计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.210200127

基于 3D CNNS 的深度伪造视频篡改检测

邢 豪 李 明
太原理工大学大数据学院 山西 晋中 030600
(923136917@qq.com)

摘 要 近年来,“Deepfake”视频引起了广泛的关注。人们很难区分 Deepfake 视频。这些篡改的视频将给社会带来巨大的潜在威胁,如被用来制作假新闻等。因此,目前需要找到一种有效识别这些合成视频的方法。针对上述问题,提出了一种基于 3D CNNS 的深度伪造视频检测模型。该模型注意到 Deepfake 视频的时域特征和空域特征的不一致,而 3D CNNS 可以有效捕获 Deepfake 视频的这一特征。实验结果表明,基于 3D CNNS 的模型在 Deepfake 检测挑战数据集和 Celeb-DF 数据集上具有较高的准确率和较强的鲁棒性,准确率可达 96.25%,*AUC* 值可达 0.92,同时该模型解决了泛化性差的问题。通过与现有的 Deepfake 检测模型进行对比,所提模型在检测准确率和 AUC 取值方面均优于现有模型,验证了该模型的有效性。

关键词:篡改视频;Deepfake 检测;时域特征;空域特征;三维卷积网络
中图法分类号 TP391.41

Deepfake Video Detection Based on 3D Convolutional Neural Networks

XING Hao and LI Ming
College of Data Science, Taiyuan University of Technology, Jinzhong, Shanxi 030600, China

Abstract In recent years, "Deepfake" has attracted widespread attention. It is difficult for people to distinguish Deepfake videos. However, these forged videos will bring huge potential threats to our society, such as being used to make fake news. Therefore, it is necessary to find a method to identify these synthetic videos. In order to solve the problem, a Deepfake video detection model based on 3D CNNS for deepfake detection is proposed. This model notices the inconsistency of temporal and spatial features in the Deepfake video, and 3D CNNS can effectively capture temporal and spatial features of deepfake video. The experimental results show that models based on 3D CNNS have high accuracy rate, and strong robustness on the Deepfake-detection-challenge dataset and Celeb-DF dataset. The detection accuracy of the proposed model reaches 96.25%, and the *AUC* value reaches 0.92. This model also solves the problem of poor generalization. By comparing with the existing Deepfake detection models, the proposed model is superior to the existing models in terms of detection accuracy and AUC value, which verifies the effectiveness of the proposed model.

Keywords Synthetic videos, Deepfake detection, Temporal features, Spatial features, 3D CNNS

1 引言

当前,随着人工智能技术的飞速发展,人工智能的能力在很多方面都超过了普通人。如果我们不合理地使用人工智能技术,那么其负面影响难以估计,而 Deepfake[1] 就是其中的代表。“Deepfake”是“deep machine learning”(深度学习)和“fake”(伪造)的英语组合,又被称为“深度伪造”,它是一种基于深度学习的人脸图像合成技术。Deepfake 通常在图片和视频制作过程中用于合成新的图像或视频。随着 GAN (Generative Adversarial Networks)[2] 的出现,Deepfake 视频变得越来越容易制作。非专业人士可以使用 GitHub 上的代码快速制作深度伪造视频和图像。Deepfake 视频在互联网上广泛传播的同时,会给社会带来巨大的潜在威胁,如互联网上的许多深度伪造视频都带有色情内容。荷兰网络安全初创公司 Deeptrace 在 2019 年 10 月发布的一份报告中估计,所有在线 Deepfake 中有 96%是色情的。除了用于制作虚假的色情视频外,这一技术还被用来在视频中歪曲知名的政客。因此,找到有效检测 Deepfake 视频的方法非常重要。为了应对 Deepfake 技术的威胁,AWS,Facebook,Microsoft 以及学者共同组成了 Deepfake 检测挑战赛[3](DFDC)。

目前,关于 Deepfake 视频的检测模型更多的是使用针对每一帧人脸特征的 CNN 方法以及关注帧与帧图像之间特征

到稿日期:2021-02-22 返修日期:2021-04-29 本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
基金项目:国家自然科学基金项目(11771321);山西省科技厅社会发展科技攻关计划项目(201703D321032)
This work was supported by the National Natural Science Foundation of China(11771321) and Shanxi Province Plan Project on Science and Technology of Social Development(201703D321032).
通信作者:李明(lm13653600949@126.com)

的 CNN＋LSTM 的方法。现有的检测模型具有高度的领域特异性，在跨领域部署中可能会导致性能大幅下降，尤其是在测试领域差距较大的情况下，现在的方法最大的问题就是泛化性能不足。而在近几年的视频动作识别领域，基于 3D CNNS 的方法效果显著，可以有效捕捉视频中人的运动特征。

本文提出了一种基于 3D CNNS 的深度伪造视频检测方法。本文的贡献如下：

(1)比较了不同的 3D CNNS 模型，并且改进了 3D CNNS 的结构。所得模型具有比现在的方法更好的泛化性能。

(2)在数据预处理阶段，本文不是只提取人脸这一小块信息，而是提取了一个包围盒，其中包括整个时间的整个脸部，并在包括每个帧的感兴趣区域中创建了一个视频。这种处理方式的优点是大大减少了所提模型对于人脸区域的误报。

2　相关工作

2.1　Deepfake 的生成方法

Deepfake 面向公众的第一个产品是 FakeApp[4]，它是 Reddit 用户使用 autoencoder-decoder[5] 结构开发的。该方法中，自动编码器提取经过压缩的人脸图像的潜在特征，然后使用解码器重建人脸图像，让解码器学习如何从含有噪声的输入中恢复人脸图像。为了在源图像和目标图像之间交换人脸，需要两个编码器/解码器对，其中每对用于在一个人脸数据集上进行训练，并且编码器的参数在两个网络对之间共享。利用自动编码器/解码器结构进行人脸操纵的过程如图 1 所示。DeepFaceLab[6]，DFaker[7] 也使用了相同的方法。

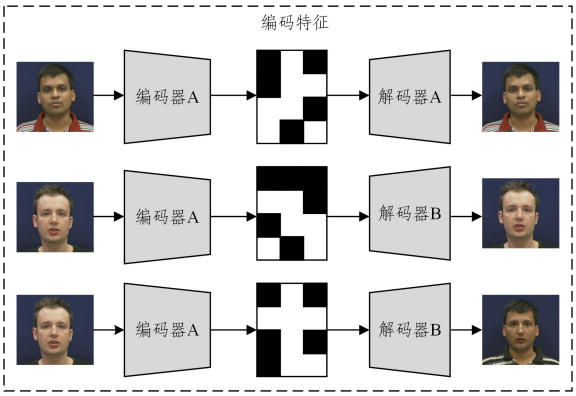

图 1　人脸操纵过程

Fig. 1　Process of face manipulation

随着生成对抗网络的提出，Deepfake 生成技术得到了提高。通过将 VGGFace[8] 中实现的对抗性损失和感知损失添加到编码器-解码器体系结构中，再加入了生成对抗网络的判别器结构，即得到深度伪造的改进版本 Faceswap-gan[9]。它增加了 VGGFace 感知损失，使眼球运动更加逼真，与输入人脸保持一致，从而获得更高质量的输出视频。Nirki 等[10] 提出一个全卷积神经网络的人脸分割算法来实现换脸。其首先利用 2D 人脸关键特征点来拟合一个 3D 模型，然后经过一个预训练的全卷积神经网络(FCN)从背景和遮挡处分割人脸，并且覆盖到对应的 3D 人脸上，最后融合到目标人脸上以实现换脸。

Thies 等提出了 Face2face[11] 技术，通过使用一种密集光度一致性方法(dense photometric consistency measure)来实时跟踪源和目标视频中的面部表情，将源人物表情传递给目标人物，使得目标人物可以实时重现源人物的表情特征，这让深度伪造技术得到了进一步发展。

如今 Deepfake 技术主要分为面部身份替换和面部表情重现。人脸操纵的不同类别示例如图 2 所示。

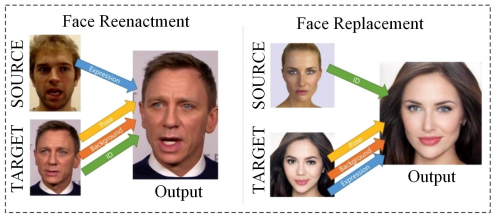

图 2　不同类别的人脸操纵示例

Fig. 2　Examples of different types of face manipulation

2.2　深度伪造视频的检测方法

目前，国内外对 Deepfake 的检测技术研究正在不断增多。Deepfake 视频检测可以分为两种类型：一种是基于单帧的图像特征，另一种是基于不同帧之间的时间特征。

基于单帧的图像特征。Yang 等[12] 提出利用头部姿势不一致的特征来检测深度伪造视频。该方法的依据是基于神经网络合成算法不能保证原始人脸和合成人脸具有一致的人脸标志点。标志点位置的误差可能不是人眼直接可见的，但可以通过不同人脸真实和虚假部分的 2D 标志点估计的头部姿势(即头部方向和位置)来揭示。通过得到的人脸头部标志点训练一个基于支持向量机的分类器来区分原始视频和深度伪造视频。深度伪造视频通常创建的分辨率有限，需要用面部扭曲的方式来匹配原始面部，而扭曲的面部区域和周围环境之间的分辨率不一致，这个过程留下了可以被 CNN 模型检测到的视觉伪影，我们可以使用 VGG16，ResNet50，ResNet101 和 ResNet152 来检测。Li 等[13] 通过检测面部扭曲伪影来鉴别深度伪造的视频。Nguyen 等[14] 提出使用基于胶囊网络[15-16] 的模型来检测深度伪造图像和伪造视频。胶囊网络采用从 VGG-19 网络中获得的特征来区分假图像或视频与真实图像或视频。Rossler 等[17] 实现了 4 种基于单帧人脸图像的检测，并在 FaceForensics＋＋基准数据集上进行了评估，最后得出结论，XceptionNet 架构是 4 种方法和压缩扰动中性能最好的检测器。Afchar 等[18] 提出了一种具有少量层的 CNN 网络 Mesonet，分别在 Deepfake 数据集和 Face2Face 数据集上进行了实验。

基于帧之间的时间特征。Sabir 等[19] 使用了递归卷积模型，首先通过 CNN 网络对其中视频帧提取空间特征，然后利用 RNN 对这些特征的序列进行建模。CNN 网络采用 DenseNet[20]，RNN 采用双向 GRU[21] 结构。Guera 等[22] 探索了另一种 RNN 方法。该模型与 Sabir 等的模型非常相似，但是他们使用 Inception-v3 作为卷积分量，并使用 LSTM[23] 作为循环单元。

FakeCatcher[24] 利用隐藏在视频中的生物信号的差异来区分假视频和真实视频。Li 等[25] 提出利用眨眼频率来识别 Deepfake 视频。最近，Li 等[26] 提出利用人脸 X 射线来检测假人脸边界区域周围的篡改痕迹。

另外，可以根据深度伪造图像的传统特征和视频成像原理进行有效检测。Zhang 等[27] 提出了一种基于图像传统特

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视频的真伪。Li 等[28]提出了一种基于光照方向一致性的伪造视频检测方法，针对真伪视频在外部成像环境上和在二维光照方向上的差异进行检测。Hu 等[29]提出了计算人脸交互比的方法，然后运用图像分割网络进行真伪检测。

近几年，在视频动作识别领域，3D CNNS 的方法取得了显著效果。Mänttäri 等[30]在动作识别的背景下提供了 RNN 和 3D CNNS 体系结构的比较。他们发现，在需要理解方向和不对称性的任务中，基于 RNN 的模型预测视频的性能优于 3D CNNS。这与按顺序处理视频的 RNN 一致。但是，在其他任务中，整个序列方向都不携带任何信息，基于 3D CNNS 的模型优于基于 RNN 的模型。Deepfake 视频始终是人们面部的记录。因此，与动作数据相比，人脸数据是非常单调的，且视频中发生的方向或事件与分类视频是真实视频还是操纵视频并没有明显关系。因此，基于 Deepfake 视频数据的性质，本文使用基于 3D CNNS 的模型来处理深度伪造视频。

3 基于 3D CNNS 的深度伪造视频检测模型

深度伪造检测模型主要由数据预处理模块和深度学习模块构成。整体的检测模型框架如图 3 所示。

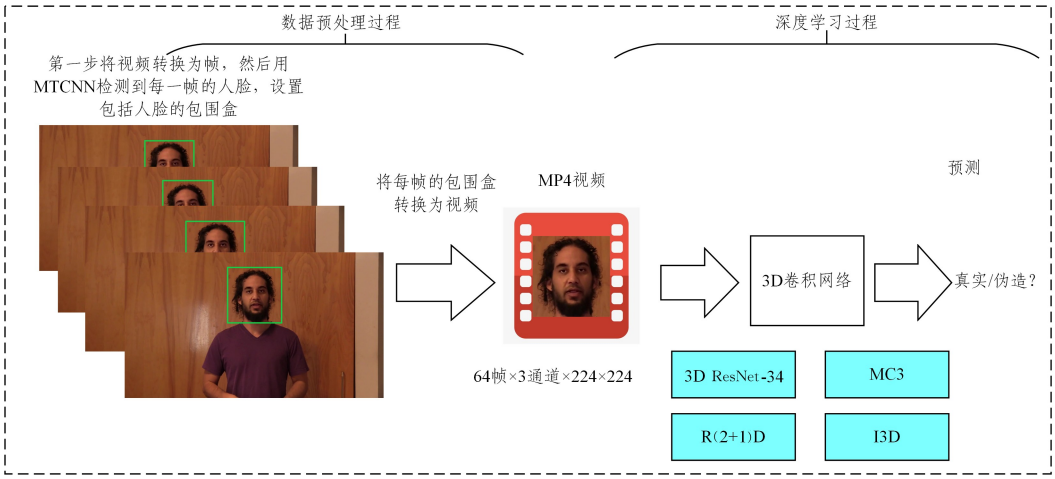

图 3 检测模型的总体架构与流程

Fig. 3 Architecture and pipeline of the detection model

3.1 数据预处理模块

对于深度伪造视频，我们主要关注人的脸部运动，而背景信息会对模型的识别结果产生干扰，因此需要对原始数据集进行预处理。人脸视频的时间长度为 10s，我们设定每 10 帧通过 MTCNN[31](Multi-task Cascaded Convolutional Networks)算法检测出人脸，然后使用边界框坐标来确定视频中的人脸位置。在视频中连续重叠的脸部被假定为一个人的脸部在时间中移动。之后，我们提取一个包围盒，该包围盒随时间的推移逐渐包括整个脸部，最后在该包围盒区域创建视频。将一个 10s 深度伪造视频分解为两个 5s 视频，每个视频只包含人脸的运动区域。原始的数据集由 50 个文件块组成，每个文件块由一定数量的真实视频和伪造视频组成，并且还包含一个 json 文件，描述了视频的标签。对于每个输入的视频，我们采用固定的 64 帧，小于 64 帧的视频则不进行训练，再将每帧的人脸视频裁剪为 224×224 大小。经过上述步骤后，将得到的人脸视频按照真假类别分别放置在不同的文件夹中以供训练。

3.2 深度学习模块

本文采用了 3D CNNS 网络，其三维卷积网络可以有效地学习到深度伪造人脸视频的空间特征和时间特征。检测模型主要采用了 I3D[32](Two-Stream Inflated 3D ConvNet)，MC3[33](Mixed 3D-2D Convolution)，R(2+1)D[33]，3D ResNet-34[34]这几种三维卷积网络，用于进行比较。

3.2.1 I3D 网络

I3D 模型采用了视频分析问题的双流法，一个通道进行 RGB 数据处理，一个通道进行光流数据处理，其双流模型结构如图 4 所示。

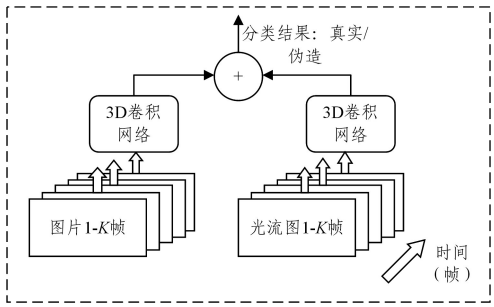

图 4 双流模型结构图

Fig. 4 Two stream model structure diagram

光流是空间运动物体在观察成像平面上的像素运动的瞬时速度。光流法是利用图像序列中像素在时间域上的变化以及相邻帧之间的相关性来找到上一帧与当前帧之间存在的对应关系，从而计算出相邻帧之间物体的运动信息的一种方法。该模型由 Inception-v1 网络拓展而来，将 Inception-v1 网络二维模型中的滤波器和池化层添加一个时间维度便可拓展为 3D-Inception 模型，其可以更好地提取视频的时序信息。Inception 模块使用 3 个不同大小的滤波器(1×1，3×3，5×5)对输入数据执行卷积操作，另外执行最大池化。将 RGB 数据和光流图数据分开训练，最后对它们各自的预测结果进行融合取平均值。

I3D 的网络模型如图 5 所示。我们对 I3D 网络进行了改进，最后一层使用 softmax 函数来输出分类结果。根据人脸

视频数据的需要，网络输入的数据为 64 帧的经过裁剪的人脸视频，每帧图像的分辨率为 224×224，每次输入 26 个视频。激活函数选用 RELU 函数，*batch size* 为 12，而损失函数采用交叉熵损失函数：

$$L=\frac{1}{N}\sum_i -[y_i \cdot \log(p_i)+(1-y_i)\log(1-p_i)]$$

其中，y_i 表示样本 i 的 label，正类为 1，负类为 0；p_i 表示样本 i 预测为正的概率。

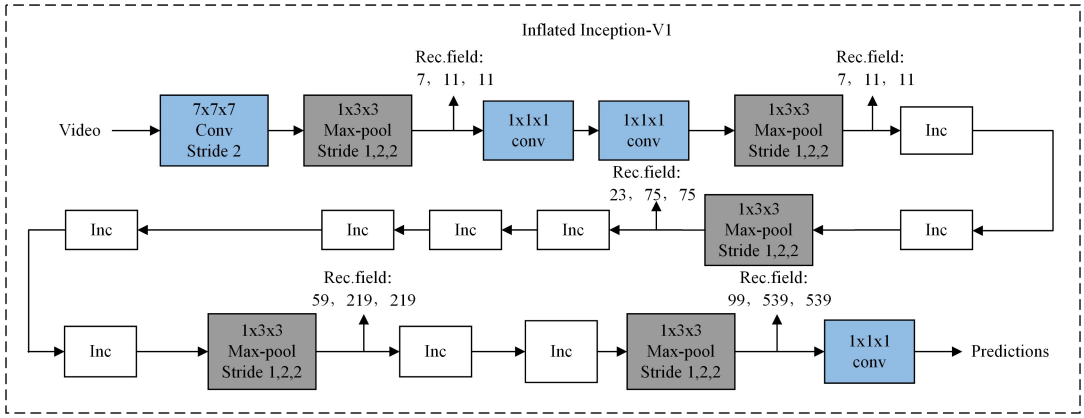

图 5　Inception-V1 网络结构图

Fig. 5　Inception-V1 Network structure diagram

数据归一化的方式设置为：

$x=(x/255)*2-1$

在网络训练初期，二分类准确率没有达到理想效果。为了提高准确率，本文使用了数据增强技术。数据增强技术主要是在训练数据上增加微小的扰动或者变化，其一方面可以增加人脸训练数据，从而提升模型的泛化能力；另一方面可以增加噪声数据，从而增加模型的鲁棒性。主要的数据增强方法有：翻转变换、随机裁剪、色彩抖动和旋转变换。

3.2.2　MC3 网络

MC3 网络融合了 2D 卷积网络和 3D 卷积网络，在网络的浅层使用三维卷积，而在顶层使用二维卷积。该结构源自 R3D 模型[34]。R3D 网络与普通残差网络的区别是其采用了 3D 卷积和 3D 池化。训练图像的分辨率分别采用 224×224 和 112×112。损失函数采用交叉熵损失函数，激活函数选用 RELU 函数，*batch size* 设置为 10。

3.2.3　R(2+1)D 网络

R(2+1)D 卷积块是一种新的时空卷积块，类似于 3D 卷积块，R(2+1)D 卷积块将 3D 卷积分解为两个独立且连续的操作：2D 空间卷积和 1D 时间卷积，其网络结构图 6 所示。

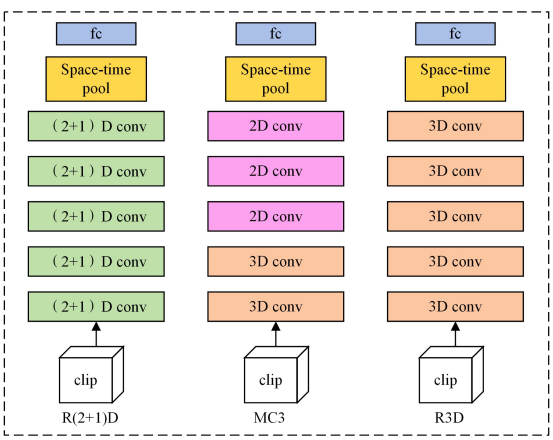

图 6　不同的残差 3D 卷积结构

Fig. 6　Different residual network architectures

两种卷积过程的对比如图 7 所示。卷积过程中的输入为包括单个特征通道的时空体积，t 代表时间范围，d 是空间宽度和高度。训练图像的分辨率采用 224×224。损失函数采用交叉熵损失函数，激活函数选用 RELU 函数，*batch size* 设置为 32。

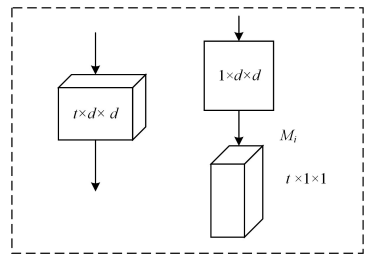

图 7　3D 卷积与(2+1)D 卷积

Fig. 7　3D convolution and (2+1)D convolution

3.2.4　3D ResNet-34 网络

深度残差网络在静态图像上表现效果较好，3D ResNet-34 网络是将 ResNet34 网络映射为三维的卷积网络，再加上时间的维度，其具体结构不发生改变。在动作识别领域上，3D ResNet 在相同深度上的表现明显优于二维的 ResNet 网络。3D ResNet-34 网络结构如图 6 的 R3D 结构。训练图像的分辨率同样采用了 224×224。损失函数采用交叉熵损失函数，激活函数选用 RELU 函数，*batch size* 设置为 16。

4　实验结果与分析

4.1　深度伪造视频数据集

实验采用的数据集为 Deepfake 视频的第二代数据集，其中包括 DFDC 数据集和 Celeb-DF[35] 数据集。

DFDC 数据集包括 119 197 个视频，每个视频时长都为 10 s，但是帧率从 15 fps 到 30 fps 不等，分辨率也从 320×240 到 3840×2160 不等。训练视频中有 19197 个视频，是由大约 430 名演员真实拍摄的片段，剩余 100 000 个视频是由真实视频生成的假脸视频。伪造人脸生成技术包括 DeepFakes，Face2Face，Faceswap，NeuralTextures 等多种生成算法，使数据集包含尽可能多的假脸视频。

Celeb-DF 数据集包括从 YouTube 收集的 590 个原始视频，这些视频中的人物有不同年龄、种族和性别，该数据集还包括生成的 5639 个相应的高质量的 Deepfake 视频。所有视频的平均长度约为 13 s，标准帧速率为每秒 30 帧。在 Celeb-DF 中合成的 Deepfake 视频的整体视觉质量得到了极大的改善，显示的视觉伪像明显减少。两种深度伪造视频数据集的相关信息如表 1 所列。

表 1 深度伪造视频数据集

Table 1 Deepfake datasets

Datasets	Tampering ratio	Number of videos	Source
DFDC	1:0.28	5214	Actors
Celeb-DF	1:0.51	5514	Youtube

4.2 实验平台设置

本文的实验平台是 Windows10 操作系统，显卡使用的是 GTX 1080TI 16GB，Python 环境为 3.7，Pytorch 为 1.3 版本。我们使用在 ImageNet 上预先训练的权重来初始化模型。该模型使用 Adamw 梯度下降法进行了优化，初始化学习率为 0.001，*batch size* 设置为 20。在 2000 次迭代后，损失开始收敛。

4.3 实验分析

为了验证所提出模型的准确性和泛化性，本文针对性地进行了对比实验，分别将不同的 3D CNNS 方法和 Mesonet，LRCN 等前沿算法进行了比较。实验采用统一的数据集进行训练。我们采用准确率和 AUC 指标来评价模型的效果，准确率(Accuracy)是指正确分类的样本占总样本的比例，是对总体准确率的评估。准确率的计算公式为：

$$Accuracy=\frac{(TP+TN)}{(TP+FN+FP+TN)} \tag{1}$$

其中，真正(True Positive，TP)是被模型预测为正的正样本；假正(False Positive，FP)是被模型预测为正的负样本；假负(False Negative，FN)是被模型预测为负的正样本；真负(True Negative，TN)是被模型预测为负的负样本。

AUC(Area Under Curve)被定义为 ROC 曲线下与坐标轴围成的面积，显然这个面积的数值不会大于 1。ROC 曲线的横坐标是假正类率，纵坐标是真正类率。由于 ROC 曲线一般都处于 $y=x$ 这条直线的上方，所以 AUC 的取值范围在 0.5 和 1 之间。$AUC=1$ 时，该模型是完美分类器，AUC 越接近于 1，检测方法的真实性越高；$AUC=0.5$ 时，模型与随机预测一样，其真实性最低，无应用价值。

表 2 的实验结果显示，在三维卷积网络中 Inception3D 的方法有着较高的准确率，相对于 CNN 方法有很大的提高。另外 3 种三维卷积网络的准确率与 CNN 方法相差不大。三维卷积网络比 CNN+LSTM 能更有效地捕捉时间和空间上的特征信息。

表 2 不同模型在 DFDC 数据集上的性能

Table 2 Performance of different models on DFDC dataset

Model	Accuracy/%	AUC
Mesonet	87.27	0.86
SE-ResNext-50	83	0.83
CNN+LSTM	95.1	0.91
3D ResNet-34	85	0.86
MC3	83.2	0.85
R(2+1)D	84.4	0.88
Inception3D	**96.25**	**0.92**

为了更好地评估本文模型的鲁棒性，我们把在 DFDC 数据集上训练好的模型放到 Celeb-DF 数据集上进行测试，通过比较准确率和 AUC 指标，来比较模型的泛化性能。

表 3 的实验结果显示，虽然 CNN+LSTM 模型在 DFDC 数据集上的准确率可以达到 90%以上，但是其泛化性能不如基于 3D CNNS 的网络。采取三维卷积网络的 4 个模型的泛化性能普遍优于现有模型。针对单帧图像特征的普通 CNN 模型只捕捉到了 DFDC 数据集伪造人脸方法的特征，在伪造人脸视频的方法有较大差异的 Celeb-DF 数据集上收效甚微。

表 3 不同模型的泛化性能比较

Table 3 Generalization performance of different models

Model	Accuracy/% (DFDC)	AUC (DFDC)	Accuracy/% (celeb DF)	AUC (celeb DF)
Mesonet	87.27	0.86	73	0.73
SE-ResNext-50	83	0.83	68	0.65
CNN+LSTM	95.1	0.91	74	0.71
3D ResNet-34	85	0.86	70	0.70
MC3	83.2	0.85	75	0.74
R(2+1)D	84.4	0.88	77	0.76
Inception3D	**96.25**	**0.92**	**83**	**0.88**

我们认为在单帧图像检测模型中，目前的深度伪造视频生成技术只能将真实图像和虚假图像的差异降到比较低的水平，所以一些伪造视频无法被检测出来。CNN+LSTM 方法虽然可以同时捕捉到伪造视频时间和空间上的特征，但是在 CNN 特征提取阶段丢失了一些低级特征，因此效果并不如基于 3D CNNS 的方法好。

I3D 网络同时考虑了 RGB 数据和光流数据，比传统的基于 3D CNNS 的方法更加出色。以上结论表明，卷积网络可以同时提取空间特征和时间特征，空间特征可以学习到不同深度伪造视频方法的特定属性，时间特征和空间特征的结合可以学习到深度伪造视频方法之间的共同属性。

5 当前挑战

虽然本文提出的基于 3D CNNS 的模型解决了泛化性问题，但还是有一些局限性。第一，3D 卷积网络相对于二维卷积网络参数量过大，训练速度比较缓慢，效率比较低，更容易发生过拟合现象。第二，训练和测试选用的数据集是经过专业机构制作的深度伪造视频，缺乏真实世界的视频，需要建立一个完整的公共数据库。第三，虽然本文模型的泛化性能得到了提高，但是准确率仍然不足 95%，无法真正地完全检测出真伪视频。

结束语 本文提出了一种基于 3D CNNS 的网络模型对深度伪造视频进行检测，取得了一定的效果，解决了现在已有检测模型中泛化性差的问题。该模型可以有效检测到深度伪造视频在时间和空间特征上的人脸差异。在 DFDC 和 Celeb-DF 数据集上的实验结果表明，本文提出的模型有着较好的性能，准确率可以达到 96%左右，AUC 可以达到 0.92。在 DFDC 数据集进行训练，在 Celeb-DF 数据集进行测试，所提

模型的准确率可以达到 83%，*AUC* 为 0.88。

随着未来人工智能技术的继续发展，深度伪造技术会不断提高，这给人脸篡改视频检测带来了很大挑战，因此后续我们会继续完善该模型。在未来的工作中，我们会尝试将 LSTM 和 3D CNNS 进行融合，即将 3D 卷积块集成在 LSTM 内，以进一步提高模型的准确率和泛化性能。

参 考 文 献

[1] Deepfake[EB/OL]. http://github.com/deepfakes/faceswap Accessed October 29,2019.

[2] GOODFELLOW I,POUGET-ABADIE J,BENGIO Y,et al. Generative adversarial nets[C]// Neural Information Processing Systems (NeurIPS'14). 2014:2672-2680.

[3] DOLHANSKY B,HOWES R,PFLAUM B,et al. The deepfake detection challenge preview dataset[J]. arXiv: 1910.08854, 2019.

[4] FakeApp[EB/OL]. http://www.malavida.com/en/soft/fakeapp.

[5] BADRINARAYANAN V,KENDALL A,CIPOLLA R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2017,39(12):2481-2495.

[6] DeepFaceLab[EB/OL]. http://github.com/iperov/DeepfaceLab.

[7] DFaker[EB/OL]. http://github.com/dfaker/df.

[8] JOSEPH I V,ZHOU Z,ZHANG C,et al. Facial Recognition via Transfer Learning:Fine-Tuning Keras_vggface[C]// 2017 International Conference on Computational Science and Computational Intelligence (CSCI). 2017.

[9] Faceswap-gan[EB/OL]. http://github.com/shaoanlu/faceswap-GAN.

[10] NIRKIN Y,MASI I,TUAN A T,et al. On face segmentation, face swapping,and face perception[C]// 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018). IEEE,2018:98-105.

[11] THIES J,ZOLLHOFER M,STAMMINGER M,et al. Face2-Face:Real-Time Face Capture and Reenactment of RGB Videos [C]// 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV, USA. Piscataway, NJ: IEEE,2016:2387-2395.

[12] YANG X,LI Y,LYU S. Exposing deepakes using inconsistent head poses[C]// 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019: 8261-8265.

[13] LI Y,LYU S. Exposing deepfake videos by detecting face warping artifacts[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. 2019:46-52.

[14] NGUYEN H H,YAMAGISHI J,ECHIZEN I. Use of a capsule network to detect fake images and videos[J]. arXiv: 1910.12467,2019.

[15] HINTON G E,KRIZHEVSKY A,WANG S D. Transforming auto-encoders[C]// International Conference on Artificial Neural Networks (ICANN). Springer,2011.

[16] SABOUR S,FROSST N,HINTON G E. Dynamic routing between capsules[C]// Conference on Neural Information Processing Systems (NIPS). 2017.

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

[18] 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). IEEE,2018:1-7.

[19] SABIR E,CHENG J,JAISWAL A,et al. Recurrent Convolutional Strategies for Face Manipulation Detection in Videos[J]. Interfaces (GUI),2019,3:1.

[20] HUANG G,LIU Z,VAN DER MAATEN L,et al. Densely Connected Convolutional Networks[C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017: 2261-2269.

[21] CHO K,VAN MERRIENBOER B,GULCEHRE C,et al. Learning phrase representations using rnn encoder-decoder for statistical machine translation[J]. arXiv:1406.1078,2014.

[22] GUERA D,DELP E J. Deepfake Video Detection Using Recurrent Neural Networks[C]// 15th IEEE International Conference on Advanced Video and Signal-Based Surveillance, Institute of Electrical and Electronics Engineers Inc. doi:10.1109/AVSS.2018.8639163,2019.

[23] HOCHREITER S,SCHMIDHUBER J. Long short-term memory[J]. Neural Computation,1997,9:1735-1780.

[24] CIFTCI U,DEMIR I. FakeCatcher:Detection of Synthetic Portrait Videos using Biological Signals[J]. arXiv: 1901.02212, 2019.

[25] LI Y,CHANG M,LYU S. In ictu oculi:Exposing ai created fake videos by detecting eye blinking [C]// WIFS. 2018.

[26] LI L,BAO J,ZHANG T,et al. Face X-Rayfor More General Face Forgery Detection[C]// CVPR,2020.

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

[28] LI J C,LIU B B,HU Y J,et al. Deepfake Video Detection Based on Consistency of Illumination Direction[J]. Journal of Nanjing University of Aeronautics & Astronautics,2020,52(5):90-97.

[29] HU Y J,GAO Y F,LIU B B. Deepfake Videos Detection Based on Image Segmentation with Deep Neural Networks[J]. Journal of Electronics & Information Technology, 2021, 43(1): 162-170.

[30] MÄNTTÄRI J,BROOMÉ S,FOLKESSON J,et al. Interpreting video features:a comparison of 3D convolutional networks and convolutional LSTM networks[J]. arXiv:2002.00367,2020.

[31] ZHANG K,ZHANG Z,LI Z,et al. Joint face detection and alignment using multitask cascaded convolutional networks

[C]// IEEE Signal Processing Letters. 2016: 1499-1503.
[32] CARREIRA J, ZISSERMAN A. Quo Vadis, action recognition? a new model and the kinetics dataset[C]// proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 6299-6308.
[33] TRAN D, WANG H, TORRESANI L, et al. A closer look at spatio-temporal convolutions for action recognition[C]// Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. 2018: 6450-6459.
[34] HARA K, KATAOKA H, SATOH Y. Learning spatio-temporal features with 3d residual networks for action recognition[C]// Proceedings of the IEEE International Conference on Computer Vision Workshops. 2017: 3154-3160.
[35] LI Y, YANG X, SUN P, et al. Celeb-DF: A Large-Scale Challenging Dataset for DeepFake Forensics[C]// 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2020.

XING Hao, born in 1994, master. His main research interests include computer vision and artificial intelligence.

LI Ming, born in 1982, Ph.D, professor, Ph.D supervisor. His main research interests include computer vision and artificial intelligence.