计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com221100109
DOI:10.11896/jsjkx.221100109

基于非关键掩码和注意力机制的深度伪造人脸篡改视频检测方法

俞 洋 袁家斌 蔡纪元 查可可 陈章屿 戴加威 冯煜翔
南京航空航天大学计算机科学与技术学院 南京 211106
(yu_yang@nuaa.edu.cn)

摘 要 自深度伪造技术(Deepfake)被提出以来,其非法应用对个人、社会、国家安全造成了恶劣影响,存在巨大隐患,因此针对人脸视频的深度伪造检测是计算机视觉领域中的热点及难点问题。针对上述问题,提出了一种基于非关键掩码和CA_S3D模型的深度伪造视频检测方法。该方法首先将人脸图像划分为关键区域和非关键区域,通过对非关键区域掩码的处理,提高了深度神经网络对人脸图像关键区域的关注程度,减少了无关信息对深度神经网络的影响和干扰;接着在S3D网络中引入上下文注意力模块,增强了对样本数据信息长程依赖的捕获能力,提高了对关键通道和特征的关注程度。实验结果表明,该方法在DFDC数据集上得到了明显的性能提升,准确率从83.85%提升到了90.10%,AUC值从0.931提升到了0.979;同时与现有的深度伪造视频检测方法进行了对比,所提方法的表现优于现有方法,验证了该方法的有效性。

关键词:深度伪造;Deepfake检测;图像掩码;三维卷积网络;注意力机制
中图法分类号 TP391.41

Deepfake Face Tampering Video Detection Method Based on Non-critical Masks and Attention Mechanism

YU Yang, YUAN Jiabin, CAI Jiyuan, ZHA Keke, CHEN Zhangyu, DAI Jiawei and FENG Yuxiang
College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China

Abstract Since the introduction of Deepfake technology, its illegal application has caused a bad impact on individuals, society and national security, and there are huge hidden dangers. Therefore, deep fake detection for face video is a hot and difficult problem in the field of computer vision. In view of the above problems, this paper proposes a deepfake video detection method based on non-critical mask and CA_S3D Model. It firstly divides the face image into key areas and non-critical regions, and improves the attention of the deep neural network to the key areas of the face image through the mask processing of the non-critical areas, and reduces the influence and interference of irrelevant information on the deep neural network. Then it introduces the contextual attention module in the S3D network, which enhances the ability to capture the long-range dependence of sample data information and improves the attention to key channels and features. Experimental results show that the proposed method improves the performance of the deep neural network on the DFDC dataset, the accuracy rate increases from 83.85% to 90.10%, and the AUC value increases from 0.931 to 0.979. By comparing with the existing deepfake video detection methods, the performance of the proposed method is better than that of the existing methods, which verifies its effectiveness.

Keywords Deepfake, Deepfake detection, Image mask, 3D CNNs, Mechanism of attention

1 引言

随着数字数据的快速增长以及深度学习技术的飞速发展,大量深度伪造(Deepfake)技术所产生的包括面部信息的虚假图像和视频在互联网上传播,这已经成为公众非常关注的问题,特别是使用Deepfake方法制作的伪造视频。“Deepfake”这一术语是指一种基于深度学习的技术,它能够通过将一个人的脸换成另一个人的脸来创建虚假视频。在深度学习盛行之前,由于缺乏便捷的视频编辑工具,视频编辑需要专业领域的相关知识,并且所涉及的过程复杂耗时,因此面部修改作品的规模和危害没有达到引起大家重视的程度。然而,随着深度学习在过去几年里的迅速发展,自动合成现实中不存在的面孔或在图像和视频中修改一个人的真实面容变得越来越容易。原因归结于:1)获取大规模公共数据的便利;2)深度学习技术的发展消除了许多人工编辑步骤,如自动编码器(Autoencoder,AE)[1]和生成对抗网络(Generative Adversarial Networks,GAN)[2];3)开放的软件和移动应用程序如ZAO和FaceApp等,使得制作虚假图像和视频的难度进一步降低。

由于深度伪造视频的制作门槛低,仿真度高且欺骗性强,因此该技术可能被滥用。就个人而言,包含其人脸的伪造

到稿日期:2022-11-14 返修日期:2023-03-09
基金项目:国家自然科学基金(62076127)
This work was supported by the National Natural Science Foundation of China(62076127).
通信作者:袁家斌(jbyuan@nuaa.edu.cn)

视频在互联网上传播可能侵犯其名誉和隐私；就社会而言，深度伪造技术可通过制造虚假新闻引起不同程度的社会混乱；就国家而言，若深度伪造技术被用于制造政治矛盾，传播极端思想或煽动不安情绪，将会对国家安全造成巨大威胁。此外，深度伪造技术还可能被不正当使用以牟取私利，比如伪造图像或视频来进行诈骗或敲诈他人等。因此，对深度伪造的检测技术进行研究是有必要的。

目前已有一些针对深度伪造的检测技术的研究，比如有些检测技术[3-5]以独立的视频帧为检测目标，但并没有利用视频拥有时序信息的特性；有些检测技术利用深度生成网络所生成数据的特征来进行检测，但目前出现了消除这些特征的方法，在面对经过特征消除的伪造视频时，这些技术的检测效果出现了下降。

本文提出了一种基于非关键掩码和注意力机制的深度伪造视频检测方法，主要贡献如下：

(1)提出了一种基于人脸关键点的非关键掩码方法，该方法通过对非关键区域掩码的处理，使得检测网络更加关注涉及伪造的部分；

(2)利用上下文注意力机制对 S3D 网络进行了改进，提出了 CA_S3D 网络(Channel Attentive Separatable 3D Net)，使模型更加关注于有价值的通道信息和特征，提高了对深度伪造视频的检测性能；

(3)在公开的深度伪造数据集 Deepfake 检测挑战赛(Deepfake Detection Challenge，DFDC)上进行了训练和测试，所提方法取得了 AUC 值为 0.979 的结果，优于大多数主流检测方法，并且对深度神经网络进行了可解释性分析，结果验证了所提方法确实能够使网络更聚焦于伪造区域。

2　相关工作

2.1　Deepfake 生成方法

Deepfake 生成人脸的方案主要有两种：变分自动编码器(VAE)和生成对抗网络(GAN)。

基于 VAE 的方案使用了两个编码器-解码器对，通常使用编码器从图像中提取人脸的潜在特征，然后使用解码器重建人脸。为了实现原始图像和目标图像之间人脸的交换，需要两对编码器和解码器，其中编码器先后在原始图像和目标图像上训练，而解码器则分别在原始图像和目标图像上单独训练。训练完成后，交换两组解码器，使用原始图像的原始编码器组合目标图像的解码器，重新生成具有原始图像特征的目标图像。生成的图像能达到给目标图像换上原始图像面容的目的，同时保持目标图像的面部表情。图 1 是一个深度伪造的例子，其中人脸 A 的潜在表示与解码器 B 连接，从原始的人脸 A 重建人脸 B。使用深度神经网络的人脸交换技术有许多公开的实现，如 FaceSwap[6]，DFaker[7] 和 DeepFaceLab[8] 等。

最近，大量研究都集中在使用基于 GANs 的模型[9-10]上。Korshunova 等[11]提出了一种基于卷积神经网络的方法，该方法接受输入图像的面部姿势、面部表情和照明条件等语义内容，然后在另一幅图像中重建这些内容。而文献[9]则提出了一种使用 GAN 的改进深度伪造方法，该方法在自动编码器架构[3]实现的 VGGface 中增加了对抗损失和感知损失，这使得生成图像中眼球的方向与输入图像更加一致，也使得伪造过程中出现的人工痕迹更加平滑，从而得到高质量的输出视频。与现有的自动编码器-解码器方法不同，这些基于 GAN 的方法在工作时无须对特定的图像进行显式训练。

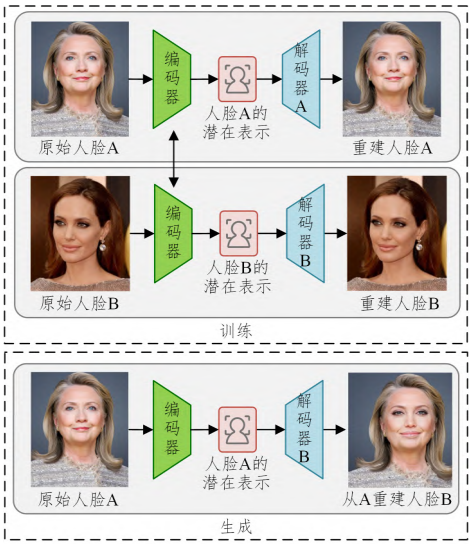

图 1　VAE 人脸伪造过程

Fig. 1　Process of face manipulation

2.2　Deepfake 检测方法

现有的检测方法主要分为两种，一种利用伪造视频在生成过程中遗留的时空篡改伪影来进行检测，另一种则利用数据驱动来进行分类。空间篡改伪影包括不一致性[12]、背景异常[13]和 GAN 指纹[14]。时间篡改伪影包括检测人的行为变化[15]、生理信号[16]和视频帧同步[17]。与之不同的是，数据驱动的方法往往通过分类[18]或异常识别[19]来检测操作。在检测的一般流程中，在特征提取环节，所有的深度伪造检测方法都采用了基于手工特征的方法或基于深度学习的方法。

在基于手工特征的 Deepfake 检测技术方面，一些研究者[6]采用了利用传统的图像伪造识别方法进行深度伪造检测的思路。Zhang 等[4]利用 SURF 算子对图像进行特征提取，然后利用提取的特征训练支持向量机进行分类，最后在高斯模糊图像集上对该技术进行了测试。Korshunov 等[5]使用图像质量度量特征以及主成分分析和线性判别分析进行特征提取，然后训练支持向量机将视频内容分类为真假。他们还指出，现有的人脸识别技术如 Facenet[20]和视觉几何组(Visual Geometry Group，VGG)[21]无法检测深度伪造。此外，单纯的基于口型同步的算法无法检测 GAN 生成的视频；SVM 分类器具有较好的深度伪造检测性能，然而它们在高质量的视觉内容上表现不佳。

基于手工特征的方法通常可以很好地检测静态数字图像中的变化，但对于深度伪造视频可能表现不好，一个原因是其忽视了视频帧与帧之间的时序特征具有连续性，另一个原因是压缩技术的使用会丢弃非常重要的视觉信息。为了克服这些问题，基于深度学习的方法正在受到广泛研究。Guera 等[17]使用 CNNs 在视频帧级提取特征，然后在提取的特征集上训练 RNN，在视频层面检测深度伪造。该工作实现了良好的检测性能，但得到的模型只具有短期记忆，仅适用于 2 s 或更短的视频。Nguyen 等[22]提出了一种多任务的 CNNs

模型，其可以同时从视频中检测和定位被篡改的内容，但是在未知的场景下，模型的评估精度会下降。Amerini 等[23]提出了一种基于光流的方法来检测数字视频中的伪造人脸，其使用 PWC-Net[24]计算每个视频帧的光流[25]，然后利用光流训练 VGG16 和 ResNet50 对真假内容进行分类。该方法具有较好的深度伪造检测性能，但是存在实时性不够的问题。

3 基于非关键掩码和 CA_S3D 模型的深度伪造视频检测方法

3.1 方法思路

Luo 等[26]观察到深度神经网络在给出准确预测的同时可能对图像中部分信息投入了不必要的关注度，由于深度伪造中图像被操作的区域只在面部区域周围，因此他们假设往模型中输入人脸图像上的小块区域就足够了，而不用输入整个图像信息。这种做法在保障模型效果的情况下，显著地缩减了输入数据的规模。受此项工作的启发，本文提出了如下假设：如果使用完整的图像，模型可能学到很多并非与视频人脸伪造确切相关的特征，因此可以将人脸各个标志（如五官等）专门提取出来用于训练。可以认为通过上述的特征处理，人为地为模型引入了更多的先验知识，同时删除了干扰信息。基于此假设，本文提出了作用于人脸图像的非关键掩码方法，后续章节中会详细介绍此方法。

除此之外，本文结合人类辨别深度伪造视频的特征及方法：通常从图像缺陷和人脸动作前后不一致两个方面进行分析，即视频单帧图像的空间信息和视频多帧之间的时序信息。鉴于此，本文选择了可以同时有效地学习到深度伪造视频中的空间信息和时序信息的三维卷积神经网络（3D CNNs）。另外，为了使网络模型能够更好地关注人脸图像中涉及篡改的区域，本文使用了上下文注意力机制对其进行了改进。

本文提出的整体检测框架如图 2 所示。

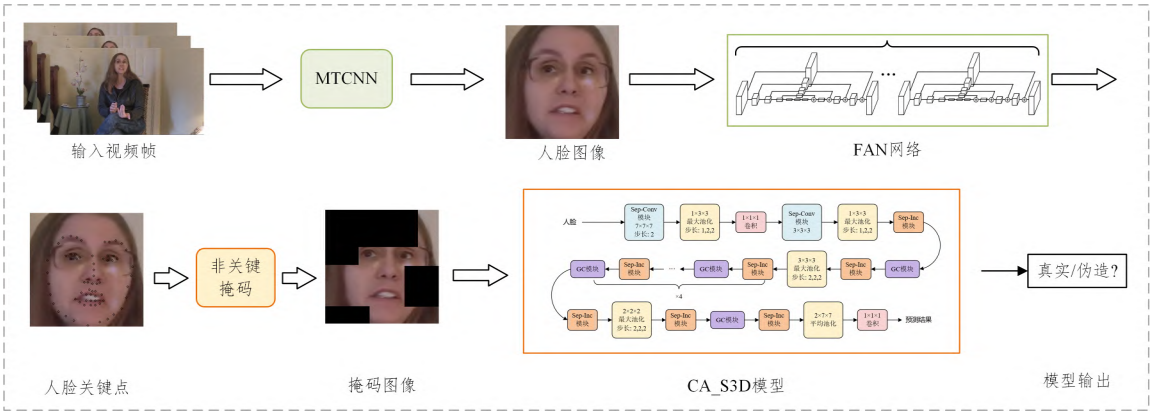

图 2 整体检测框架

Fig. 2 Architecture of the detection model

3.2 非关键掩码

基于 3.1 节中的思路，本文设计了基于人脸关键点检测的非关键掩码（Non-critical Masks）。

人脸关键点（Face Landmark）是人脸各个部位的重要特征点，通常是轮廓点与角点，其中最关键的 5 个点分别为左右两个嘴角、两个眼的中心以及鼻子。人脸关键点能够反映各个部位的脸部特征。随着技术的发展和对精度要求的提高，人脸关键点的数量经历了从最初的 5 个点到如今超过 200 个点的发展历程。本文选择了现今最通用的 68 点标注方案，具体内容如图 3 所示，其中各点的坐标依次记为 $(H_i, W_i)(i=0,1,\cdots,66,67)$，而完整人脸图像的像素高度记为 H_p，像素宽度记为 W_p。

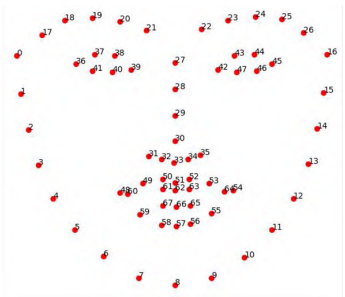

图 3 人脸关键点

Fig. 3 Face landmarks

通过对深度伪造视频的观察，可以看到深度伪造视频中人脸篡改基本集中于五官区域，而图像中的其余区域则没有明显改动，具体样例如图 4 所示。鉴于此观察结果，结合 3.1 节中的假设思路，本文将人脸区域划分为 4 个关键区域和 8 个非关键区域。其中 4 个人脸关键区域分别为眼睛、鼻子和嘴巴所在的区域；在考虑几何规整、灵活性和最终效果等因素后，将其余区域划分为 8 个非关键区域。

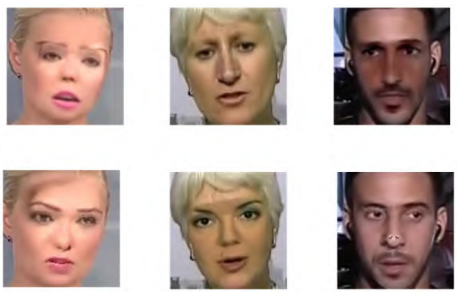

图 4 深度伪造视频中的人脸篡改示例

Fig. 4 Examples of face tampering in deep fake videos

如图 2 所示，对视频帧的人脸图像进行人脸关键点检测后，得到 68 个人脸关键点，利用这些关键点将图片分为 4 个人脸关键区域和 8 个非关键区域。具体地，非关键区域的边界各点坐标如表 1 所列。在划分人脸关键区域时，区域的上、下、左、右各扩大了 10%，最终结果如图 5 所示。

表 1　非关键区域边界各点坐标

Table 1　Coordinates of points on the boundary of non-critical areas

区域	左上角坐标	右上角坐标	左下角坐标	右下角坐标
区域 1	$(0,0)$	$(0,W_{36})$	$(H_{41},0)$	(H_{41},W_{36})
区域 2	$(0,W_{36})$	$(0,W_{45})$	(H_{37},W_{36})	(H_{44},W_{45})
区域 3	$(0,W_{45})$	$(0,W_p)$	(H_{46},W_{45})	(H_{46},W_p)
区域 4	$(H_{41},0)$	(H_{41},W_{48})	$(H_{57},0)$	(H_{57},W_{48})
区域 5	(H_{46},W_{54})	(H_{46},W_p)	(H_{57},W_{54})	(H_{57},W_p)
区域 6	$(H_{57},0)$	(H_{57},W_{48})	$(H_p,0)$	(H_p,W_{48})
区域 7	(H_{57},W_{48})	(H_{57},W_{54})	(H_p,W_{48})	(H_p,W_{54})
区域 8	(H_{57},W_{54})	(H_{57},W_p)	(H_p,W_{54})	(H_p,W_p)

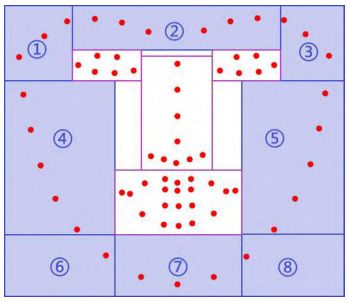

图 5　人脸关键点与区域

Fig. 5　Face landmarks and regions

之后，在人脸图像的非关键区域中随机选择数个区域用黑色填充，进行掩码处理。随后，将掩码图像作为输入数据送入网络模型进行训练。进行了掩码处理的图像可以让模型更加关注被篡改的五官区域，而不是无关的其他非关键区域。为了尽可能地保留视频帧之间的时序信息，对每个视频中的所有视频帧采用相同的随机区域，而不同视频之间的随机区域则不尽相同。

3.3　CA_S3D 网络模型

视频理解领域中的 S3D 网络[27]在视频识别、动作识别等领域表现优秀。S3D 网络是在 I3D[28]网络的基础上改进而来，由三维卷积块和 Inception[29]模块组成，并用最大池化层进行信息汇集。S3D 网络将深度可分离卷积的思想应用到了 I3D 网络中，把 I3D 网络中的三维卷积块替换成了时空可分离卷积，是更加轻量化的 I3D 网络，在减少了网络参数和计算量的同时，提升了准确率。

原 S3D 网络是针对视频理解等问题而设计的，主要考虑视频的整体特征，而在面对深度伪造视频检测问题时则容易忽略细粒度的特征，出现关注非关键特征的问题。因此，本文采用上下文注意力网络 GCNet[30]对原网络进行了改进，提出了 CA_S3D 网络模型。

通过对已有上下文注意力网络的分析，将网络抽象成了全局上下文建模框架。该框架由上下文建模模块、变换模块和融合模块组成，其定义公式为：

$$z_i = F(x_i, \delta(\sum_{j=1}^{N_p} \alpha_j x_j)) \tag{1}$$

其中，i 为查询元素的编号，j 为列举所有可能元素的编号，N_p 为特征图的元素数目（比如，对于图像，$N_p = H \cdot W$，对于视频，$N_p = H \cdot W \cdot T$），x 为输入元素，z 为输出元素；$\sum_j \alpha_j x_j$ 为上下文编码模块，它可以通过利用权重为 α_j 的加权平均将所有元素的特征汇聚起来；$\delta(\cdot)$ 为特征变换模块，它可以捕获通道级别的特征依赖关系；$F(\cdot,\cdot)$ 为融合模块，它可以将全局上下文特征汇聚到每个元素的特征当中。

在此基础上，设计并实现了如图 6 所示的具体网络结构，其定义公式为：

$$z_i = x_i + W_{v2}\,\mathrm{ReLU}\left(\mathrm{LN}\left(W_{v1}\sum_{j=1}^{N_p}\frac{e^{W_k x_j}}{\sum_{m=1}^{N_p} e^{W_k x_m}} x_j\right)\right) \tag{2}$$

其中，W_{v1}，W_{v2} 和 W_k 为线性变换矩阵，e 为自然底数，LN 为 Layer Normalization 层，ReLU 为 ReLU 激活函数；$\alpha_j = \frac{e^{W_k x_j}}{\sum_m e^{W_k x_m}}$ 为上下文编码模块中汇聚全局注意力的权重，$\delta(\cdot) = W_{v2}\,\mathrm{ReLU}(\mathrm{LN}(W_{v1}(\cdot)))$ 为特征变换模块。

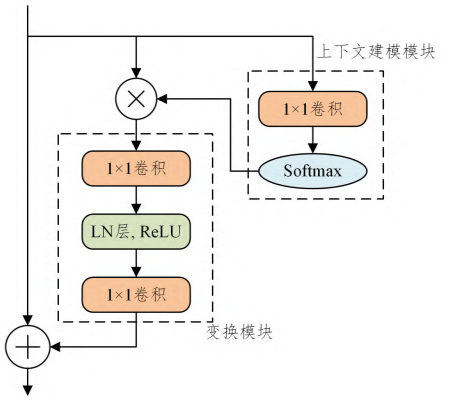

图 6　GC 模块结构图

Fig. 6　Architecture of GCNet

本文在原网络的 Inception 模块后加入了上下文注意力层，该层通过捕捉全局上下文信息的长程依赖，向网络中加入了注意力机制，为特征图中的不同通道赋予了不同的权重，使得网络在训练过程中更加关注那些重要的通道，从而关注对预测结果影响更大的特征，提高最终的性能。CA_S3D 网络模型的具体结构如图 7 所示。

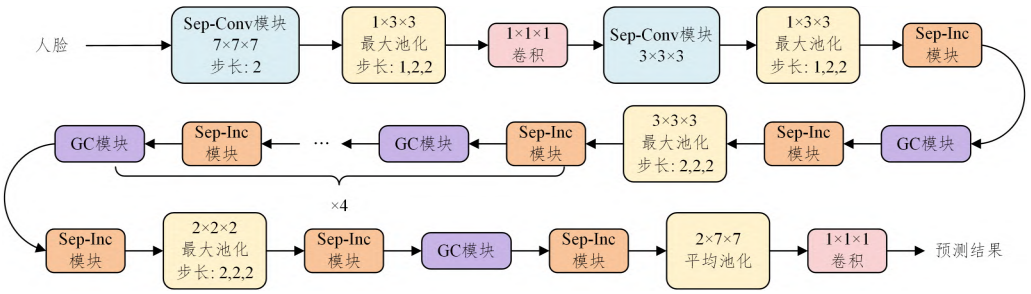

图 7　CA_S3D 网络架构图

Fig. 7　Architecture of CA_S3D

4 实验结果与分析

4.1 数据集介绍

DFDC 数据集是 Facebook 与微软、亚马逊和麻省理工学院等其他公司和学术机构合作发布的最新公共数据集之一，包括 119 197 个视频，时长均为 10 s，帧率分布为 15 fps 到 30 fps，分辨率分布为 320×240 到 3 840×2 160。其中有 19 197个视频是由大约 430 名演员拍摄的真实片段，剩余 100 000个视频则是由真实视频生成的假脸视频。值得注意的是，与其他流行的数据集不同，这个数据集没有使用公开的数据或来自社交媒体网站的数据，并且伪造视频是用两种不同的未知方法制作出来的。另外，DFDC 数据库考虑了不同的采集场景（室内和室外）、光照条件（白天和夜晚等）、人到相机的距离以及姿势变化等条件，是一个更复杂、更接近真实视频的数据集。

本文从 DFDC 数据集中随机选择了 700 个真实视频和 700 个伪造视频，并按照 5∶1∶1 的比例将其划分为训练集、验证集和测试集，即训练集包含 500 个真实视频和 500 个伪造视频，验证集和测试集则均包含 100 个真实视频和 100 个伪造视频。

4.2 实验设置

预处理：本文使用多任务级联卷积网络（Multi Task Cascaded Convolutional Networks，MTCNN）作为人脸检测器，从视频帧中提取人脸区域坐标，截取人脸图像（该方法在文献[31-33]中均有使用），然后将人脸图像的像素尺寸统一为 224×224。为了验证所提出方法的效果，本文并没有采取数据增强方法。

环境设置：本文实验基于 Pytorch 深度学习框架搭建，在 2 块 NVIDIA GeForce RTX 3060 12 GB 显卡上训练。

训练参数：对于每一个视频，本文以 10 帧的间隔共采样 20 帧视频帧作为输入数据。GC 模块中压缩率 $r=1/16$。模型的 batchsize 值为 24，损失函数为 BCEWithLogitsLoss，最大训练 epoch 值为 400，初始学习率为 0.002。模型采用 Adam 算法作为优化器，权重衰减值为 0.000 000 1，实验中还采用了余弦学习率衰减算法[34]来调整学习率。学习率调度算法如式(3)所示：

$$\eta_t=\frac{1}{2}\left(1+\cos\left(\frac{t\pi}{T}\right)\right)\eta \tag{3}$$

其中，t 为当前 epoch 数，T 为总 epoch 数，η 为初始学习率，η_t 为对应当前 epoch 数的学习率。

评价指标：由于深度伪造视频检测问题是一个二分类问题，本文采用 sigmoid 函数作为最后预测结果的映射函数，分类结果判别阈值设置为 0.55。此外，本文采用准确率（Accuracy，Acc）和 AUC 指标来评估模型的分类效果，采用二分类交叉熵损失函数（Binary Cross Entropy Loss，BCELoss）来评估模型预测结果与实际标签之间的差距。

准确率的计算公式如式(4)所示：

$$Acc=\frac{TP+TN}{TP+TN+FP+FN} \tag{4}$$

其中，TP 为真正样本数，即被预测为正的正样本数；TN 为真负样本数，即被预测为负的负样本数；FP 为假正样本数，即被预测为正的负样本数；FN 为假负样本数，即被预测为负的正样本数。

AUC 指标的全称是 Area Under Curve，即 ROC 曲线与坐标轴形成的面积，取值范围为[0，1]，取值越接近 1 说明分类器的效果越好。ROC 曲线全称为受试者工作特征曲线（Receiver Operating Characteristic Curve），它是根据一系列不同的二分类方式（分界值或决定阈），以真正率（True Positive Rate，TPR）为纵坐标、假正率（False Positive Rate，FPR）为横坐标绘制的曲线。

二分类交叉熵损失函数的计算公式如式(5)所示：

$$l_n=-[y_n\cdot\log x_n+(1-y_n)\cdot\log(1-x_n)] \tag{5}$$

其中，n 为样本编号，x_n 为模型的预测结果，y_n 为样本实际标签。

4.3 结果分析

4.3.1 实验结果

为了验证所提出的深度伪造视频检测方法的效果，本文进行了针对性的消融实验，通过比较各种方案进行训练来考查模型的性能，验证了所提方法的效果。同时，通过选择不同的非关键掩码区域的个数进行实验，得到了性能最佳的掩码区域数目，具体地说，使用的区域个数分别为 6，8，4 和 2。实验结果如表 2 所列，最优结果用下划线标注。图 8 给出了各个方案的 ROC 曲线。

表 2 不同方案的性能对比

Table 2 Performance comparison of different schemes

Model	*Acc*/%	*AUC*	*BCELoss*
S3D	83.85	0.931	0.575
CA_S3D	89.58	0.965	0.545
CA_S3D+m6	<u>90.10</u>	0.968	0.544
CA_S3D+m8	47.92	0.494	0.732
CA_S3D+m4	88.54	<u>0.979</u>	<u>0.536</u>
CA_S3D+m2	89.06	0.973	0.544

从表 2 所列的实验结果可知，改进后的 CA_S3D 网络与原网络相比，*Acc* 从 83.85%提升到了 89.58%，*AUC* 值从 0.931 提升到了 0.965，*BCELoss* 从 0.575 提升到了 0.545，表明 CA_S3D 网络具有更好的分类效果，其预测结果与样本的实际标签差距更小。同时，选择使用合适的非关键掩码方案可以使模型的性能得到明显的提升，其中选择 4 个掩码区域的方案为最优，与未使用非关键掩码的 CA_S3D 模型相比，*AUC* 值从 0.965 提升到了 0.979，*BCELoss* 从 0.545 提升到了 0.536。另外，从非关键掩码不同方案的比较可以看出，选择过多的掩码区域个数会导致模型的性能出现明显的下降，其中选择全部 8 个掩码区域的方案性能下降得最为明显，仅仅拥有 47.92%的 *Acc* 和 0.494 的 *AUC* 值，而 *BCELoss* 也下降到了 0.732。关于这部分结果，本文认为可能是在过多的区域进行掩码操作，使得人脸图像丢失了过多的信息，导致模型学习到的特征不够丰富，因此模型的泛化能力出现了明显下降。

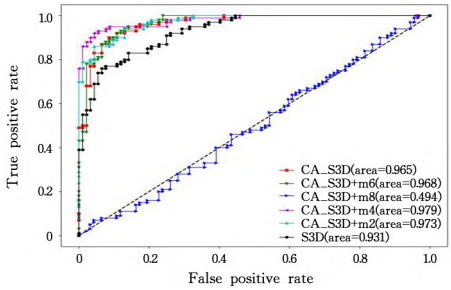

图 8　不同方案的 ROC 图像

Fig. 8　ROC images of different schemes

4.3.2　网络模型参数量对比

本文还将所提出的网络模型与其他主流方法进行了参数量对比，其结果如表 3 所列。比较结果说明，本文提出的网络模型具有更少的参数，因此在处理视频数据时计算量更小。

表 3　本文模型与其他模型的参数量对比

Table 3　Comparison of parameters between the proposed model and other models

Model	Number of parameters
X-ception[35]	20.8×10^6
I3D[28]	12.3×10^6
R3D[36]	33.2×10^6
ip-CSN-152[37]	32.2×10^6
ResNet+MS-TCN[38]	24.8×10^6
CA_S3D(Ours)	8.1×10^6

4.3.3　鲁棒性分析

为了进一步分析所提方法的鲁棒性，对测试集数据进行了模糊、添加噪声和压缩这 3 种处理，其中噪声又分为高斯噪声和传感器噪声，结果如表 4 所列，使用准确率 Acc 作为评估标准。实验结果表明，视频中可能存在的不确定性因素对本文方法确实有所影响，但影响并不大，其中一些方案甚至在不确定性因素的影响下具有更好的性能，这可能是因为不确定性因素破坏了非关键特征而保留了关键特征。由此可见，本文方法对于检测视频中可能存在的干扰性因素有不错的鲁棒性。

表 4　本文方法的鲁棒性分析(Acc)

Table 4　Robustness analysis of the proposed method(Acc)

(单位：%)

Model	Blur	GaussianNoise	Compression	ISONoise
CA_S3D	89.06	89.58	89.58	91.15
CA_S3D+m6	88.02	89.58	90.10	88.02
CA_S3D+m4	90.63	89.06	89.06	88.02
CA_S3D+m2	88.54	90.63	89.58	89.06

4.3.4　可解释性分析

本文还使用了 Captum[39] 工具对所提方法进行了可解释性分析，具体地说，采用了遮挡可解释性分析的方法。利用 Captum 工具，可以通过小滑块滑动遮挡图像上不同区域的方法，观察哪些区域被遮挡后会显著地影响模型的分类决策，由此可以分析得到模型在进行分类预测时会更加关注输入数据的哪些区域。

结果如图 9 所示。其中每个方案对应的左侧图像为遮掩区域分析结果，阴影区域代表该区域被遮掩时会对模型的预测结果产生影响，阴影区域颜色越深则表示影响越大；右侧图像为原图和结果图重叠生成的图像。从结果可以看出，当未使用所提方法时，模型关注的区域较为分散，且有些区域为非关键区域；而使用所提方法后，模型关注的区域会逐渐转移到人脸的关键特征等更可能包含伪造伪影的区域。可以看到，其中非关键掩码方法可以在 CA_S3D 模型的基础上使模型更加关注重要区域。

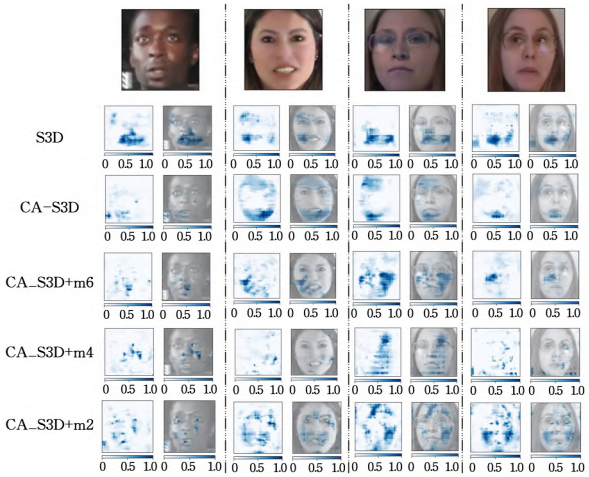

图 9　不同方案中模型的遮挡可解释性分析结果

Fig. 9　Occlusion interpretability analysis results for models in different schemes

4.3.5　与其他方法的对比

为了进一步说明所提方法的有效性，本文选取其他较为前沿的深度伪造视频检测方法进行了比较。由于仅使用准确率作为评价指标受测试集中数据分布的影响较大，为了保证比较结果的科学性和公平性，此部分选择使用不易受数据分布影响的 AUC 来作为结果评价指标。

与其他方法的比较结果如表 5 所列，结果说明了本文所提出的深度伪造视频检测方法能够使模型在 DFDC 数据集上取得优秀的性能，优于其他方法。

表 5　本文方法与其他方法的对比

Table 5　Comparison between the proposed method and other methods

Method	Backbone	Train Set	Test Set(*AUC*)
			DFDC
PCL+I2G[40]	ResNet-34	DFDC	0.944
Nehate 等[31]	ViT&EfficientNetV2	DFDC	0.952
Coccomini 等[32]	Conv. Cross ViT	DFDC	0.951
Bondi 等[41]	EfficientNetB4	DFDC	0.922
Our Method	CA_S3D	DFDC	0.979

结束语　针对深度神经网络在预测时存在对数据产生不相关关注从而影响性能的问题，本文提出了一种基于非关键掩码和 CA_S3D 模型的深度伪造视频检测方法。该方法首先通过对人脸图像中的非关键区域进行掩码处理，提高了深度神经网络对更可能包含深度伪造信息的关键区域的关注程度。同时，为了能够更好地利用视频数据中包含的时序信息，本文提出了 CA_S3D 模型，使用上下文注意力模块 GC 模块对 S3D 网络进行改进，使其能更好地捕获样本数据信息的长程依赖，关注重要的通道和特征。实验结果表明，本文提出的

方法在 DFDC 数据集上取得了最佳为 90.10%的准确率、AUC 值为 0.979 和 BCELoss 为 0.536 的不错性能。遮挡可解释性分析的结果也验证了本文提出的非关键掩码的有效性。

虽然本文的方法取得了不错的效果，但是由于其以人脸图像为基础，因此当深度伪造视频中出现侧向人脸或不完整人脸等问题时，其效果会受到影响。未来会尝试将本文方法与其他检测方法进行融合，减少对人脸图像的依赖程度，以进一步提高对深度伪造视频的检测性能。

参 考 文 献

[1] KINGMA D P, WELLING M. Auto-Encoding Variational Bayes [EB/OL]. https://arxiv.org/abs/1312.6114.

[2] GOODFELLOW I J, POUGET-ABADIE J, MIRZA M, et al. Generative Adversarial Nets[C]// Proceedings of the 27th International Conference on Neural Information Processing Systems—Volume 2. Cambridge, MA, USA: MIT Press, 2014: 2672-2680.

[3] AKHTAR Z, DASGUPTA D. A Comparative Evaluation of Local Feature Descriptors for DeepFakes Detection[C]// 2019 IEEE International Symposium on Technologies for Homeland Security(HST 2019). 2019: 1-5.

[4] ZHANG Y, ZHENG L, THING V L L. Automated face swapping and its detection[C]// 2017 IEEE 2nd International Conference on Signal and Image Processing(ICSIP). 2017: 15-19.

[5] KORSHUNOV P, MARCEL S. DeepFakes: a New Threat to Face Recognition? Assessment and Detection[EB/OL]. [2021-08-23]. http://arxiv.org/abs/1812.08685.

[6] Faceswap[CP/OL]. https://github.com/deepfakes/faceswap.

[7] DFaker[CP/OL]. https://github.com/dfaker/df.

[8] DeepFaceLab[CP/OL]. https://github.com/iperov/DeepFaceLab.

[9] Faceswap-GAN[CP/OL]. https://github.com/shaoanlu/faceswap-GAN.

[10] NIRKIN Y, KELLER Y, HASSNER T. FSGAN: Subject Agnostic Face Swapping and Reenactment[C]// 2019 IEEE/CVF International Conference on Computer Vision (ICCV). 2019: 7183-7192.

[11] KORSHUNOVA I, SHI W, DAMBRE J, et al. Fast Face-Swap Using Convolutional Neural Networks[C]// 2017 IEEE International Conference on Computer Vision (ICCV). 2017: 3697-3705.

[12] 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). IEEE, 2019.

[13] 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). 2020: 3204-3213.

[14] YU N, DAVIS L, FRITZ M. Attributing Fake Images to GANs: Learning and Analyzing GAN Fingerprints[C]// 2019 IEEE/CVF International Conference on Computer Vision (ICCV). 2019: 7555-7565.

[15] AGARWAL S, FARID H, GU Y, et al. Protecting World Leaders Against Deep Fakes[C]// CVPR Workshops. 2019: 38-45.

[16] LI Y, CHANG M C, LYU S. In Ictu Oculi: Exposing AI Created Fake Videos by Detecting Eye Blinking[C]// 2018 IEEE International Workshop on Information Forensics and Security (WIFS). 2018: 1-7.

[17] GUERA D, DELP E J. Deepfake Video Detection Using Recurrent Neural Networks[C]// 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). IEEE, 2018.

[18] LI Y, LYU S. Exposing DeepFake Videos By Detecting Face Warping Artifacts[C]// IEEE Conference on Computer Vision and Pattern Recognition Workshops(CVPRW). 2019.

[19] CHUGH K, GUPTA P, DHALL A, et al. Not Made for Each Other-Audio-Visual Dissonance-Based Deepfake Detection and Localization[C]// Proceedings of the 28th ACM International Conference on Multimedia. New York, NY, USA: Association for Computing Machinery, 2020: 439-447.

[20] 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). IEEE, 2015.

[21] O'TOOLE A J, JONATHON PHILLIPS P, JIANG F, et al. Face Recognition Algorithms Surpass Humans Matching Faces Over Changes in Illumination[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2007, 29(9): 1642-1646.

[22] NGUYEN H H, FANG F, YAMAGISHI J, et al. Multi-task Learning for Detecting and Segmenting Manipulated Facial Images and Videos[C]// 2019 IEEE 10th International Conference on Biometrics Theory, Applications and Systems(BTAS). 2019: 1-8.

[23] AMERINI I, GALTERI L, CALDELLI R, et al. Deepfake Video Detection through Optical Flow Based CNN[C]// 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW). IEEE, 2019.

[24] SUN D, YANG X, LIU M Y, et al. PWC-Net: CNNs for Optical Flow Using Pyramid, Warping, and Cost Volume[C]// 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2018: 8934-8943.

[25] ALPARONE L, BARNI M, BARTOLINI F, et al. Regularization of optic flow estimates by means of weighted vector median filtering[J]. IEEE Transactions on Image Processing, 1999, 8(10): 1462-1467.

[26] LUO Z, KAMATA S I, SUN Z. Transformer and Node-Compressed Dnn Based Dual-Path System For Manipulated Face Detection[C]// 2021 IEEE International Conference on Image Processing(ICIP). 2021: 3882-3886.

[27] XIE S, SUN C, HUANG J, et al. Rethinking Spatiotemporal Feature Learning: Speed-Accuracy Trade-Offs in Video Classification[C]// Computer Vision (ECCV 2018): 15th European Conference. Munich, Germany, 2018: 318-335.

[28] CARREIRA J,ZISSERMAN A. Quo Vadis,Action Recognition?A New Model and the Kinetics Dataset[C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017:4724-4733.

[29] SZEGEDY C,LIU W,JIA Y,et al. Going deeper with convolutions[C]// 2015 IEEE Conference on Computer Vision and Pattern Recognition(CVPR). 2015:1-9.

[30] CAO Y,XU J,LIN S,et al. GCNet:Non-local Networks Meet Squeeze-Excitation Networks and Beyond[EB/OL]. [2022-09-28]. http://arxiv.org/abs/1904.11492.

[31] NEHATE C,DALIA P,NAIK S,et al. Exposing DeepFakes using Siamese Training[C]// 2022 IEEE India Council International Subsections Conference(INDISCON). 2022:1-6.

[32] COCCOMINI D A,MESSINA N,GENNARO C,et al. Combining EfficientNet and Vision Transformers for Video Deepfake Detection[C]// Image Analysis and Processing(ICIAP 2022). 2022:219-229.

[33] WANG Z,LI X,NI R,et al. Attention Guided Spatio-Temporal Artifacts Extraction for Deepfake Detection[C]// Pattern Recognition and Computer Vision. 2021:374-386.

[34] HE T,ZHANG Z,ZHANG H,et al. Bag of Tricks for Image Classification with Convolutional Neural Networks[C]// 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). 2019:558-567.

[35] ROSSLER A,COZZOLINO D,VERDOLIVA L,et al. FaceForensics++:Learning to Detect Manipulated Facial Images [C]// 2019 IEEE/CVF International Conference on Computer Vision(ICCV). IEEE,2019.

[36] HARA K,KATAOKA H,SATOH Y. Learning Spatio-Temporal Features with 3D Residual Networks for Action Recognition [C]// 2017 IEEE International Conference on Computer Vision Workshops(ICCVW). 2017:3154-3160.

[37] TRAN D,WANG H,FEISZLI M,et al. Video Classification With Channel-Separated Convolutional Networks[C]// Proceedings of the IEEE International Conference on Computer Vision. 2019:5551-5560.

[38] HALIASSOS A,VOUGIOUKAS K,PETRIDIS S,et al. Lips Don't Lie:A Generalisable and Robust Approach to Face Forgery Detection[C]// 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). 2021:5037-5047.

[39] Captum[CP/OL]. https://captum.ai/.

[40] ZHAO T,XU X,XU M,et al. Learning Self-Consistency for Deepfake Detection[C]// 2021 IEEE/CVF International Conference on Computer Vision(ICCV). 2021:15003-15013.

[41] BONDI L,DANIELE CANNAS E,BESTAGINI P,et al. Training Strategies and Data Augmentations in CNN-based DeepFake Video Detection[C]// 2020 IEEE International Workshop on Information Forensics and Security(WIFS). 2020:1-6.

YU Yang, born in 1995, postgraduate. His main research interests include deep learning and deepfake detection.

YUAN Jiabin, born in 1968, Ph.D, professor, Ph.D supervisor, is a member of China Computer Federation. His main research interests include high-performance computing, quantum computing, deep learning, medical image processing, etc.

（责任编辑:柯颖）