融合全局时序和局部空间特征的伪造人脸视频检测方法

陈 鹏[1,2], 梁 涛[1,2], 刘 锦[1,2], 戴 娇[1], 韩冀中[1]

[1]中国科学院信息工程研究所 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, 深度学习在人工智能领域表现出优异的性能。基于深度学习的人脸生成和操纵技术已经能够合成逼真的伪造人脸视频, 也被称作深度伪造, 让人眼难辨真假。然而, 这些伪造人脸视频可能会给社会带来巨大的潜在威胁, 比如被用来制作政治虚假新闻, 从而引发政治暴力或干扰正常选举等。因此, 亟需研发对应的检测方法来主动发现伪造人脸视频。现有的方法在制作伪造人脸视频时, 容易在空间上和时序上留下一些细微的伪造痕迹, 比如纹理和颜色上的扭曲或脸部的闪烁等。主流的检测方法同样采用深度学习, 可以被划分为两类, 即基于视频帧的方法和基于视频片段的方法。前者采用卷积神经网络(Convolutional Neural Network, CNN)发现单个视频帧中的空间伪造痕迹, 后者则结合循环神经网络(Recurrent Neural Network, RNN)捕捉视频帧之间的时序伪造痕迹。这些方法都是基于图像的全局信息进行决策, 然而伪造痕迹一般存在于五官的局部区域。因而本文提出了一个统一的伪造人脸视频检测框架, 利用全局时序特征和局部空间特征发现伪造人脸视频。该框架由图像特征提取模块、全局时序特征分类模块和局部空间特征分类模块组成。在 FaceForensics++数据集上的实验结果表明, 本文所提出的方法比之前的方法具有更好的检测效果。

关键词 伪造人脸; 深度伪造; 人脸检测; 视频检测; 时序特征; 空间特征
中图法分类号 TP37 **DOI 号** 10.19363/J.cnki.cn10-1380/tn.2020.02.06

Forged Facial Video Detection Based on Global Temporal and Local Spatial Feature

CHEN Peng[1,2], LIANG Tao[1,2], LIU Jin[1,2], DAI Jiao[1], HAN Jizhong[1]

[1] 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 Nowadays, deep learning has shown impressive performance in the field of artificial intelligence. Face generation and manipulation techniques based on deep learning have enabled the creation of sophisticated forged facial video, also known as Deepfakes, which is indistinguishable by human eyes. However, these forged videos will bring huge potential threats to our society, such as being used to make fake political news, which will incite political violence or sabotage elections. Therefore, there is an urgent need to develop effective methods for forged facial video detection. When producing a forged facial video, existing methods are prone to exhibit some spatial and temporal subtle traces, such as distortion of color and texture or facial temporal flickering. The mainstream detection methods also adopt deep learning, which can be divided into frame-based and fragment-based. The former exploits Convolutional Neural Network (CNN) to find spatial traces in single frame. The latter combines Recurrent Neural Network (RNN) to capture temporal traces between frames. These methods are based on the global information of image to make decision, but the minor traces generally exist in the local area of face. Thus, we propose a unified detection framework for forged facial video, which exploits global temporal and local spatial feature to discover manipulated facial videos. It consists of an image feature extractor, a global temporal feature classifier and a local spatial feature classifier. The experimental results on FaceForensics++ demonstrate that our proposed method achieves better performance than previous methods.

Key words forged face; Deepfakes; face detection; video detection; temporal feature; spatial feature

1 引言

近年来, 深度学习在计算机视觉领域表现出优异的性能, 被广泛地应用在各种任务中, 其中, 以生成对抗网络[1](Generative Adversarial Networks, GANs)为代表的深度生成模型(Deep Generative

通讯作者: 戴娇, 博士, 高级工程师, Email: daijiao@iie.ac.cn。
收稿日期: 2020-1-15; 修改日期: 2020-2-29; 定稿日期: 2020-3-13

Model, DGM)可以被用来生成虚假的图像或视频内容, 使人眼难辨真假。近期, 一位网民利用深度学习技术将女明星的脸替换到色情视频中, 吸引了大量的关注, 并随着相应代码的开源, 引起了一股深度伪造热潮。荷兰的一家研究机构统计了互联网上 14678 条伪造人脸视频, 发现其中色情视频占 96%, 政治人物类和其他类占 4%。当这些伪造人脸视频在社交网络上大量传播时, 会给社会带来巨大的潜在威胁。这类技术除了被用来制作虚假的色情视频, 还会被滥用到其他行为, 一是可能会被不法分子用来进行诈骗、勒索等违法行为, 二是可能会被用来制作政治假新闻, 抹黑政治人物, 干扰正常选举, 甚至破坏外交关系。因此, 伪造人脸图像或视频等伪造内容的发现, 对维护网络空间乃至政治和经济社会的稳定具有极为重要的现实意义。

传统的数字图像内容伪造[2]包括移动-复制(Copy-move)、拼接(Splicing)和修饰(Retouching)。移动-复制指从图像中拷贝一个区域粘贴到图像中的另一个区域。拼接指合并两个或多个图像, 与移动-复制有点相似, 不同的是拼接的图像来自其他图像。修饰指去除图像的一些缺陷。视频中的伪造可以分为视频帧之间和视频帧内部两种类型[3]。视频帧之间的伪造只改变视频帧间的顺序, 而保持视频帧中的内容不变, 包括删除帧, 插入帧和打乱帧顺序。视频帧内部的伪造只改变视频帧中的内容, 包括拼接、移动-复制、修饰。数字图像和视频中的人脸伪造则是一种针对人脸的新型内容伪造方式, 相对传统的伪造方式具有更强的欺诈性和危害性。

数字图像中人脸伪造可以分为人脸合成(Face Synthesis)和人脸操纵(Face Manipulation)两种类型。人脸合成指生成整个人脸图像, 包括五官区域, 头发以及头部姿势等, 合成的人脸在现实中是不存在的。人脸操纵指修改真实人脸图像的部分人脸信息。根据所修改的信息不同, 人脸操纵可以进一步分为人脸身份替换(Facial Identity Replacement)、人脸表情扮演(Facial Expression Reenactment)和人脸属性修改(Facial Attribute Modification)。如图 1 所示, 人脸表情替换指将源域的人脸五官区域替换到目标域的脸部, 改变了目标域的身份信息而保持表情和光照等其他条件不变。人脸表情扮演指根据源域的脸部表情, 同步修改目标域的脸部表情, 而保持身份信息和光照等其他条件不变。人脸属性修改指改变年龄、性别、肤色等属性信息或者添加眼镜、耳环等饰品。视频中人脸伪造主要是上面提到的人脸操纵, 其中, 人脸属性修改这种伪造技术一般应用在抖音等娱乐性的手机应用中, 而人脸身份替换和人脸表情扮演则更多地被滥用来制作具有恶意的伪造内容。因此, 本文主要针对人脸身份替换和人脸表情扮演的伪造人脸视频进行检测。

在深度学习技术出现之前, 数字图像中人脸操纵有着较高的门槛, 需要人工进行操作, 要求操作者具备高超的领域知识, 使用专业级编辑工具, 并花费大量的时间和精力。视频中人脸操纵则难度更大, 因为需要对视频中每帧图像都要进行处理, 并要考虑图像之间的时序连贯性。随着 DGM 展现出强大的生成能力, 其逐渐被用到人脸操纵中, 只需足够的训练数据和计算资源, 一旦模型完成训练, 就能够自动化地处理图像中人脸, 而不需要人工干预, 这也使得视频中的人脸操纵变得简单可行。

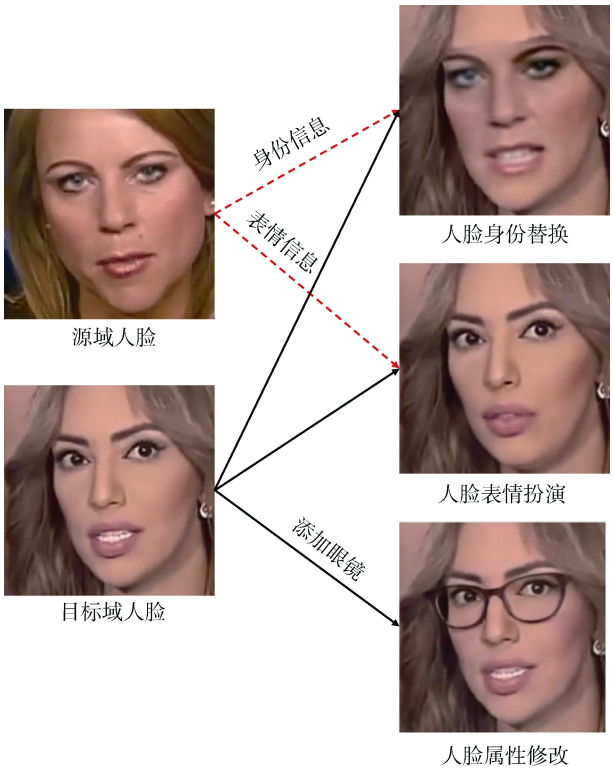


图 1 人脸操纵示例

Figure 1 Examples of face manipulation

然而, 现有的人脸操纵方法生成伪造人脸视频还没有达到天衣无缝的程度, 依然存在着缺陷, 容易在视觉上展现出细微的伪造痕迹。从单帧图像的角度来看, 伪造的人脸会留下一些纹理和颜色上的扭曲, 例如额头上的拼接痕迹。从连续多帧图像的角度来看, 由于现有的操纵方法都是对视频帧依次进行处理, 视频帧彼此之间缺少上下文信息, 加上不同的光照条件等因素的影响, 会导致脸部相同区域出现闪烁等伪造特征。这些视觉上的伪造痕迹对伪

造人脸视频检测具有重要的作用。

根据处理的对象不同, 伪造人脸视频的检测方法可以分为两种, 即基于视频帧的检测方法[23-31]和基于视频片段的检测方法[33-36]。两者所利用的鉴别特征有所不同, 前者将视频检测问题转换成图像检测问题, 只关注图像中的空间特征, 以期发现真实人脸图像和伪造人脸图像的不同。后者则更多关注视频片段中的时序特征, 捕捉连续多帧图像中的时序伪造痕迹。这些方法都是基于图像的全局信息进行决策, 然而伪造痕迹很细微, 且一般存在于五官的局部区域。

基于上述分析, 本文提出了一个统一的伪造人脸视频检测框架, 能够结合视频中的时序信息和空间信息来发现伪造痕迹从而鉴别人脸视频真伪。该框架包含全局时序特征分类模块和局部空间特征分类模块两个支路, 并共用图像特征提取模块提取图像级别特征。全局时序特征分类模块发掘连续多帧图像中的时序伪造痕迹进行决策, 与此同时, 局部空间特征分类模块则重点关注眼睛、鼻子和嘴巴等五官区域的细节信息, 从而实现更优的检测效果。

本文接下来的章节将按照以下内容进行组织: 第二章主要介绍人脸视频的操纵方法以及伪造人脸视频检测方法的相关工作。第三章主要介绍本文提出的检测框架。第四章是实验部分, 主要与前沿算法进行对比并对比不同模块的有效性和准确性。第五章是结论, 总结了本文的工作并展望未来的工作。

2 相关工作

2.1 人脸视频的操纵方法

本节主要介绍视频中的人脸身份替换和人脸表情扮演两种类型的相关工作。

人脸身份替换旨在驱动源域的人脸替换到目标域的人脸上, 改变目标域人脸的身份信息而保留表情和光照其他条件不变。Blanz 等人[5]提出了一种基于三维形变模型[6](3D Morphable Model, 3DMM)的换脸方法, 能够根据人脸的形状、纹理和光照条件进行处理, 然而该方法需要人工干预, 标注人脸 7 个关键点和发际线进行参考。Kowalski[8]提出了 FaceSwap 方法, 通过检测人脸关键点、拟合 3DMM、利用图像纹理最小化关键点和投影形状的差异, 最后将渲染好的模型融合到人脸中达到人脸身份替换的目的。Korshunova 等人[9]借鉴了图像风格迁移的思想, 将人脸的身份和表情分别映射成图像的风格和内容, 进行特定人物的人脸身份替换。DeepFakes-AE[10]方法采用了自动编码器(Auto-Encoder, AE)模型, 对于源域的人脸和目标域的人脸共享同一个编码器, 但使用了两个不同的解码器来重建源域的人脸和目标域的人脸, 通过人脸关键点定位并裁剪出人脸五官区域, 将其送进模型变换, 再利用泊松融合[11]方法将变换后的图像贴回目标域的人脸上。FaceSwap-GAN[12]直接使用 GANs 对源域的人脸和目标域的人脸进行转换, 能够生成更为逼真的细节。为了更好地处理人脸遮挡的情况, Nirkin 等人[13]提出了一个统一的人脸替换框架 FSGAN, 将人脸表情扮演、人脸分割、面部图像修复和图像融合流程整合到同一个模型进行处理, 能够更便捷地进行人脸替换。

人脸表情扮演旨在根据源域的人脸表情来操纵目标域的人脸表情, 同时保留目标域人脸的身份信息和光照等其他条件不变。Thies 等人[14]提出的 Face2Face 模型通过拟合源域人脸和目标域人脸的 3DMM, 得到人脸身份、表情和光照参数, 然后用源域人脸的表情参数替换目标域人脸的表情参数, 再通过面部渲染得到修改后的人脸, 并对唇部图像检索从而改进嘴部的生成细节。Suwajanakorn 等人[15]通过控制人脸关键点和 3DMM 操作奥巴马说话时的唇部动作, 能够根据任意一段语音合成奥巴马逼真的演讲视频。Averbuch-Elor 等人[16]提出一种方法, 能够根据一段说话视频, 利用人脸关键点进行 2D 变换, 从而驱动任意的单张人脸图像“动”起来。随着深度学习在计算机视觉领域展现出优异的性能, 深度生成模型逐渐被利用在人脸扮演任务中。Wiles 等人[17]提出的 X2Face 将身份信息和表情信息分别编码成向量, 拼接后再通过解码器生成人脸扮演图像。Pumarola 等人[18]提出的 GANimation 摒弃了使用人脸关键点作为表情的表示, 转而运用面部动作单元(Facial Action Units)作为表情的表示, 并在生成人脸阶段根据特定的驱动人脸表情信息生成注意力掩码, 从而专注于面部特定区域的图像生成。NeuralTextures[19]方法利用人脸模型进行追踪和渲染对应的 UV 掩码, 这些掩码带有驱动人脸的身份信息和原始人脸的表情信息, 此后通过编码器解码器网络生成最终的人脸。Zakharov 等人[20]将自适应实例规范化(Adaptive Instance Normalization, AdaIN)[21]应用到人脸表情变换中, 能够更好地表达人脸身份信息, 并解决了目标域人脸训练样本少的问题。MarioNETte[22]方法则利用三维面部关键点解耦、图像注意力机制、人脸姿势对齐等模块完成了人脸表情扮演。

2.2 伪造人脸视频的检测方法

按照处理对象的不同, 伪造人脸视频检测方法

可以分为基于视频帧的检测方法和基于视频片段的检测方法两类。

Bayar 等人[23]提出利用 CNN 对伪造图像进行检测, 并设计一个卷积层用来抑制图像的语义内容, 而关注像素级别的信息。Cozzolino 等人[24]提出利用 CNN 去学习局部残差特征进行伪造图像检测。Zhou 等人[25]提出一个双流模型来检测伪造人脸图像, 其中一个支路使用 CNN 捕捉图像中的视觉伪造痕迹, 另一个支路提取底层的噪声残差特征并使用支持向量机进行分类, 最后融合两个支路的分数进行决策。Afchar 等人[26]认为对于伪造人脸视频, 图像底层的噪声信息会因为视频的压缩而严重退化, 而图像高层的语义信息则过于相似, 这两者都不适合伪造人脸视频的检测, 进而提出浅层的 CNN 来抽取中间层特征进行决策。Yang 等人[27]提出了基于脸部方向和头部姿态不一致的检测方法。深度学习模型需要大量的数据进行驱动, 然而伪造人脸数据的获取代价极高。为解决数据量不足的问题, Li 等人[28]发现人脸身份替换操作会留下变形操作的痕迹, 因而仿真这种变形操作生成伪造数据来训练分类器。Stehouwer 等人[29]采用注意力机制让分类器更加关注伪造区域的信息, 并且使用注意力特征图定位具体的伪造区域。Nguyen 等人[30]利用多任务学习的思想将真假分类和伪造区域定位统一到一个模型中。此外, Nguyen 等人[31]基于胶囊网络[32](Capsule Network)设计一个分类器, 相较与传统的网络使用了更少的参数而达到相近的效果。然而, 上述的工作都只考虑单个视频帧的空间信息, 而未注意视频中丰富的时序信息对伪造人脸视频检测同样重要。

一些方法发现真实人脸和伪造人脸在某些生物统计特征上存在不同, 并以此为作为鉴别的依据, 例如眨眼频率[33]与 rPPG[34]等, 但是随着伪造的人脸越来越逼真, 这些方法逐渐失去其效果。Güera 等人[35]发现现有的伪造人脸视频生成方法都是逐帧进行处理的, 视频帧缺乏上下文信息, 会在视频中留下闪烁等伪造痕迹, 因此, 提出用预训练的 CNN 提取视频帧的视觉特征, 然后使用这些特征训练一个循环神经网络 RNN 进行分类。Sabir 等人[36]扩展了文献[35]中的方法, 对模型进行端到端的训练, 并且对视频进行预处理, 只裁取人脸区域进行检测。

3 基于全局时序和局部空间特征的伪造人脸视频检测方法

3.1 数据预处理

伪造人脸视频与其他形式伪造视频明显的不同是, 其篡改区域主要在人脸上, 因此, 可以通过人脸检测算法定位人脸区域, 从而缩小处理范围。这样做, 对于模型的学习有两点好处, 一是能够聚焦人脸区域, 学习到更多细微的鉴别特征, 二是能够减少最后的特征图中来自背景的噪声信息。

对于一条包含人脸的视频, 首先被划分成连续不重叠的视频片段集合, 每个视频片段包含 T 帧图像。然后, 如图 2 所示, 对于每帧图像, 使用 Dlib[37]工具包中的人脸定位和对齐算法检测出 68 个人脸关键点, 再利用眉毛、眼睛、鼻子和唇部区域的 51 个关键点得到一个包裹矩形框, 将得到的矩形框扩大 1.6 倍作为裁剪的尺寸, 并将裁剪后的人脸区域缩放至 224×224 大小, 由此得到裁剪后的视频片段集合 $\{C_k\}$。为方便阐述, 下文中所提的视频片段皆指经过裁剪后的视频片段。

图 2 人脸裁剪示意图

Figure 2 Illustration of face cropping

3.2 模型框架

图 3 展示了本文提出的基于全局时序和局部空间特征的检测模型, 整个网络架构包括图像特征提取模块, 全局时序特征分类模块和局部空间特征分类模块 3 个部分。

3.2.1 图像特征提取模块

本文采用 CNN 作为图像特征提取模块, CNN 是一种前馈神经网络, 能够提取丰富的视觉表征。本文中的图像特征提取模块可以选用图像分类模型的卷积层。具体的, 采用 VGG16 网络[38], 将其全连接层和最后一个池化层去掉, 并在每个卷积层后添加批标准化(Batch Normalization, BN)层[39]。BN 层能够加速模型的收敛, 并降低模型对初始化权重的敏感性。给定一条视频片段 $\{I_1,I_2,\cdots,I_t\}, t\in[1,T]$, 将其每个视频帧输入到图像特征提取模块中, 输出图像级别特征 $\{F_1,F_2,\cdots,F_t\}, t\in[1,T]$, 其中 F_t 的维度大小为 $[512,w,h]$, w 和 h 由输入图像大小决定, 本文中输入图像大小为 224×224, 因此, w 和 h 都为 14。

3.2.2 全局时序特征分类模块

由于现有的伪造人脸视频生成方法是逐帧对视频进行处理的, 在连续多个视频帧之间会存在着细

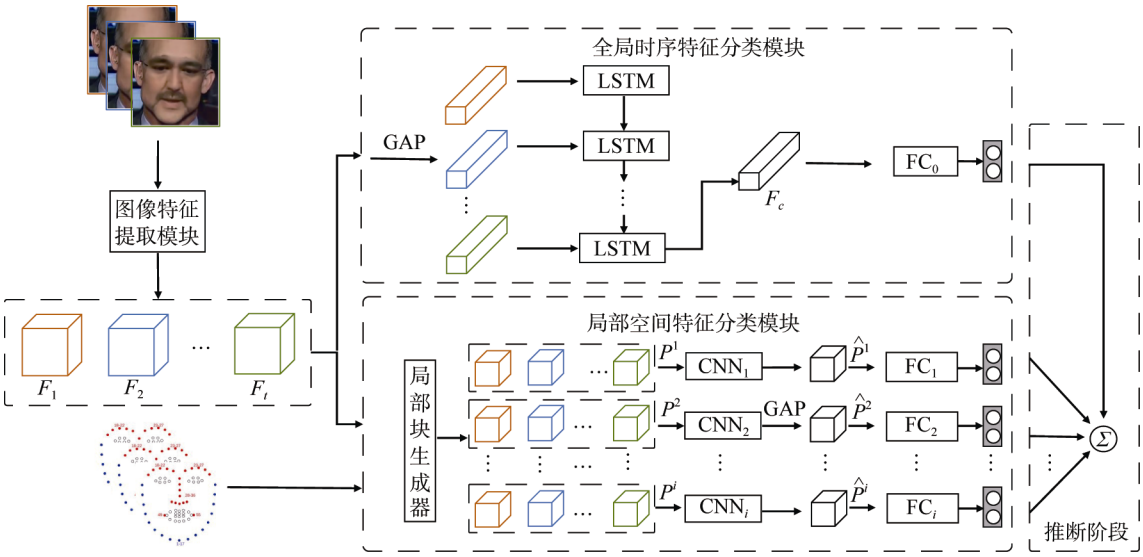


图 3　本文方法整体框架示意图

Figure 3　An overview of our proposed framework

微的时序伪造痕迹，例如脸部区域会出现细微的闪烁。为了能够捕捉此类丰富的时序信息进行决策，本文构建了全局时序特征分类模块，将图像级别特征输入到 RNN 中提取片段级别时序特征。全局时序特征分类模块采用了长短时记忆(Long Short Term Memory, LSTM)网络[40], LSTM 网络能够利用记忆单元和门限的更新有效地学习数据的时序特征。对于图像特征提取模块提取的图像级别特征 $\{F_1,F_2,\cdots,F_t\}, t\in[1,T]$，使用一个 14×14 的全局平均池化(Global Average Pooling, GAP)层进行空间维度压缩，得到 512 维的特征向量，再按照序列逐步输入到 LSTM 网络中，并根据如下公式逐步更新网络:

$$f_t=\sigma(W_f x_t+U_f h_{t-1}+b_f), \tag{1}$$

$$i_t=\sigma(W_i x_t+U_i h_{t-1}+b_i), \tag{2}$$

$$o_t=\sigma(W_o x_t+U_o h_{t-1}+b_o), \tag{3}$$

$$c_t=f_t\odot c_{t-1}+i_t\odot\tanh(W_u x_t+U_c h_{t-1}+b_c), \tag{4}$$

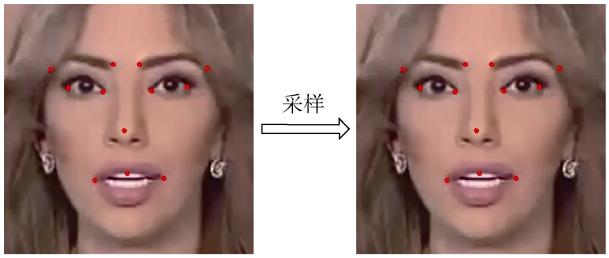
$$h_t=o_t\odot\tanh(c_t), \tag{5}$$

其中，x 表示输入序列，i,f,o 和 c 分别表示式输入门、遗忘门、输出门和记忆单元，$\odot$ 表示元素相乘，σ 表示 Sigmoid 激活函数，W,U 和 b 是 LSTM 网络中待学习的参数，h 是隐藏层状态。将 LSTM 网络最后一步的隐藏层状态作为视频片段级别特征 F_c，再使用 512 维的全连接(Fully Connected, FC)层进行二分类，经过 Softmax 归一化操作，最后得到全局时序特征分数 S_c。

3.2.3　局部空间特征分类模块

通过观察伪造人脸视频，我们发现伪造人脸在五官区域会存在细微的纹理或颜色上的扭曲，如果只使用全局信息进行预测，在经过 GAP 操作时，这种局部的细节信息会有所损失。基于上述观察，本文提出的局部空间特征分类模块将重点关注五官区域的局部细节进行决策。

如 3.1 节介绍，对于一个视频片段，可以得到 T 帧图像中的 68 个人脸关键点坐标，将 T 帧图像的人脸关键点坐标取均值作为视频片段的人脸关键点坐标。视频中的一些人脸可能会受遮挡、光照等条件影响，无法被检测出关键点，我们将统计视频中平均人脸关键电脑作为替代。为了减少计算量，可以从 68 个人脸关键点中挑选 N 个关键点作为代表子集 $\{(x_i,y_i)\}, i=1,2,\cdots,N$。如图 4 所示，本文的方法选择了 12 个关键点作为五官轮廓的代表子集，包括眉毛的端点、鼻子的中心点、嘴巴的端点以及唇部的中心点和端点。



(a) 68个人脸关键点　(b) 12个五官轮廓关键点

图 4　人脸关键点示意图

Figure 4　Illustration of facial landmarks

对于一个视频片段，在得到其图像级别特征 $\{F_1,F_2,\cdots,F_t\}, t\in[1,T]$ 后，使用局部块生成器(Local

Patch Generator, LPG)以五官区域的关键点为中心从图像级别特征 F_t 中采样局部特征块。局部块生成器的设计启发于空间变换网络[41](Spatial Transformer Networks, STN),两者都需要构建仿射变换矩阵 $\boldsymbol{\theta}$, 再利用 $\boldsymbol{\theta}$ 去计算一个采样网格(Sampling Grid), 该采样网格标记着 F_t 中的哪些位置点会被采样到特征块中, 最后根据采样网格利用双线性插值算法从 F_t 中采样得到特征块, 其流程如图 5 所示。两者不同的是, STN 使用一个定位网络去学习 $\boldsymbol{\theta}$, 而局部块生成器直接根据人脸关键点计算得到 $\boldsymbol{\theta}$。本方法中所构建的仿射变换矩阵 $\boldsymbol{\theta}$ 如公式(6)所示:

$$\boldsymbol{\theta}=\begin{bmatrix} s_h & 0 & a_x \\ 0 & s_w & a_y \end{bmatrix}, \tag{6}$$

其中, s_h 和 s_w 表示特征块的大小, 其值固定为 F_t 大小的 0.15 倍, a_x 和 a_y 表示特征块在 F_t 中的平移, 其值由人脸关键点坐标计算得到。

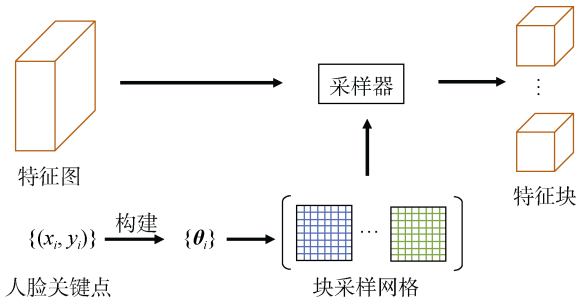


图 5 局部块生成器示意图

Figure 5 Illustration of local patch generator

在图像级别特征 $\{F_1,F_2,\cdots,F_t\}, t\in[1,T]$ 经过块生成采样后, 得到局部特征块集合 $\{P_t^i\}, t=1,2,\cdots,T$, $i=1,2,\cdots,N$, P_t^i 表示第 t 张图像中的第 i 个局部特征块。对于基于每个关键点采样的局部特征块, 先使用1×1卷积层将通道压缩至 256 维, 然后使用 GAP 操作压缩空间维度, 再对 T 个局部特征块进行池化, 最后送进 256 维的全连接层进行分类。考虑到每个关键点在特征图上采样的特征块具有不同语义信息, 例如眼睛和鼻子等, 本文设置了不同的卷积层 $\{\text{Conv}_i\}, i=1,2,\cdots,N$ 和全连接层 $\{\text{FC}_i\}, i=1,2,\cdots,N$ 对其进行处理, 彼此之间不共享参数。算法 1 为局部空间特征分类模块进行决策的具体过程。

算法 1. 局部空间特征分类模块的分类方法

输入: 一个视频片段的图像特征 $F_t, t=1,2,\cdots,T$

五官轮廓关键点 $(x_i,y_i), i=1,2,\cdots,N$

输出: 局部空间特征的决策分数 $S_i, i=1,2,\cdots,N$

(1) 根据 $(x_i,y_i), i=1,2,\cdots,N$ 构建采样局部特征的变换矩阵 $\boldsymbol{\theta}_i, i=1,2,\cdots,N$。

(2) 根据上一步中得到的变换矩阵 $\boldsymbol{\theta}_i$ 从图像特征 $F_t, t=1,2,\cdots,T$ 中采样局部特征块 $P_t^i, t=1,2,\cdots,T$, $i=1,2,\cdots,N$。

(3) 将 P_t^i 送入对应的卷积层 Conv_i 压缩通道维数至 256, 并在空间维度进行全局平均池化, 得到压缩的特征向量 $\hat{P}_t^i$。

(4) 将 $\{\hat{P}_t^i\}, t=1,2,\cdots,T$ 进行图像级别的池化, 得到视频片段级别的局部特征 $\hat{P}^i$。

(5) 将 $\hat{P}^i$ 送入对应的全连接层 FC_i 计算概率, 并进行 Softmax 归一化, 得到最后的决策分数 $S_i, i=1,2,\cdots,N$。

3.3 损失函数

本文中使用的损失函数由全局时序分类损失和局部空间分类损失两部分组成, 两者都是用交叉熵(Cross Entropy, CE)损失进行计算。

全局时序分类损失的计算公式如下:

$$L_{global}=-(y\log(p)+(1-y)\log(1-p)), \tag{7}$$

其中, y 是真实标签($y=0$ 表示真类; $y=1$ 表示假类), p 是预测概率。

局部空间分类损失的计算公式如下:

$$L_{local}=\frac{1}{N}\sum_{i}^{N}-(y\log(p_i)+(1-y)\log(1-p_i)), \tag{8}$$

其中, y 是真实标签, p_i 是第 i 个局部块的预测概率, N=12 是局部分类器的个数。

最终的损失函数为全局时序分类损失和局部空间分类损失的结合, 计算公式如下:

$$L_{total}=L_{global}+\alpha L_{local}, \tag{9}$$

其中, α 代表局部空间分类损失的权重, 本文中其值设置为 1。

3.4 决策融合

在推断阶段, 对全局时序分类模块和局部空间分类模块的结果进行决策融合, 计算公式如下:

$$S_{final}=\lambda S_c+(1-\lambda)\frac{1}{N}\sum_{i}^{N}S_i, \tag{10}$$

其中, λ 和 $1-\lambda$ 分别代表全局时序决策分数和局部空间决策分数的权重, 本文中 λ 的值设置为 0.3, 具体参数调整参见 4.6 节实验。

4 实验结果与分析

4.1 数据集介绍

为了评价所提出方法的性能, 我们在FaceForensics++[4]公开数据集进行了实验。FaceForensics++是一个大规模的人脸伪造视频数据集, 包含 1000 条从视频网站上收集的原始视频, 每条视频时长约 15 秒, 并利用四种最新的人脸伪造方法进行伪造, 每种方法各生成 1000 条虚假视频, 其中训练集占 720 条, 验证集和测试集各占 140 条。数据集中包含人脸身份替换(FaceSwap 和 DeepFakes)和人脸表情扮演(Fac2Face 和 NeuralTextures)两种类型, 其中 FaceSwap 和 Face2Face 是基于计算机图形学的方法, Deepfakes 和 NeuralTextures 是基于深度学习的方法。数据集为了更好地适应互联网上的视频质量, 将所有视频按照 H.264 编码方式进行压缩, 一共分成三个版本, 其中 Raw 表示未经压缩, HQ 表示压缩参数为 23, LQ 表示压缩参数为 40。由于在数据集 Raw 版本各种方法的检测效果都很好, 本文选取视频质量适中的版本 HQ 进行实验。

4.2 实验设置

对于 FaceForensics++的训练集和测试集, 将其划分成不重叠的视频片段, 每个片段包含 8 个连续视频帧。每类抽取的视频片段数量如表 1 所示。对于每个视频片段中的视频帧, 根据 3.1 节中的方法裁剪出人脸区域。为了实现公平的比较, 同文献[4]一样, 针对于每类伪造方法, 分别训练一个二分类器, 并使用相同的评价指标, 即准确率(Accuracy, Acc), 其计算公式表示为:

$$Acc = \frac{TP + TN}{TP + TN + FP + FN}, \tag{11}$$

其中, *TP* (True Positive)表示将假类样本正确分类为假类样本的数量, *TN* (True Negative)表示将真类样本分类为真类样本的数量, *FP* (False Positive)表示将真类样本错误分类为假类样本的数量, *FN* (False Negative)表示将假类样本错误分类为真类样本的数量。

4.3 实现细节

本文实验平台为 Ubuntu 16.04 操作系统, 配置两块 Tesla P100 显卡, 搭配 Intel Xeon E5-2682 v4 @ 2.50GHz 16 核处理器, 所有代码都是在 PyTorch 框架下实现。我们使用在 ImageNet 上预训练的权重对图像特征提取模块进行初始化。模型的优化使用随机梯度下降方法进行, 其中 momentum 和 weight decay 分别设置为 0.9 和 5×10^{-4}, 学习率初始化为 0.001, 每 1000 步衰减至原来的 0.9。训练集的批次大小设置为 12, 迭代的总次数设置为 30000。

表 1 FaceForensics++ 中每个类别提取的片段数量

Table 1 Numbers of clips for each category in FaceForensics++

	Pristine	DeepFakes	Face2Face	FaceSwap	NeuralTextures
训练集	45445	45445	45446	36026	36068
测试集	9137	9137	9138	7374	7376

4.4 不同视频片段长度对检测效果的影响

为了探究连续多帧图像中的时序信息对伪造人脸视频检测效果的影响, 本实验对输入到模型中的视频片段进行了不同的设置, 其长度 T 分别设置为 1,2,4 和 8, 当 $T=1$ 时表示不使用全局时序特征分类模块, 而使用一个 512 维的全连接层进行替代。实验比对结果如表 2 所示。

表 2 不同视频片段长度下, 本文方法在 FaceForensics++ HQ 数据集上的准确率(%)

Table 2 Accuracy(%) of our method with different clip length on FaceForensics++ HQ

Clip Length	DeepFakes	Face2Face	FaceSwap	NeuralTextures
T=1	99.16	99.08	99.61	93.62
T=2	99.43	99.22	99.67	94.57
T=4	99.55	99.40	99.66	94.98
T=8	**99.6**	**99.44**	**99.72**	**95.5**

从实验结果可以观察到, 随着视频片段的增加所有类别上准确率都有不同程度的提升, 并在 $T=8$ 时取得最佳的检测效果, 验证了连续多帧图像中的时序信息对伪造人脸视频检测的有效性。

4.5 不同分类模块的检测效果

为了验证所提出方法的合理性和有效性, 本文针对性地进行了消融实验, 以图像特征提取模块为基础网络, 逐步添加全局时序特征分类模块和局部空间特征分类模块, 分别计算其在 FaceForensics++ HQ 数据集上的检测准确率。本实验设置了五种不同的对比模型。Baseline 表示直接使用图像特征提取模块, 加上全局池化层和全连接层进行分类, 并只选取视频片段的第一帧为代表进行测试。Baseline_T8 表示使用和 Baseline 同样的网络结构, 但测试时, 分别计算视频片段每一帧的预测分数, 再将 8 帧的预测分数取平均作为视频片段的决策结果。GLNet_G 表示使用图像特征提取模块和全局时序特征分类模块, GLNet_L 表示使用图像特征提取模块和局部空

间特征分类模块, GLNet 表示使用所有的模块, 即本文提出的方法, 三者都是在整个视频片段进行训练和测试。实验对比结果如表 3 所示。

表 3　在 FaceForensics++ HQ 数据集上消融实验结果
Table 3　Results of ablation study on FaceForensics++ HQ

	DeepFakes	Face2Face	FaceSwap	NeuralTextures
Baseline	99.14	98.93	99.45	93.16
Baseline_T8	99.45	99.16	99.64	94.25
GLNet_G	99.31	99.28	99.62	94.7
GLNet_L	99.44	99.41	99.7	93.43
GLNet	**99.6**	**99.44**	**99.72**	**95.5**

从实验结果可得到以下几点观察: (1)根据 Baseline 和 Baseline_T8 的结果对比, 可以发现简单融合多帧的预测结果能够提升对视频片段的检测准确率。(2)根据 Baseline、Baseline_T8 和 GLNet_G 的结果对比, 可以发现 GLNet_G 的检测效果相对于 Baseline 有一定的提升, GLNet_G 在 Face2Face 和 NeuralTextures 上比 Baseline_T8 的效果更优, 而在其他两个类别效果要差, 需要注意的是 GLNet_G 利用 LSTM 最后的隐藏层进行判断, 其中包含了整个视频片段的时序特征, 而 Baseline_T8 还是依靠单帧图像中空间特征进行判断, 两者有着本质上的区别。(3)GLNet_L 在 DeepFakes, Face2Face 和 FaceSwap 上比 GLNet_G 检测效果更好, 而 GLNet_G 在 NeuralTextures 上比 GLNet_G 检测效果要好, 从侧面说明了不同伪造方法在空间上和时序上展现的伪造痕迹是具有一定的差异。(4)GLNet 的检测效果比 GLNet_G 和 GLNet_L 两个单分支模型的检测效果都要好, 证明了融合全局时序特征分类模块和局部空间特征分类模块进行决策的有效性。(5)根据 Baseline_T8 和 GLNet 的结果比对, 可以发现虽然简单融合多帧的预测分数能够提升一些检测效果, 但本文提出的方法能够学习到丰富的全局时序特征和局部空间特征进行决策, 从而达到了更优的检测效果。

4.6　决策融合实验

为了探究在融合全局空间特征分类分数和局部空间特征分类分数进行决策时, 参数 λ 的取值对最终决策结果的影响, 对每类检测器进行了不同参数 λ 的对比实验, 结果如图 6 所示。

从实验结果可知, 参数 λ 的不同取值对检测器的效果有一定的影响, 尤其在 Face2Face 和 NeuralTextures 两个类别上。总体来看, 随着参数 λ 的增大, 检测准确率先提升再下降, 并在 λ=0.3 附近取得最优的结果, 表示在本文所提模型中, 局部细节特征比全局时序特征要重要。

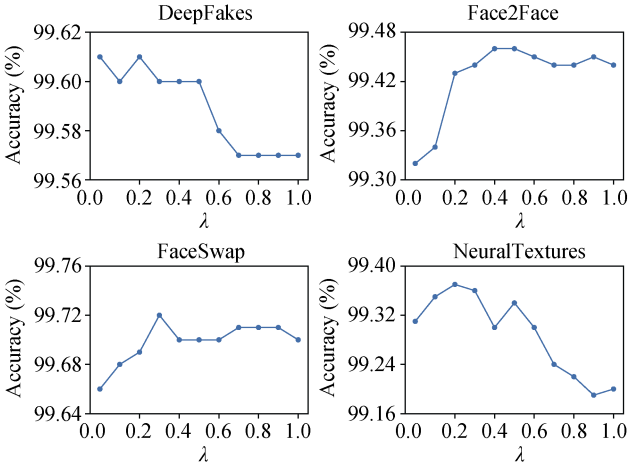


图 6　不同参数 λ 检测性能
Figure 6　Detection performance with different setting of parameter λ

4.7　与其他算法的比较

本实验选取同样基于深度学习的前沿检测算法进行比较, 包括 Bayar and Stamm[23], Cozzolino et al.[24], MesoNet[26]和 Inception[4], 其中 Inception 是 FaceForensics++数据集上的基线方法。GLNet 表示本文提出的方法, 包含图像特征提取模块、全局时序特征分类模块和局部空间特征分类模块, 并使用整个视频片段进行训练和测试。由于其他方法都是基于单个图像的检测结果, 为了更加公平的比较, 可以参照表 3 中 Baseline 结果进行比较, 后者是基于 VGG16 网络实现的。实验比对结果如表 4 所示。

表 4　各方法在 FaceForensics++ HQ 上的准确率(%)
Table 4　Accuracy(%) of manipulation-specific forgery detectors on FaceForensics++ HQ

	DeepFakes	Face2Face	FaceSwap	NeuralTextures
Bayar and Stamm[23]	90.18	94.93	93.14	80.95
Cozzolino et al.[24]	81.78	85.32	85.69	80.60
MesoNet[26]	95.26	95.84	93.43	85.96
Inception[4]	98.85	98.36	98.23	94.5
GLNet_T8	**99.6**	**99.44**	**99.72**	**95.36**

根据 Baseline 和 Inception 的结果比对, 可以发现 Inception 在 NeuralTextures 上的检测效果要优于 Baseline,而在其他三类上检测效果要差, 这可能是由于网络模型和人脸裁剪方案不同造成的。从整个实验结果来看, 与其他基于深度学习的伪造人脸视频

的检测方法相比，本文所提出的方法在FaceForensics++ HQ数据集上的检测效果最好，充分验证了其有效性。

4.8 决策依据分析

为了进一步探究分类模型在鉴别伪造人脸时，具体关注哪些重要区域进行决策，本文使用梯度加权类激活映射[42](Gradient-weighted Class Activation Mapping, Grad-CAM)对FaceForensics++数据集中几种伪造方法生成的样本进行了分析。Grad-CAM能对CNNs模型的决策提供视觉可解释性，通过目标类别的梯度，在最后一个卷积层产生一个粗略的热力图(Heatmap)，突出显示图像中用于预测目标类别的重要区域。为了得到某个卷积层的Grad-CAM，需要对其所有的激活特征图求权加和。具体地，定义第 k 个特征图对类别 c 的权重为 α_k^c，其计算公式如下:

$$\alpha_k^c = \frac{1}{Z}\sum_i\sum_j \frac{\partial y^c}{\partial A_{ij}^k}, \tag{12}$$

其中，Z 表示第 k 个特征图中像素点的个数，y^c 表示类别 c 的预测分数，A_{ij}^k 表示第 k 个特征图中像素点 (i,j) 的激活值。在得到每个特征图的权重后，最终的Grad-CAM可根据如下公式计算:

$$L_{Grad-CAM}^c = ReLU\left(\sum_k \alpha_k^c A^k\right), \tag{13}$$

其中，a_k^c 表示第 k 个特征图对类别 c 的权重，A^k 表示第 k 个特征图的激活值。

具体的，对于FaceForensics++的每种伪造方法，直接使用4.5节中训练好的Baseline模型结合Grad-CAM生成热力图，结果如图7所示，第一列是原始样例，其余列分别对应四种伪造方法的样例。从图7可以观察，Pristine样例的激活区域主要集中在唇部区域。DeepFakes样例的激活区域主要集中在眉毛附近，这与实际情况也相符合，DeepFakes方法需将转换后的人脸贴回到目标域的人脸上，可能会在眉毛附近区域留下贴合的痕迹。Face2Face样例的激活区域主要集中在鼻子区域，该方法是基于计算机图形学的方法实现的，对五官区域的形变支持不好，尤其会在鼻梁区域留下阴影或鼻孔区域造成较大的扭曲。FaceSwap同样是基于计算机图形学的方法实现，对于五官的形变扭曲控制较差，因此，其样例的激活区域集中在唇部和眼睛区域。NeuralTextures样例的激活区域除了聚集在五官区域，同样会分布在脸颊区域，这可能由该方法基于纹理渲染合成伪造人脸造成的。

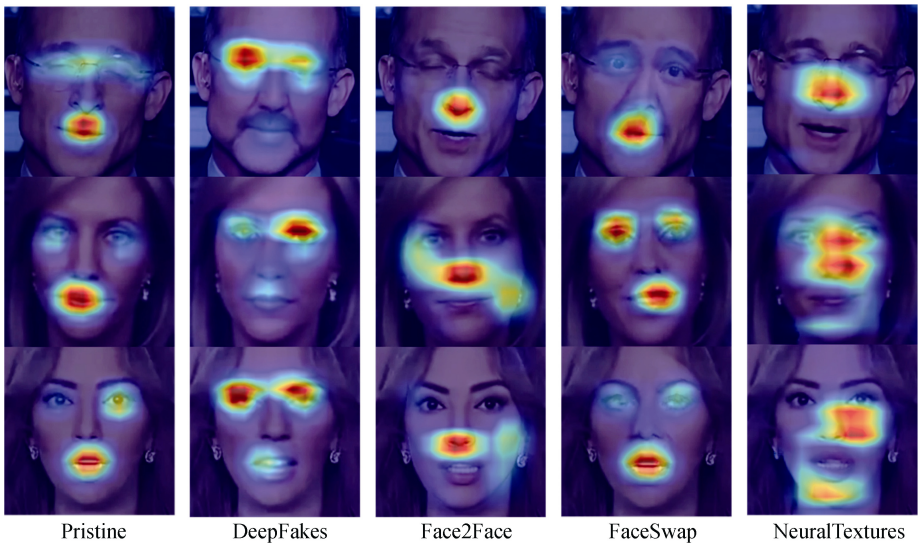


图7 FaceForensics++中伪造方式的Grad-CAM热力图实例

Figure 7 Grad-CAM heatmaps of manipulated methods in FaceForensics++

由此，可以观察到不管基于何种方式伪造人脸，人脸五官区域的细节信息对于鉴别真伪都很重要。本文提出的局部空间特征分类模块正是基于上述观察，以人脸关键点为中心采样五官区域的局部信息进行决策。此外，局部空间特征分类模块和图像特征提取模块联合进行端到端的训练，也会引导图像特征提取模块重点关注五官区域的特征，从而学习到更具有鉴别性的特征。

5 结论

本文提出了一个统一的伪造人脸视频检测框架，该模块包含图像特征提取模块、全局时序特征分类

模块和局部空间特征分类模块, 融合视频中的全局时序信息和五官区域的局部细节信息进行决策, 提升了对伪造人脸视频的检测效果。实验结果表明, 本文提出的方法在检测准确率上比其他前沿算法取得了更好的效果。

未来主要工作包括: (1)设计更加通用的检测模型来应对未知伪造方式的人脸视频检测; (2)设计统一的模型同时进行人脸定位和人脸伪造鉴别任务来提升伪造人脸视频的检测效率。

参考文献

[1] Goodfellow I., Pouget-Abadie J., Bengio Y, et al. Generative adversarial nets[C]. *Neural Information Processing Systems (NeurIPS'14)*, 2014: 2672-2680.

[2] Walia S., Kumar K., Digital image forgery detection: a systematic scrutiny[J]. *Australian Journal of Forensic Sciences*, 2019, 51(5): 488-526.

[3] Johnston P., Elyan E. A review of digital video tampering: from simple editing to full synthesis[J]. *Digital Investigation*, 2019, 29:67-81.

[4] Rössler A., Cozzolino D., Verdoliva, L. Riess, C., et al. Faceforensics++: Learning to detect manipulated facial images[C]. *IEEE International Conference on Computer Vision (ICCV'19)*, 2019:1-11.

[5] Blanz V., Scherbaum K., Vetter T., et al. Exchanging faces in images[C]. *Computer Graphics Forum*. 2004. Oxford, UK and Boston, USA: Blackwell Publishing, 2004:669-676.

[6] Blanz, V., Vetter, T., Rockwood, A. A Morphable Model for the Synthesis of 3D Faces[C]. *ACM Siggraph*, 2002:187-194.

[7] Cao C, Weng Y L, Zhou S, et al. FaceWarehouse: A 3D Facial Expression Database for Visual Computing[J]. *IEEE Transactions on Visualization and Computer Graphics*, 2014, 20(3): 413-425.

[8] 3D face swapping implemented in Python, https://github.com/MarekKowalski/FaceSwap, 2015.

[9] Korshunova I., Shi W., Dambre J., et al. Fast face-swap using convolutional neural networks[C], *IEEE International Conference on Computer Vision (ICCV'17)*, 2017:3677-3685.

[10] Deepfakes Software For All, https://github.com/deepfakes/ faceswap, 2017.

[11] Pérez P, Gangnet M, Blake A. Poisson Image Editing[J]. ACM *Transactions on Graphics*, 2003, 22(3): 313.

[12] faceswap-GAN, https://github.com/shaoanlu/faceswap-GAN, 2018.

[13] Nirkin Y., Keller Y., Hassner T., FSGAN: Subject Agnostic Face Swapping and Reenactment[C], *IEEE International Conference on Computer Vision (ICCV'19)*, 2019: 7184-7193.

[14] 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.

[15] Suwajanakorn S., Seitz S. M., Kemelmacher-Shlizerman I., Synthesizing obama: learning lip sync from audio[C]. *ACM Trans. Graphics*, 2017: 95-97.

[16] Averbuch-Elor H., Cohen-Or D., Kopf J., et al. Bringing portraits to life[C]. *ACM Trans. Graphics*, 2017:196-199.

[17] Wiles O, Koepke A S, Zisserman A. X2Face: A Network for Controlling Face Generation by Using Images, Audio, and Pose Codes[EB/OL]. 2018: arXiv:1807.10550[cs.CV]. https://arxiv.org/abs/1807.10550.

[18] Pumarola A, Agudo A, Martinez A, et al. GANimation: Anatomically-aware Facial Animation from a Single Image[EB/OL]. 2018: arXiv:1807.09251[cs.CV]. https://arxiv.org/abs/1807.09251.

[19] Thies J., Zollhöfer M., Nießner M., Deferred Neural Rendering: Image Synthesis using Neural Textures[EB/OL]. 2019: arXiv preprint arXiv:1904.12356.

[20] Zakharov E, Shysheya A, Burkov E, et al. Few-Shot Adversarial Learning of Realistic Neural Talking Head Models[EB/OL]. 2019: arXiv:1905.08233[cs.CV]. https://arxiv.org/abs/1905.08233.

[21] Huang X, Belongie S. Arbitrary Style Transfer in Real-time with Adaptive Instance Normalization[EB/OL]. 2017: arXiv:1703.06868[cs.CV]. https://arxiv.org/abs/1703.06868.

[22] Ha S., Kersner M., Kim B., et al. MarioNETte: Few-shot Face Reenactment Preserving Identity of Unseen Targets[EB/OL]. 2019: arXiv preprint arXiv:1911.08139.

[23] Bayar B., Stamm M. C., A deep learning approach to universal image manipulation detection using a new convolutional layer[C]. *ACM Workshop on Information Hiding and Multimedia Security (IH&MMSEC'16)*,2016: 5-10.

[24] Cozzolino D, Poggi G, Verdoliva L. Recasting Residual-based Local Descriptors as Convolutional Neural Networks: An Application to Image Forgery Detection[EB/OL]. 2017: arXiv:1703.04615 [cs.CV]. https://arxiv.org/abs/1703.04615.

[25] Zhou P., Han X., Morariu V. I., et al. Two-stream neural networks for tampered face detection[C]. *IEEE Computer Vision and Pattern Recognition Workshops (CVPRW'17)*, 2017:1831-1839.

[26] Afchar D., Nozick V., Yamagishi J. et al. Mesonet: a compact facial video forgery detection network[C]. *IEEE International Workshop on Information Forensics and Security (WIFS'18)*, 2018: 1-7.

[27] 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.

[28] Li Y., Lyu S. Exposing deepfake videos by detecting face warping artifacts[C]. *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW'19)*, 2019: 46-52.

[29] Stehouwer J., Dang H., Liu F., et al. On the Detection of Digital Face Manipulation[EB/OL]. 2019: arXiv preprint arXiv:1910.01717.

[30] Nguyen H H, Fang F M, Yamagishi J, et al. Multi-task Learning for Detecting and Segmenting Manipulated Facial Images and Videos[EB/OL]. 2019: arXiv:1906.06876[cs.CV]. https://arxiv.org/abs/1906.06876.

[31] Nguyen H. H., Yamagishi J., Echizen I. Use of a Capsule Network to Detect Fake Images and Videos[EB/OL]. 2019: arXiv preprint arXiv:1910.12467.

[32] Sabour S., Frosst N., Hinton G. E. Dynamic routing between capsules[C]. *Neural Information Processing Systems (NeurIPS'17)*, 2017: 3856-3866.

[33] Li Y Z, Chang M, Lyu S W. In Ictu Oculi: Exposing AI Generated Fake Face Videos by Detecting Eye Blinking[EB/OL]. 2018: arXiv:1806.02877[cs.CV]. https://arxiv.org/abs/1806.02877.

[34] Ciftci U A, Demir I. FakeCatcher: Detection of Synthetic Portrait Videos Using Biological Signals[EB/OL]. 2019: arXiv:1901.02212[cs.CV]. https://arxiv.org/abs/1901.02212.

[35] Güera D., and Delp E. J. Deepfake video detection using recurrent neural networks[C]. *IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS'18)*, 2018: 1-6.

[36] Sabir E., Cheng J., Jaiswal A., et al. Recurrent Convolutional Strategies for Face Manipulation Detection in Videos[C]. *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW'17)*, 2019: 80-87.

[37] King D. E. Dlib-ml: A machine learning toolkit[J], *Journal of Machine Learning Research*, 2009, 10(7): 1755-1758.

[38] Simonyan K. , Zisserman A. Very deep convolutional networks for large-scale image recognition[EB/OL]. 2019: arXiv preprint arXiv:1409.1556.

[39] Ioffe S. , Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift[EB/OL], 2017: arXiv preprint arXiv:1502.03167.

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

[41] Jaderberg M, Simonyan K, Zisserman A, et al. Spatial Transformer Networks[EB/OL]. 2015: arXiv:1506.02025[cs.CV]. https://arxiv.org/abs/1506.02025.

[42] Selvaraju R. R., Cogswell M., Das A., et al. Grad-cam: Visual explanations from deep networks via gradient-based localization[C]. *IEEE International Conference on Computer Vision (ICCV'17)*, 2017:618-626.

陈鹏 于 2016 年在山东大学(威海)计算机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第三研究室攻读博士学位。研究领域为计算机视觉、人工智能安全。研究兴趣包括: 目标检测、对抗样本以及深度伪造的生成和检测。Email: chenpeng@iie.ac.cn

梁涛 于 2018 年在重庆理工大学计算机科学与工程学院获得学士学位。现在中国科学院信息工程研究所第三研究室攻读硕士学位。研究领域为计算机视觉、深度伪造检测。研究兴趣包括: 深度学习、图神经网络、视频检索。Email: liangtao0305@iie.ac.cn

刘锦 于 2018 年在北京交通大学信息安全(保密技术)专业获得学士学位。现在中国科学院信息工程研究所第三研究室攻读博士学位。研究领域为计算机视觉、图像生成。研究兴趣包括: 深度伪造的生成和检测。Email: liujin@iie.ac.cn

戴娇 于 2019 年在中国科学院(计算机系统结构)专业获得博士学位。现任中国科学院信息工程研究所高级工程师。研究领域为多媒体信息处理、人工智能安全。研究兴趣包括: 图像检索、图像深度伪造、对抗防御。Email: daijiao@iie.ac.cn

韩冀中 于 2001 年在中国科学院计算技术研究所获得博士学位, 现为中国科学院信息工程研究所第三研究室正研级高工、博士生导师。主要研究领域为大数据存储与管理、多媒体信息智能化处理。研究兴趣: 多媒体内容理解、深度伪造。Email: hanjizhong@iie.ac.cn