邻像素作为初始像素, 以顺时针为方向提取特征, 则该像素提取出的二进制模式为 00000011, 则 00000011 为该中心像素的二进制模式。(2)对某一区域所有像素的 LBP 二进制模式进行直方图统计, 作为整个区域的 LBP 特征。由于 8 位二进制数可以表示成从 0 到 255 共计 256 个不同的数字, 所以在半径为 1, 采样点位 8 的情况下, 在图像一个区域中提取的 LBP 特征共有 256 维。(3)将图像沿横、纵划方向分为 $m\times n$ 的区域, 如图 4 所示, 对于每一个区域, 我们都会提取一个 256 维的 LBP 特征, 将所有 $m\times n$ 个区域所提出的特征连接起来后得到的 $m\times n\times 256$ 维特征作为整张人脸的 LBP 特征。

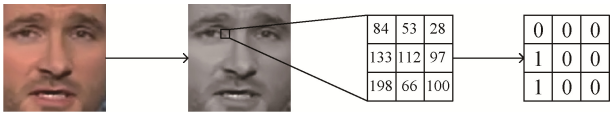

图 3 计算图像像素的 LBP 二进制模式
Figure 3 The calculation of LBP binary mode of image pixel

在图像一个区域内 n 个采样点的 LBP 特征的维度是 2^n, 当 n 很大的时候, LBP 特征的维数会大的惊人。后来研究者发现, LBP 二进制模式主要集中在 0 向 1 或 1 向 0 跳变两次及两次以下的模式中, 并称满足这种条件的二进制模式为 LBP 特征等价模式类。

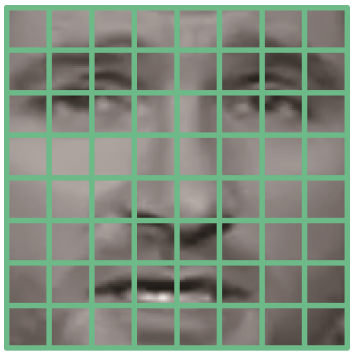

图 4 提取 LBP 特征时对图像进行划分
Figure 4 Image dividing in LBP feature extraction

例如二进制模式 10001111 与 00011100 均属于等价模式类。值得注意的是, 二进制模式 00011111 也属于等价模式类, 因为如果我们将该模式的八位二进制数首尾相接后进行观察, 该模式也恰好只有两次跳变。由于跳变 0 次、跳变 1 次、跳变 2 次的二进制模式数分别为 2、0、56, 所以 LBP 特征等价模式类中的二进制模式共有 58 种。为了降低 LBP 特征的维度, 需要适当地对 LBP 特征进行合并。以半径为 1, 采样点位 8 的 LBP 特征为例, 该情况下的 LBP 特征等价模式类中 58 种二进制模式各为一类, 而不属于等价模式类的二进制模式由于出现的次数很少, 所以可以被合并成一类。通过这种合并方式, 半径为 1, 采样点位 8 的 LBP 特征的维数将会由 256 降至 59。对于采样点数更多的 LBP 特征, 用这种方式来进行降维效果更为明显。称用这种合并方式精简后的 LBP 特征为等价模式 LBP 特征。

3.2 HOG 特征的计算

方向梯度直方图(Histogram of oriented gradient, HOG)[12]特征是一种广泛应用于图像处理、人脸识别和行人检测等任务的描述子, 该特征可以反映图像中不同区域梯度的幅度和方向的分布。

HOG 特征的提取流程如图 5 所示。具体流程如下: (1)首先对待检测图像进行灰度化, 以方便提取特征。(2)再对图像进行伽马校正, 伽马校正可以对整张图像进行有效调节, 减少光照变化对于图像信息的影响。(3) 定义 k 个方向来对图像中每个像素点的梯度方向和大小进行计算。(4)将图像分为一个一个的单元,对每个单元中像素的梯度进行统计。每个像素梯度的方向决定统计直方图中哪个分量进行增加, 而像素梯度的大小决定统计直方图中对应分量增加的幅度。(5)以一定步长对图像提取重叠的图像块, 每个图像块含有一定数目的单元。将每个单元的特征连接起来作为整个图像块的特征。(6)将所有图像块

的特征连接起来, 作为整张图像的特征。例如, 在一张尺寸为64×64 的图像中, 定义 9 个方向来对像素梯度的大小和方向进行计算, 每8×8 个像素构成一个单元, 每 2×2 个单元构成一个图像块, 提取图像块的步长为 8。则每个图像块的特征维度为9×2×2=36。而整张图像会提取出 7×7=49 个图像块。即整张图像的 HOG 特征维度为 36×49=1764 维。

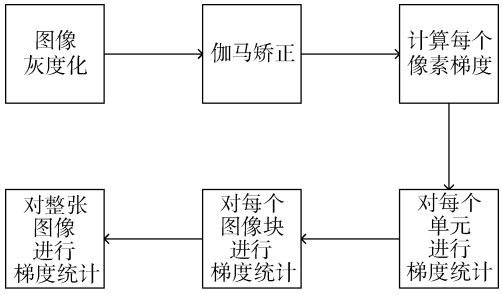

图 5 HOG 特征提取流程图

Figure 5 Flow diagram of HOG feature extraction

3.3 LightCNN 的结构

MFM(Max-feature-map)是文献[7]提出来的一种新的激活层, 这种激活层可以对神经网络中的信息过滤和筛选, 以获取更为良好的特征表示, 同时去除掉网络中存在的冗余信息, 使神经网络中的信息更为简洁和紧凑。MFM 分为两种结构, 分别称为 MFM2/1 结构和 MFM2/3 结构。

对于 MFM2/1 结构, 假设输入是一个卷积层 $x^n \in \Re^{H\times W}$, 其中 H 和 W 分别代表该卷积层高度方向和宽度方向的尺寸, 假设该层的通道数为 $2N$, 即 n={1,2,⋯,2N} 则该卷积层经过 MFM2/1 后的输出 $\hat{x}$ 如公式(2)所示。

$$\hat{x}_{ij}^{k} = \max(x_{ij}^{k}, x_{ij}^{k+N}) \tag{2}$$

在公式(2)中, max 代表取最大值, $1 \leqslant k \leqslant N$, $1 \leqslant i \leqslant H$, $1 \leqslant j \leqslant W$。

对于 MFM3/2 结构, 假设输入是一个一个卷积层 $x^n \in \Re^{H\times W}$, 其中 H 和 W 分别代表该卷积层高度方向和宽度方向的尺寸, 假设该层的通道数为 $3N$, 即 n={1,2,⋯,3N} 则该卷积层经过 MFM3/2 后的输出 $\hat{x}$ 如公式(3)和公式(4)所示。

$$\hat{x}_{ij}^{k_1} = \max(x_{ij}^{k}, x_{ij}^{k+N}, x_{ij}^{k+2N}) \tag{3}$$

$$\hat{x}_{ij}^{k_2} = \text{median}(x_{ij}^{k}, x_{ij}^{k+N}, x_{ij}^{k+2N}) \tag{4}$$

在公式(3)中, max 代表取最大值, 在公式(4)中, median 代表取中值, 在公式(3)和公式(4)中 $1 \leqslant k \leqslant N$, $1 \leqslant i \leqslant H$, $1 \leqslant j \leqslant W$。

LightCNN 的网络结构中大量使用了 MFM 激活层。基于 MFM 的特性, 在人脸识别相关研究任务中, LightCNN 可以在拥有较少参数的情况下依然能够有很高的检测准确率, 是近年来人脸识别方向非常重要的一个研究突破。

3.4 Inception 网络的结构

随着深度学习的不断发展, 简单的网络难以满足研究者对提取深层次特征的需要。随后, 人们不断设计出性能更强的网络, 以适应更高的需求。在其他的研究者着力于通过增加深度来提升网络性能时, Inception 网络[45-48]通过并行化的方式来优化神经网络的性能。

Inception 网络的一个基础模块如图 6 所示。Inception 基础模块由四条分支组成, 分别通过1×1 卷积层、3×3 卷积层、5×5 卷积层和3×3 最大池化层来从输入中提取不同尺度的特征, 再将这些不同尺度的特征连接起来作为这一模块的输出。然而, 由于涉及到5×5 的卷积, 这种 Inception 的基本模块会面临计算量过大的问题, 于是研究者在 Inception 基础模块中加入一些1×1 卷积层, 来降低网络的计算复杂度, 改进后的 Inception 模块如图 7 所示。

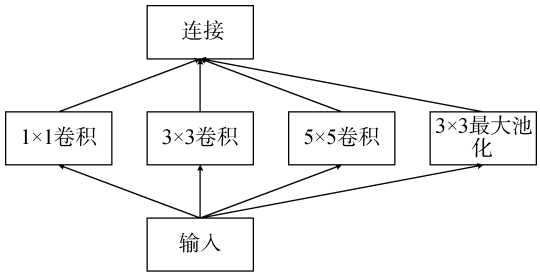

图 6 Inception 网络基础模块

Figure 6 Basic module of Inception net

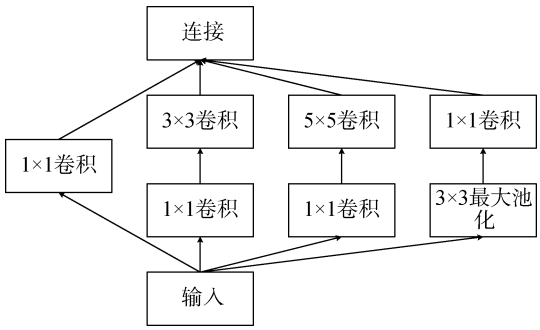

图 7 改进后的 Inception 网络基本模块

Figure 7 Improved basic module of Inception net

3.5 残差网络的结构

由于神经网络的梯度在反向传播的过程中会不

断衰减, 如果一个神经网络过深的话就会出现梯度消失现象, 导致该网络的性能急剧下降。

残差网络(Residual network, ResNet)[49]可以有效解决以上问题, 残差网络一个单元的结构如图 8 所示。在该图中可以看到残差网络在普通的神经网络中增加了一个跃层的连接, 使得网络的输出从原来的 $H(x)=F(x)$, 变成 $H(x)=F(x)+x$。这样做的好处是使得残差网络在每一个局部模块的训练目标转化成残差值 $H(x)-x$, 这也是残差网络中“残差”这一名字的来源。当网络比较深的时候, 对网络进行训练以使得每个模块的残差 $H(x)-x$ 尽量向 0 靠拢, 此时网络每个模块的输出基本与输入相同, 即 $H(x)=x$。基于这样的跃层连接模块, 残差网络中浅层输出能够直接传递到更深的层次,以使神经网络的性能不会因为层数的增长而急剧下降。通过这种跃层连接模式, 我们可以利用更深的网络得到输入信号的深层次特征。

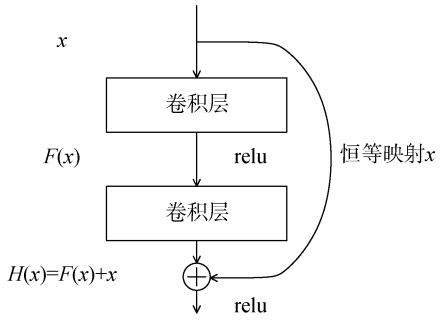

图 8 残差网络一个单元的结构图

Figure 8 Diagram of the structure of one unit in Residual network

3.6 孪生网络的结构

孪生网络(Siamese network)[8-9]是一种有着广泛应用的神经网络, 其结构如图 9 所示。孪生网络从两分支的输入样本中提取特征后, 该网络的对比损失函数会依据相似度标签来不断优化在这两分支的输入样本中提取的特征。训练好的孪生网络能够对两个分支输入样本的相似程度进行判别, 两个分支的相似程度可以用欧氏距离、曼哈顿距离或余弦距离等来度量。

孪生网络的两个分支结构完全一样, 且其参数为共享的(同时更新)。因此在训练阶段, 如果我们将两个分支的输入对调, 对网络的参数完全没有影响; 在测试阶段, 如果我们将两个分支的输入对调, 对测试结果也没有影响。

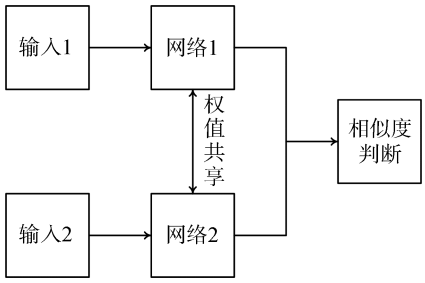

图 9 孪生网络结构图

Figure 9 Diagram of the structure of Siamese network

4 基于帧间差异的人脸篡改视频检测

本节首先指出未被篡改的视频和人脸篡改视频帧间差异的不同; 然后讲述提出的基于帧间差异的人脸篡改视频检测框架; 最后描述在该框架下的两种具体的检测方法: 即基于 LBP/HOG 特征的检测方法和基于孪生网络的检测方法。

4.1 未篡改视频和人脸篡改视频帧间差异的不同

首先, 我们从视频人脸篡改的原理上的来说明基于帧间差异的检测方法能够作为区分人脸篡改视频和真实视频的一种方式。Deepfake 和 Face2face 等很多视频篡改方法均是先将视频拆解成帧序列, 再逐帧对视频进行篡改, 最后简单地将篡改后的视频帧序列连接起来。在这个过程中, 这些视频篡改方法不会对视频帧间的连续性进行优化。这就意味着篡改后视频帧间的连续性很难达到原视频的程度, 篡改后视频人脸区域的帧间差异将会明显大于真实视频中人脸区域的帧间差异。

其次, 我们从视觉效果上来说明对于人脸篡改视频检测, 帧间的差异是一个很好的线索。视频帧间的不连续性往往比视频帧内存在的瑕疵更容易被察觉到。举个例子, 对于一段制作效果精良的 Deepfake 篡改视频, 如果我们将视频暂停在某一帧后去进行观察, 往往会误以为该视频中的人脸没有被篡改过。而一旦我们开始连续播放这些视频, 视频中人脸区域局部的亮度跳变和形状突变都会很容易地被注意到, 而这些帧与帧之间明显的差异在真实视频中是几乎不可能出现的。在 Face2face 篡改视频中也有类似的帧间瑕疵存在, 该视频中篡改人脸的嘴部区域经常会出现不自然的形状突变以及突发的局部模糊。以上所列举的 Deepfake 和 Face2face 篡改视频中帧与帧之间存在的视觉瑕疵, 都会很容易暴露这些视频的篡改事实。

4.2 基于帧间差异的人脸篡改视频检测框架

根据 4.1 节所描述的内容, 人脸篡改视频帧与帧之间差异会比真实视频帧与帧之间差异更为显著, 因此利用帧间差异来进行对人脸篡改视频进行检测的方法可以作为区分人脸篡改视频和真实视频的一个线索。在此基础之上, 我们提出了一种基于视频帧间差异的人脸篡改视频检测框架, 如图 10 所示。该框架的大致检测流程如下: (1)首先将视频分解为帧序列, 再利用 DLIB 库[50]定位出每一个帧中人脸的 68 个标志点, 并以此截取出人脸图像。(2)提取特征对截取的人脸图像进行表示。(3)设计合适的方式来对视频相邻帧中人脸图像的差异进行表示, 并以此来判断该视频是否被篡改过。

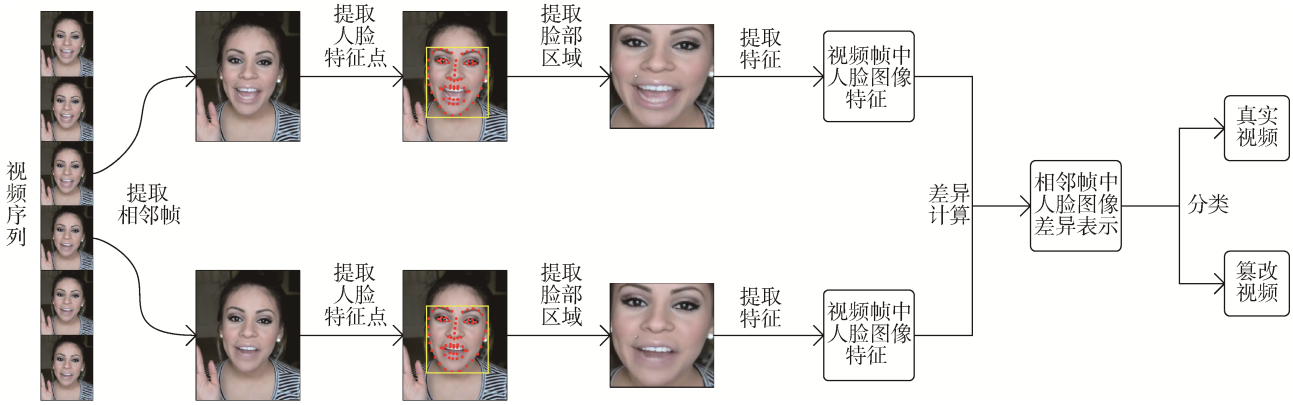

图 10 提出的基于帧间差异的人脸篡改视频检测框架流程图

Figure 10 Flow diagram of the proposed detection framework for human-face-tampered videos based on interframe difference

在基于帧间差异的人脸篡改视频检测框架下, 我们需要找到合适的特征来描述视频帧序列中人脸图像, 并找到有效的方式来对视频相邻帧中人脸图像的差异进行表示。

4.3 基于帧间差异的人脸篡改视频检测框架下基于 LBP/HOG 特征的检测方法

如果要利用帧间差异来区分真实的视频和人脸篡改视频, 需要提取有效的特征来反映视频帧间人脸图像的差异性。考虑到 LBP 特征可以反映图像局部纹理的统计特性, 而 HOG 特征可以反映图像局部梯度的统计特性, 我们使用以上两种特征来对视频帧序列中的人脸图像进行表示。

在我们提取出视频帧序列中人脸图像的特征后, 下一步要进行的是如何利用所提取的特征来表示视频帧间人脸图像的差异性。我们使用特征相减并取绝对值的方式来表示在视频相邻帧中人脸图像中提取的特征 $X=(x_1,x_2,\cdots,x_n)$ 和 $Y=(y_1,y_2,\cdots,y_n)$ 之间的差异 $diff(X,Y)$, 计算方法如公式(5)所示。

$$diff(X,Y)=(|x_1-y_1|,|x_2-y_2|,\cdots,|x_n-y_n|) \quad (5)$$

4.4 基于帧间差异的人脸篡改视频检测框架下基于孪生网络的检测方法

随着越来越多的研究者使用深度学习作为工具来推进他们的研究进展, 在这里我们也使用孪生神经网络对视频帧中人脸图像进行特征提取。我们所用的孪生网络以 LightCNN、Inception 网络和残差网络为主体, 以对比损失函数作为为网络的损失函数。

4.4.1 所使用的 LightCNN 结构:

LightCNN 共有三种典型的结构, 分别称为 LightCNN-4(如表 1 所示)、LightCNN-9(如表 2 所示)和 LightCNN-29(如表 3 所示)。表 1、表 2、表 3 中的 MFM 激活层均为 MFM2/1 结构, 表 3 中的 “$\begin{bmatrix}3\times3/1,1\\3\times3/1,1\end{bmatrix}$” 代表一个残差模块。

表 1 LightCNN-4 的结构

Table 1 Structure of LightCNN-4

层类型	卷积核尺寸/步长	输出尺寸
Conv1	9×9/1	120×120×96
MFM1	-	120×120×48
Pool1	2×2/2	60×60×48
Conv2	5×5/1	56×56×192
MFM2	-	56×56×96
Pool2	2×2/2	28×28×96
Conv3	5×5/1	24×24×256
MFM3	-	24×24×128
Pool3	2×2/2	12×12×128
Conv4	4×4/1	9×9×384
MFM4	-	9×9×192
Pool4	2×2/2	5×5×192
fc1	-	512
MFM_fc1	-	256

表 2 LightCNN-9 的结构

Table 2 Structure of LightCNN-9

层类型	卷积核尺寸/步长, 填充	输出尺寸
Conv1	5×5/1,2	128×128×96
MFM1	-	128×128×48
Pool1	2×2/2	64×64×48
Conv2a	1×1/1	64×64×96
MFM2a	-	64×64×48
Conv2	3×3/1,1	64×64×192
MFM2	-	64×64×96
Pool2	2×2/2	32×32×96
Conv3a	1×1/1	32×32×192
MFM3a	-	32×32×96
Conv3	3×3/1,1	32×32×384
MFM3	-	32×32×192
Pool3	2×2/2	16×16×192
Conv4a	1×1/1	16×16×384
MFM4a	-	16×16×192
Conv4	3×3/1,1	16×16×256
MFM4	-	16×16×128
Conv5a	1×1/1	16×16×256
MFM5a	-	16×16×128
Conv5	3×3/1,1	16×16×256
MFM5	-	16×16×128
Pool4	2×2/2	8×8×128
fc1	-	512
MFM_fc1	-	256

表 3 LightCNN-29 的结构

Table 3 Structure of LightCNN-29

层类型	卷积核尺寸/步长, 填充	输出尺寸
Conv1	5×5/1,2	128×128×96
MFM1	-	128×128×48
Pool1	2×2/2	64×64×48
Conv2_×	$\begin{bmatrix}3\times3/1,1\\3\times3/1,1\end{bmatrix}\times1$	64×64×48
Conv2a	1×1/1	64×64×96
MFM2a	-	64×64×48
Conv2	3×3/1,1	64×64×192
MFM2	-	64×64×96
Pool2	2×2/2	32×32×96
Conv3_×	$\begin{bmatrix}3\times3/1,1\\3\times3/1,1\end{bmatrix}\times2$	32×32×96
Conv3a	1×1/1	32×32×192
MFM3a	-	32×32×96
Conv3	3×3/1,1	32×32×384
MFM3	-	32×32×192
Pool3	2×2/2	16×16×192
Conv4_×	$\begin{bmatrix}3\times3/1,1\\3\times3/1,1\end{bmatrix}\times3$	16×16×192
Conv4a	1×1/1	16×16×384
MFM4a	-	16×16×192
Conv4	3×3/1,1	16×16×256
MFM4	-	16×16×128
Conv5_×	$\begin{bmatrix}3\times3/1,1\\3\times3/1,1\end{bmatrix}\times4$	16×16×128
Conv5a	1×1/1	16×16×256
MFM5a	-	16×16×128
Conv5	3×3/1,1	16×16×256
MFM5	-	16×16×128
Pool4	2×2/2	8×8×128
fc1	-	512
MFM_fc1	-	256

4.4.2 所使用 Inception 网络结构

在 Inception 网络的第一个版本 Inception-v1[45] 面世后, 研究者又在此基础上陆续提出了几种改进的版本。Inception-v2[46]在 Inception-v1 的基础上, 引入了批标准化(batch normalization, BN)层, BN 层可以在一定程度上使网络避免出现梯度消失现象并提升网络收敛速度。Inception-v3[47]对 Inception 网络中的各模块进行了进一步的分解, 加快了网络的运算速度。

4.4.3 所使用的残差网络结构:

残差网络[49]分为 ResNet-18、ResNet-34、ResNet-50、ResNet-101 和 ResNet-152 等结构。其中 ResNet-50、ResNet-101 和 ResNet-152 三种网络结构由于有较大的深度, 故而检测性能比 ResNet-18、ResNet-34 两种网络更好。

ResNet-50、ResNet-101 和 ResNet-150 三种网络结构如表 4 所示。在表 4 中, $\begin{bmatrix}1\times1,64\\3\times3,64\\1\times1,256\end{bmatrix}$ 模块所代表的结构如图 11 所示。

4.4.4 所使用的孪生网络的结构

孪生神经网络的两个分支可以分别从输入样本中提取特征, 在训练阶段, 该网络的对比损失函数会依据相似度标签来不断优化在这两分支的输入样本中提取的特征; 在测试阶段, 孪生神经网络可以对两分支中提取的特征进行相似度度量。因此, 我们将孪生网络融入到所提出的基于帧间差异的人脸篡改视频检测框架中来进行篡改检测。基于帧间差异的人脸篡改视频检测框架下基于孪生网络的检测方法的流程如图 12 所示。

由于对比损失函数可以根据相似度标签动态地优化孪生网络两个分支所提取的特征, 因此我们使用对比损失函数来作为孪生网络的损失函数。另外, 在我们使用的对比损失函数中, 使用欧氏距离来衡量两个分支所提取特征的差异。考虑到 LightCNN 在人脸识别领域, Inception 网络和残差网络在图像分类与检测领域都有着良好的表现, 我们以这三种网络作为孪生网络的主干网络。另外, 除了孪生网络中本身带有的以欧氏距离来表示差异的方式外, 我们还额外使用特征相减并取绝对值的方式对视频相邻帧中人脸图像的差异进行表示。

表 4 ResNet-50、ResNet-101 和 ResNet-152 网络的结构

Table 4 Structure of ResNet-50, ResNet-101 and ResNet-152

	ResNet-50	ResNet-101	ResNet-152
卷积层	卷积核尺寸 7×7, 64, 步长 2		
池化层	池化核尺寸 3×3,最大池化, 步长 2		
模块 1	$\begin{bmatrix}1\times1,64\\3\times3,64\\1\times1,256\end{bmatrix}\times3$	$\begin{bmatrix}1\times1,64\\3\times3,64\\1\times1,256\end{bmatrix}\times3$	$\begin{bmatrix}1\times1,64\\3\times3,64\\1\times1,256\end{bmatrix}\times3$
模块 2	$\begin{bmatrix}1\times1,128\\3\times3,128\\1\times1,512\end{bmatrix}\times4$	$\begin{bmatrix}1\times1,128\\3\times3,128\\1\times1,512\end{bmatrix}\times4$	$\begin{bmatrix}1\times1,128\\3\times3,128\\1\times1,512\end{bmatrix}\times8$
模块 3	$\begin{bmatrix}1\times1,256\\3\times3,256\\1\times1,1024\end{bmatrix}\times6$	$\begin{bmatrix}1\times1,256\\3\times3,256\\1\times1,1024\end{bmatrix}\times23$	$\begin{bmatrix}1\times1,256\\3\times3,256\\1\times1,1024\end{bmatrix}\times36$
模块 4	$\begin{bmatrix}1\times1,512\\3\times3,512\\1\times1,2048\end{bmatrix}\times3$	$\begin{bmatrix}1\times1,512\\3\times3,512\\1\times1,2048\end{bmatrix}\times3$	$\begin{bmatrix}1\times1,512\\3\times3,512\\1\times1,2048\end{bmatrix}\times3$
池化层	平均池化		
全连接	输出尺寸 1000 维		
分类层	softmax		

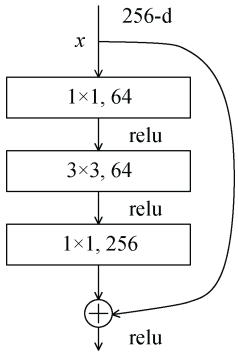

图 11 残差网络中一个单元的结构图

Figure 11 Diagram of the structure of one unit in Residual network

使用欧氏距离和特征相减并取绝对值的方式来表示在视频相邻帧人脸图像中提取的特征 $X=(x_1,x_2,\cdots,x_n)$ 和 $Y=(y_1,y_2,\cdots,y_n)$ 之间的差异 $diff(X,Y)$，计算方法分别如公式(6)和公式(5)所示。

$$diff(X,Y)=\sqrt{(x_1-y_1)^2+(x_2-y_2)^2+\cdots+(x_n-y_n)^2} \tag{6}$$

4.4.5 孪生网络所使用的对比损失函数

在基于孪生网络的检测方法中，我们使用对比损失来作为网络的损失函数。一对输入样本的对比损失 L 如公式(7)所示，如果输入中含有多对样本，则最终的损失函数值为各样本对损失值的平均值。

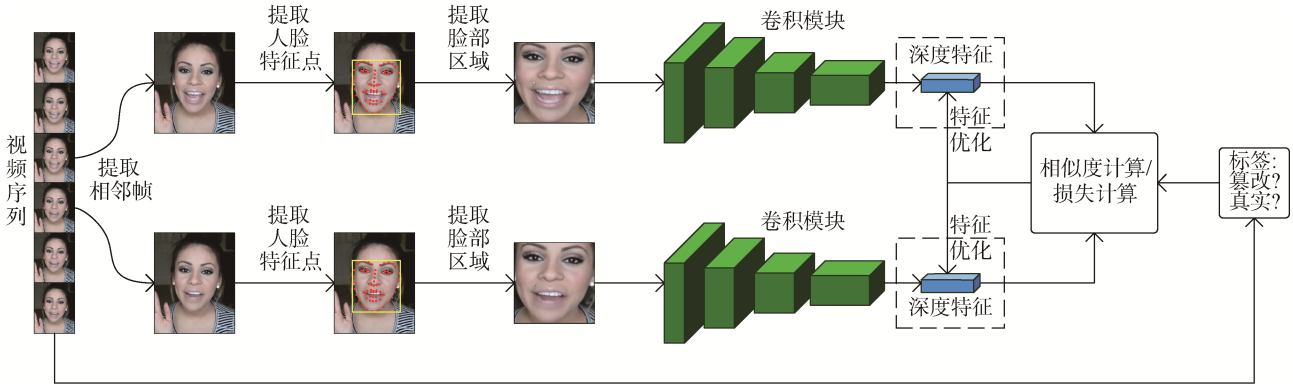

图 12 基于帧间差异的人脸篡改视频检测框架下基于孪生网络的检测方法流程图

Figure 12 Flow diagram of the Siamese-network-based detection method under the interframe-difference-based detection framework for human-face-tampered videos

(注: 虚线框内的为使用孪生网络提取的深度特征)

$$L=\frac{1}{2}(YD^2+(1-Y)(\max(m-D,0))^2) \tag{7}$$

在公式(7)中，Y 为两分支输入样本的相似度标签，其中当两个输入相似时，标签 Y 等于 1；而当两个输入不相似时，标签 Y 等于 0；D 为孪生网络两分支所提取特征的相似程度，在这里我们用欧氏距离来表示两分支所提取特征的相似程度，如公式(8)所示；而 m 代表相似度阈值；

$$\begin{aligned} D &= D(A,B)=\|A-B\|_2 \\ &=\sqrt{(a_1-b_1)^2+(a_2-b_2)^2+\cdots+(a_n-b_n)^2} \end{aligned} \tag{8}$$

在公式(8)中，两输入分支所提取的特征分别为 $A=(a_1,a_2,\cdots a_n)$ 和 $B=(b_1,b_2,\cdots b_n)$，n 为特征 A 与 B 的特征维度。

对比损失函数的具体作用如下：(1)当两输入的样本相似时，标签 $Y=1$，在这种情况下公式(7)中损失函数 L 的后一项为 0，只剩下前一项。如果网络两个分支所提取特征的距离较大，表明网络所提取的特征并不理想，此时网络的损失函数值会增加以改进所提取的特征。(2)而当两输入的样本不相似时，标签 $Y=0$，在这种情况下公式(7)中损失函数 L 的前一项为 0，只剩下后一项。如果网络两个分支所提取特征的距离较小，表明网络所提取的特征不理想，此时网络的损失函数值同样会增加以改进所提取的特征。

5 人脸篡改视频的检测实验

本节描述的是人脸篡改视频检测方法的实验部分。包括数据集的介绍、实验评价标准的介绍、实验实现细节的说明和实验结果的展示与分析。

5.1 实验所用数据集

我们进行的篡改视频检测实验主要使用两种数据集，FaceForensics++数据集[10]和 TIMIT 数据集[51-52]，下面分别介绍这两种数据集的具体情况。

5.1.1 FaceForensics++数据集

FaceForensics++[10]是一个公开的视频篡改检测数据集，该数据集包含 4000 个不同的视频，分别是从互联网上下载的 1000 个真实视频和由这些视频合成的的 Deepfake 篡改视频、Face2face 篡改视频和 Faceswap[53] 篡改视频各 1000 个。FaceForensics++数据集中的大部分视频都来自于 Youtube 网站，该数据集中的视频都经过了一定筛选，目的是使这些视频较少出现脸部被遮挡等无法检测出人脸的状况。图 13 展示的是从 FaceForensics++数据集中选出的视频截图，其中第一行是真实视频截图，第二行是与第一行对应的 Deepfake 篡改视频截图，第三行为与第一行对应的 Face2face 篡改视频截图。

图 13 FaceForensics++数据集概况

Figure 13 Overview of the FaceForensics++ dataset

FaceForensics++数据库中的所有 4000 个视频均有三种不同压缩率的版本，依次为无压缩的版本(记为 C0)、轻微压缩的版本(压缩量化参数为 23，记为 C23)和严重压缩的版本(压缩量化参数为 40，记为 C40)。图 14(a)(b)(c)分别展示了从 FaceForensics++数据集里 C0、C23、C40 三种版本的视频中截取的人脸图像。

(a) (b) (c)

图 14 FaceForensics++数据集中三种压缩率的视频

Figure 14 Videos of three compression rates in FaceForensics++ dataset

5.1.2 TIMIT 数据集

TIMIT 数据集由 VidTIMIT 数据集和 DeepfakeTIMIT 数据集构成。VidTIMIT 数据集[51]全部由真实视频组成，该数据集中的视频记录了人物面向镜头朗读短句的场景，每个视频中只有一个人出现。该数据集涉及 43 个不同人物，每个人物有十多段视频。DeepfakeTIMIT[52]是另外一个公开的人脸篡改视频数据集，该数据集的所有视频都是以 VidTIMIT 数据集为基础制备的，且都是 Deepfake 篡改视频。DeepfakeTIMIT 数据集的开发者从 VidTIMIT 数据集所涉及的 43 个人中选取出 32 个人

组成 16 对, 并让每对中的两个人互相换脸来制备篡改视频。对于 DeepfakeTIMIT 数据集中的 16 对人物, 每对中的两个人的外表都比较相似, 这样可以使制备出的篡改视频效果更加逼真。在 DeepfakeTIMIT 数据集中, 每个人物各有 10 段视频。DeepfakeTIMIT 数据集的开发者使用高分辨率、低分辨率两种 GAN 模型来制作 Deepfake 篡改视频, 高分辨率 GAN 模型输出图像的尺寸为128×128 ,而低分辨率 GAN 模型的删除图像尺寸为 64×64 。因此, DeepfakeTIMIT 数据集中的所有篡改视频均有两种版本。

从 TIMIT 数据集中选取的视频截图如图 15 所示, 其中第一行是从 VidTIMIT 数据集中选出的真实视频截图, 第二行是从 DeepfakeTIMIT 数据集中选出的与第一行截图对应的 Deepfake 篡改视频截图。

图 15 TIMIT 数据集概况
Figure 15 Overview of the TIMIT dataset

值得注意的是, TIMIT 数据集与 FaceForensics++数据集的压缩率不同, 所用的 Deepfake 工具的版本和参数设置也不尽相同。所以, 使用在 FaceForensics++数据集上训练的模型来对 TIMIT 数据集进行检测是一个比较有难度任务。

5.2 实验评价标准

在我们的实验中, 对于基于孪生网络的检测方法, 使用图像级的检测准确率和视频级的检测准确率两种评价标准。图像级的检测准确率指的是对视频中提取的每一对帧样本单独预测相似度标签并以视频帧为单位统计准确率; 视频级的检测准确率指的是用每段视频中所有提取样本对的相似度均值来计算该视频的预测标签, 最后以视频为单位统计准确率。

而对于基于 LBP/HOG 特征的方法, 无法进行视频级准确率的计算, 所以只统计图像级的检测准确率。

5.3 实验的实现细节

在 FaceForensics++数据集中, 我们从 1000 个真实视频中随机选取 700 个作为训练集, 剩下的 300 个作为测试集; 对于 1000 个篡改视频, 我们选取与真实视频对应的 700 个作为训练集, 剩下的 300 个作为测试集; 在这种情况下, 最终训练集有 1400 个真伪视频, 而测试集有 600 个真伪视频。在我们的实验中, 我们均匀地从 FaceForensics++数据集的每个视频中截取 50 对相邻帧, 作为训练和测试阶段所用的样本。三种不同压缩率的视频(C0、C23 和 C40)都以前面所讲的方式进行处理来制作训练和测试所用样本。

在 TIMIT 数据集中, 我们以 DeepfakeTIMIT 数据集作为负样本, 并从 VidTIMIT 数据集中选出与 DeepfakeTIMIT 数据集涉及的 32 个人物对应的视频作为正样本。在 TIMIT 数据集中, 我们从每段视频中均匀截取 20 对相邻帧来进行实验。对于在 TIMIT 数据集上训练和测试的实验, 我们从 32 个人物中随机挑选 24 个人物, 以这 24 个人物的有关视频作为训练集, 以剩下的 8 个人物的有关视频作为测试集。我们只用 DeepfakeTIMIT 数据集中低分辨率 GAN 模型制作出的篡改视频来进行实验。

我们使用的孪生神经网络是基于 Tensorflow 来搭建的。我们使用梯度下降法来训练网络并利用 Adam 优化器来对网络参数进行优化。每个训练批次有 30 个样本, 初始学习率为 1e-4。总共训练轮次为 12000, 每过 3000 轮后, 学习率降到之前的 10%。孪生网络使用以欧氏距离来作为距离度量的对比损失函数, 对比损失的相似度阈值为 0.5。在使用孪生网络提取特征的过程中, 我们分别使用以 LightCNN-4 为主干的孪生网络(记为 Siamese-LightCNN-4)、以 LightCNN-9 为主干的孪生网络(记为 Siamese-LightCNN-9)、以 LightCNN-29 为主干的孪生网络(记为 Siamese-LightCNN-29)、以 Inception-v1 为主干的孪生网络(记为 Siamese-Inception-v1)、以 Inception-v2 为主干的孪生网络(记为 Siamese-Inception-v2)、以 Inception-v3 为主干的孪生网络(记为 Siamese- Inception-v3)、以 ResNet-50 为主干的孪生网络(记为 Siamese-ResNet-50)、以 ResNet-101 为主干的孪生网络(记为 Siamese-ResNet-101)和以 ResNet-152 为主干的孪生网络(记为 Siamese-ResNet-152)共六种网络结构来进行实验。在我们使用的孪生网络的结构中, 每个分支最终分别输出 64 维特征来对输入图像进行表示。在使用特征相减并取绝对值的方式来对相邻帧中人脸图像的差异进行表示时, 我们使用径向基函数(Radial basis function, RBF)核的 SVM 来作为分类器对样本进行分类, 其中参数 γ 为 1 或 2, 惩罚因子 C 为 0.8。在所有数据集中, 我们将提取的人脸统一

缩放到128×128 尺寸来进行实验。

5.4 实验结果展示与分析

以下对基于 LBP/HOG 特征的检测方法和基于孪生网络的检测方法的实验结果进行展示与分析。

5.4.1 基于 LBP/HOG 特征的检测方法的实验结果与分析

表 5 展示的是在基于 LBP/HOG 特征的检测方法中, 当对 Deepfake 和 Face2face 篡改视频帧中人脸图像提取不同的 LBP 或 HOG 特征时的图像级检测准确率比较。实验使用 FaceForensics++数据集作为训练集和测试集, 使用特征相减并取绝对值的方式对相邻帧中人脸图像差异进行表示。提取的 LBP 特征为等价模式 LBP 特征, 在提取 LBP 特征时, 将人脸图像分割成 8×8 共 64 个图像块; 在提取 HOG 特征时, 定义 12 个梯度方向, 每个像素每个单元大小为 16×16 像素, 每个图像块大小为 2×2 单元, 图像块步长为 16 个像素。

在表 5 中可以看到, 在使用等价模式 LBP 特征对压缩程度为 C0 和 C23 的 Deepfake 视频进行检测时, 检测准确率分别可以达到92.49%和86.32%;而对压缩程度为 C0 和 C23 的 Face2face 视频进行检测时, 检测准确率分别可以达到 88.83%和 78.84%, 这也初步验证了我们提出的基于视频帧间差异的检测框架的有效性。另外, 虽然在 C40 压缩程度的数据集中, 基于 LBP/HOG 特征的检测方法准确率不高, 距离实用还有一定差距, 但是也能说明基于帧间差异这样的研究思路和技术路线是正确的。在本文中, 我们研究的重点在于使用有关深度学习的方法, 即基于孪生网络的方法, 来提高检测准确率。

表 5 在基于 LBP/HOG 特征的检测方法中, 当对 Deepfake 和 Face2face 篡改视频帧中人脸图像提取不同的 LBP/HOG 特征时的图像级检测准确率比较(%)

Table 5 Comparison of image level detection accuracies of Deepfake and Face2face tampered videos among different types of LBP/HOG features extracted in human face images in video frames in LBP/HOG-feature-based detection method(%)

特征	特征维度	人脸篡改视频类型	γ	FaceForensics++ 数据集压缩程度		
				C0	C23	C40
LBP 特征	3776	Deepfake	1	91.91	85.37	76.73
			2	92.49	86.32	77.83
		Face2face	1	88.08	78.08	69.89
			2	88.83	78.84	70.63
HOG 特征	2352	Deepfake	1	82.09	79.41	73.61
			2	78.33	76.69	72.07
		Face2face	1	80.16	73.53	66.32
			2	77.35	70.93	63.60

5.4.2 基于孪生网络的检测方法的实验结果与分析

图 16 展示的是在基于孪生网络的检测方法中, 图像级检测准确率和训练轮次的关系。其中图 16(a)(b)分别展示的是在 Deepfake 篡改视频集上和在 Face2face 篡改视频集上进行训练与检测的实验结果。实验使用 Siamese-ResNet-50 网络进行训练和测试, 使用 FaceForensics++数据集作为训练集和测试集, 使用欧氏距离对视频相邻帧中人脸图像差异进行表示。

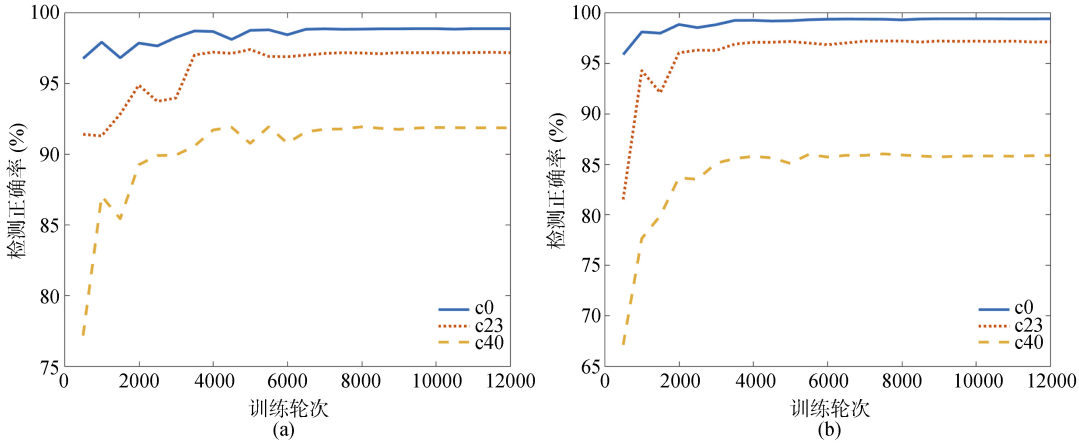

图 16 检测准确率和训练轮次的关系折线图

Figure 16 Line chart of the relationship between detection accuracy and training iteration

在图 16 中可以看出当使用训练轮次在 4000 轮之前的模型进行检测时, 准确率有比较大的波动; 当使用训练轮次在 4000 轮之后的模型进行检测时, 准确率基本趋于稳定。可见使用基于孪生网络的方

法对人脸篡改视频进行训练与检测, 最终都能得到一个稳定的检测准确率值。

表 6 展示的是在基于孪生网络的检测方法中, 当对 Face2face 篡改视频帧中人脸图像采用几种不同的孪生网络进行特征提取时的图像级检测准确率比较, 实验使用 FaceForensics++数据集作为训练集和测试集, 使用欧氏距离对相邻帧中人脸图像差异进行表示。

表 6 在基于孪生网络的检测方法中, 当对 Face2face 篡改视频帧中人脸图像采用几种不同的孪生网络进行特征提取时的图像级检测准确率比较(%)

Table 6 Comparison of image level detection accuracies of Face2face tampered videos among different Siamese networks via which features of human face images in video frames are extracted in Siamese-network-based detection method (%)

方法	FaceForensics++ 数据集压缩程度		
	C0	C23	C40
Fridrich & Kodovsky[54]	99.40	75.87	58.16
Cozzolino et al.[55]	**99.60**	79.80	55.77
Bayar & Stamm[56]	99.53	86.10	73.63
Rahmouni et al.[57]	98.60	88.50	61.50
Raghavendra et al.[58]	97.70	93.50	82.13
Meso-4[42]	94.60	92.40	83.20
MesoInception-4[42]	96.80	93.40	81.30
Nguyen et al.[59]	98.80	96.10	76.40
Capsule-Forensics[43]	99.13	97.13	81.20
Capsule-Forensics-Noise[43]	99.37	96.50	81.00
Siamese-LightCNN-4	95.76	90.68	83.93
Siamese-LightCNN-9	98.37	95.19	84.90
Siamese-LightCNN-29	97.93	93.08	85.71
Siamese-Inception-v1	98.17	95.80	84.29
Siamese-Inception-v2	98.98	**97.22**	**87.76**
Siamese-Inception-v3	98.48	94.69	86.00
Siamese-ResNet-50	99.34	97.10	85.85
Siamese-ResNet-101	99.12	96.59	85.06
Siamese-ResNet-152	98.96	96.51	85.58

(注: 粗体表示每列数据的最高值)

在表 6 中可以看到, 和 10 种其他文献中提出的方法相比, 我们所使用的九种孪生网络均能达到较高的检测准确率。其中 Siamese-Inception-v2 网络在 C23 和 C40 两种压缩率的视频集上有最高的检测准确率, 这些结果说明所提出的基于帧间差异的检测框架和基于孪生网络的检测方法是有效的。在 C40 压缩率的视频集上, Siamese-Inception-v2 网络的检测准确率明显超过表 6 中所列的其他文献方法, 这说明我们使用的基于孪生网络的检测方法在高压缩率的环境下有不错的表现。除此以外, 我们使用的基于孪生网络的检测方法在C0压缩率的视频集中也有较好的表现。

表 7 展示的是在基于孪生网络的检测方法中, 当对 Face2face 篡改视频帧中人脸图像采用几种不同的孪生网络进行特征提取时的视频级检测准确率比较。实验使用 FaceForensics++数据集作为训练集和测试集, 使用欧氏距离对相邻帧中人脸图像差异进行表示。

表 7 在基于孪生网络的检测方法中, 当对 Face2face 篡改视频帧中人脸图像采用几种不同的孪生网络进行特征提取时的视频级检测准确率比较(%)

Table 7 Comparison of video level detection accuracies of Face2face tampered videos among different Siamese networks via which features of human face images in video frames are extracted in Siamese-network-based detection method (%)

方法	FaceForensics++ 数据集压缩程度		
	C0	C23	C40
Meso-4[42]	-	95.30	-
MesoInception-4[42]	-	95.30	-
Capsule-Forensics[43]	99.33	98.00	82.00
Capsule-Forensics-Noise[43]	99.33	96.00	83.33
Siamese-LightCNN-4	97.33	94.67	89.17
Siamese-LightCNN-9	99.33	97.83	89.50
Siamese-LightCNN-29	99.17	94.83	90.33
Siamese-Inception-v1	99.33	97.50	88.33
Siamese-Inception-v2	99.50	99.00	**91.83**
Siamese-Inception-v3	99.00	96.83	89.50
Siamese-ResNet-50	**99.83**	**99.33**	90.83
Siamese-ResNet-101	**99.83**	**99.33**	89.00
Siamese-ResNet-152	99.33	98.83	89.50

(注: 粗体表示每列数据的最高值)

在表 7 中可以看到, 相比于 4 种其他文献提出的方法, 我们所使用的九种孪生网络均能达到较高的检测准确率。其中 Siamese-ResNet-50 和 Siamese-ResNet-101 网络同时在 C0 和 C23 两种压缩率的视频集上有最高的检测准确率; 而我们所用的 Siamese-Inceptoinv2 网络在 C40 压缩率的视频集上有最高的检测准确率, 这些结果说明所提出的基于帧间差异的检测框架和基于孪生网络的检测方法是有效的。在 C40 压缩率的视频集上,基于孪生网络方法的视频级检测准确率更是明显超过 Capsule-Forensics[43]和 Capsule-Forensics-Noise[43], 这说明我

们使用的基于孪生网络的检测方法在高压缩率的环境下有不错的表现。

表 8 展示的是在基于孪生网络的检测方法中, 当对 Deepfake 篡改视频帧中人脸图像采用几种不同的孪生网络进行特征提取时的图像级检测准确率比较。实验使用 FaceForensics++数据集作为训练集和测试集, 使用欧氏距离对相邻帧中人脸图像差异进行表示。

表 8 在基于孪生网络的检测方法中, 当对 Deepfake 篡改视频帧中人脸图像采用几种不同的孪生网络进行特征提取时的图像级检测准确率比较(%)

Table 8 Comparison of image level detection accuracies of Deepfake tampered videos among different Siamese networks via which features of human face images in video frames are extracted in Siamese-network-based detection method (%)

网络	FaceForensics++数据集压缩程度		
	C0	C23	C40
Siamese-LightCNN-4	94.25	91.44	87.27
Siamese-LightCNN-9	95.75	92.53	88.42
Siamese-LightCNN-29	96.09	92.99	88.90
Siamese-Inception-v1	97.08	95.26	91.28
Siamese-Inception-v2	98.50	96.10	**92.16**
Siamese-Inception-v3	**99.08**	96.01	91.46
Siamese-ResNet-50	98.86	**97.08**	91.84
Siamese-ResNet-101	98.83	96.39	91.57
Siamese-ResNet-152	98.79	96.46	91.08

(注: 粗体表示每列数据的最高值)

在表 8 中可以看到, 我们所使用的基于孪生网络的检测方法中, 使用九种网络结构在三种压缩程度的视频集中均能达到较高的检测准确率, 这些结果验证了所提出的基于帧间差异的检测框架和基于孪生网络的检测方法的有效性。其中在高压缩率的视频集(C40)中能有高达 92.16%的图像级检测准确率, 这说明基于孪生网络的检测方法可以较为有效地抵御压缩处理。

表 9 展示的是在基于孪生网络的检测方法中, 当对 Deepfake 篡改视频帧中人脸图像采用几种不同的孪生网络进行特征提取时的视频级检测准确率比较, 实验使用 FaceForensics++数据集作为训练集和测试集, 使用欧氏距离对相邻帧中人脸图像差异进行表示。

在表 9 中可以看到, 我们所使用的基于孪生网络的检测方法中, 使用九种网络结构在三种压缩程度的视频集中均能达到较高的检测准确率, 这些结果验证了所提出的基于帧间差异的检测框架和基于孪生网络的检测方法的有效性。其中在高压缩率的视频集(C40)中能有高达 95.33%的视频级检测准确率, 这说明基于孪生网络的检测方法可以较为有效地抵御压缩处理。

表 10 展示的是在基于孪生网络的检测方法中, 在基于孪生网络的检测方法中, 在 TIMIT 数据集上进行实验的图像级和视频级检测准确率比较。实验使用 TIMIT 数据集作为训练集和测试集, 使用欧氏距离对相邻帧中人脸图像差异进行表示。

表 9 在基于孪生网络的检测方法中, 当对 Deepfake 篡改视频帧中人脸图像采用几种不同的孪生网络进行特征提取时的视频级检测准确率比较(%)

Table 9 Comparison of video level detection accuracies of Deepfake tampered videos among different Siamese networks via which features of human face images in video frames are extracted in Siamese-network-based detection method (%)

网络	FaceForensics++数据集压缩程度		
	C0	C23	C40
Siamese-LightCNN-4	95.33	93.50	92.17
Siamese-LightCNN-9	96.83	95.50	91.17
Siamese-LightCNN-29	96.67	95.33	91.83
Siamese-Inception-v1	98.50	97.83	**95.33**
Siamese-Inception-v2	98.83	98.00	**95.33**
Siamese-Inception-v3	**99.00**	97.83	**95.33**
Siamese-ResNet-50	98.83	**98.50**	95.00
Siamese-ResNet-101	98.83	97.50	94.83
Siamese-ResNet-152	**99.00**	96.83	94.50

(注: 粗体表示每列数据的最高值)

表 10 在基于孪生网络的检测方法中, 在 TIMIT 数据集上的图像级和视频级检测准确率比较(%)

Table 10 Comparison of image level and video level detection accuracies on TIMIT dataset in Siamese-network-based detection method (%)

网络	图像级检测准确率	视频级检测准确率
Siamese-LightCNN-4	90.29	94.38
Siamese-LightCNN-9	97.50	**100.00**
Siamese-LightCNN-29	97.54	**100.00**
Siamese-Inception-v1	90.74	91.88
Siamese-Inception-v2	97.58	98.13
Siamese-Inception-v3	93.77	93.75
Siamese-ResNet-50	95.70	98.13
Siamese-ResNet-101	**98.07**	99.38
Siamese-ResNet-152	93.98	95.63

(注: 粗体表示每列数据的最高值)

在表 10 中可以看到, 我们所使用的基于孪生网络的检测方法中, 所用的九种网络结构在 TIMIT 数据集上都能达到较高的检测准确率。尤其是 Siamese-LightCNN-9 和 Siamese-LightCNN-29 结构, 达到了 100%的视频级检测准确率。

表 11 展示的是在基于孪生网络的检测方法中, 当对 Deepfake 和 Face2face 篡改视频相邻帧中人脸图像差异使用两种方式进行表示时的图像级检测准确率比较, 实验使用的是 Siamese-ResNet-50 网络,使用 FaceForensics++数据集作为训练集和测试集。在使用特征相减并取绝对值的方式来对相邻帧中人脸图像的差异进行表示时, 所使用的 RBF 核的 SVM 的参数 γ 值为 2。

从表 11 中可以看出基于孪生网络的检测方法在使用两种不同差异表示方式的情况下都能保持较高的检测准确率。

表 11　在基于孪生网络的检测方法中, 当对 Deepfake 和 Face2face 篡改视频相邻帧中人脸图像差异使用两种方式进行表示时的图像级检测准确率比较(%)

Table 11　Comparison of image level detection accuracies of Deepfake and Face2face tampered videos among two types of ways to express differences of human face images in adjacent frames of videos in Siamese-network-based detection method (%)

相邻帧中人脸图像差异表示方式	维度	人脸篡改视频类型					
		Deepfake			Face2face		
		FaceForensics++ 数据集压缩程度			FaceForensics++ 数据集压缩程度		
		C0	C23	C40	C0	C23	C40
欧式距离	1	**98.86**	97.08	**91.84**	**99.34**	97.10	85.85
特征相减并取绝对值	64	98.82	**97.17**	91.82	99.32	**97.17**	**85.90**

(注: 粗体表示每列数据的最高值)

表 12 展示的在基于孪生网络的检测方法中, 当创建训练集和测试集的生成模型不相同时对 Deepfake 篡改视频图像级的检测准确率比较, 实验使用 FaceForensics++数据集作为训练集, 使用 TIMIT 数据集作为测试集; 使用欧氏距离对相邻帧中人脸图像差异进行表示。

表 12　在基于孪生网络的检测方法中, 当创建训练集和测试集的生成模型不相同时对 Deepfake 篡改视频图像级的检测准确率比较(%)

Table 12　Comparison of image level detection accuracies of Deepfake tampered videos when the training dataset and the testing dataset are created on different generation models in Siamese-network-based detection method (%)

网络	FaceForensics++数据集压缩程度		
	C0	C23	C40
Siamese-ResNet-50	67.65	75.15	69.00
Siamese-ResNet-101	**68.49**	**79.56**	66.95
Siamese-ResNet-152	68.30	77.16	**72.61**

(注: 粗体表示每列数据的最高值)

在表 12 中我们可以看到基于孪生网络的检测方法具有一定的泛化能力, 该方法在在当训练集和测试集生成模型不相同的情况下, 对 Deepfake 篡改视频的检测准确率最高可以到达 80%左右。

表 13 展示的是在基于孪生网络的检测方法中, 当对 Deepfake 篡改视频提取不同间隔的帧对来计算帧间差异时的图像级检测准确率比较。对视频提取相邻帧对进行帧间差异计算、提取间隔一帧的帧对进行帧间差异计算、提取间隔两帧的帧对进行帧间差异计算三种方式分别如图 17、图 18、图 19 所示。

表 13　在基于孪生网络的检测方法中, 当对 Deepfake 篡改视频提取不同间隔的帧对来计算帧间差异时的图像级检测准确率比较(%)

Table 13　Comparison of image level detection accuracies of Deepfake tampered videos among different intervals at which pairs of frames are extracted for calculating interframe differences in Siamese-network-based detection method (%)

方法	FaceForensics++ 数据集压缩程度		
	C0	C23	C40
对视频提取相邻帧对进行帧间差异计算	98.86	97.08	91.84
对视频提取间隔一帧的帧对进行帧间差异计算	98.70	**97.31**	**92.15**
对视频提取间隔两帧的帧对进行帧间差异计算	**98.88**	97.10	91.77

(注: 粗体表示每列数据的最高值)

实验使用 Siamese-ResNet-50 网络进行训练和测试, 使用 FaceForensics++数据集作为训练集和测试集, 使用欧氏距离对相邻帧中人脸图像差异进行表示。

在表 13 中可以看到, 在基于孪生网络的方法中, 不但使用相邻帧对比较视频帧间人脸图像差异可以很好地区分真实视频和篡改视频, 使用间隔一帧的帧对和使用间隔两帧的帧对来比较视频帧间人脸图像差异同样可以有较高的检测准确率。

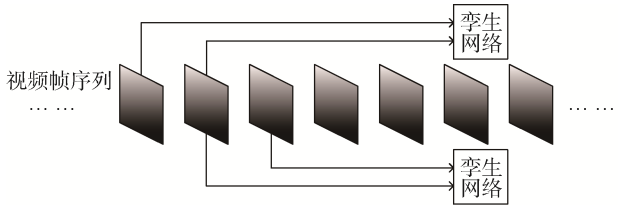

图 17　对视频提取相邻帧对进行帧间差异计算的示意图

Figure 17　Diagram of calculation of interframe differences between pairs of adjacent frames extracted from videos

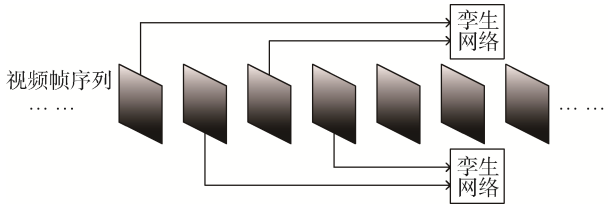

图 18　对视频提取间隔一帧的帧对进行帧间差异计算的示意图

Figure 18　Diagram of calculation of interframe differences between pairs of frames at interval of one frame extracted from videos

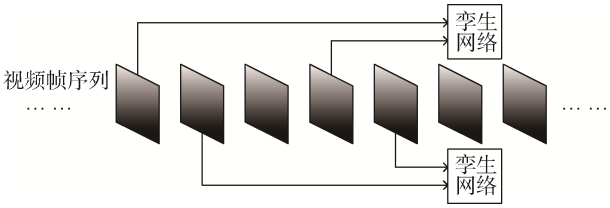

图 19　对视频提取间隔两帧的帧对进行帧间差异计算的示意图

Figure 19　Diagram of calculation of interframe differences between pairs of frames at interval of two frames extracted from videos

表 14 展示的是在基于孪生网络的检测方法中, 当对 Deepfake 篡改视频提取不同间隔的帧对来计算帧间差异时的视频级检测准确率比较。对视频提取相邻帧对进行帧间差异计算、提取间隔一帧的帧对进行帧间差异计算、提取间隔两帧的帧对进行帧间差异计算三种方式分别如图 17、图 18、图 19 所示。实验使用 Siamese-ResNet-50 网络进行训练和测试, 使用 FaceForensics++数据集作为训练集和测试集, 使用欧氏距离对相邻帧中人脸图像差异进行表示。

表 14　在基于孪生网络的检测方法中, 当对 Deepfake 篡改视频提取不同间隔的帧对来计算帧间差异时的视频级检测准确率比较(%)

Table 14　Comparison of video level detection accuracies of Deepfake tampered videos among different intervals at which pairs of frames are extracted for calculating interframe differences in Siamese-network-based detection method (%)

方法	FaceForensics++ 数据集压缩程度		
	C0	C23	C40
对视频提取相邻帧对进行帧间差异计算	98.83	98.50	95.00
对视频提取间隔一帧的帧对进行帧间差异计算	**99.00**	**98.83**	**95.33**
对视频提取间隔两帧的帧对进行帧间差异计算	**99.00**	**98.83**	94.83

(注: 粗体表示每列数据的最高值)

在表 14 中可以看到, 在基于孪生网络的方法中, 不但使用相邻帧对比较视频帧间人脸图像差异可以很好地区分真实视频和篡改视频, 使用间隔一帧的帧对和使用间隔两帧的帧对来比较视频帧间人脸图像差异同样可以有较高检测准确率。

表 15 展示的是在基于孪生网络的检测方法中, 当训练集和测试集压缩率不同时对 Deepfake 篡改视频的图像级检测准确率比较。实验使用 FaceForensics++数据集作为训练集和测试集, 但训练集和测试集样本的压缩率不同; 使用欧氏距离对相邻帧中人脸图像差异进行表示。

在表 15 中可以看到, 只有当使用压缩程度为 C0 的视频集所训练的模型去对压缩程度为 C40 的视频集进行检测时准确率较低, 在其他情况下使用基于孪生网络的检测方法都有较高的跨压缩率检测准确率。

表 16 展示的是在基于孪生网络的检测方法中, 当训练集和测试集压缩率不同时对 Face2face 篡改视频的图像级检测准确率比较。实验使用 FaceForensics++数据集作为训练集和测试集, 但训练集和测试集样本的压缩率不同; 使用欧氏距离对相邻帧中人脸图像差异进行表示。

表 15　在基于孪生网络的检测方法中，当训练集和测试集压缩率不同时对 Deepfake 篡改视频的图像级检测准确率比较(%)

Table 15　Comparison of image level detection accuracies of Deepfake tampered videos when the training dataset and the testing dataset have different compression rates in Siamese-network-based detection method (%)

网络	FaceForensics++数据集压缩程度					
	训练 C0		训练 C23		训练 C40	
	测试 C23	测试 C40	测试 C0	测试 C40	测试 C0	测试 C23
Siamese-ResNet-50	81.22	58.73	**97.49**	82.49	**93.69**	**93.83**
Siamese-ResNet-101	83.25	60.48	97.09	83.18	93.63	93.51
Siamese-ResNet-152	**83.67**	**61.48**	96.88	**84.99**	92.78	92.71

(注: 粗体表示每列数据的最高值)

表 16　在基于孪生网络的检测方法中，当训练集和测试集压缩率不同时对 Face2face 篡改视频的图像级检测准确率比较(%)

Table 16　Comparison of image level detection accuracies of Face2face tampered videos when the training dataset and the testing dataset have different compression rates in Siamese-network-based detection method (%)

网络	FaceForensics++数据集压缩程度					
	训练 C0		训练 C23		训练 C40	
	测试 C23	测试 C40	测试 C0	测试 C40	测试 C0	测试 C23
Siamese-ResNet-50	90.18	**60.02**	**98.43**	**71.70**	**90.58**	**89.67**
Siamese-ResNet-101	90.85	59.60	97.89	71.61	89.43	88.49
Siamese-ResNet-152	**92.11**	59.60	97.92	68.93	89.99	89.07

(注: 粗体表示每列数据的最高值)

在表 16 中可以看到，只有当使用压缩程度为 C0 或 C23 的视频集所训练的模型去对压缩程度为 C40 的视频集进行检测时准确率较低，在其他情况下使用基于孪生网络的检测方法都有较高的跨压缩率检测准确率。

表 17 展示的是在基于孪生网络的检测方法中，不同网络参数数量的对比。

表 17　在基于孪生网络的检测方法中，不同网络参数数量的对比

Table 17　The comparison of number of parameters for different networks in Siamese-network-based detection method

网络	参数数量
Siamese-LightCNN-4	**4146k**
Siamese-LightCNN-9	5519k
Siamese-LightCNN-29	12611k
Siamese-Inception-v1	7690k
Siamese-Inception-v2	12251k
Siamese-Inception-v3	22817k
Siamese-ResNet-50	23744k
Siamese-ResNet-101	42710k
Siamese-ResNet-152	58330k

(注: 粗体表示每列数据的最低值)

在表 17 中可以看到，Siamese-LightCNN-4 和 Siamese-LightCNN-9 用有最少的参数数量，这也意味着这两种网络所需要的存储空间较小，相比其他网络更有优势。

表 18 展示的是对 Deepfake 篡改视频不同方法的训练时间对比，表 19 展示的是对 Face2face 篡改视频不同方法的训练时间对比。所有实验使用 FaceForensics++数据集。对于基于 LBP/HOG 特征的方法，使用特征相减并取绝对值的方式对相邻帧中人脸图像差异进行表示，RBF 核 SVM 分类器的参数 γ 为 1。提取的 LBP 特征为等价模式 LBP 特征，在提取 LBP 特征时，将人脸图像分割成 8×8 共 64 个图像块；在提取 HOG 特征时，使用 12 个梯度方向，每个像素每个单元大小为 16×16 像素，每个图像块大小为 2×2 单元，图像块步长为 16 个像素。对于基于

通讯作者： 赵险峰，博士，研究员，zhaoxianfeng@iie.ac.cn
本课题得到国家重点研发计划课题(No. 19QY2202, No.19QY(Y)0207)；中国科学院信息工程研究所攀登计划项目
收稿日期: 2019-12-20；修改日期: 2020-03-09；定稿日期: 2020-03-10

孪生网络的方法, 使用欧氏距离对相邻帧中人脸图像差异进行表示。对于基于 LBP/HOG 特征的方法, 训练时间包括特征提取时间和训练 SVM 分类器所用时间两个部分。

在表 18 和表 19 中可以看到, 基于孪生网络检测方法的训练时间远远低于基于 LBP/HOG 特征方法的训练时间, 可见深度学习方法(基于孪生网络方法)相比于传统的人工设计特征的方法(基于 LBP/HOG 特征的方法)还是有明显的优势。另外, 在基于孪生网络的各方法中, Siamese-Inception-v1 拥有最快的速度, 其次是 Siamese-LightCNN-4 和 Siamese-LightCNN-9。

表 18　对 Deepfake 篡改视频不同方法的训练时间对比(秒)

Table 18　Comparison of training time of different methods for Deepfake tampered videos (s)

方法	FaceForensics++数据集压缩程度		
	C0	C23	C40
LBP	25588.19	35580.03	44603.80
HOG	41213.88	42004.62	45682.85
Siamese-LightCNN-4	2224.62	2395.06	2385.40
Siamese-LightCNN-9	2705.70	2548.23	2699.71
Siamese-LightCNN-29	4218.95	4281.51	4225.70
Siamese-Inception-v1	**1971.15**	**2024.69**	**1938.86**
Siamese-Inception-v2	3230.15	3235.27	2987.01
Siamese-Inception-v3	3158.05	3230.19	3388.81
Siamese-ResNet-50	2973.12	2708.79	2895.00
Siamese-ResNet-101	4115.10	3927.86	4120.49
Siamese-ResNet-152	5138.22	5142.36	5345.51

(注: 粗体表示每列数据的最低值)

表 19　对 Face2face 篡改视频不同方法的训练时间对比(秒)

Table 19　Comparision of training time of different methods for Face2face tampered videos (s)

方法	FaceForensics++数据集压缩程度		
	C0	C23	C40
LBP	33909.77	40886.64	46393.58
HOG	42719.78	47132.13	49718.70
Siamese-LightCNN-4	2389.10	2366.29	2164.15
Siamese-LightCNN-9	2644.46	2625.21	2667.41
Siamese-LightCNN-29	4280.06	4275.12	4293.32
Siamese-Inception-v1	**1931.25**	**2233.80**	**1992.70**
Siamese-Inception-v2	3025.71	3246.76	3206.91
Siamese-Inception-v3	3290.03	3419.52	3258.08
Siamese-ResNet-50	2883.53	2774.60	2734.70
Siamese-ResNet-101	4063.20	4098.18	4187.54
Siamese-ResNet-152	5300.67	5191.24	5180.10

(注: 粗体表示每列数据的最低值)

表 20 展示的是对 Deepfake 篡改视频不同方法的测试时间对比, 表 21 展示的是对 Face2face 篡改视频不同方法的测试时间对比。在表 20 和表 21 中, 统计的是每个图像对的检测时间。所有实验使用 FaceForensics++数据集。对于基于 LBP/HOG 特征的方法, 使用特征相减并取绝对值的方式对相邻帧中人脸图像差异进行表示, RBF 核 SVM 分类器的参数 γ 为 1。提取的 LBP 特征为等价模式 LBP 特征, 在提取 LBP 特征时, 将人脸图像分割成 8×8 共 64 个图像块; 在提取 HOG 特征时, 使用 12 个梯度方向, 每个像素每个单元大小为 16×16 像素, 每个图像块大小为 2×2 单元, 图像块步长为 16 个像素。对于基于孪生网络的方法, 使用欧氏距离对相邻帧中人脸图像差异进行表示。对于基于 LBP/HOG 特征的方法, 测试时间包括特征提取时间和用分类器进行分类所需时间两个部分。

表 20　对 Deepfake 篡改视频不同方法的测试时间对比(毫秒)

Table 20　Comparision of testing time of different methods for Deepfake tampered videos (ms)

方法	FaceForensics++数据集压缩程度		
	C0	C23	C40
LBP	238.4172	309.8581	390.1255
HOG	258.6365	270.6588	290.3943
Siamese-LightCNN-4	3.8691	3.8983	3.8779
Siamese-LightCNN-9	4.1242	4.1060	4.1219
Siamese-LightCNN-29	5.3720	5.3403	5.3047
Siamese-Inception-v1	**3.5386**	**3.5085**	**3.5177**
Siamese-Inception-v2	3.9279	3.7713	3.7245
Siamese-Inception-v3	4.0120	3.9441	4.0327
Siamese-ResNet-50	4.1798	4.0847	4.0552
Siamese-ResNet-101	4.8047	4.8283	4.8997
Siamese-ResNet-152	5.5888	5.6053	5.6054

(注: 粗体表示每列数据的最低值)

在表 20 和表 21 中可以看到, 基于孪生网络检测方法的测试时间明显低于基于 LBP/HOG 特征检测方法的测试时间, 可见深度学习方法(基于孪生网络方法)比传统的人工设计特征的方法(基于 LBP/HOG 特征的方法)的测试效率更高。除此以外, 我们还可以看到 Siamese-Inception-v1、Siamese-Inception-v2 和 Siamese-LightCNN-4 三种网络对每一对视频相邻帧的测试时间都可以低至 4 毫秒之内。

6 结论与未来工作

由于人脸篡改视频制作技术所产生的威胁日益增强, 我们急需有效的方法来对这些人脸篡改视频进行检测。在本文中, 我们提出了一种基于视频帧间差异的人脸篡改视频检测框架, 该框架可以反映视频中的人脸在被篡改前后帧间差异的不同。此外, 我们使用基于 LBP/HOG 特征的检测方法和基于孪生网络的检测方法来验证所提出的人脸篡改视频检测框架的有效性。在 FaceForensics++数据集上, 基于LBP/HOG 特征的检测方法可以有较高的检测准确率, 而基于孪生网络的检测方法能达到更高的检测准确率,且该方法有较强的鲁棒性;在这里鲁棒性是指一种检测方法可以在三种不同情况下达到较高的检测准确率, 这三种情况是: 对视频相邻帧中人脸图像差异用两种不同方式进行表示、提取三种不同间隔的帧对来计算帧间差异以及训练集与测试集压缩率不同。

表 21 对 Face2face 篡改视频不同方法的测试时间对比 (毫秒)

Table 21 Comparision of testing time of different methods for Face2face tampered videos (ms)

方法	FaceForensics++ 数据集压缩程度		
	C0	C23	C40
LBP	302.5017	392.1125	438.0663
HOG	276.6914	305.0928	330.7491
Siamese-LightCNN-4	3.9498	3.9216	3.8575
Siamese-LightCNN-9	4.2518	4.1828	4.1751
Siamese-LightCNN-29	5.3340	5.4202	5.2993
Siamese-Inception-v1	**3.5536**	**3.7009**	**3.6953**
Siamese-Inception-v2	3.8026	3.8222	3.7983
Siamese-Inception-v3	3.9423	3.9860	3.9677
Siamese-ResNet-50	4.1776	4.1641	4.1588
Siamese-ResNet-101	4.8845	4.8444	4.7913
Siamese-ResNet-152	5.6505	5.5947	5.6520

(注: 粗体表示每列数据的最低值)

目前有很多人脸篡改视频检测方法能够在训练集和测试集由同参数同类型的模型生成的情况下有较好的检测效果, 而一旦制备训练集和测试集所用的生成模型有较大差异, 检测效果会大打折扣。在将来, 研究者可以着力于开发更复杂更有效的人脸篡改视频检测网络以提高在训练集和测试集用不同模型生成情况下的检测准确率, 在这个过程中可以借鉴一些效果良好的人脸识别网络, 如 FaceNet、LightCNN 等。另外, 如果未来 Deepfake、Face2face 等篡改工具的开发者考虑到了在视频人脸篡改过程中对帧间关系的影响并对这些人脸篡改视频帧间的连续性进行优化, 我们依然可以利用在人脸篡改过程中左右脸出现的不对称来对这些人脸篡改视频进行检测。

致 谢 在此向本文成文中给予指导的老师、提供帮助的同学和给本文提出建议的评审专家表示感谢。

参考文献

[1] I.J. Goodfellow, J. Pouget-Abadie, M. Mirza, et al. Generative adversarial nets[C]. *International Conference on Neural Information, Processing Systems*, 2014: 34-42.

[2] Deepfake-Faceswap, https://github.com/deepfakes/faceswap.

[3] Faceswap-GAN, https://github.com/shaoanlu/faceswap-GAN.

[4] Thies J, Zollhofer M, Stamminger M, et al. Face2Face: Real-Time Face Capture and Reenactment of RGB Videos[C]. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 27-30, 2016. Las Vegas, NV, USA. Piscataway, NJ: IEEE, 2016: 2387-2395.

[5] Taigman Y, Yang M, Ranzato M, et al. DeepFace: Closing the Gap to Human-Level Performance in Face Verification[C]. *2014 IEEE Conference on Computer Vision and Pattern Recognition*, June 23-28, 2014. Columbus, OH, USA. Piscataway, NJ: IEEE, 2014: 1701-1708.

[6] Schroff F, Kalenichenko D, Philbin J. FaceNet: A Unified Embedding for Face Recognition and Clustering[C]. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 7-12, 2015. Boston, MA, USA. Piscataway, NJ: IEEE, 2015: 815-823.

[7] Wu X, He R, Sun Z N, et al. A Light CNN for Deep Face Representation with Noisy Labels[J]. *IEEE Transactions on Information Forensics and Security*, 2018, 13(11): 2884-2896.

[8] J. Bromley, I. Guyon, Y. LeCun,et al. Signature verification using a" siamese" time delay neural network[C]. *Advances in neural information processing systems*, 1994: 737-744.

[9] Chopra S, Hadsell R, LeCun Y. Learning a Similarity Metric Discriminatively, with Application to Face Verification[C]. *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, San Diego, CA, USA. Piscataway, NJ: IEEE, 2005 : 105-110.

[10] A. Rössler, D. Cozzolino, L. Verdoliva, et al. FaceForensics++: Learning to detect manipulated facial images[OL/IE],2019: arXiv preprint arXiv:1901.08971.

[11] Ojala T, Pietikainen M, Maenpaa T. Multiresolution Gray-scale and

Rotation Invariant Texture Classification with Local Binary Patterns[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2002, 24(7): 971-987.

[12] Dalal N, Triggs B. Histograms of Oriented Gradients for Human Detection[C]. *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, San Diego, CA, USA. Piscataway, NJ: IEEE, 2005: 177-181.

[13] M. Mirza, S. Osindero, Conditional generative adversarial nets[OL/IE], 2014: arXiv preprint arXiv:1411.1784.

[14] Isola P, Zhu J Y, Zhou T H, et al. Image-to-Image Translation with Conditional Adversarial Networks[C]. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 21-26, 2017. Honolulu, HI. Piscataway, NJ: IEEE, 2017: 1125-1134.

[15] Zhu J Y, Park T, Isola P, et al. Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks[C]//2017 IEEE International Conference on Computer Vision (ICCV), October 22-29, 2017. Venice. Piscataway, NJ: IEEE, 2017: 2223-2232.

[16] M.Y. Liu, T. Breuel, J. Kautz. Bayesian Image Super-resolution, Continued[M]. Advances in Neural Information Processing Systems 19. The MIT Press, 2007 .

[17] Huang R, Zhang S, Li T, et al. Beyond Face Rotation: Global and Local Perception GAN for Photorealistic and Identity Preserving Frontal View Synthesis[C]. *2017 IEEE International Conference on Computer Vision (ICCV)*, October 22-29, 2017. Venice. Piscataway, NJ: IEEE, 2017: 2439-2448.

[18] Y. Choi, M. Choi, M. Kim, et al. "Stargan: Unified generative adversarial networks for multi-domain image-to-image translation," *IEEE Conference on Computer Vision and Pattern Recognition*, 2018: 8789-8797.

[19] Choi Y, Choi M, Kim M, et al. StarGAN: Unified Generative Adversarial Networks for Multi-domain Image-to-Image Translation[C]. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 18-23, 2018. Salt Lake City, UT. Piscataway, NJ: IEEE, 2018: 8789-8797.

[20] Y. Jo, J. Park, SC-FEGAN: Face Editing Generative Adversarial Network with User's Sketch and Color[OL/IE].2019: arXiv preprint arXiv:1902.06838.

[21] Y. Li, S. Lyu. Exposing DeepFake Videos By Detecting Face Warping Artifacts. 2018: arXiv preprint arXiv:1811.00656.

[22] Dirik A E, Memon N. Image Tamper Detection Based on Demosaicing Artifacts[C]. *2009 16th IEEE International Conference on Image Processing (ICIP)*, November 7-10, 2009. Cairo, Egypt. Piscataway, NJ: IEEE, 2009: 1497-1500.

[23] Ferrara P, Bianchi T, de Rosa A, et al. Image Forgery Localization Via Fine-Grained Analysis of CFA Artifacts[J]. *IEEE Transactions on Information Forensics and Security*, 2012, 7(5): 1566-1577.

[24] X.L. Zhang, Z. Fang, X.P. Zhang, Forgery Detection via Inter-channel Correlation of CFA Images[J]. *JOURNAL OF APPLIED SCIENCES—Electronics and Information Engineering*, 2015, 33(1): 87-94.
(张晓琳, 方针, 张新鹏, 利用通道间相关性的 CFA 图像盲取证[J]. *应用科学学报*, 2015, 33(1): 87-94.)

[25] W.X. Su, Z. Fang, Identifying Image Authenticity Based on CFA Inconsistency of Interpolation Characteristics[J]. *JOURNAL OF APPLIED SCIENCES—Electronics and Information Engineering*, 2019, 37(1): 33-40.
(苏文煊, 方针, 基于 CFA 插值特性不一致的图像真伪鉴别[J]. *应用科学学报*, 2019, 37(1): 33-40.)

[26] S. Peng, Y.Y. Peng, C.Y. Xiao, Image tampering detection algorithm based on CFA interpolation[J]. *Transducer and Microsystem Technologies*, 2015, 34(6): 141-144.
(彭双, 彭圆圆, 肖昌炎, 基于 CFA 插值的图像篡改检测算法[J]. *传感器与微系统*, 2015, 34(6): 141-144.)

[27] Chierchia G, Parrilli S, Poggi G, et al. PRNU-based Detection of Small-size Image Forgeries[C]. *2011 17th International Conference on Digital Signal Processing (DSP)*, July 6-8, 2011. Corfu, Greece. Piscataway, NJ: IEEE, 2011: 1-6.

[28] Chierchia G, Poggi G, Sansone C, et al. A Bayesian-MRF Approach for PRNU-Based Image Forgery Detection[J]. *IEEE Transactions on Information Forensics and Security*, 2014, 9(4): 554-567.

[29] Lin X F, Li C T. Refining PRNU-based Detection of Image Forgeries[C]. *2016 Digital Media Industry & Academic Forum (DMIAF)*, July 4-6, 2016. Santorini, Greece. Piscataway, NJ: IEEE, 2016: 222-226.

[30] Bahrami K, Kot A C. Image Tampering Detection by Exposing Blur Type Inconsistency[C]. *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 4-9, 2014. Florence, Italy. Piscataway, NJ: IEEE, 2014: 2654-2658.

[31] Bahrami K, Kot A C, Li L D, et al. Blurred Image Splicing Localization by Exposing Blur Type Inconsistency[J]. *IEEE Transactions on Information Forensics and Security*, 2015, 10(5): 999-1009.

[32] Johnson M K, Farid H. Exposing Digital Forgeries through Chromatic Aberration[C]. *Proceeding of the 8th workshop on Multimedia and security - MM&Sec '06*, September 26-27, 2006. Geneva, Switzerland. New York, USA: ACM Press, 2006: 48-55.

[33] Mayer O, Stamm M C. Accurate and Efficient Image Forgery Detection Using Lateral Chromatic Aberration[J]. *IEEE Transactions on Information Forensics and Security*, 2018, 13(7): 1762-1777.

[34] Y.Z. Chen, Z. Fang, Detection of Digital Image Forgery Based on Chromatic Aberration[J]. *Journal of Applied Sciences*, 2015, 33(6): 604-614.
(陈竺益, 方针, 基于色像差特性的图像篡改检测[J]. *应用科学学报*, 2015, 33(6): 604-614.)

[35] Vazquez-Padin D, Comesana P, Perez-Gonzalez F. An SVD Approach to Forensic Image Resampling Detection[C]. *2015 23rd*

European Signal Processing Conference (EUSIPCO), August 31-September 4, 2015. Nice. Piscataway, NJ: IEEE, 2015: 2067-2071.

[36] Vazquez-Padin D, Perez-Gonzalez F, Comesana-Alfaro P. A Random Matrix Approach to the Forensic Analysis of Upscaled Images[J]. *IEEE Transactions on Information Forensics and Security*, 2017, 12(9): 2115-2130.

[37] Y. Li, M.C. Chang, S. Lyu, In ictu oculi: Exposing ai generated fake face videos by detecting eye blinking. 2018: arXiv preprint arXiv:1806.02877.

[38] Yang X, Li Y Z, Lyu S W. Exposing Deep Fakes Using Inconsistent Head Poses[C]. *ICASSP 2019 - 2019 IEEE International Conference on Acoustics*, Speech and Signal Processing (ICASSP), May 12-17, 2019. Brighton, United Kingdom. Piscataway, NJ: IEEE, 2019: 8261-8265.

[39] R. Wang, L. Ma, F. Juefei-Xu, et al. FakeSpotter: A Simple Baseline for Spotting AI-Synthesized Fake Faces.2019: arXiv preprint arXiv:1909.06122.

[40] H. Li, B. Li, S. Tan et al. Detection of deep network generated images using disparities in color components. 2018: arXiv preprint arXiv:1808.07276.

[41] X. Yang, Y. Li, H. Qi et al. Exposing GAN-synthesized Faces Using Landmark Locations. 2019: arXiv preprint arXiv:1904. 00167.

[42] Matern F, Riess C, Stamminger M. Exploiting Visual Artifacts to Expose Deepfakes and Face Manipulations[C]. *2019 IEEE Winter Applications of Computer Vision Workshops (WACVW)*, January 7-11, 2019. Waikoloa Village, HI, USA. Piscataway, NJ: IEEE, 2019: 83-92.

[43] 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 (WIFS)*, December 11-13, 2018. Hong Kong, China. Piscataway, NJ: IEEE, 2018: 1-7.

[44] 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 (ICASSP)*, May 12-17, 2019. Brighton, United Kingdom. Piscataway, NJ: IEEE, 2019: 2307-2311.

[45] E. Sabir, J. Cheng, A. Jaiswal, et al. Recurrent Convolutional Strategies for Face Manipulation Detection in Videos[J]. *Interfaces (GUI)*, 2019, 3(1):23-28.

[46] C. Szegedy, W. Liu, Y. Jia, et al. Going deeper with convolutions[C]. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015: 1-9.

[47] S. Ioffe, C. Szegedy, Batch normalization: Accelerating deep network training by reducing internal covariate shift. 2015: arXiv preprint arXiv:1502.03167.

[48] Szegedy C, Vanhoucke V, Ioffe S, et al. Rethinking the Inception Architecture for Computer Vision[C]. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 27-30, 2016. Las Vegas, NV, USA. Piscataway, NJ: IEEE, 2016: 2818-2826.

[49] C. Szegedy, S. Ioffe, V. Vanhoucke et al. Inception-v4, Inception-ResNet and the impact of residual connections on learning[C]. *Thirty-First AAAI Conference on Artificial Intelligence*, Feb. 2017:456-461.

[50] K. He, X. Zhang, S. Ren et al. “Deep residual learning for image recognition,” *IEEE conference on computer vision and pattern recognition*, 2016. 770-778.

[51] 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 (CVPR)*, June 27-30, 2016. Las Vegas, NV, USA. Piscataway, NJ: IEEE, 2016: 770-778.

[52] D.E. King, Dlib-ml: A machine learning toolkit[J]. *Journal of Machine Learning Research*, 2009, 10:1755-1758.

[53] Sanderson C, Lovell B C. Multi-Region Probabilistic Histograms for Robust and Scalable Identity Inference[M]. Advances in Biometrics. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009: 199-208.

[54] P. Korshunov, S. Marcel, Deepfakes: a new threat to face recognition? assessment and detection. 2018: arXiv preprint arXiv:1812. 08685.

[55] Faceswap, https://github.com/MarekKowalski/FaceSwap/.

[56] Fridrich J, Kodovsky J. Rich Models for Steganalysis of Digital Images[J]. *IEEE Transactions on Information Forensics and Security*, 2012, 7(3): 868-882.

[57] Cozzolino D, Poggi G, Verdoliva L. Recasting Residual-based Local Descriptors as Convolutional Neural Networks[C]. *Proceedings of the 5th ACM Workshop on Information Hiding and Multimedia Security*-IHMMSec '17, June 20-22, 2017. Philadelphia, Pennsylvania, USA. New York, USA: ACM Press, 2017: 159-164.

[58] 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, June 20-22, 2016. Vigo, Galicia, Spain. New York, USA: ACM Press, 2016: 5-10.

[59] Rahmouni N, Nozick V, Yamagishi J, et al. Distinguishing Computer Graphics from Natural Images Using Convolution Neural Networks[C]. *2017 IEEE Workshop on Information Forensics and Security (WIFS)*, December 4-7, 2017. Rennes. Piscataway, NJ: IEEE, 2017: 1-6.

[60] Raghavendra R, Raja K B, Venkatesh S, et al. Transferable Deep-CNN Features for Detecting Digital and Print-Scanned Morphed Face Images[C]. *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, July 21-26, 2017. Honolulu, HI, USA. Piscataway, NJ: IEEE, 2017: 1822-1830.

[61] Nguyen H H, Tieu TN D, Nguyen-Son H Q, et al. Modular Convolutional Neural Network for Discriminating between Computer-Generated Images and Photographic Images[C]. *Proceedings of the 13th International Conference on Availability, Reliability and Security* - ARES 2018, August 27-30, 2018. Hamburg, Germany. New York, USA: ACM Press, 2018: 1-7.

张怡暄 于 2013 年在华中科技大学光信息科学与技术专业获得学士学位, 现在中国科学院信息工程研究所信号与信息处理专业攻读硕士学位。研究兴趣包括人工智能、多媒体取证技术。Email: zhangyixuan@iie.ac.cn

曹纭 于 2012 年在中国科学院软件研究所获得博士学位。现任中国科学院信息工程研究所副研究员。研究领域为多媒体内容安全。研究兴趣包括: 隐写与隐写分析、数字内容取证等。Email: caoyun@iie.ac.cn

李根 于 2019 年在中国航天科工第二研究院计算机科学与技术专业获得工学硕士学位。现在中国科学院信息工程研究所网络空间安全专业攻读博士学位。研究领域为人工智能安全。研究兴趣包括: 多媒体信息取证技术。Email: ligen1@iie.ac.cn

赵险峰 中国科学院信息工程研究所研究员, 中国科学院大学网络空间安全学院教授, 博士生导师。2003 年于上海交通大学获博士学位, 研究方向为信息隐藏、多媒体取证与内容安全分析等。任 IJDCF、IWDW 等期刊、会议的编委、主席或委员, 任中国电子学会通信与信息安全专委会、中国图象图形学会多媒体取证与安全专委会等学术组织的委员。曾承担国家自然科学基金、国家重点研发计划、中科院战略性先导专项、部委专项等任务 40 余项, 在 IEEE TIFS、ACM IH & MMSec 等本领域重要刊物和会议上发表论文 150 余篇, 获得与申请专利 29 项, 撰写或参与撰写著作 5 部, 主持研制的系统有重要应用, 获保密科学技术奖(部级)一等奖、中科院“朱李月华”优秀教师、ACM IH & MMSec 最佳论文奖等荣誉。Email: zhaoxianfeng@iie.ac.cn