基于多级时空域 3D 卷积的换脸视频检测方法

包 晗[1,2], 符皓程[1,2], 曹 纭[1,2], 赵险峰[1,2], 汤 朋[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 近年来, 视频换脸技术发展迅速。该技术可被用于伪造视频来影响政治行动和获得不当利益,从而给社会带来严重危害, 目前已经引起了各国政府和舆论的广泛关注。本文通过分析现有的主流视频换脸生成技术和检测技术, 指出当前主流的生成方法在时域和空域中均具有伪造痕迹和生成损失。而当前基于神经网络检测合成人脸视频的算法大部分方法只考虑了空域的单幅图像特征, 并且在实际检测中有明显的过拟合问题。针对目前检测方法的不足, 本文提出一种高效的基于时空域结合的检测算法。该方法同时对视频换脸生成结果在空域与时域中的伪造痕迹进行捕捉, 其中, 针对单帧的空域特征设计了全卷积网络模块, 该模块采用 3D 卷积结构, 能够精确地提取视频帧阵列中每帧的伪造痕迹; 针对帧阵列的时域特征设计了卷积长短时记忆网络模块, 该模块能够检测伪造视频帧之间的时序伪造痕迹; 最后, 根据特征分类设计特征网络金字塔网络结构, 该结构能够融合不同尺寸的时空域特征, 通过多尺度融合来提高分类效果, 并减少过拟合现象。与现有方法相比, 该方法在训练中的收敛效果和分类效果方面有明显优势。除此之外, 我们在保证检测准确率的前提下采用较少的参数, 相比现有结构而言训练效率更高。

关键词 视频换脸; 神经网络检测; 卷积长短时记忆网络; 特征网络金字塔
中图法分类号 TP309.2、TP391.4 **DOI 号** 10.19363/J.cnki.cn10-1380/tn.2022.09.03

Multi-scale Time-Spatial Domain Detection of Fabricated Face Video Based on 3D Convolution

BAO Han[1,2], FU Haocheng[1,2], CAO Yun[1,2], ZHAO Xianfeng[1,2], TANG Peng[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 Recently, Deepfake technology has developed rapidly, which can be used to forge videos. The abuse of such fake videos has caused serious harm to society and has now attracted widespread attention from governments and public opinion. Based on a thorough investigation, this paper figures out that the current mainstream generation methods have forgery traces and generation losses in both the temporal and spatial domains. However, most of the current algorithms for detecting fabricated face videos based on neural networks only consider the features of a single image in the spatial domain, and have overfitting problems, resulting in low accuracy in actual detection. In order to solve the mentioned shortcomings, this paper evaluates the state-of-the-art detection algorithms of the Deepfake face and proposes an effective detection algorithm based on the combination of spatial and temporal features. Our network considers both spatial and temporal features of the fabricated face video. As for the single frame in the video, we present a fully convolutional network to extract the spatial feature. This module adopts a 3D convolution structure, which can accurately extract the forgery traces of each frame in the video frame array. As for frame array, we build a module based on a convolutional network with Long Short-Term Memory (LSTM) for temporal feature extraction. This module is able to detect timing forgery traces between fake video frames. At last, we apply Feature Pyramid Networks (FPN) to improve the accuracy of face classification. This structure can fuse Time-Spatial features of different sizes. It can improve the classification effect through multi-scale fusion and reduce overfitting. Comparative experiments have demonstrated that the proposed method is more effective in terms of the performance of training convergence and classification accuracy. In addition, we adopt fewer parameters and achieve high detection accuracy, resulting in higher training efficiency compared with the existing methods.

Key words deepfake videos; neural network detection; convolutional long and short-term memory; feature pyramid networks

通讯作者: 赵险峰, 研究员, Email: zhaoxianfeng@iie.ac.cn。
本课题得到国家重点研发计划课题(No. 2019QY2202, No. 2020AAA0140000)的资助。
收稿日期: 2019-12-31; 修改日期: 2020-04-01; 定稿日期: 2022-07-14

1 背景

视频换脸技术近年来发展迅速, 特别是随着深度学习技术的迅猛发展, 以 Deepfake 为代表的人工智能视频换脸技术已经可以用于生成以假乱真伪造视频[1]。该技术具有高度真实性、泛在普适性和快速演化性等特征。随着该技术的普及, 端到端的视频换脸软件层出不穷, 使得换脸技术的门槛越来越低, 这也使得该技术被恶意使用的可能性大大增加。利用视频换脸技术生成的伪造视频传播虚假信息的行为不仅会侵犯公民的合法权益、破坏社会稳定与国家安全, 还可能消解社会共同的信任基础[2]。

目前人脸生成方法主要包括假脸的生成替换和面部表情迁移两大类别。假脸的生成替换一般采用生成网络生成源视频的人脸, 生成的人脸的面部表情和目标视频的表情一致, 使用融合方法将生成源视频人物的脸替换为目标视频的人脸, 达到换脸的效果, 代表方法为 DeepFake 和 FaceSwap, 可用的生成软件为 FaceSwap[3], DeepFaceLab[4]等。该类方法以文献[5]为基础, 即编码-解码器的结构可以将人脸图像分为结构和风格两部分, 通过保留源图像结构并将风格替换为目标图像的方法, 实现人脸的风格转化。面部表情迁移则是通过神经网络或其他建模方法获得源视频面部表情动作特征, 更改目标视频的面部表情与之匹配。与假脸生成替换相比, 这种方法输出的目标视频的人脸仍为原始目标视频的人脸图像, 但其面部动作与源视频一致, 从而达到操控人物面部的效果, 其代表方法为 Face2Face 和 NeuralTexture。Face2Face 最早由 Thies 等人[6]提出, 采用 3D 建模的方法, 重建关于源人物的面部模型, 并通过特征匹配使模型具有与目标人物的表情。Thies 等人[7]修改神经网络生成模型, 基于 Gatys 等人[8]的层间损失函数和 Risser 等人[9]的内容损失函数, 提出了 NeuralTexture 方法。该方法被应用于人脸合成图像和视频中, 能够让生成的人脸视频更加富有细节, 其分辨率比 Face2Face 更高。2014 年 Goodfellow 等人提出生成对抗网络[10](Generative Adversarial Nets, GAN)为视频风格转换方法提供新方向。其中 Cycle-GAN[11]和 Recycle-GAN[12]是典型代表。传统的 DeepFake 和 Face2Face 系列结合了 GAN 模型可以用于优化生成效果。目前, FaceSwap 已经推出了基于最小二乘生成对抗网络[13](Least Squares Generative Adversarial Networks, LSGAN)的 FaceSwap-GAN[14]版本。而 Face2Face 模型结合了 Recycle-GAN 的生成策略, 相比传统的 3D 模型, Recycle-GAN 能更好地在时域方面实现风格转化。

最初人工生成人脸伪造检测使用通用的图像伪造检测方法[15-17], 利用神经网络提取图像的伪造或篡改痕迹, 其中篡改类型包括图像复制, 拼接, 润饰和增强等。随着生成方法的不断发展, 逐渐出现了人脸生成的专有检测方法, 包括生物特征的提取和神经网络自动分类等。典型的生物特征检测算法为眨眼检测, Li 等人[18]通过统计自然人像视频和伪造生成视频的眨眼频率的差异, 进行换脸伪造视频检测。典型的网络检测结构为 XceptionNet 和 MesoNet。XceptionNet[19] 将原来 InceptionNet[20] 的核心 Inception 模块改为深度可分离卷积(Depthwise Separable Convolution)的形式, 减少了网络规模。XceptionNet 在实现图像分类上非常实用, 使用 ImageNet 预处理模型, 并对其尾部进行处理, 可以对自然人脸和生成人脸进行分类。MesoNet 系列是 Afchar 等人[21]设计的包括 4 层卷积层和 2 层全连接层的轻量级分类网络。Meso-Inception-v4 网络将 MesoNet 前两层普通的卷积层替换为 Inception 模块, 在人脸伪造分类检测中具有不错的效果。

通过分析现阶段的分类检测网络, 我们发现了其中存在的一些不足:

(1) 大部分网络都是针对单幅图像进行检测, 并没有考虑时序特征, 具有局限性;

(2) 现在大部分网络存在较为严重的过拟合问题, 需要让神经网络的分类模块忽略整体内容而对异常部分更加敏感。

本文的主要贡献在于:

(1) 结合当前视频换脸技术的发展现状和当前检测网络的局限性提出一种新的检测模型(Multi-scale Time-Spatial domain 3D convolution network, MTS3DCNet)整体流程如图 1 所示, 该模型包括输入预处理、时空域结合的特征提取结构和多级预测结构, 可以提供端到端的分类结果;

(2) 使用特征金字塔结构缓解人脸伪造检测训练中过拟和的问题。

本文后续章节安排如下: 第二节介绍本文提出的网络的相关技术; 第三节具体介绍网络的结构; 第四节给出实验结果及分析; 第五节给出结论以及未来工作方向。

2 相关工作

2.1 3D 卷积结构

随着深度学习的发展, 卷积神经网络(Convolutional Neural Network, CNN)在图像领域应

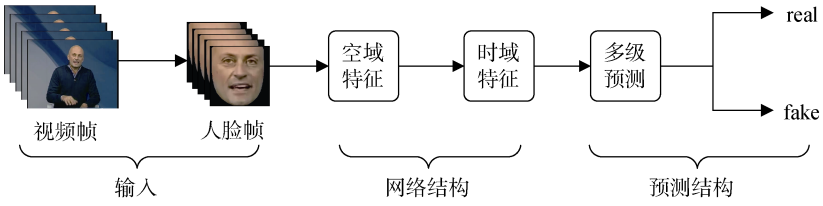


图 1　检测流程图

Figure 1　Detection process

用非常普遍。与 2D 卷积结构不同, 3D 卷积结构[22](3D convolutional neural networks)适合处理多幅图像的特征。卷积操作的维度为 3, 即卷积核增加时域方向上的维度, 通过堆叠多个连续的帧组成一个立方体, 然后在立方体中运用 3D 卷积核。在 3D 卷积中, 卷积核的结果中每一个特征体都是连续帧阵列相同位置的特征分布。卷积核在时间, 图像长和宽三个维度上移动卷积得到特征结果, 故能够得到帧阵列的时域特征。

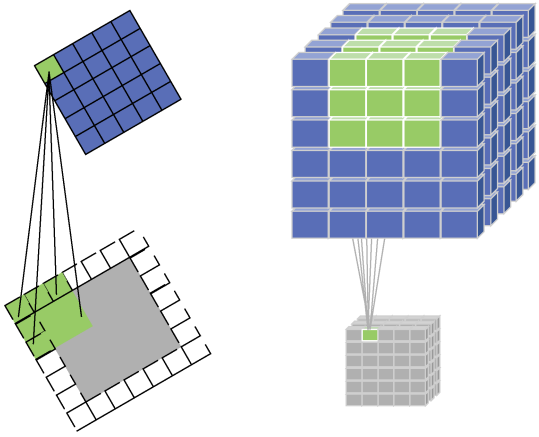


图 2　2D 卷积和 3D 卷积对比

Figure 2　Comparison of 2D convolution and 3D convolution

2.2　卷积长短时记忆网络

卷积长短时记忆网络[23](Convolutional Long short-term memory, ConvLSTM)是一种时间循环神经网络。由于一般的循环神经网络(Recurrent neural network, RNN)存在的长期依赖问题和梯度消失的问题, 长短时记忆网络[24](Long short-term memory, LSTM)应运而生。LSTM 具有遗忘门功能, 用于控制前向单元对于当前单元的影响程度。该网络结构能够很好地预测具有时域特征的数据中每个单元的状态, 得到预测结果。虽然 LSTM 在处理一维时序数据时具有优势, 但在处理二维或高维时序数据时, 其运算规则并不适合高维操作。故在 LSTM 的基础上加上卷积操作, 能够更有效地提取图像特征。

ConvLSTM 内部结构和 LSTM 相似, 如图 3 所示, 其状态更新公式(1~5)中 $i,f,\mathcal{C},o,\mathcal{H}$ 均为三维张量。

$$i_t=\sigma\left(\begin{matrix}W_{xi}*\mathcal{X}_t+W_{hi}*\mathcal{H}_{t-1}\\+W_{ci}\circ\mathcal{C}_{t-1}+b_i\end{matrix}\right) \tag{1}$$

$$f_t=\sigma\left(\begin{matrix}W_{xf}*\mathcal{X}_t+W_{hf}*\mathcal{H}_{t-1}\\+W_{cf}\circ\mathcal{C}_{t-1}+b_f\end{matrix}\right) \tag{2}$$

$$\mathcal{C}_t=f_t\circ\mathcal{C}_{t-1}+i_t\circ\tanh\left(\begin{matrix}W_{xc}*\mathcal{X}_t\\+W_{hc}*\mathcal{H}_{t-1}+b_c\end{matrix}\right) \tag{3}$$

$$o_t=\sigma\left(\begin{matrix}W_{xo}*\mathcal{X}_t+W_{ho}*\mathcal{H}_{t-1}\\+W_{co}\circ\mathcal{C}_t+b_o\end{matrix}\right) \tag{4}$$

$$\mathcal{H}_t=o_t\circ\tanh\left(\mathcal{C}_t\right) \tag{5}$$

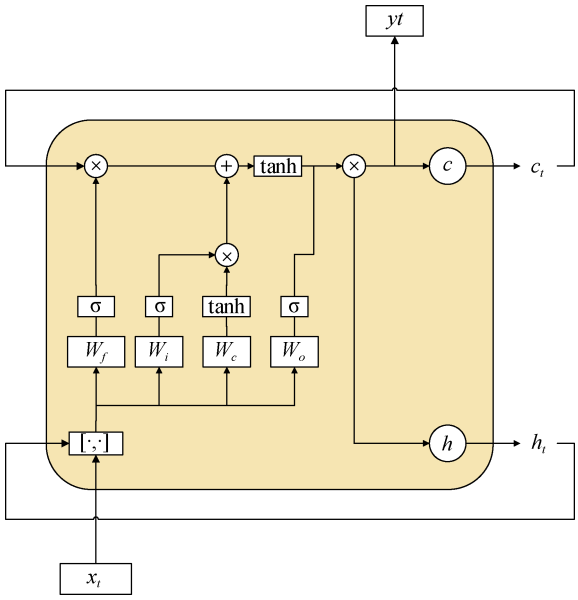


图 3　单个 ConvLSTM 结构

Figure 3　Single ConvLSTM structure

2.3　图像特征金字塔网络

图像特征金字塔网络[25](Feature pyramid networks, FPN)结构是为了解决不同尺度图像的目标检测和超分辨率重建等问题。由于神经网络的输入大小固定, 为了使网络能够识别经过缩放的不同尺寸

的图片, FPN 借助传统的图像金字塔模型, 构建了特征金字塔网络。该网络将图像特征经过多次降采样得到层级特征, 再将降采样的结果反向经过上采样融合, 得到融合特征, 最后使用融合特征进行预测, 如图 4 所示。实验中证明该网络具有能够选择不同层次中有用的特征, 提高分类检测的准确度。

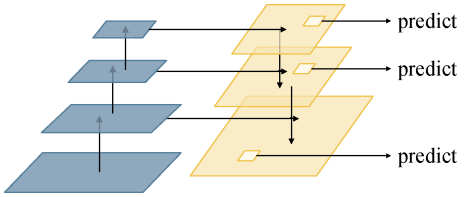


图 4 特征分类金字塔结构

Figure 4 Feature pyramid structure of classification

2.4 OpenFace 人脸识别库

本文网络所使用的预处理方法是 OpenFace 实现的。OpenFace[26]是由卡耐基梅隆大学开发的一套通用的人脸识别库, 该库基于 dlib 人脸识别库[27]和 MTCNN[28]人脸识别神经网络实现人脸识别, 面部特征提取, 面部位置标定三大功能, 其中输入源包括图像和视频。人脸识别模块可以识别出视频或图像中所有出现的人脸, 进行编号并标记出判断为同一人脸的置信度。面部特征提取模块能够提取 dlib 特征点的位置, 确定五官在视频中的位置, 并依据面部动作编码系统[29](Facial Action Coding System, FACS)定义了 20 多种动作单元[30], 用于描述人物的面部表情特征(主要体现在眼部和嘴部)。该模块设置 0 和 1 表示相应动作单元是否存在, 设置范围 0~5 表示该动作单元的存在强度。面部位置标定模块可以推测拍摄视频的摄像机的位置, 从而计算跟踪头部姿势, 包括世界坐标系下头部偏离 X, Y, Z 轴的角度。另外, 利用其面部蒙版提取模块, 可以得到去除背景的纯面部特征, 并能可视化输出面部的方向梯度直方图[31] (Histogram of Oriented Gradient, HOG)特征。

3 提出方法

3.1 现有伪造方法缺陷分析

自然视频真实地记录人物面部表情和动作状态, 每一帧的画面可以被认为是一个连续变化的采样结果, 其运动变化范围符合面部表情特性。而生成视频相当于拟合一个采样结果, 该结果的变化范围包含两种噪声, 分别是生成器生成的面部表情和原有表情之间的差异和面部覆盖中的损失。其中生成面部表情和原有表情之间的差异包括了生成质量损失和表情迁移损失, 如图 5 和图 6 所示。生成质量损失是生成网络结构和迭代次数等因素导致人物的五官或脸型出现走样的情况, 而表情迁移损失是网络在获得两个表情域间映射的映射点位置的差异性。而面部覆盖中的损失, 包括了匹配损失和融合损失, 如图 7 所示。匹配损失是在面部覆盖时特征点对应出现的误差, 融合损失包括了由于生成的脸部边缘和

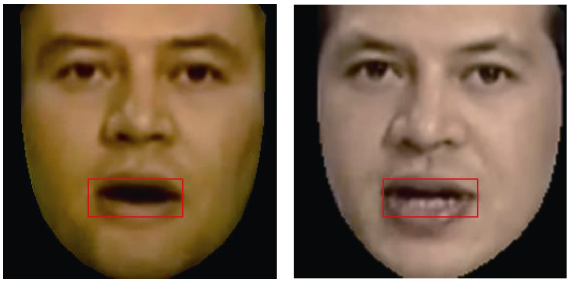


图 5 生成质量损失

Figure 5 Generate quality loss

(注: 左图为 DeepFake 生成的人脸图像, 右图是真实的人脸图像, 可以看到在牙齿的生成方面有瑕疵)

图 6 表情迁移损失(抽样)

Figure 6 Expression transfer loss (sampling)

(注: 上层人脸为真实人脸, 下层人脸为替换后的人脸, 4 帧为每隔 25 帧抽样的结果。可以看到人物的表情并不完全一致。)

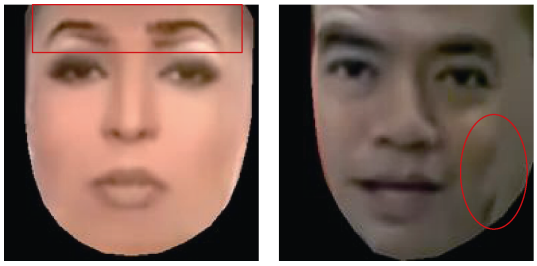


图 7 匹配损失和融合损失

Figure 7 Matching loss and fusion loss

(注: 左图为匹配损失, 可以看到由于眉毛部分的特征点匹配误差导致人脸的瑕疵; 右图为融合损失, 可以看到覆盖后的人脸和原有人脸的肤色差异)

头发以及脸部以外的背景色调和亮度使用融合方法导致的误差, 即使使用融合方法(例如泊松融合)也会出现像素值和正常相机拍摄的分布不一致的情况。

现有的检测方法采用的网络以降采样形式的 CNN 为主, 该类型网络结构集中体现在检测面部生成损失中的生成质量损失和面部覆盖损失, 其中浅层网络能够捕捉到生成图像内容的差异性, 深层次网络可以捕捉到经过模糊或融合带来的像素分布的差异性。但是对于不加入残差模块的检测网络, 其分类结果受融合损失带来的影响十分严重。对于表情迁移损失, 从单帧的角度观察和自然人脸无明显差异。这是由于主流的人脸生成替换方法基于单幅图像, 在视频上表现为单帧替换, 导致人物在面部动作变化时具有独立性, 即使经过例如卡尔曼滤波等操作或构建类似 Recycle-GAN 的时序性模型, 在实际操作和生成中只能保证相邻帧之间的连续性, 而从较长的视频帧阵列上分析, 这两种误差的随机性仍然使得面部运动在时域方面不符合人类运动习惯。以 CNN 为基础的网络检测结果则不能获得到时域特征, 即使加入了 LSTM 等时序检测结构, 其输入的时序特征为固定的某一层级的特征, 并不能完整反映多个层次的误差, 在训练中容易产生过拟合的现象。

根据对现有生成方法的局限性和换脸视频的生成特点的分析, 我们旨在构造一个多级网络, 能够提取人脸生成视频的多级图像的特征和多级时序变化特征, 并达到二分类的效果。网络结构的整体结构如图 8 所示, 该网络共分为三个模块, 首先使用 3D 卷积结构对单帧图像进行特征特征提取, 并在时序维度对特征进行基础融合, 然后将多个维度特征结合的结果分别送入时序检测结构, 得到层级的预测特征热图, 最后使用特征金字塔结构对结果进行分类融合, 通过一个全连接层得到最终的分类结果。该网络的具体细节将在下文阐述。

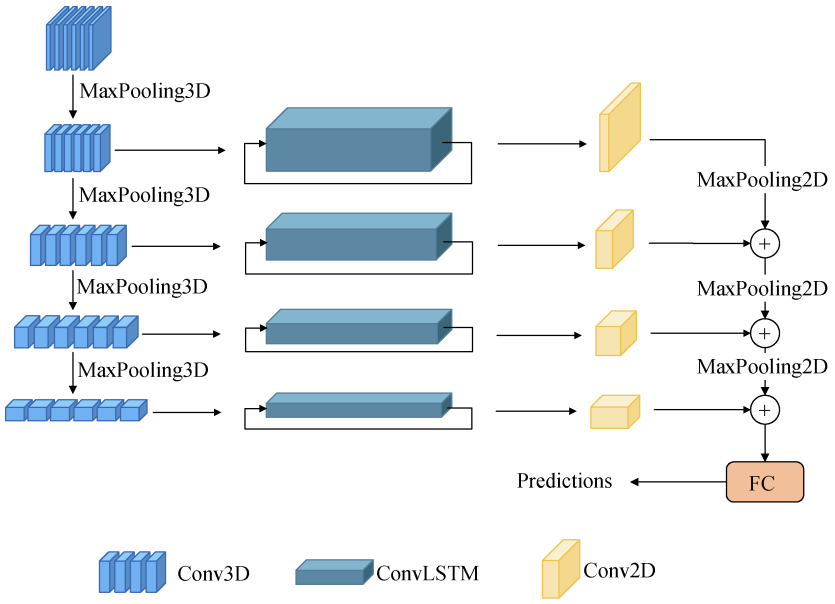


图 8 MTS3DCNet 网络检测整体结构

Figure 8 MTS3DCNet structure of detection

3.2 MTS3DCNet 网络结构

3.2.1 输入和输出维度

与一般深度神经网络不同, 我们需要提取真假人脸视频的运动不一致性特征, 运动的估计将由视频帧阵列表达, 故输入层为维, 包括帧数, 单幅帧图像宽, 单幅帧图像高, 图像颜色通道数。网络使用交叉熵(Cross entropy loss)作为损失函数, 故输出维度为 1 维, 向量长度为 2, 输出为二分类的预测结果。

3.2.2 图像特征提取结构

对于生成的单帧人脸伪造图像而言, 单帧中的特征提取仍需要卷积结构, 而在时序上, 所得到的损失特征直接送入时序检测结构中效果表现不佳, 故需要提取单帧空域和多帧时域的共同特征。另外, 生成人脸视频帧在不同深度的特征表现不同, 需要具有多种深度的神经网络才能更好的提取到不一致性特征, 故使用 3D 卷积结构是一个比较周全的选择。在图像特征提取模块我们设计一个连续降采样的3D卷积CNN特征提取层, 单层结构如图 9 所示, 将输入的帧阵列经过两个 3D 卷积结构得到特征图, 并将结果经过 MaxPooling3D 层降采样得到下层输入, 共 5 层级联, 每一层提取一个级别的特征。每个提取层将输出一个特征阵列, 连接到相应层的预测网络对该层进行预测。实验表明, 相比使用二维卷积, 三维卷积不仅能够实现在单幅图像的长宽和通道数中提取特征, 在时域方面通过控制卷积核大小达到将提取时域变化特征的效果, 对时序特征进行初步提取, 有助于后面时序结构的分类预测。

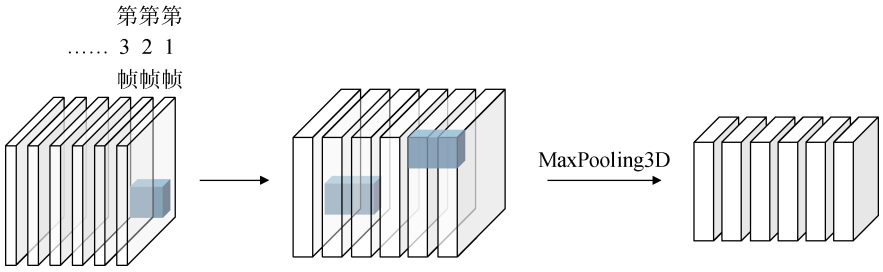


图 9 单层级联特征提取结构

Figure 9 Single-level cascade feature extraction structure

3.2.3 时序特征提取结构

为了提取表情迁移特征等在时序中的不一致性, 我们使用 ConvLSTM 将上一节介绍的特征提取网络结构提取到的层级特征阵列送入 ConvLSTM 中进行预测, 得到预测特征图。每层的输出对应一个独立的 ConvLSTM 的输入, 最终会得到不同尺寸的特征图。以 DeepFake 方法为例, 实验表明, 时序特征提取结构是该网络能够区分人脸生成与否的重要组成部分, 去掉该部分会使网络的普适性降低, 导致分类失败。

表 1 LSTM 结构对 DeepFake 生成方法的分类检测结果的影响(分类正确率)

Table 1 Effect of LSTM structure on classification detection results of DeepFake generation method(Classification accuracy)

网络结构 \ CRF	0	23	40
MTS3DCNet _frame 5 Without LSTM	50.18	97.48	49.96
MTS3DCNet_frame 10 Without LSTM	50.15	50.01	49.69

3.2.4 预测结构

相比普通的分类网络只关注最后一层的预测结果, 不同尺度的多层预测能够提高预测的鲁棒性, 减缓过拟合现象发生, 并能在速度和准确率之间进行权衡。实验表明, Meso-Inception-v4 和单层预测模型在预测生成比较真实或区域替换的面部模型时, 训练的第一次迭代就会出现 8.7%~15.9%的过拟合现象。这是由于网络在经过多次卷积后得到结果很大程度上为图像内容, 在分类区域替换或生成较为真实的人脸时往往在图像内容差别不大, 而例如模糊度等比较低级的分类特征能更好地反映正常人脸和生成人脸之间的差异。故使用单层预测不能正确把握分类特征, 而使用四层预测结果往往会在测试中的正确率高于训练值。由于我们需要关注每一层结果的表征状态, 而权衡特征之间的分类关系, 我们采用类似 FPN 的预测结构, 但是不同于图像内容分类识别, 网络需要对图像纹理的高层特征敏感, 故我们修改了特征金字塔网络的结构, 如图 10 所示, 将自顶向上的预测顺序改为自顶向下, 突出卷积后高层次的预测特征并融合低维特征进行层间预测, 再将结果送入全连接层进行激活分类, 得到最终的结果。以 NeuralTexture 生成方法为例, 实验表明, 金字塔结构在促进网络收敛和减少网络训练过拟合方面有着良好的表现效果, 表 2 表现了在训练第一个轮次(epoch)后其检测损失的对比, 可以看到在加入金字塔结构后网络更容易向正确的分类方向收敛。

4 实验结果及分析

4.1 实验数据来源

本文采用 FaceForensics++ 数据集[32], 是德国慕尼黑工业大学视觉计算组构造的一个大型换脸数据

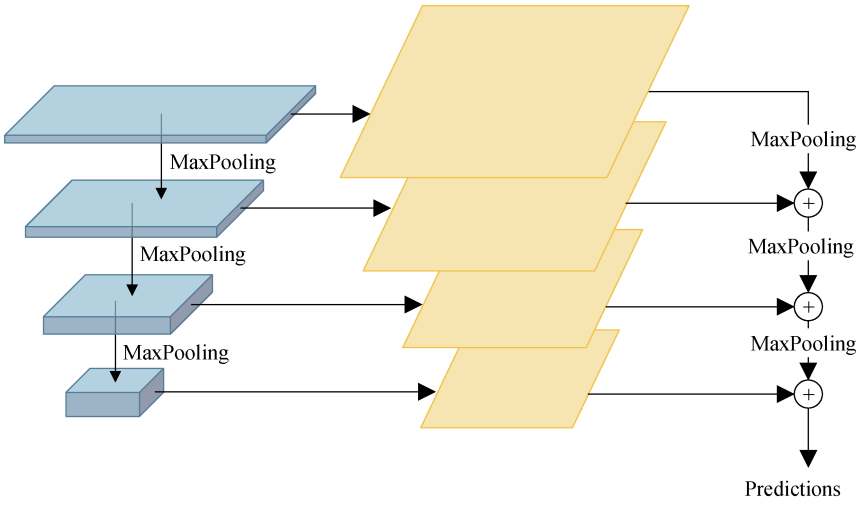


图 10　修改后的特征金字塔结构

Figure 10 Modified feature pyramid structure of classification

表 2　金字塔结构对 NeuralTexture 生成方法的检测损失的影响

Table 2　Effect of pyramid structure on detection of loss of NeuralTexture generation method

网络结构 \ CRF	0	23	40
MTS3DCNet _frame 5 Without pyramid	0.1193	0.6855	0.6862
MTS3DCNet_frame 10 Without pyramid	0.2505	0.5976	0.6875
MTS3DCNet _frame 5	**0.0849**	**0.5546**	**0.4077**
MTS3DCNet_frame 10	**0.1490**	**0.4946**	**0.5542**

库。该数据集在互联网上收集了 1000 个人的短视频作为正样本，每个视频中只包含单个人像，并且人物的头部动作比较平缓，更注重表达人物的面部动作表情变化。

负样本为 1000 个样本间的面部替换，替换人物的编号已经标注。到发稿为止，FaceForensics++ 数据库针对这 1000 样本采用了 4 种面部替换方法，包括 DeepFake、Face2Face、FaceSwap 和 NeuralTexture 等。另外，针对每一种替换方法和原始视频样本，FaceForensics++ 数据集提供了不同视频质量的版本，包括了原版本的高清视频，即固定码率因子(Constant Rate Factor, CRF)为 0，以及 CRF 为 23 和 40 的压缩视频集。我们在三种不同的质量因子上分别测试网络对于不同生成方法的分类效果。实验中将 FaceForensics++ 数据集的视频比例划分为 7∶2∶1 用于训练，训练测试和预测结果。

4.2　训练方法

4.2.1　预处理方法

为了去除背景等干扰分类的因素，我们使用 OpenFace 库[33]对帧图像进行预处理，处理流程如图 11 所示。首先确定视频帧中人脸的位置，将面部提取，然后去掉背景和头发等干扰因素，通过平移，旋转等操作将面部对齐，设定面部占比为 0.7。最后将面部归一化到 256×256 大小比例一致，保证模型的通用性。输入的类别标签应使用独热编码(One-Hot)进行预处理。

图 11　网络预处理流程

Figure 11 Pre-processing process of the network

4.2.2　训练参数选择

在训练该模型采用自适应 Adam 优化器[34]调节梯度。其中训练针对 DeepFake 和 FaceSwap 的检测模型时设置初始学习率为 5e-5，训练针对 Face2Face 和 NeuralTexture 的检测模型时设置初始学习率为 1e-5。训练时每种方法迭代 20 个轮次后查看测试结果。

4.3　实验结果

我们测试了输入不同长度的视频帧对正确率的影响，选择帧阵列长度为 5 帧和 10 帧作为第一个输入维度进行训练和测试，分别测试了在 CRF 为 0, 23 和 40 的条件下 4 种生成方法的二分类检测效果，并

用我们的方法与 Xception 和 Meso-Inception-v4 网络分类效果进行比较。其中 Xception 和 Meso-Inception-v4 网络的得到结果为单帧结果，而我们的网络输入为 5 或 10 的帧阵列，但每个帧阵列均来自同一视频，故可得到结果为 5 或 10 帧均为真或假的结果。与单帧相比，相当于对帧阵列输出了平均结果，在最终计算准确率时也是相当于计算了所有帧输出的平均结果，故帧阵列和单帧得到的正确率是具有可比性的。该网络对 4 种生成方法的检测结果如下:

表 3　网络对 DeepFake 生成方法的分类检测结果

Table 3　Results of classification detection of DeepFake generation method by networks

网络结构 \ CRF	0	23	40
Meso-Inception-v4[32]	98.41	95.26	89.52
XceptionNet[32]	99.59	98.85	94.28
MTS3DCNet _frame 5	**99.45**	**98.96**	**92.55**
MTS3DCNet_frame 10	**99.79**	**99.21**	**91.94**

表 4　网络对 Face2Face 生成方法的分类检测结果

Table 4　Results of classification detection of Face2Face generation method by networks

网络结构 \ CRF	0	23	40
Meso-Inception-v4[32]	97.96	95.84	84.44
XceptionNet[32]	99.14	98.36	91.56
MTS3DCNet_frame 5	**94.44**	**98.53**	**91.06**
MTS3DCNet_frame 10	**94.34**	**95.82**	**90.97**

表 5　网络对 FaceSwap 生成方法的分类检测结果

Table 5　Results of classification detection of FaceSwap generation method by networks

网络结构 \ CRF	0	23	40
Meso-Inception-v4[32]	96.07	93.43	83.56
XceptionNet[32]	99.61	98.23	93.70
MTS3DCNet_frame 5	**99.65**	**98.48**	**91.51**
MTS3DCNet_frame 10	**99.69**	**98.27**	**90.86**

表 6　网络对 NeuralTexture 生成方法的分类检测结果

Table 6　Results of classification detection of NeuralTexture generation method by networks

网络结构 \ CRF	0	23	40
Meso-Inception-v4[32]	97.05	85.96	75.74
XceptionNet[32]	99.36	94.50	82.11
MTS3DCNet_frame 5	**98.66**	**92.58**	**84.93**
MTS3DCNet_frame 10	**99.38**	**93.75**	**84.33**

相比 Mesonet 和 Xception 两种检测方法，我们的方法不仅在正确率上能够达到更优秀的效果，而且在网络搭建中，使用的参数量小于 Xception 网络，这表明我们在训练时使用更少的时间能够得到更加快速收敛效果，而 Xception 网络由于其庞大的参数可能会导致参数冗余。Mesonet 系列的网络参数量较少，但其深度不够会导致其特征提取畸形，这体现在训练 Xception 网络和 Mesonet 时具有较深程度的过拟合现象。

表 7　三种网络结构参数量对比

Table 7　Comparison of parameters of three networks

网络结构	参数量
XceptionNet	20811050
Meso-Insception-v4	28742
MTS3DCNet	14168386

总体来看，我们的网络结构在检测压缩的人脸生成视频有更好的表现。而在实际环境中，人脸生成视频更多的是在互联网和社交媒体中传播。在主流的视频网站(如 YouTube)和社交媒体网站(如 Facebook)上传视频中一般需要经过压缩，所以提高低质量下的检测效果是十分必要的。

另外，根据结果来看，我们的网络结构在检测更为逼真的换脸算法时更加有效。文献[32]给出了人工检测者在检测 4 种生成方法和原始视频的正确率，如图 12 所示。可以看到 Face2Face 和 NeuralTexture 两种生成方法在视觉上更真实。而我们的检测网络在检测这两种方法中有着较大的提高，而对于 DeepFake 和 FaceSwap 生成方法，我们的检测网络也能得到不错的效果。

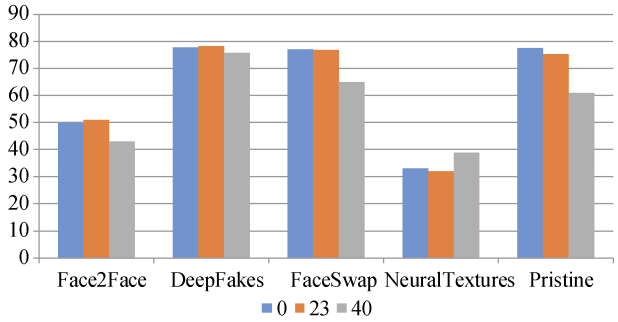


图 12　人工检测者检测生成方法的正确率

Figure 12　Accuracy of manual detector detection and generation method

5　结论

随着神经网络和深度学习的不断发展，人工生成的人脸视频的滥用现象越来越严重，本文提出了

一种神经网络检测结构, 该结构与当前主流的神经网络检测方法相比, 有如下优点:

(1) 同时考虑了时域和空域的特征, 提取特征更加丰富;

(2) 使用类似金字塔的层级预测结构缓解了分类检测的过拟合现象;

(3) 对于目前互联网和社交媒体中实用的压缩视频效果有较大的提升;

(4) 对于更真实的生成方法有更好的检测效果。

本文的后续工作将针对每种生成方法优化模型训练参数和预处理的方式, 从整体上提高检测的正确率, 并提高模型的泛化能力。

致 谢 在此向本文成文中给予的指导老师, 提供帮助的同学和给本文提出建议的评审专家表示诚挚的感谢。

参考文献

[1] Suwajanakorn S, Seitz S M, Kemelmacher-Shlizerman I. Synthesizing Obama[J]. *ACM Transactions on Graphics*, 2017, 36(4): 1-13.

[2] L.S. Wang, On the Integrated Regulation of “Deep Forgery” Intelligent Technology—Talking from “Yang Mi’s Face Changing Video”. Oriental Law, https://doi.org/10.19404/j.cnki.dffx.. Nov. 2019.

[3] Faceswap, github, https://github.com/deepfakes/faceswap.

[4] DeepFaceLab, github, https://github.com/iperov/DeepFaceLab.

[5] Huang X, Liu M Y, Belongie S, et al. Multimodal unsupervised image-to-image translation[C]. *Proceedings of the European conference on computer vision (ECCV)*, 2018: 172-189.

[6] Thies J, Zollhöfer M, Stamminger M, et al. Demo of Face2Face: Real-Time Face Capture and Reenactment of RGB Videos[M]. ACM SIGGRAPH 2016 Emerging Technologies, 2016: 1-2.

[7] Thies J, Zollhöfer M, Nießner M. Deferred Neural Rendering: Image Synthesis Using Neural Textures[J]. *ACM Transactions on Graphics*, 2019, 38(4): 66.

[8] Gatys L, Ecker A S, Bethge M. Texture synthesis using convolutional neural networks[J]. *Advances in neural information processing systems*, 2015, 28.

[9] E. Risser, P. Wilmot, and C. Barnes. Stable and controllable neural texture synthesis and style transfer using histogram losses[EB/OL]. 2017: ArXiv Preprint ArXiv:1701.08893.

[10] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets[J]. *Advances in neural information processing systems*, 2014, 27.

[11] Zhu J Y, Park T, Isola P, et al. Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks[J]. *2017 IEEE International Conference on Computer Vision*, 2017: 2242-2251.

[12] Bansal A, Ma S, Ramanan D, et al. Recycle-gan: Unsupervised video retargeting[C]. *Proceedings of the European conference on computer vision (ECCV)*, 2018: 119-135.

[13] Mao X D, Li Q, Xie H R, et al. Least Squares Generative Adversarial Networks[C]. *2017 IEEE International Conference on Computer Vision*, 2017: 2813-2821.

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

[15] Cozzolino D, Poggi G, Verdoliva L. Recasting Residual-Based Local Descriptors as Convolutional Neural Networks: An Application to Image Forgery Detection[C]. *The 5th ACM Workshop on Information Hiding and Multimedia Security*, 2017: 159-164.

[16] Bayar B, Stamm M C. A Deep Learning Approach to Universal Image Manipulation Detection Using a New Convolutional Layer[C]. *The 4th ACM Workshop on Information Hiding and Multimedia Security*, 2016: 5-10.

[17] Rahmouni N, Nozick V, Yamagishi J, et al. Distinguishing Computer Graphics from Natural Images Using Convolution Neural Networks[J]. *2017 IEEE Workshop on Information Forensics and Security*, 2017: 1-6.

[18] Li Y Z, Chang M C, Lyu S W. In Ictu Oculi: Exposing AI Generated Fake Face Videos by Detecting Eye Blinking[EB/OL]. 2018: ArXiv Preprint ArXiv: 1806.02877.

[19] Chollet F. Xception: Deep Learning with Depthwise Separable Convolutions[J]. *2017 IEEE Conference on Computer Vision and Pattern Recognition*, 2017: 1800-1807.

[20] Szegedy C, Liu W, Jia Y Q, et al. Going Deeper with Convolutions[C]. *2015 IEEE Conference on Computer Vision and Pattern Recognition*, 2015: 1-9.

[21] 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*, 2018: 1-7.

[22] Ji S W, Xu W, Yang M, et al. 3D Convolutional Neural Networks for Human Action Recognition[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2013, 35(1): 221-231.

[23] Shi X, Chen Z, Wang H, et al. Convolutional LSTM network: A machine learning approach for precipitation nowcasting[J]. *Advances in neural information processing systems*, 2015, 28: 802-810.

[24] Hochreiter S, Schmidhuber J. Long Short-Term Memory[J]. *Neural Computation*, 1997, 9(8): 1735-1780.

[25] Lin T Y, Dollár P, Girshick R, et al. Feature Pyramid Networks for Object Detection[C]. *2017 IEEE Conference on Computer Vision and Pattern Recognition*, 2017: 936-944.

[26] Amos B, Ludwiczuk B, Satyanarayanan M. Openface: A general-purpose face recognition library with mobile applications[J]. *CMU School of Computer Science*, 2016, 6(2): 20.

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

[28] Yin X, Liu X M. Multi-Task Convolutional Neural Network for Pose-Invariant Face Recognition[J]. *IEEE Transactions on Image Processing*, 2018, 27(2): 964-975.

[29] Friesen E, Ekman P. Facial action coding system: a technique for the measurement of facial movement[J]. *Palo Alto*, 1978, 3(2): 5.

[30] P. Ekman, and E. Friesen. FACS - Facial Action Coding System, Carnegie Mellon School of Computer Science, https://www.cs.cmu.edu/~face/facs.htm.

[31] Dalal N, Triggs B. Histograms of Oriented Gradients for Human Detection[J]. *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2005, 1: 886-893vol.1.

[32] Rössler A, Cozzolino D, Verdoliva L, et al. FaceForensics++: Learning to Detect Manipulated Facial Images[C]. *2019 IEEE/CVF International Conference on Computer Vision*, 2019: 1-11.

[33] Openface, github, https://github.com/cmusatyalab/openface.

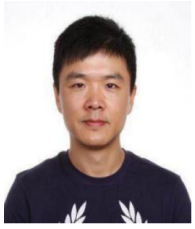
[34] D.P. Kingma, J. Ba. Adam: A method for stochastic optimization[EB/OL]. 2014: ArXiv Preprint ArXiv:1412.6980.

包晗 于 2018 年在湖南大学获得本科学士学位, 现在中国科学院大学信息工程研究所软件工程专业攻读硕士学位, 研究领域为多媒体安全, 研究兴趣包括视频处理、多媒体取证、多媒体异常检测、图像生成等。Email: baohan@iie.ac.cn

符皓程 于 2018 年 6 月在上海大学通信工程专业获得工学学士学位, 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。研究领域为多媒体信息安全, 研究兴趣包括: 隐写与隐写分析、图像取证等。Email: fuhaocheng@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。

汤朋 于 2018 年在西南交通大学获得软件工程专业本科学位, 现在中国科学院大学信息工程研究所软件工程专业攻读硕士学位, 研究兴趣包括隐写与隐写分析、数字水印、多媒体取证技术。Email: tangpeng@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:数字内容取证、隐写与隐写分析等。Email: caoyun@iie.ac.cn