###### Deepfake detection

1. **(HeadPose) Exposing Deep Fakes Using Inconsistent Head Poses --- ICASSP 2019**
2. 主要内容

从deepfake制作流程中固有限制入手，发现合成的人脸拼接到原图片中会产生3D头部姿态误差，即在伪造图像中使用整个脸部的landMark和仅用中心合成区域的landMark估计的头部姿势存在不一致，而真实图片中差异不大，所以可将这种差异作为特征向量输入至SVM中，判断真假图片。

2）实验细节

数据处理：Dlib提取68个2D关键点，OpenFace提取3D关键点模型，确定整张和中间关键点序号。

数据集：UADFV（49真视频+49假视频，35个用于训练，294\*500）、DARPA GAN（241真实图片+252伪造图片）

训练集：从35真实视频+35伪造视频中提取21694个图片。

测试集：剩余14真实视频+14伪造视频中的11058帧图片，以及DARPA GAN数据集中全部图片。

评估指标：AUC、ROC

对比方法：无，只进行消融实验。

优化器：训练RBF核的SVM分类器，使用5倍交叉验证grid搜索超参数。

此外，将每个视频作为一个单元进行评估，通过平均此视频上帧图片的预测结果。

1. **VA(VA-MLP)Exploiting Visual Artifacts to Expose Deepfakes and Face Manipulations --- WACVw 2019**
2. 主要内容

此工作利用眼睛，牙齿，面部轮廓的特征/差异描述伪影，用来检测人脸修改。针对三种人脸操纵利用不同的特征进行检测：①ProGAN生成/合成的人脸。利用左、右眼睛颜色差异检测，包含六个维度特征：HSV颜色域3个分量间距离+RGB颜色域三个通道的相关性。分类器：KNN分类，欧式距离。②Deepfakes。利用眼睛和牙齿区域缺失的反射以及缺失的细节进行检测。对眼睛和牙齿区域提取纹理特征（每个像素9个纹理特征），最后为每个样本生成特征向量：对眼睛、牙齿、整个图片的所有像素的9个为纹理特征平均。分类器：3层神经网络分类器MLP+逻辑回归模型③Face2Face。利用脸部边缘和鼻尖的特征检测。分类器同2。

对于以上不同操纵方法，选择不同特征伪影，主要依据全局一致性、光照估计、面部几何估计三方面的分析。①ProGAN生成/合成的人脸可以发现在左眼和右眼之间有很大的颜色差异（异色症）②Deepfake操作可以发现镜面反射：眼睛的反射不见了或者简化成一个白色斑点，这种伪影导致眼睛的整体浑浊；发现几何图形缺失，牙齿通常不会被建模，是一个白色斑点，而不是单个牙齿。③Face2Face操作可以发现阴影效应：在鼻子的区域，其中一侧被渲染得太暗；发现由几何形状不精确估计而出现沿掩模边界的伪影，通常在鼻子区域、脸部咬合边界周围和眉毛变得明显。

1. 实验细节

数据集：①ProGAN训练集（真假各2000）、测试集（真假各1000）；Glow生成1000个随机人脸用于评估泛化；②Deepfake训练集（真2440假2070）、测试集（真375假445）；Afchar提出的数据集用于评估泛化；

③Face2Face从FF中取的自驱动的样本，已提供crop的视频帧；训练集（10000帧）、测试集（5000样本）；

视频提取单帧；分辨率各不同但检测时resize为固定大小；

优化器：ADAM；无对比方法，只有另一数据集上泛化能力评估

评价指标：ROC、AUC

1. **(X-ray)Face x-ray for more general face forgery detection --- CVPR 2020**
2. 主要内容

此方法训练集不包含伪造生成的图片，使用真实的混合图片（数据增强）进行训练。可以完成检测和确定混合位置这两个任务。基于发现：多数人脸修改方法都有共同步骤，即将修改后的人脸混合到已有背景图像中。X射线是一个灰度图像，揭示输入图像是否可以分解成来源不同的2张图片的混合。其中，伪造图片的X射线会有混合边界，而真实图像不存在，X射线利用此差异进行区分检测。另外，X射线仅假设混合步骤存在，不依赖伪影且与具体伪造技方法无关，所以在未见示例上表现好，泛化能力强。

1. 实验细节

数据集：训练集FF++（1000真实视频，由真实混合的图片），测试集FF++/DFD/DFDC/Celeb-DF（论文中有介绍）

评估指标：AUC、AP (Average Precision)、EER (Equal Error Rate)

对比方法：Xception/FWA/LAE/FT/MTDS，另外消融实验证明两种数据增强的作用。

优化器：Adam

1. **(FWA) Exposing deepfake videos by detecting face warping artifacts --- CVPRw 2019**

1）主要内容

此方法训练集不包含伪造生成的图片作为负样例，将人脸仿射扭曲中的伪影作为区别真实和虚假图片的显著特征。基于发现，目前deepfake算法只能生成固定分辨率的图像，需要进一步扭曲以匹配源视频中的原始人脸（尺寸），这样的转换会留下特定的伪影（分辨率不一致导致的），并已证明其可以被CNN有效捕获。此方法两个优点：①负样例直接通过图像操作得到，不用伪造算法生成，节省了资源和时间②此伪影在各种伪造技术中存在，更具鲁棒性。负样本生成：检测人脸，提关键点，对齐人脸的转换矩阵，仿射扭曲回原图片，不同尺寸下扭曲以增加数据多样性。

2）实验细节

数据集：UADFV（图片评估取视频中所有帧图片；视频评估时每个视频作为分析单元，对所有帧评估，结果只返回top1/3的平均分数）；DeepfakeTIMIT（假视频包含LQ64\*64和HQ128\*128，对应原始数据集VidTIMIT和DeepfakeTIMIT中所有假视频用于验证）

评估指标：AUC

对比方法：HeadPose/Two-stream/Meso-4/MesoInception-4

优化器：SGD

**(DSP-FWA) DSP-FWA:** **Dual Spatial Pyramid for Exposing Face Warp Artifacts in DeepFake Videos**

**https://github.com/danmohaha/DSP-FWA. (Accessed on 12/18/2019).**

1. **(Xception-c23)** **Faceforensics++: Learning to detect manipulated facial images**

1）主要内容

主要贡献：对于人脸伪造检测研究提供不同压缩率下的benchmark；公开大规模数据集1000真实+4种伪造方法产生的视频；为检测器提供广泛的评估；提出一种伪造检测方法：尝试各种先现有learning-based方法，比较了6种不同网络架构，发现基于XceptionNet的分类在检测方面优于其他所有变体。

2）实验细节

将数据集分成固定的训练、验证和测试集，分别由720、140和140个视频组成；H.264编解码器压缩视频，raw/LQ/HQ

1. **(Mesonet)** **Mesonet: a compact facial video forgery detection network --- IFSw 2018**

1）主要内容

特别关注Deepfake换脸和Face2Face重现两种视频伪造技术。由于压缩会极大地降低数据质量，传统的图像取证技术通常不适用于视频，再此之前没有方法专门用于检测视频伪造技术。本文采用了深度学习的方法，提出了两个层数低、参数少的网络，以聚焦于图像的mesoscopic性质（少量层学习中间表示）。通过对相当复杂的网络结构逐渐进行简化，直到得到相同结果，但是更高效的方法：①Meso-4使用4层卷积+pooling提特征，1层隐藏层的dense网络用于分类。此外为了提高泛化性能，卷积层引入relu非线性和batch nom来正则化输出，防止梯度消失效应；全连接层使用dropout正则化提高鲁棒性。②MesoInception-4使用inception模块替换Meso-4前两层，使用3\*3替换5\*5卷积以避免高层语义。这里只替换2层，实验验证了替换太多层/高层语义，结果不太好。

2）实验细节

数据集：网上搜集deepfake换脸视频（因为没有数据集提供换脸视频，且训练一对人脸时间长；真实数据也是网上收集，数量基本一致）、FaceForensics中Face2Face方法（1000真实+1000伪造视频（300用于训练，150用于测试），此数据集中提供H.264的三种压缩方式（0lossless compression, 23 light compression, 40strong compression）

训练集：Deepfake（假视频5111帧，真实7250帧），Face2Face（假视频4500帧，真实4500帧）。

测试集：Deepfake（假视频2889帧，真实4259帧），Face2Face（假视频3000帧，真实3000帧）。

评估指标：检测准确率（真、假）

对比方法：与FaceForensics中结果比较，即Xception。

优化器：Adam

1. **(Multi-Task)** **Multi-Task Learning for Detecting and Segmenting Manipulated Facial Images and Videos --- [BTAS](https://ieeexplore.ieee.org/xpl/conhome/9179990/proceeding) 2019**

1）主要内容

多任务学习方法同时检测被操纵的图像和视频，并在每次查询中定位被操纵的区域。任务之间的信息共享，提高不同任务的性能。采用半监督学习方法，网络包括编码器和Y形解码器：编码特征的激活用于二分类，解码器输出一个分支用于分割修改区域，另一个分支用于重建输入，有助于提高整体性能。另外，只需要在少量数据上fine-tuning以处理未见示例。

图像中操作区域定位常用两种方法：分割整个图像和滑动窗口执行二分类。本文使用分割方法定位，不同之处在于只考虑面部区域而非整张图片（这也是FWA方法和resampling detection的不同之处）。

1. 实验细节

数据集：FaceForensics（自驱动和源-目标驱动的重现视频各1004个，真假各704用于训练，150测试，150验证）、FaceForensics++（真实1000假视频3000，每个生成方法下720用于训练，140测试，140验证）

实验设置：实验只是用23压缩下的视频，只用FaceForensics源-目标驱动中的704\*2训练，其余划分下测试。关于帧处理，使用Forensictransfer.’s settings每个训练视频中取200帧，每个测试和验证视频取10帧。由于没有对选取规则的描述，这里选择前200或者10帧。每个数据集归一化的均值= (0:485;0:456;0:406)和标准差= (0:229;0:224;0:225)

评估指标：分类任务accuracy、equal error rate (EER)；分割任务Acc

对比方法：变化Y形自编码器的网路深度等等

优化器：Adam

1. **(CapsuleNet) Capsule-forensics: Using capsule networks to detect forged images and videos --- ICASSP 2019**

1）主要内容

此工作第一次将胶囊网络用于数字取证问题，针对表示攻击和深度伪造两个域的检测（视频和图片都可以）。步骤：①预处理，包括视频提取帧（平均概率得最终结果）和VGG-19提取特征（maxpooling的输出）②胶囊网络，由3个主要胶囊网络+2输出胶囊（分别针对真/假图片）组成，VGG-19提特征分布到3个网络，再动态路由到输出胶囊。

2）实验细节

数据集：表示攻击检测Idiap REPLAY ATTACK；换脸检测Afchar数据集；人脸再现检测FF中Face2Face对应数据；

对比方法：Meso-4、Xception、Two-stream等；表示攻击检测方法

评价指标：主要是ACC，还有HTER

1. **(DefakeHop) DefakeHop A Light-Weight High-Performance Deepfake Detector --- ICME 2021**

1）主要内容

非深度方法DefakeHop包括三个模块：①Kuo利用连续子空间学习SSL开发了PixelHop++理论，之前工作将其用于从低分辨率人脸图像中进行特征学习，本文首次将PixelHop++用于从高分辨率图像中提取的patches的特征提取（通过c/w Saab变换，PCA的一种变形）②提出了一种有效的特征蒸馏模块，使用空间降维去除人脸的空间相关性，并使用软分类器包含每个通道的语义，来进一步降低特征维数得到更简洁的人脸表示。③最后集成不同的区域和帧。区域集成，由于每个人脸区域对于不同的Deepfake技术可能有不同的优势和劣势，将不同区域（左眼，右眼和嘴）的概率集成在一起。多帧集成，对于每一帧，将当前帧和相邻的6帧(前3和后3帧)连接起来，以合并时间信息。最后通过平均来自同一视频的所有帧概率计算整个视频被伪造的概率。也可以使用不同方法聚合帧级概率来决定。

1. 实验细节

图像预处理：crop/align/normalize。OpenFace2提取68个关键点，resize128\*128并旋转到特定的坐标，使所有样本保持一致，没有不同的头部姿态或人脸大小。最后，从不同部位(如左眼、右眼和嘴巴)裁剪出大小32\* 32的patches，作为PixelHop++模块的输入数据。

数据集：一代UADFV、FaceForensics ++ (c23和c40两种质量的测试)、二代Celeb-DF v1和Celeb-DF v2

训练测试：UADFV（80%用于训练，20%用于测试），FF++和Celeb-DF提供测试集，其余视频用于训练。

实验细节：在各个数据集与其他方法的对比，AUC值从celeb-df和survey中直接摘取。FF++中两种质量下的评估。模型参数量的比较。文中提出方法在各个数据集上的ROC曲线、帧级和视频级的AUC、在各个人脸区域的AUC值。

评估指标：AUC、ROC

对比方法：InceptionV3、Meso4、FWA(ResNet-50)、HeadPoseb (SVM)、VA-MLP、Xception-raw、Multi-task、CapsuleNet、DenseNet+RNN、DSP-FWA (SPPNet)、Xception、DefakeHop (Frame)、DefakeHop (Video)时间顺序？

1. **(CViT) Deepfake Video Detection Using Convolutional Vision Transformer --- arxiv 2021.02**

1）主要内容

卷积视觉Transformer由两个组成部分：CNN和视觉Transformer(ViT)。CNN(没有全连接层的VGG架构，不进行分类)提取可学习的特征，而ViT将学习到的特征作为输入，并用注意机制进行分类。此工作基于Deepfake检测方法的两个弱点：数据预处理和（检测）泛化性。预处理包含：人脸提取224\*224和数据增强。本文方法具有泛化能力的原因：①CNN+Transformer学习局部和全局特征。②在训练和测试时对数据都进行预处理。③更大、多样性的数据集训练模型以检测不同设置和生成方法下的伪造。

2）实验细节

预处理：人脸检测BlazeFace，人脸提取MTCNN和face recognition。为了提高准确率，三个库一起使用。保存为224\*224的jpeg格式。Albumentations库用于数据增强，90%的图片进行增强，数据增加近一倍。每个batch32张图片进行归一化，均值= (0:485;0:456;0:406)和标准差= (0:229;0:224;0:225)。

数据集：162174张图片，其中训练、验证、测试的划分比例为70：15：15，真假图片数量相等。90%的压缩率。

UADFV、 FaceForensics++（还包含了faceshifter）；DFDC仅用于测试（400个未见视频，每个视频取30张图片）。

评估指标：ROC、AUC

优化器：Adam

阈值：<0.5为真实，>0.5为假。

1. **(F3-Net) Thinking in frequency Face forgery detection by mining frequency-aware clues**
2. 主要内容

该方法利用频率感知检测伪造。基于观察：频率可以很好地描述细微的伪影或压缩错误。该方法利用两个互补的频域感知线索，提出双流协同学习框架：①FAD通过频率感知图像分解来学习微妙的伪造模式。②LFS用于提取局部频率统计信息。③MixBlock交叉注意模块融合以上2个分支，用于联合特征交互。

1. 实验细节

数据集：FF++（1000真实视频，720用于训练，140用于验证，140用于测试。4种伪造方法共4000假视频。真实视频增强4倍以解决真假数据不平衡问题。遵循FF++设置每个视频取270帧）

评估指标：AUC、ROC、PR

对比方法：MesoNet、Face X-ray、Xception。

Xception-ELA [27]、Xception-PAFilters [10]、Steg.Features[24]、LD-CNN [14]、Constrained Conv [6]、CustomPooling CNN [49]、Optical Flow(Deepfake video detection through optical ow based cnn)、Slowfast [20]

优化器：SGD

1. **(Celeb-DF) Celeb-DF: A Large-scale Challenging Dataset for DeepFake Forensics**
2. 主要内容

实验评估了检测方法的性能，有两个目的：①数据集的对比。使用平均检测性能作为各种DeepFake数据集的挑战级别的指标，进一步比较Celeb-DF与现有的DeepFake数据集。②检测方法的对比。调查了当前DeepFake检测方法在大量DeepFake视频上的性能，特别是在Celeb-DF高质量视频上。

合成方法是对基本DeepFake生成算法进行了细化（针对现有数据集特定伪影）：低分辨率（合成人脸的分辨率越高视觉质量越好，在适应输入时受大小调整影响小）、不匹配的颜色（通过在训练中随机扰动人脸颜色，训练数据增强）、不准确的人脸mask（合成有更多背景的脸，以插值/扭曲覆盖原来部分）、时序闪烁（平滑时间序列）。

1. 实验细节

数据集：UADFV、DF-TIMIT-LQ、DF-TIMIT-HQ、FF-DF、DFD、DFDC、Celeb-DF、

对比方法：Two-stream、MesoNet、HeadPose、FWA、VA、Xception、Multi-task、Capsule、DSP-FWA

评估指标：视频质量Mask-SSIM，检测方法ROC、AUC

1. **(M2TR)M2TR: Multi-modal Multi-scale Transformers for Deepfake Detection**
2. **(MaDD) Multi-attentional Deepfake Detection**
3. **(LRNet) Improving the Efficiency and Robustness of Deepfakes Detection through Precise Geometric Features**
4. **(DBiRNN)** **Recurrent convolutional strategies for face manipulation detection in videos** **--- CVPRW 2019**

1）主要内容

利用RNN（单流/双流/3D卷积）挖掘图像流的时间信息temporal coherence，来识别视频中人脸是否被篡改。通过实验，选择模型变化（RNN网络结构）与特定域的面部处理技术相结合的最佳策略。基于发现：人脸伪造不要求时间相关性而且是一帧帧修改，所以使用时序伪影/差异作为捕获异常人脸的方式，另外使用人脸对齐处理以消除其他因素对判断的影响。与CNN+LSTM工作相近，但是此工作检测多种人脸修改类型，不仅仅是deepfake+面部对齐处理以消除混合因素影响+使用双向递归网络。

2）实验细节

数据集：FF++（1000真实视频，720用于训练，140用于验证，140用于测试）

评估指标：AUC、ROC、PR

对比方法：无，只进行消融实验。

优化器：Adam

1. **(CNN+LSTM) Deepfake video detection using recurrent neural networks --- AVSS 2018**
2. **(Two-stream) Two-stream neural networks for tampered face detection.**
3. **(Two-branch) Two-branch recurrent network for isolating deepfakes in videos**

**数据集相关工作：**

**Celeb-DF**

Youtube上59个名人的视频片段，590真实视频，5639假视频。每个视频的帧率30，每个视频平均13秒.

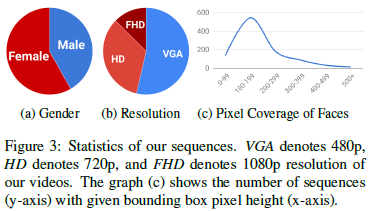
真实视频数据分布。性别：56.8%男性，43.2%女性。国籍：亚洲人5.1%，非洲裔美国人6.8%，白种人88.1%。年龄：60以上8.5%，50-60岁30.5%，40s 26.6%，30s 28%，30以下6.4%。59对象每一对交换人脸。

**FF++**

FaceForensics++: Learning to Detect Manipulated Facial Images.

FaceForensics: A Large-scale Video Dataset for Forgery Detection in Human Faces.

Forensictransfer: Weakly-supervised domain adaptation for forgery detection.



Investigating the Impact of Pre-processing and Prediction Aggregation on the DeepFake Detection Task.

**SR-DF**

基于FF++的1000个真实视频；使用dlib裁剪视频帧的人脸区域；使用DoveNet后处理/合成生成视频；

四种伪造生成方法：First/Faceshifter/IcFace/FSGAN

