视频换脸鉴伪技术调研

**一 视频换脸技术**

Deepfakes，是一种混合深度学习与造假的合成技术.视频换脸就是deepfakes技术最主要的产物,将一段视频中人脸替换为其他人的肖像,从而生成具有极高欺骗力的视频.

Deepfakes名词起源于2017年底，有用户将《神奇女侠》女主角盖尔·加朵的脸部特征完美移植到成人电影的女演员头上,受到巨大压力的成人交流社区就将上传这部影片的匿名账户“DeepFakes”以侵犯他人隐私的理由封停了！“DeepFakes”对官方的封禁极度不满，于是将制作这部片子的软件源代码，以免费公开的形式发布在了GitHub上.

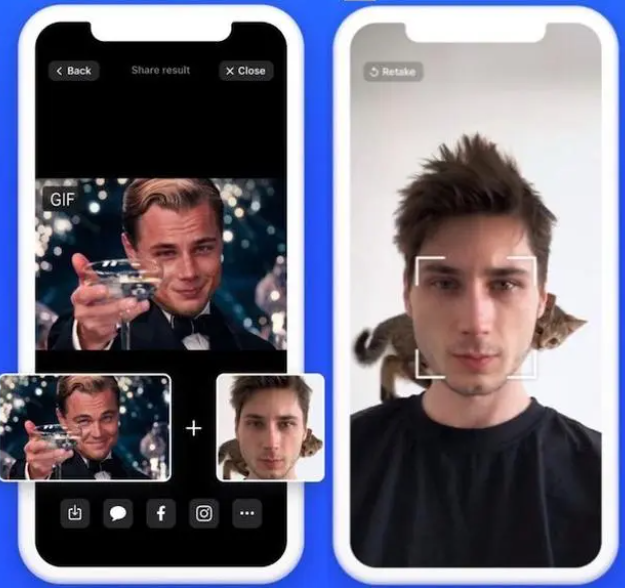
目前视频换脸主要有两大开源产品,[faceswap](https://github.com/deepfakes/faceswap)和DeepFaceLab,star量分别为43.5k和37.3k.

**二 视频换脸产品**

视频换脸是很早就有的技术,最初用于影视工业,如《爵迹2》《速度与激情7》都曾使用换脸技术对演员进行更换.但这项技术极其复杂,需要专业人士花费巨大精力才能完成.

自deepfakes开源后,视频换脸技术就进入高速发展进阶. 2018年1月，就有FakeApp应用程序问世。此应用程序使用户可以轻松创建和共享彼此交换脸部的视频.2019年更有火爆产品 ZAO发布.这款应用冷启动阶段迅速成为现象级产品，

视频换脸技术火爆原因可以归结为以下原因：

1、好玩,新鲜有趣,它激发了互联网用户的猎奇心理。

2、操作简单,相比于用于影视行业的换脸技术,deepfakes极大拉低了准入门槛,上文提到的ZAO应用简单的操作是其火爆的主要原因之一.

正因为视频换脸技术有大量受众,能产生大量互联网流量.所以在内行眼中这项技术简直就是摇钱树.因此,在各方利益推动下,其发展**远快于**对应的鉴伪技术. 同时广泛的用户群体又能为模型迭代提供大量数据支撑,依照现在技术发展趋势,更简单,更真实,人眼无法分辨的视频换脸技术一定会诞生,不过推测不会以app等应用形式出现(ZAO的经历就是一个先例).

有需求就有发展动力,目前AI换脸黑产已经在国内形成了完整的产业链，从下游成品情色视频、中游定制视频到上游软件及教程都有提供，有的卖家甚至提供全产业链产品和售后服务。这项技术在我们看不见的地方影响我们的生活.

**二 视频换脸对社会影响**

除了好玩,猎奇.人脸更担心视频换脸技术对社会产生的不良影响.毕竟deepfakes技术的起源就是负能量开始.比如将嫁公众人物脸型嫁接到到不良视频上，或者政治人物脸型嫁接到重要事件讲话但态度完全相反的视频上，会产生恶劣影响.前者会损坏当然公众形象,将会浪费大量的公共资源去调查视频是否为真.而后者则可能引发暴乱，危及公共安全。这些行为已经发生过,对当事人产生巨大影响.

印度记者Rana Ayyub，因揭露了古吉拉特暴乱中某些黑幕，受到报复.不法分子利用DeepFakes技术，将Rana的脸与色情片女主角的脸调换，并将视频发布到网络大肆传播，在Rana的社交账号下，充满各种不堪入目的骚扰和辱骂,对当事人造成巨大心理伤害。另种,对政治人物的换脸则会造成更严重后果.2018 年，由 Deepfake 合成的加蓬总统 Ali Bongo 新年致辞视频曾引起了兵变.2019年9月播出了一段deepfake意大利前总理马泰奥·伦齐（Matteo Renzi）侮辱其他意大利政客的视频。2022年3月,乌克兰总统被伪造投降视频，在视频中，呼吁乌克兰士兵放下武器.这个视频作为战争状态下**第一个**针对一国元首的Deepfake视频，预示着一种新的舆论战形式。当公众人物遭到deepfake技术攻击时,能调动社会资源去维权,为自己辩护.那么当普通人遇到这类事件该如何维护自身权益呢? 这是deepfake技术潜在风险,鉴于此,整个社会对deepfake技术大都持发对态度.

**二 社会对视频换脸技术处理方式**

各个国家在立法层面已经对deepfake技术进行限制

中国：2020年印发的《法治社会建设实施纲要（2020 - 2025年）》进一步要求，对深度伪造等新技术应用，制定和完善规范管理办法。

美国：2019年正式签署生效的相关法案中包含了和 deepfake 相关的条文，要求政府向立法机构通报涉及跨国、有组织、带有政治目的的 deepfake 虚假信息行为。

欧盟：GDPR、欧盟人工智能框架提议、版权保护框架、虚假信息针对政策等高级别法律文件，都对可能和 deepfake 有关的事务实现了交叉覆盖。在成员国级别上，荷兰立法机构在2020年曾经要求政府制定打击 deepfake 成人视频的政策，以及表示会考虑将相关问题写入该国刑法。

在社交传播领域,很多有影响力的公司也deepfake的发布进行阻断, 如今Google Colab 已把deepfake拉黑,列入禁用列表.2020年，Facebook 也宣布在 Facebook 产品平台全面禁止 deepfake 类视频。Twitter 封杀了一批经常发布 deepfake 视频的账号。

五 关于视频换脸造假技术简介

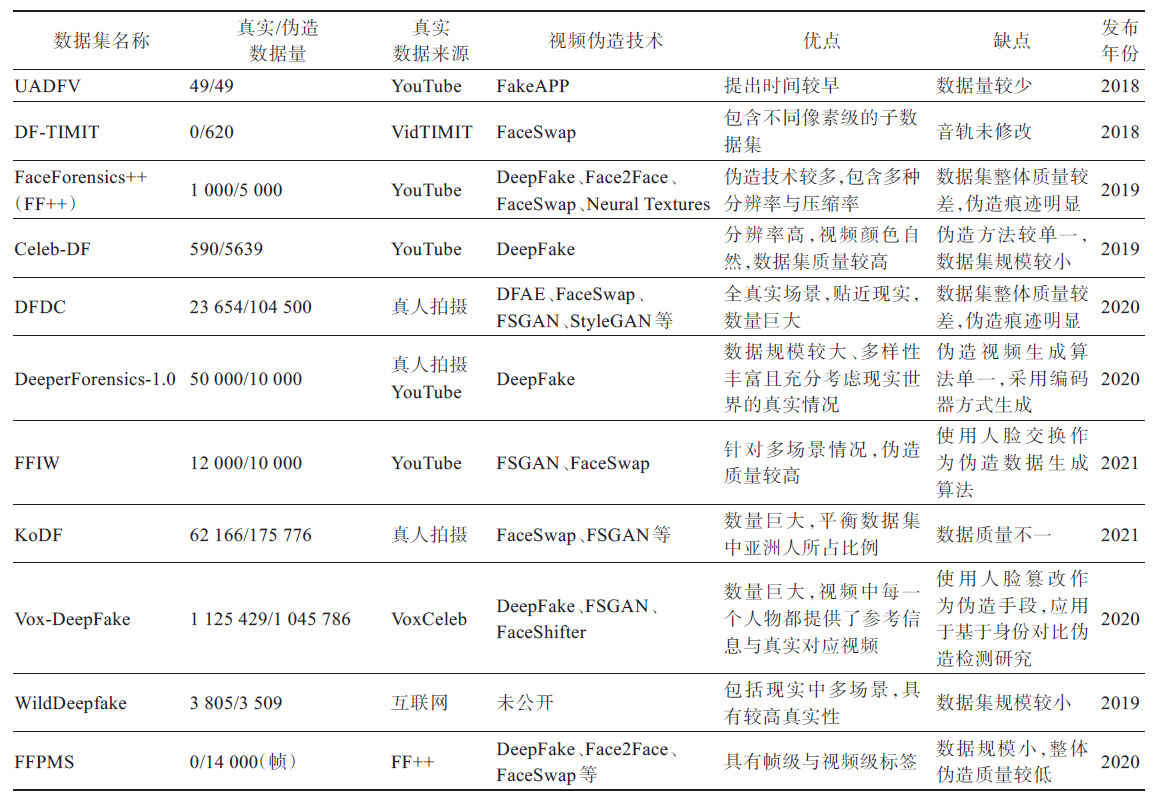
...

**五 视频鉴伪技术**

经过第二章分析,我们可以认识到视频换脸技术对社会的危害.随着deepfake技术发展和因为制作简单并大量传播,仅靠人工筛查很不现实.因此需要一套专门的方式去检查视频是否为deepfake作品.deepfake是基于深度学习的技术,那么与之同源的深度学习方法成为检测deepfake的最好选择.

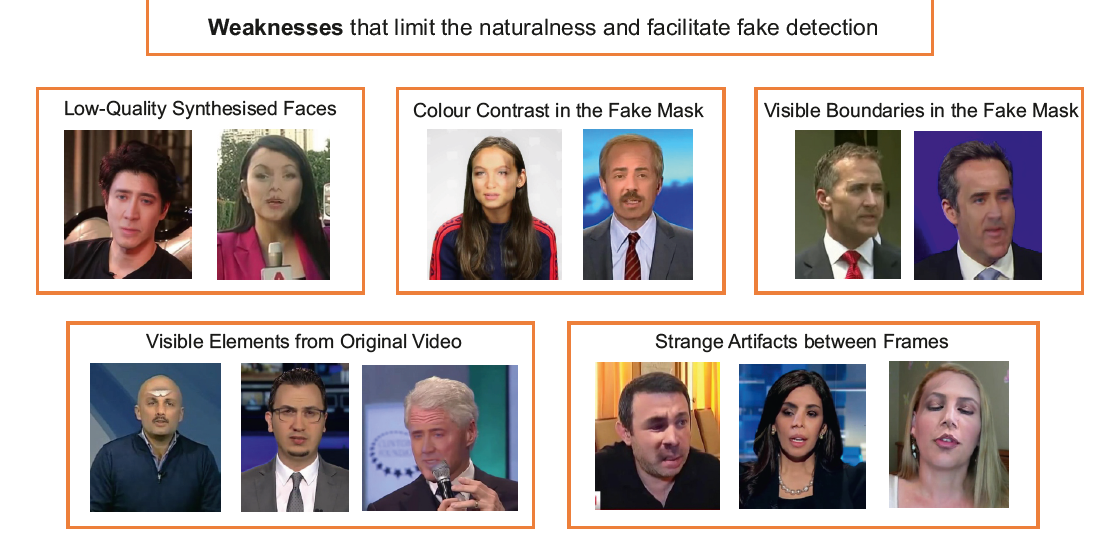
**5.1数据集介绍**

为适应deepfake检测需求,学术界与工业界构建了专门的数据集,以检测deepfake视频.这些数据集的优劣会直接影响实现最终算法模型优劣.现在对这些数据集做个简单介绍.

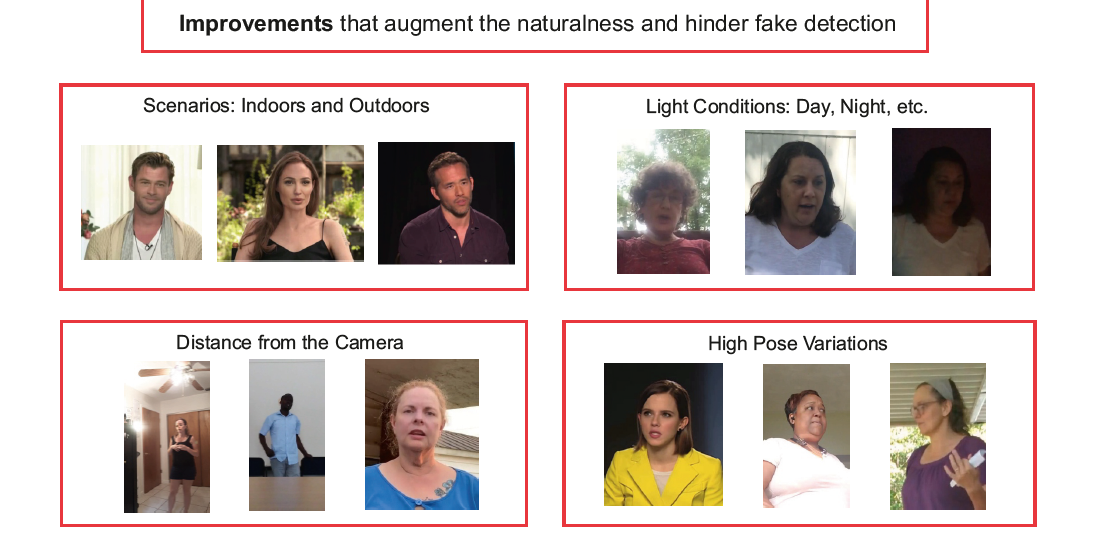


除了Celeb-DF,2019年前建立数据集都被称为第一代数据集,2020年后的是第二代. 区别为第一代数据集压缩率大,图片模糊,伪造痕迹明显.大多数鉴伪算法都能取得较好效果. 第二代数据集在伪造方法,采样质量,场景,数据量都有大幅提升,人眼更难分辨,更贴近生产环境遇到的伪造视频.

上表数据集,DFDC、FaceForencies++、Celeb-DF 为当前领域内研究常用的数据集，其中Celeb-DF 由于伪造质量较高常被用来模拟现实生活中的伪造视频，进而用于测试模型的泛化能力.尽管这两年涌现出大量优秀数据集,但数据集的建立与伪造技术的发展之间仍然存在差距，若要使模型达到工业应用标准，需要再次提升训练数据集数量与质量.



第一代数据集

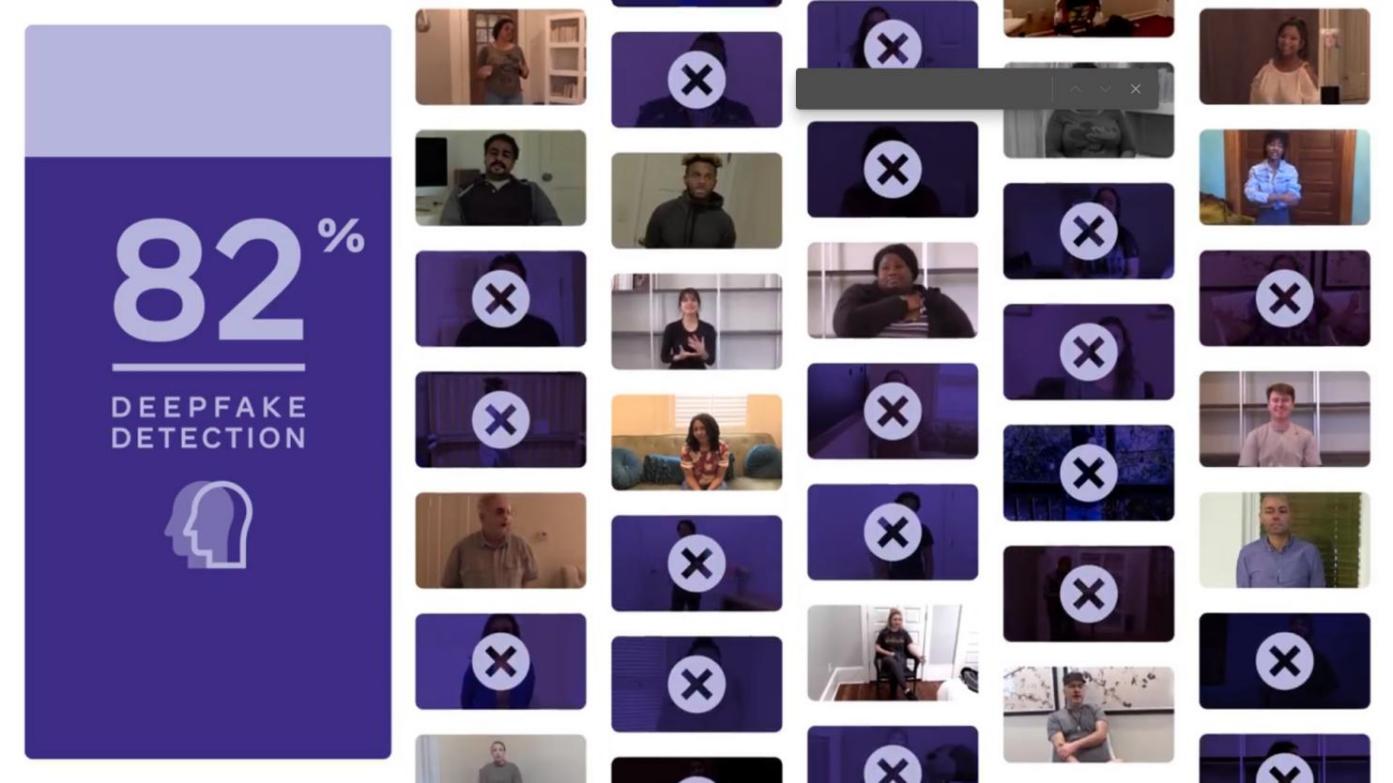


第二代数据集

**5.2 业界SOTA技术介绍**

最近的关于视频换脸鉴伪技术最新动态是ok主导,联合康奈尔，麻省理工学院，牛津大学，加州大学伯克利分校，马里兰大学，大学园区和2020年的Facebo奥尔巴尼 - 纽约州立大学的人工智能，微软和学术界的合作伙伴共同构建Deepfake检测挑战赛（DFDC）(见5.1).

**关于竞赛结果**

****

约有2 114名参赛者向比赛提交了35000多个检测算法，其中最优模型准确率高达82.56%。然而，当这些算法对包含4000多个视频短片的黑盒数据集数据集进行测试时，**得分最好的模型准确率为65.18%,** 结果表明模型泛化能力极差.竞赛获奖者表示,兼容识别新视频是对准确率影响最大的部分，因为新视频可能包括用不同技术交换的不同面孔,Facebook表示不打算在其网站上使用任何获奖模型. 这表明在野外检测深度造假视频是一个非常具有挑战性的问题。

**提升准确率方法**

他认为，提高检测正确率的一个方法就是关注视频帧之间的转换。Seferbekov说:“即使是非常高质量的deepfakes 作品也会在画面之间出现闪烁。”人类很擅长发现这些不一致的地方，尤其是在面部镜头中。但是要自动捕捉这些标志性的缺陷将需要更大、更多样的训练数据和更多的计算能力。Seferbekov尝试过追踪这种帧转换，但最后没有成功。他说：“CPU是真正的瓶颈。”

**五 效果指标**

在人脸视频深度伪造检测方面，常用的指标是ACC 与AUC。

ACC（accuracy）:准确率，正确预测数量占全部测试集数量的比值；

AUC（area under curve）为ROC 曲线所围出图形的面积。ROC称为受试者工作特征曲线，它是根据一系列不同的二分类方式（通常为阈值），以真阳性率为纵坐标，假阳性率为横坐标绘制的曲线。AUC指数通过计算ROC所围图形的面积来衡量分类器学习与分类效果好坏。

**六 关于视频换脸技术最新研究动态**

当前对deepfake的鉴伪技术还没达到可以工业应用的程度.尽管如此,学术界依然对相关技术做着不懈努力.下面挑选最近一些最新的,有代表性的研究成果进行介绍.梳理下研究思路.总结目前最流行的解决问题的思路.只对有较好测试效果的方法才做详细介绍.

6.1 空间特征法

视频本质是图片帧的连续移动,伪造视频也是对单帧图片伪造,然后再处理回连续视频.既然是对图片的伪造,必然对图片空域、频域分布产生扰动，所以基于这类方法的研究思路为提取为局部与整体的不一致性,这种差异可作为模型训练与判别的特征,于是就将视频人脸鉴伪转换为对单帧图片的分类任务.

但是这类方法仅关注单张图片,无法提取视频中相邻帧之间由于伪造所带来的时序上不一致。同时,如果伪造视频质量不好,就会损失大量空域,频域信息,遇到这类情况,这类基于图片本身的方法会变的很差.不过,这类方法简单,易于操作,适用所有数据集研究.所以有大量研究者使用此类方法尝试鉴别deepfake伪造视频.根据研究方式,可分为以下几类:

6.11 基于图片空间信息

① 联合YCbCr 与RGB两种色彩空间,分别提取YCbCr 色彩空间中Cb 和Cr 分量的图像边缘信息，提取RGB 色彩空间中G 分量的图像边缘二阶梯度信息，输入深度网络进行分类.

② 原始视频帧应用于EfficientNet和ViT（vision transformer）及Cross-ViT进行分类.

③ 使用分组卷积替代普通卷积,以提取更丰富特征.

④ 引入胶囊网络，基于动态路由的胶囊网络能有效避免图像平移、扭曲、旋转而带来的误差，同时能够使用更少的训练数据最大化地学习到有效信息.

⑤ 融合人脸图像原始特征与基于块级别隐藏特征的双流网络，在图像空间域特征基础上结合了色彩滤波阵列（color filterarray，CFA）、局部噪声残差这样的低级别相机特征，共同进行训练与分类

⑦ 采用对抗网络同时训练生成器与判别器，用判别器进行视频鉴伪。

⑧ 真实视频图像的局部之间是具有一致性的,换脸后的图片一致性被打破，通过检测图像像素之间的不一致性便可以判断视频真伪.该方法在跨库测试中取得超过90%的准确率，是利用空间域特征进行人脸视频深度伪造检测研究中泛化能力最好的模型之一.下面对这个方法重点介绍:

...

6.12 基于图片频域信息

视频伪造都要经过上采样过程,而上采样之后的图像在频域上与自然图片会有明显差异,基于这种差异作为鉴伪的判定条件.

已知的研究成果有,高频信号祛除了颜色纹理,比低频信号更能够有效区分真实与伪造视频.

分析频域信息基本思路为将图片帧特征图各个特征通道高频与低频信号进行分离后重新组合，再通过卷积与线性池化操作有效提取频率特征.

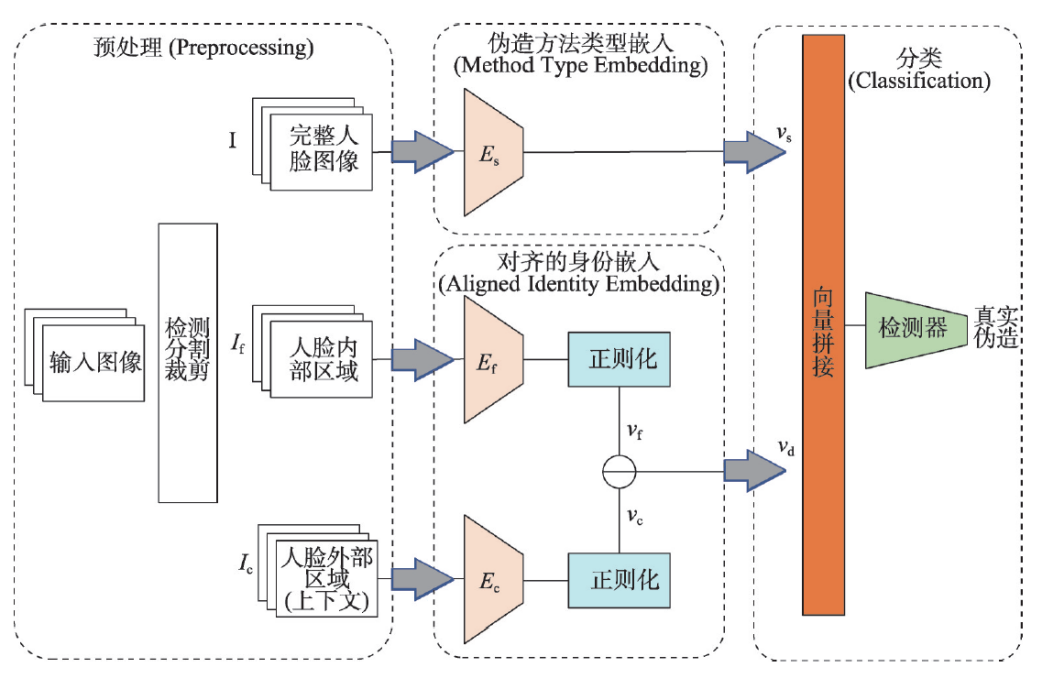
单独使用频域信息会忽略了原始图片空间的像素特征, 进阶操作是结合空间信息,从频域特征与空域特征两方面计算两两区域之间的差异，以判断视频真伪.

6.13 基于图片整体关联性

视频换脸只是对视频帧中人脸部分进行篡改，而画面中的其他部位（如人的躯干、图片背景）是不做更改

的。因为被篡改和未被篡改的区域之间在理论上存在着必然的不同，所以利用这种不同进行人脸视频深度伪造的检测（即基于图像上下文空间的检测方法）也是近几年提出的重要方法。视频换脸目标人脸裁剪下来,经过编辑后放到源人物脸上,并定存在图片融合过程,那么这个过程一定会产生融合边界.边界是否存在的判定就可以作为真伪标准.，

如以下设计方法中,分别对应视频帧中的人脸部位和扣除上下文背景信息部分，两个网络模型的输出作差便是人脸与上下文之间的差异信息。之后再配合第三个视觉网络，以待检测的真伪视频为训练数据，将三个网络的输出进行融合用作最后的分类。



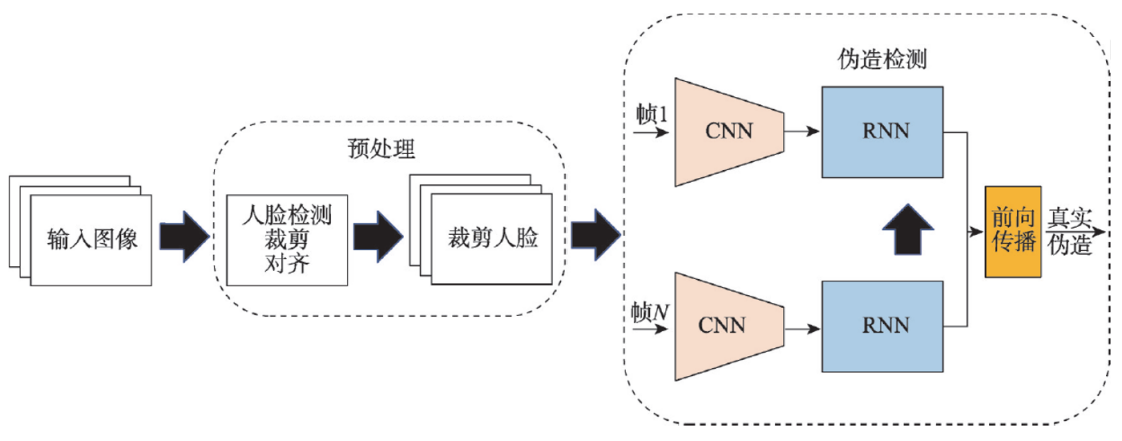
6.2 基于时空特征融合的方法

视频本质便是帧的快速切换,因此相邻视频帧在背景,人物动作上有关联性.特点是视频帧之间的特征提取弥补了单纯利用空间特征提取造成的特征缺失问题.

根据使用方法不同,可以分为以下三个子方法

6.21 基于RNN网络

Rnn本用于自然语言方向提取上下文关联性的网络.在视频处理领域,也可用来挖掘相邻帧之间的相关性.基本思路为，先利用CNN 进行每一视频帧特征提取，再使用RNN挖掘相邻帧之间的时序关系。关系结构如下图:



有很多变种,比如用LSTM替代rnn提取时序信息.但处理反式都大同小异,基于这个架构的方法并没有结合深度伪造视频的独有特点,因此在库内及跨库检测中并没有体现出太高的准确率.

另一项新颖的研究是从伪造视频的生成阶段入手，认为视频在伪造的最后阶段对每一帧进行压缩编码时，在生成I 帧、B 帧、P 帧的过程中带来预测误差，可以以帧间的预测误差作为特征输入，经过CNN的特征提取与LSTM的时序提取，最终进行分类。

6.22 基于卷积的时空融合特征检测方法

基于卷积的时空融合特征检测效果依赖于卷积核的设计,通常的方法是对视频帧卷积核的时间维度进行设计，以提取帧间的连续性与相关性等特征。部分研究方向如下:

1. 以多个连续视频帧为一组,输入到3D卷积中提取特征.但3D卷积对时间维度注意力不足,影响检测效果.
2. DeepFakes 是单帧篡改的,导致相邻帧之间会存在抖动,以用于用文本分类里的1 - d 卷积处理,最后得到时空实例，用于最终检测.

6.23 基于像素位移的时空融合方法

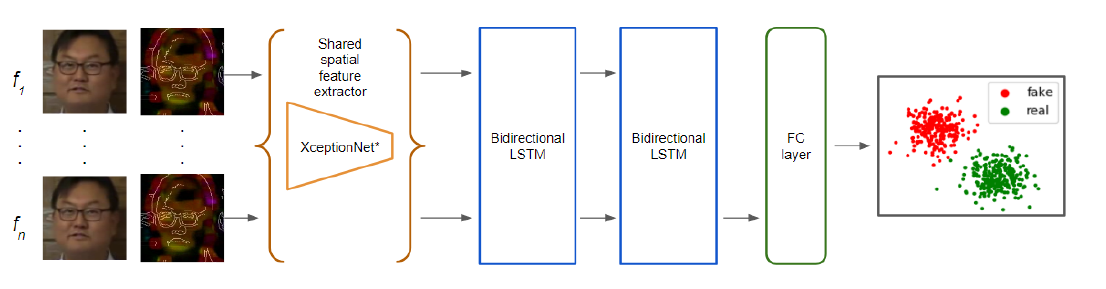
又称为光流法,是用图像序列中像素在时间域上的变化以及相邻帧之间的相关性来找到上一帧跟当前帧之间存在的对应关系，从而计算相邻帧之间物体运动信息的一种方法.

真实视频与伪造视频在所形成的光流的大小、方向、分布等方面存在差异，而这个差异可以被CNN获取与分析,进而判断视频真伪.

使用此方法获得最好性能的算法设计如下:

利用OpenCV的Canny 边缘检测方法获取每一帧的边缘信息，形成图像的“边-流特征图”。再与图像的原始RGB 以不同方式进行融合，共同形成输入模型的最终特征组合。根据融合方式不同选择不同的Xception 结构，并接入Bi-LSTM网络挖掘帧间的关联性与相关性，最后使用全连接与Sigmoid 进行二分类。

文章在常用的FaceForensics++、DFDC 等数据集上进行训练、验证与测试，训练与测试使用相同数据集时平均准确率最高达到97.94%，使用不同数据集时最高达到81.29%，表示了模型具有一定的泛化能力。



6.3 基于生物特征的检测方法

换脸鉴伪归根到底还是对人的判断. 所有基于图片或视频的空间、时间、频率、像素等方面的检测只是对于载体图片的检测，而基于生物特征的检测则是基于人的，是不以伪造技术、承载介质不同而出现不同的.这种方法对数据集视频压缩率、是否包含音频等方面有一定的要求.

依据是基于生物特征的检测主要依赖于两个假设：

一是不同身份的人所表现出来的行为方式、说话习惯等是不同的，因此可以作为鉴定视频中所出现人物是否符合其所表现出来的身份的标准；

二是真实的人与计算机生成和伪造的人相比，在行为表现上是不同的。

这类方法有局限性,主要用于对重要人物伪造视频的检测,用途很窄,相当于一个定制化算法.

6.31 基于人物生理信号的检测方法

主要原理是，分析其行为特征、生物特征生理特征，以识别视频中人物是否具有真实的人的生物特征的方法。如真实人物对象平均会在6 s 出现一次眨眼行为，而伪造视频则不会在伪造过程中对此进行建模，就可通过检测这种差异判断视频真伪。

又如有研究认为，血液在流过脸部时会引起皮肤颜色的微小变化，这种变化肉眼无法看到，但通过视频中帧的像素点变化可以检测到，因此推测假的视频中的心率变化与真的视频中的心率变化不一.

以上两个例子都需要收集当事人的生物特征才能实现,是完全没有泛化行.

6.32 基于音画特征的检测方法

这种方式检测的场景优限,仅适用于说话的场景.没语言表达的场景就用不到,如视频换脸的原视频在跳舞就搞不定,不能用于通用检测.

**6.4 基于transform的方法**

在2021年后,有大量的研究尝试使用transform研究视频鉴伪问题.

