A Survey of Deepfake Detection

组长：吕月明
组员：樊红兴
组员：张时润
组员：管伟楠

1 本地已有数据集

到目前为止，我们申请了包括 FaceForensics++ [1]，DeepFake Detection Challenge Dataset [2]，DeeperForensics [3] 等在内的数据集.

表 1: 组内已有的视频数据集

Datasets	Fake videos	Real videos	Methods	Local path	year
FF ++ [1]	5000	1000	5	/hd1/DeepFakeDetection/	2019
Celeb-DF [4]	590	5,639	1	/hd1/fanhongxing/Celeb-DF-v2/	2019
DFDC Pre [5]	4073	1140	2	/hd1/DFDC-preview/	2019
DFDC [2]	104500	23654	2	/hd1/DFDC/	2020
DeeperForensics [3]	60000	1000	1	/data1/deeperforensics/	2020

同时，我们收集了部分 GAN 的生成图像，主要从以下两个途径获得：

1. 从网上直接下载其他研究组公开的 GAN 生成图像；

2. 我们利用公开的预训练模型权重，自己生成了一批数据。

真实人脸数据有 FFHQ, Celeba HQ 等，伪造人脸数据主要包括 FSGAN，FaceSwap，PG-GAN，VAE, StyleGan，CycleGan，WGANGP, DCGAN, Glow, BEGAN, MeGlass，StarGan，DFFD 数据集，ZAO 等。

真实人脸数据路径：

FFHQ：/data/zhangyaru/ffhq/images1024x1024/

Celeba HQ: /data3/fanhongxing/GeekPwn2020/data/celeba-1024/

伪造人脸数据路径：

各种 GAN：/data3/fanhongxing/GeekPwn2020/data/

DFFD：/data1/deepfake-dataset/dffd-dataset/ （‘-’改为下横线）

ZAO：/data1/ZAO/

2 组内的检测方法

2.1 GAN 图片取证模型泛化性研究，作者轩心升

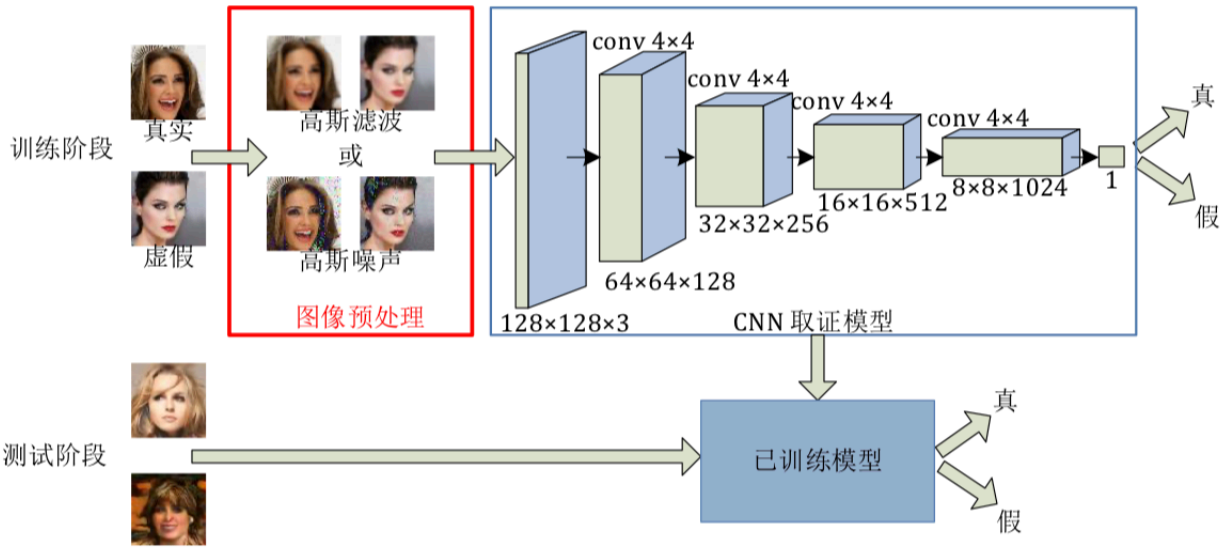

图 1: GAN 图片取证模型泛化性研究模型架构

最近，GAN 生成的人脸图像越来越逼真，高质量，甚至人眼难以察觉。另一方面，取证领域继续开发方法来检测这些生成的伪造图像，并尝试确保视觉内容的可信度。尽管研究人员已经开发出了一些方法来检测生成的图像，但很少有人探讨取证模型的泛化能力这一重要问题。随着新型 GAN 的快速出现，取证模型检测新型 GAN 图像的泛化能力绝对是必不可少的研究课题，这也是非常具有挑战性的。在本文中，我们探讨了这个问题，并建议使用预处理的图像来训练取证 CNN 模型。通过对真实和伪造图像应用相似的图像级别预处理，破坏了不稳定的低级噪声提示，并且强制取证模型学习更多固有特征以对生成的和真实的面部图像进行分类。我们的实验结果也证明了该方法的有效性。本工作所介绍出的方法与其他生成图像

取证工作的主要区别在于，本工作在训练阶段使用了图像预处理步骤，以破坏生成图像的低级不稳定伪像，并迫使取证模型专注于更多内在的取证线索。本工作使用的的图像预处理方法不同于其他的图像取证工作，一般的图像取证工作旨在增强高频像素噪声并专注于低层级的像素统计的线索。而本工作有意通过引入使用平滑滤波或者噪声的预处理步骤来破坏或抑制这些线索。另一方面，这种对真实图像和生成图像同时使用一种图像预处理的操作的方法，可以提升两种图像之间在像素层级上的统计特征相似性，可以增加模型训练的难度，迫使取证模型学习到更多与生成算法内在意义相关的特征。从机器学习的角度来研究和设计取证模型，将整个模型架构分为训练阶段和测试阶段。整个训练的流程如图 1.1 所示，从图中可知本工作在模型训练阶段增加了图像预处理操作，其目的是添加图像模糊或者增加图像噪声。在测试阶段，本工作并没有在此阶段使用图像预处理操作，而是改用为直接使用原始图像作为输入。由于本工作的主要重点是验证所提出的图像预处理操作对提升取证模型泛化能力的有效性，因此，本工作并没有设计比较复杂的 CNN 网络架构，而是直接使用了基于深度卷积生成对抗网络 DCGAN 的判别模型（CNN 网络）作为计算机生成图像检测模型。

2.2 基于空洞卷积的低质量人脸深度伪造图片检测，作者卞明运

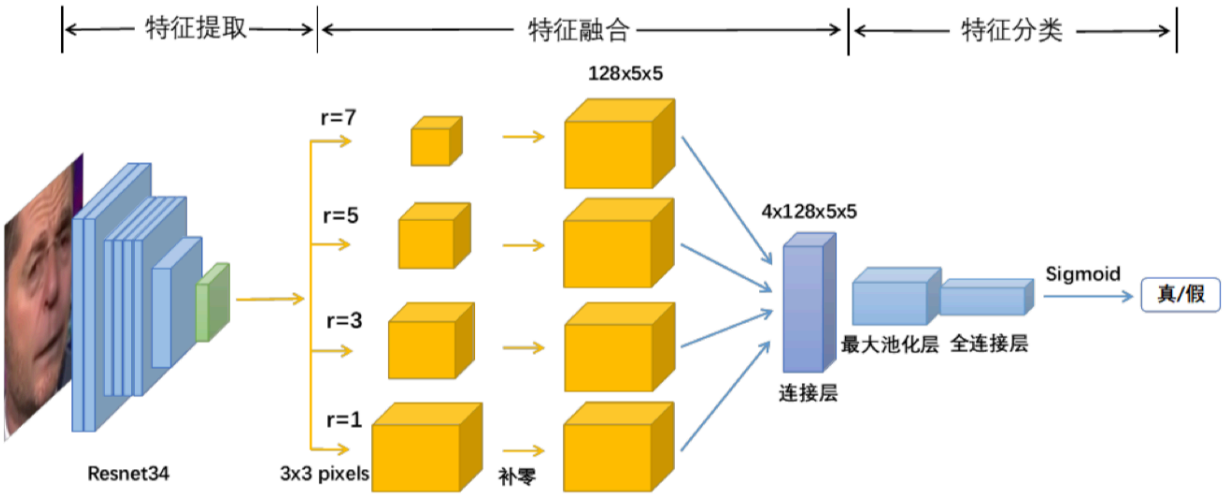

图 2: 空洞卷积的低质量人脸深度伪造图片检测模型架构

图像分类任务中，在使用卷积神经网络提取图像特征信息时，通过最大池化和下采样的操

作提高模型感受野的同时会带来特征图分辨率下降，上下文信息丢失等问题，在原本分辨率较低的人脸深度伪造图像中，局部有效特征信息的丢失会对分类准确度带来较大的影响。提出基于空洞卷积的多尺度信息融合的人脸深度伪造检测方法。该方法采用全卷积 Resnet34 模型和基于空洞卷积的多通道特征融合的结构，利用空洞卷积网络在提高网络感受野的同时不丢失特征图信息的优势，融合多尺度空间特征信息，最大程度的捕捉到图像上下文信息。实验在 Faceforensic++，Celeb-deepfakeforensics，Deepfake in the wild，DFDC previous 数据集上均取得比当前其它方法更好的效果。结果表明，提取高层语义信息后进行多尺度特征融合，可以提高分辨率低的人脸深度伪造图像的分类效果。我们所提出的人脸深度篡改检测的模型主要包括个部分：1）特征提取部分；2）特征融合部分；3）特征分类部分。对于特征提取部分，为了从篡改人脸图片学习到像素级的可区分特征，采用主流的 Resnet34 网络，并对网络输出层作了修改；对于特征融合部分，采用多层空洞卷积神经网络并行叠加方式，在保证较大感受野和信息损失较少情况下，将不同尺度的高层语义特征进行融合；对于特征分类部分，通过最大池化层提取主要特征信息，再通过 Sigmoid 层输出模型判断分数，最后和标签进行匹配得出人脸真假判断结果。网络结构如图 1.2 所示。总结来说，我们主要有三个贡献：1）针对低分辨率的人脸深度伪造图片提出一个多层次、多尺度的空洞卷积模块；2）提出了一个深度残差网络 Resnet34 和多通道的空洞卷积网络融合的人脸深度伪造检测网络；3）通过在四个相关数据集上和主流方法的结果比较，证明提出的方法的有效性。

2.3 基于面部 3D 形状信息的鲁棒人脸交换检测方法，作者管伟楠

近年来，随着对抗性网络（GAN）的快速发展，伪造的图像和视频（称为 DeepFake）的创建变得更加高效和智能。可以使用深度学习方法轻松创建高度校准的 DeepFakes，尤其是对于面部交换的 DeepFakes。换脸是最近 DeepFakes 的一种常见生成方法。但是，当前检测这种 DeepFake 的方法主要基于深度学习，这表明其可解释性和鲁棒性较弱。由于当前换脸方法对 3D 形状信息一致性的无知，提出了一种基于 3D 脸部形状的方法来检测换脸 DeepFakes。方法利用 3D 可变形模型提取面部 3D 形状信息。由于面部 3D 形状信息的提取过程受图像质

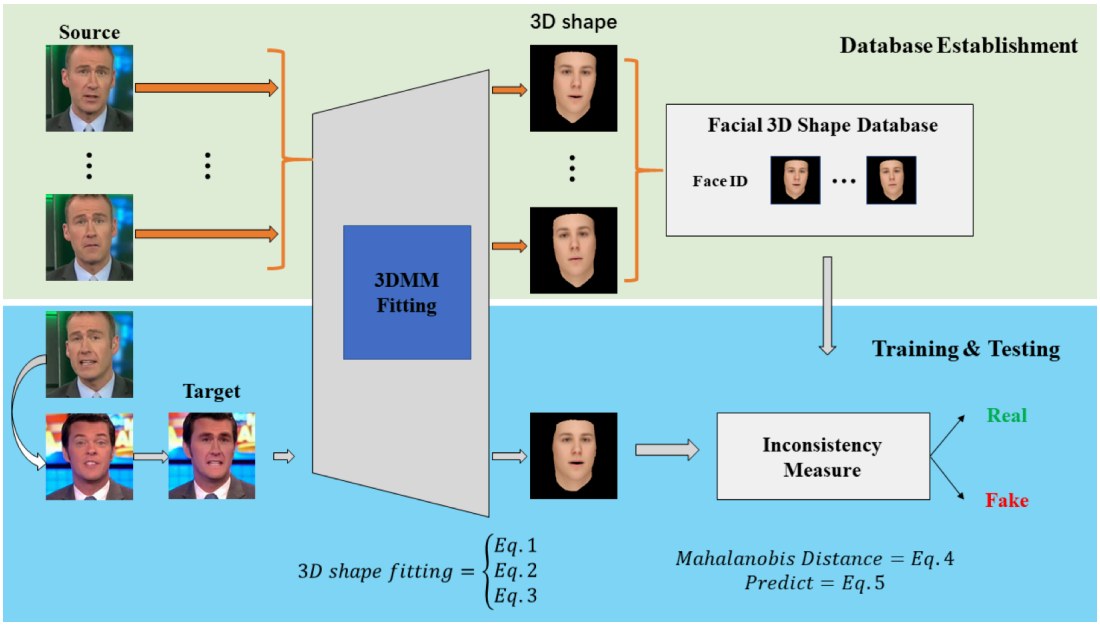

图 3: 基于面部 3D 形状信息的鲁棒人脸交换检测模型架构

量的影响较小，因此我们提出的方法对于低质量视频保持了良好的鲁棒性，同时还显示了相当好的解释性。网络结构如图 1.3 所示。

方法首先根据选取一个人物的一段视频，通过 3DMM 的方法提取头部三维形状信息，然后注册到 reference dataset，然后对于待检测样本，根据该样本的人脸来比较 reference dataset 和待检测样本的头部三维形状信息，进行判别实验。结果表明，我们的方法在大多数后处理数据上均保持良好的性能，而基于深度学习的先前方法的性能急剧下降，甚至在经过一些后处理操作后也完全丢失。此外，当通过另一种算法生成人脸交换 DeepFakes 时，我们的方法仍然保持良好的检测性能，但是深度学习方法绝对无法验证可疑样本的真实性。

2.4 基于空洞卷积和注意力机制的 Deepfake 图片检测方法，作者张时润

由于生成对抗性网络（GAN）技术的成熟，为合成 Deepfake 信息（例如图像，视频）方面带来了便利，使得在辨别图片真伪上加大了难度，给社会隐私安全带来了严重的威胁。由此提出基于注意力机制和空洞卷积的 Deepfake 图片检测方法，利用 resnet34 卷积神经网络对待检测的图像提取深层次的特征，再经过空洞卷积模块提高网络感受野，由注意力模块进行权

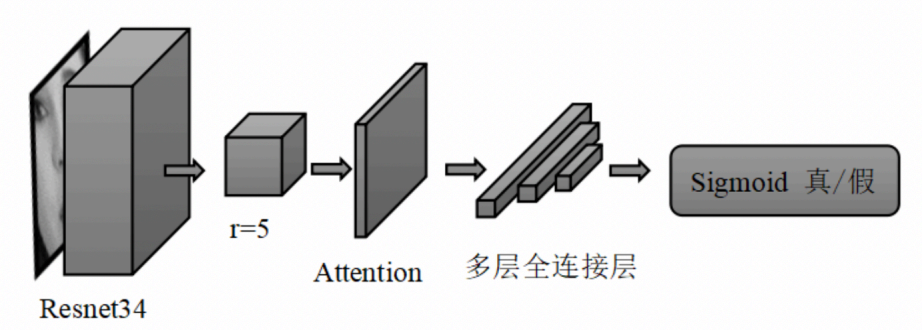

图 4: 基于空洞卷积和注意力机制的 Deepfake 图片检测模型架构

重的重加权，最后使用多层全连接神经网络对特征进行分类达到检测出篡改图片的目的。网络结构如图 1.4 所示。提出的 Deepfake 检测模型主要包括三个部分。1）特征提取模块。采用在 ImageNet 数据集上进行图像分类预训练并且对网络输出结构作了修改的 Resnet34 作为前端，来提取初步的伪造特征，从伪造图片学习到像素级的分类特征。2）特征融合权重再分配模块。特征融合再分配模块主要由空洞卷积和注意力机制组成。受图像语义分割问题中下采样会降低图片分辨率、丢失信息而提出的一种卷积思路。利用空洞卷积添加空洞来扩大感受野，让原来 3x3 的卷积核，在相同数量和计算量的情况下拥有更大的感受野。采用扩张系数为 5 的空洞卷积。用来扩大感受野，更好的捕获到人脸换脸边界伪影。再结合注意力机制，将有限的注意力集中在重要的信息上，从而节省资源，来快速获得最有效的信息，从而节省资源，来快速获得最有效的信息。3）特征分类模块。最后使用三层的全卷积网络将卷积层产生的特征图（feature map）映射成一个固定长度的特征向量，用于对前面提取的特征做加权。最后三层全连接层的维度分别是 2048、1024、512 最后输出 2 维。使用三层的全连接层是为了提高网络的非线性能力，在每个全连接层之间都加入了归一化和比率为 0.5 的 dropout 来减少网络的过拟合和参数的计算量。最后使用 Sigmoid 层输出模型判别分数进行二分类。

2.5 一种抵抗 JPEG 压缩的鲁棒伪造图像检测算法，作者樊红兴

DeepFake 的快速发展引起了人们对 Internet 信息安全的极大关注，如何有效检测虚假信息已成为当务之急。尽管现有的检测方法可以在某些情况下有效地检测合成图像，但是当

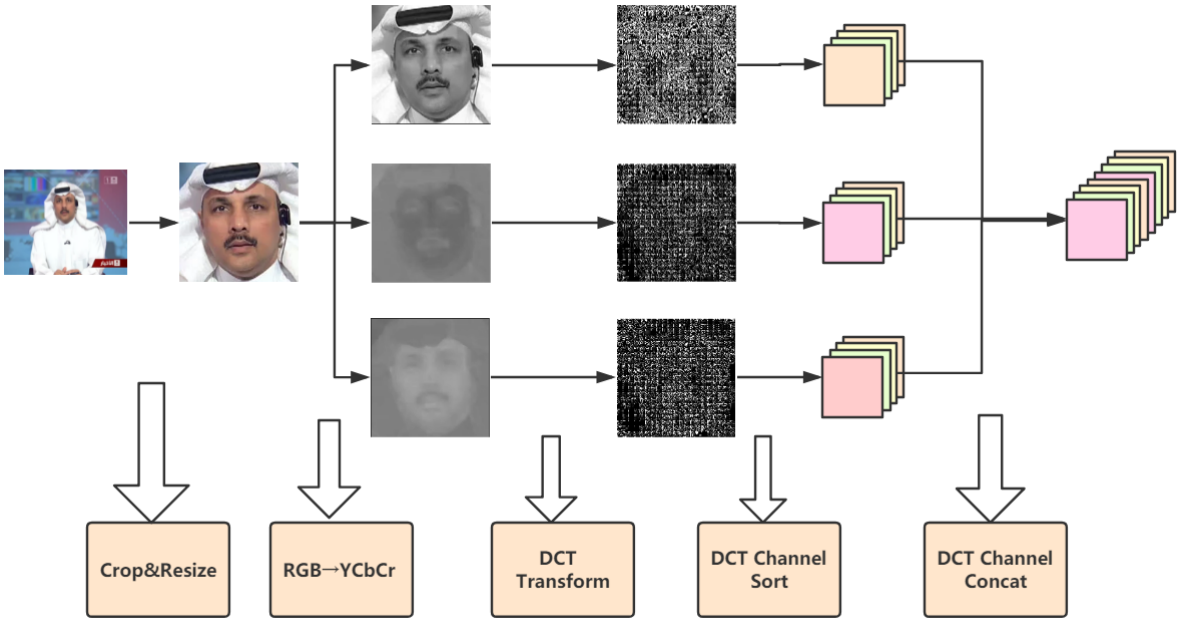

图 5: 数据预处理的整体流程

面对后期处理时，特别是在图像压缩领域中最常用的 JPEG 压缩，它们通常会失去准确性。为了解决这个问题，我们首先将 DCT 频域的学习方法引入 DeepFake 检测中，并使用离散余弦变换（DCT）系数作为卷积神经网络（CNN）的输入。使用注意力机制，我们可以从 DCT 系数中为不同的伪造方法选择不同的重要系数通道，然后在这些选定的通道上重新训练模型以获得更好的鲁棒性。我们对 FaceForensics++ 数据集中的四种伪造方法进行了实验，发现低频信息对于伪造检测更为重要。对 JPEG 压缩图像数据进行的实验表明，与空间域学习方法相比，我们提出的方法可以更好地抵抗 JPEG 压缩。

数据处理方法如图 5所示。由于 DeepFake 主要伪造脸部，因此我们裁剪出脸部区域，然后将其调整为固定大小。JPEG 使用 YCbCr 颜色空间，因此应将图像从 RGB 转换为称为 YCbCr 的颜色空间。它具有三个分量 Y，Cb 和 Cr：Y 分量表示像素的亮度，而 Cb 和 Cr 分量表示色度（分为蓝色和红色分量）。图像被分成 8×8 像素的块，并且对于每个块，Y，Cb 和 Cr 数据均进行离散余弦变换（DCT）。以上步骤与标准 JPEG 压缩相同。之后，将相同频率的所有分量归为一个通道。我们以一个例子来说明上述操作。首先输入 RGB 图像，然后进行脸部裁剪后，将脸部区域的大小调整为 448x448。当前图像形状为 448×448×3。然后将图像

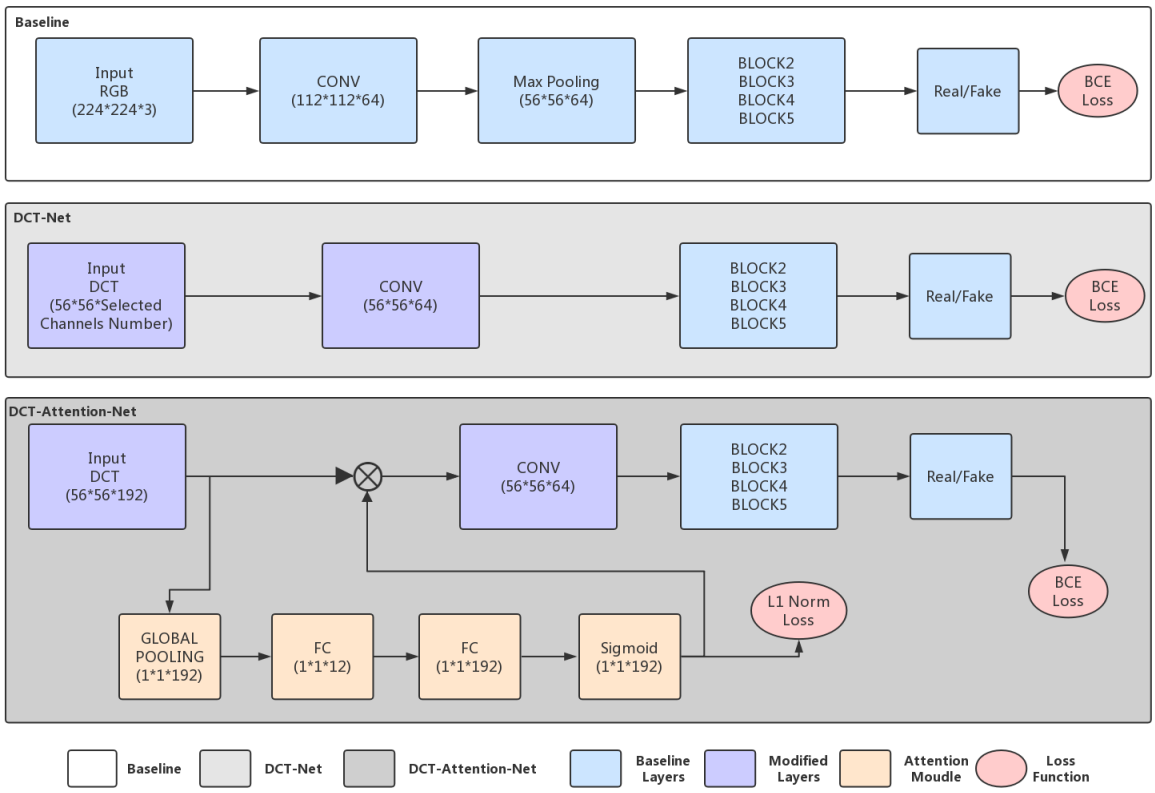

图 6: 基准网络框架是标准的 ResNet-50，我们选择 ResNet-50 作为 DCT 网络，同时用卷积层代替输入层后面的卷积层和池化层。DCT-Net 是在 DCT 网的基础上建立起来的，具有注意机制和 L_1 损失模块。

转换为 YCbCr 颜色空间并分成 8× 的块 8 个像素，并且对于每个块，Y，Cb 和 Cr 数据中的每个数据都经过 DCT。现在将图像转换到频域，形状仍然是 448×448×3。接下来，将相同频率的二维 DCT 系数分组到一个通道中，以形成三维 DCT 立方体。对于 Y 通道，当前形状 448×448×1 变为 56×56×64。因此图像的形状变为 56×56×192。

网络结构如图 6所示，基准网络框架是标准的 ResNet-50，我们选择 ResNet-50 作为 DCT 网络，同时用卷积层代替输入层后面的卷积层和池化层。DCT-Attention-Net 是在 DCT-Net 的基础上建立起来的，具有注意机制和 L_1 损失模块。DCT-Attention-Net 网络的主要作用是挑选出重要的 DCT 通道，然后根据挑选出来的重要通道，我们使用 DCT-Net 进行重新的训练，结果证明了该网络能有效的抵抗 JPEG 压缩。我们观察 DCT-Attention-Net 对这四种不同的伪造方式挑选出来的通道，发现这些通道大多是低频的通道，我们通过这个发现，

手工的挑选出来一部分低频通道，结果发现同样可以有效的抵抗 JPEG 压缩。

3 当前研究进展

3.1 深度伪造图像检测

3.1.1 图像级检测方法

对于视觉上的深度伪造生成技术，主要分为面部合成、属性编辑、人脸交换和表情变换四个方面。但是，对于图像级的伪造检测方法，大多数并不是针对特定的生成方法而提出的。图像级的深度伪造检测方法大多分为如下三种，即

a. 通用的分类网络，比如 [6] 和 [7]。简单来说，这种方法将深度伪造检测任务视为一种分类任务，根据给定输入图片将其分为真实样本和伪造样本两类，从而进行伪造样本的判别。这种方式的优点在于简单易用，让网络直接学习到可以区分真假样本的特征，适合在一些 deepfake 检测的比赛中应用。但其缺点在于，并没有关注到伪造样本的本质问题，其检测性能很大程度取决于训练数据的分布，对于未见过的伪造方式生成的数据，并不能很好的判别，这种方法训练得到的模型泛化能力比较差。

b. 基于特定线索的深度伪造检测技术。这一类别中所提出的方法，是通过寻求某些特定的伪造线索进行伪造检测，具有较强的可解释性。S. McCloskey 和 M. Albright [8] 采用了颜色特征，并以 SVM 为分类器完成伪造检测。该文章表明，GAN 网络生成的图像和真实相机拍摄的图像在颜色处理上有所不同。因此可以以颜色特征作为区分 GAN 生成图像和真实图像的判别线索。N. Yu 等 [9] 则是验证了 GAN 网络生成伪造样本时会存留下的一些模型痕迹，这是由模型本身所带来的，类似于模型所特有的特征，该方法则是利用了待检测图片的“指纹”和选用生成一系列图片的“指纹”做相关性匹配，由此进行判断。R. Wang [10] 等人认为在人脸检测过程中，真假样本在神经元的层层激活模式下会有所不同，因此可以通过监控神经元的行为以此为依据利用 SVM 进行分类判别。J. Stehouwer

等 [11] 提出，针对不同的伪造样本，每一个样本造假痕迹明显的地方不一定是同一区域，如果更加关注于造假的区域，会更易对伪造样本进行检测，由此提出了一种基于注意力机制的方法，在原有的 CNN 模型基础上，增加注意力机制，自适应关注篡改区域，提高检测准确率。P. Zhou [12] 等人提出了一种双流网络，他们考虑图像的细节伪影可以很好的帮助原有分类网络提升鉴伪性能，因此使用一个流依靠分类网络从全局角度检测是否是篡改过的人脸，另一流基于隐写特征捕捉噪声残留，最后将两个流进行融合，从而判断是否为伪造的人脸。Y. Li [13] 等人则将注意力转向了 3D 头部姿态的误差，他们认为将人脸拼接至另一个人面部时，对 3D 头部姿态的估计可能会出现误差，因此来寻找头部姿态和人脸中心区域的差异性来检测伪造样本。A. Bharati [14], 提出了一种基于受限玻尔兹曼机的深度方法。将人脸小块送进网络，用以学习真实图片和修饰过后的图片之间的区别。相应的，A. Jain [15] 等人设计了一个 CNN 网络，由 6 个卷积层，两个全连接层构成，中间有跨层连接，这个检测系统的输入和 [14] 类似，将不重叠的人脸小块送入网络加以训练，最终得到检测网络。同时，考虑到简化方法建模，F. Matern [16] 等提出了一种相对简单的模型，该模型考虑了相对简单的视觉特征，比如眼睛颜色，牙齿细节的缺失等，由此构建 Logistic 回归模型和 MLP 模型以判断伪造样本。除了时域特征，X. Zhang 等 [17] 提出了一种基于频域的方法，使用 AutoGAN 来生成图像，再转换至频域上进而训练分类器，对于 StarGAN 生成的面部属性编辑检测能够有很好的准确度。

c. 其他一些 CNN 方法。L. Nataraj [18] 等人提出一种 CNN 与共生矩阵相结合的方法，这是一种基于隐写分析和自然图像统计的检测系统，取得了不错的成果。D. Afchar [19] 等人则提出了两个层数不多，相对简单的网络，意在基于中层的语义进行检测，分别用了 Meso-4 和改进的 Meso-4 网络。因为基于图像噪声的低层（语义层级非常低）水平分析不能够应用于被压缩的视频内容中，因为这些内容的图像噪声会被衰弱的。相似的，在高层语义水平，人类的眼睛不能够非好的区分伪造的图片，尤其是当这张图片描述的是人的脸的时候，因此他们选择采用中层语义特征进行检测。H. Nguyen [20] 等人则提出了一种基于多任务学习的 CNN 方法，同时检测伪造样本并定位篡改区域，基于自编码器来构

建检测系统。S. Tariq [21] 等以 Adobe Photoshop CS6 生成面部属性编辑的伪造样本，比如上妆，眼镜，太阳镜，头发和帽子等，用 VGG16，VGG19，ResNet 和 XceptionNet 作为检测网络，进行伪造样本检测。

总而言之，当前图像级的深度伪造检测方法种类繁多，基于各种线索以及不同的分类模型等等，而当前方法所关注的问题中除了性能上的提升，检测方法的鲁棒性和可解释性更是焦点所在。

3.2 深度伪造视频检测

3.2.1 视频级数据集

UADFV 数据集 [22] 包含 49 个真实视频和 49 个虚假视频，每个视频大约持续 11 秒，总共 32752 帧。其中虚假视频基于 FakeApp，由真实视频对应生成。FFW 数据集 [23] 包含 150 个分辨率大于 480p 的高质量视频，其中 50 个视频由 FakeApp 生成。DeepFake-TIMIT 数据集 [24] 包含 640 个基于 Vid-TIMIT 数据集 [25] 和 Faceswap-GAN 生成的深度伪造视频。其被分为两个大小相等的子集，64×64 像素的 DF-TIMIT-LQ 数据集和 128×128 像素的 DF-TIMIT-HQ 数据集。FaceForensics++ 数据集 [1] 包含 1000 个真实视频和 1000 个虚假视频。其中虚假视频基于 Faceswap，由真实视频对应生成。是 FaceForensics 数据集的扩充。Google/Jigsaw 数据集 [5] 包含 3068 个虚假视频，该数据集基于 28 个不同性别, 年龄和种族的 363 个原始视频生成。伪造方式尚未公开。Celeb-DF 数据集 [4] 包含 590 个真实视频和基于真实视频生成的 5639 个虚假视频，这些视频来自 59 个不同性别、年龄、种族的采访者。伪造方式尚未公开。DFDC 数据集 [5] 包含 1131 个真实视频和基于真实视频生成的 4113 个虚假视频，这些视频来自 66 个不同性别、年龄、种族的人。伪造方式尚未公开。基于发布时间和算法的综合分析, Li 等人 [26] 提出将 UADFV 数据集 [22] 、DeepFake-TIMIT 数据集 [23] 和 FaceForensics++ 数据集 [1] 归为第一代深度视频伪造数据集, 而 Google/Jigsaw 数据集 [82], Facebook 深度伪造检测挑战数据集 [5] 和 Celeb-DF [4] 为第二代数据集。相比于第一代数据集, 第二代数据集的数量和质量均有较大的提升。DeeperForensics-1.0 数据集 [3] 包

含 60000 个视频, 其中 10000 个虚假视频, 1760 万帧图片, 在规模上是现有同类型数据集的 10 倍。Jiang 等人 [3] 召集了 100 位计算机视觉领域的专家，对当前主流深度伪造视频数据集的质量进行真实性评估。评估结果表明，相比于 UADFV [22]、FaceForensics ++ [1] 等主流数据集，Celeb-DF [4] 和 DeeperForensics-1.0 [3] 数据集更加真实。

3.2.2 视频级方法介绍

由于视频压缩带来的强烈退化，大多数的深度伪造图像检测技术不能直接用于深度伪造视频检测技术。而且，深度伪造视频检测具有随着不同视频帧变化的时间特性，因此很难被静态的深度伪造图像检测技术检测。所以深度伪造图像检测技术的很多方法并不适用于视频级的检测任务。设计并利用伪造视频的相关特征来进行检测，才能取得更好的结果。目前，主流方法大致分为两类: 一类是基于视频帧间的时序特征的检测方法, 另一类是基于视频帧内视觉伪像的检测方法。

1）基于视频帧间的时序特征的检测方法

对于单张的图像，无论真假，都不能反映出人脸的呼吸、心跳、眨眼等活体特征。但是对于视频级的数据来说，根据人脸面部说话的口型变化以及眨眼的频率等都能判断出该视频是否是真实的。而且很多深度伪造视频的获取都是基于静态的图像来训练产生的，所以可以基于生理特征的方法来判断视频是真实的还是伪造的。

基于生理特征的第一篇工作是 Li 等人 [22] 提出的基于眨眼来鉴别深度伪造视频的方法。首先检测每一帧的人脸，定位人脸关键点信息，然后利用人脸对齐算法将人脸关键点定位到统一的空间，降低人脸头部转向和移动带来的干扰。做完上述操作后，再定位提取并缩放眼睛区域的关键点，形成一段帧序列，送入长期循环卷积网络 (Long-term recurrent convolutional networks, LRCN) [27] 中，先由 CNN 网络提取人眼特征，然后在 LSTM-RNN 网络中学习序列级的特征，最后输出到全连接层中检测视频中人眼的眨眼频率。在真实视频中可以检测到 34.1/min 眨眼频率，然而在虚假视频中只有 3.4/min 眨眼频率，设定一个正常人眨眼频率阈值为 10/min，从而这区分出真假视频。该方法在 EBV [26] 等数据集上取得了较好的结

果。但是如果伪造的视频考虑到眨眼频率的真实性，特意训练具备眨眼能力的模型，则该种方法无效。从另一方面来说，无论伪造的视频是何种类型，伪造视频的生成过程基本都是一帧一帧操作的，必然会带来帧间的时间不连续或者抖动。所以将视频帧输入到时间序列网络中，然后分类器给出真假分类结果也是可行的方案。基于时间序列的第一篇工作是 Güera 等人 [26] 提出的一种端到端的时间感知的方法。作者在论文中首先证明了深度伪造视频帧间具有不一致的特性，进而基于 CNN 和 LSTM 来检测深度伪造视频。给一段视频序列，CNN 网络首先提取特征，每一帧得到一个 2048 维的向量，然后将特征向量输入到 LSTM 网络中，再将输出送入全连接层和 softmax 层，得到最终真假视频分类结果。作者从网站上收集了 300 个虚假视频，测试了不同视频帧长度的视频，在不到 2 秒的视频 (以每秒 24 帧的速度采样 40 帧的视频) 的情况下，可以准确地预测所分析的片段是否来自一个深度伪造的视频，准确率达到 97。该方法的缺点是鲁棒性不足，容易受到对抗样本的攻击。而且需要真实和伪造数据作为训练数据，比较低效。在之后，Sabir 等人 [28] 利用行为识别领域中时间信息处理视频的方法，基于递归卷积网络 (Recursive cortical network, RCN) 提出了一种基于视频流时空特征的检测方法。首先对视频序列进行预处理，人脸检测、人脸裁剪、人脸对齐，然后把每一帧输入到 RNN 网络中，进行端到端的学习。该方法在数据集 FaceForensics++ [1] 上比之前的最好结果提升了 4.55 左右的准确率。

2) 基于视频帧内视觉伪像的检测方法

基于视频帧内视觉伪像的检测方法主要是通过提取视频帧内的判别特征，并将提取到的特征送入深层或浅层分类器中进行训练，从而实现真假数据分类。所以此类方法与深度伪造图像的检测技术是相通的，使得有些方法既可以检测虚假伪造图像，也可以检测虚假伪造视频。其中，深层分类器主要是基于神经网络模型实现，而浅层分类器主要是结合传统机器学习模型实现。1) 深层分类器 Afchar 等人 [19] 经过分析，认为由于视频压缩带来的图像噪声强烈退化，基于图像噪声的微观分析并不会起作用。同时，在高层语义层面，特别是当图像描绘的是人脸时，人眼又很难分辨出伪造的图像。所以作者进行了折中，在介于高层语义信息和低层微观信息之间，在介观分析水平上，提出了少量层的神经网络模型 MesoNet, 包括 Meso-4 和

MesoInception-4。Meso-4 网络由四个连续的卷积神经网络构成，每层后面加上批归一化和最大池化层，最后连接两个全连接层和 sigmoid 层进行分类。而 MesoInception-4 网络把 Meso-4 前面两个卷积层用 Inception 模块进行替代，将几个具有不同卷积核大小的卷积层的输出进行堆叠，从而增加函数优化空间。同时，这种方法在保证了高性能的基础上，参数量也明显少于 ResNet-50[67], XceptionNet [29] 等神经网络结构，构建了一个轻量级的检测网络。作者在自己创建的 deepfake 数据集和 FaceForensics 的 Face2Face 数据集上实验，为了提高泛化和鲁棒性，对输入进行了缩放、旋转、水平翻转、亮度和色调变化等数据增强操作。实验结果如下表所示，单独考虑每一帧，该方法对 Deepfakes 视频的平均检出率为 90，对网络上真实扩散条件下的 Face2Face 视频的平均检出率为 95。Afchar 等人同时也证明了眼睛和嘴巴部位的特征在深度伪造视频检测中具有至关重要的作用。由于目前的深度伪造生成算法只能生成有限分辨率的图像，而且需要将目标人脸通过仿射变换 (如缩放、旋转和剪切等) 匹配到原始视频中，这就会造成合成区域和原始区域之间的分辨率不一致的问题，并在伪造的视频中留下视觉伪像。于是 Li 等人 [30] 利用这一发现，提出了基于 CNN 模型的深度伪造视频检测方法。作者直接通过模拟此类仿射变换的方式简化了负样本的生成过程，提取原始图像的面部区域以及关键点坐标，并从多个尺度实现对齐处理, 再对随机选取的缩放图像应用高斯模糊并将其形变回原始图像，减少了时间消耗和资源消耗，并且具有较好的泛化性能。但是该模型未在大量压缩视频上进行性能评估。并且可能对特定分布的伪造视频过拟合，因此训练数据的多样性需要多方面的提高。2) 浅层分类器将目标人脸拼接到原始人脸的面部区域过程中，会在从二维面部图像估计三维头部姿态（比如头的方向和位置）时引入误差。Yang 等人 [13] 基于这一观察，进行实验来证明了这一现象，并且将这种特征用 SVM 分类器进行分类。作者通过两种方法来估计图像或视频中的头部姿态，一种是用检测得到的 68 个关键点来估计，另一种是只用中心区域的关键点来估计，将两种估计方法得到的头部三维单位向量比较余弦距离，实验证明在真实人脸中两种方法估计得到的余弦距离较为接近 (0-0.02)，但是虚假人脸中两种方法估计得到的余弦距离较远 (0.02-0.08)，这是因为中间的人脸区域和外部的人脸轮廓关键点来自不同的域，所以两者的误差会比较大。因此可以通过这种方法将两种分布

区分开，进而区分出真假数据。基于深度伪造视频部分区域像素关系存在突变性, Koopman 等人 [31] 提出了一种基于光响应非均匀性 (Photo response non-uniformity, PRNU) 的检测方法。PRNU 是一种噪声模式, 噪声源于数码相机的感光传感器的出厂缺陷。每个数字相机的 PRNU 都不相同, 通常被视为数字图像的指纹 [32, 33]。由于被交换面部会改变视频帧中面部区域的局部 PRNU, 所以被广泛应用于面部操作检测。该方法首先将视频转换为帧, 并裁剪有问题的面部区域。然后将裁剪的帧按照顺序分为八个组, 在其中为每个组计算平均 PRNU。为了比较这些组之间的 PRNU, 计算归一化的互相关分数, Koopman 等人创建了一个包含 10 个真实视频和 16 个通过 DeepFaceLab 制造的伪造视频的测试数据集, 分析结果表明深度伪造视频和真实视频的平均标准化互相关系数存在显著差异。但该方法未在较大的数据集上进行测试, 不能够准确区分伪造视频和真实视频之间的互相关性, 也无法确定准确的似然比。基于真实视频和伪造视频之间的帧内伪像或固有特征的区别。最后介绍一种为国家领导人和世界范围的名人（POIs）制定的深度伪造视频检测技术。世界上没有一片树叶是一样的，同样的，Agarwal 等人 [34] 认为每一个人在说话时都会展现出不一样的面部表情和头部运动，这称之为软生物特征模型。但是深度伪造的人物相关视频，则不会存在这样的软生物特征，因此基于这一观察，可以用此特征进行特定人物的真实虚假视频判断。给定一段视频片段，用 OpenFace2 开源工具追踪人脸和头部运动，面部肌肉的运动可以被编码成特定的运动单元（AU），利用 OpenFace2 提供的 AU，生成 20 个特征向量，然后用 Pearson 相关性系数测量向量之间的相似度，进而得到 190 维的特征向量，然后用 SVM 来进行真假数据的区分。

参考文献

[1] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "Faceforensics++: Learning to detect manipulated facial images," in Proceedings of the IEEE International Conference on Computer Vision, pp. 1–11, 2019.

[2] B. Dolhansky, J. Bitton, B. Pflaum, J. Lu, R. Howes, M. Wang, and C. C. Ferrer, "The

deepfake detection challenge dataset," arXiv preprint arXiv:2006.07397, 2020.

[3] L. Jiang, W. Wu, R. Li, C. Qian, and C. C. Loy, "Deeperforensics-1.0: A large-scale dataset for real-world face forgery detection," arXiv preprint arXiv:2001.03024, 2020.

[4] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-df: A new dataset for deepfake forensics," arXiv preprint arXiv:1909.12962, 2019.

[5] B. Dolhansky, R. Howes, B. Pflaum, N. Baram, and C. C. Ferrer, "The deepfake detection challenge (dfdc) preview dataset," arXiv preprint arXiv:1910.08854, 2019.

[6] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1800–1807, 2017.

[7] M. Tan and Q. V. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in International Conference on Machine Learning(ICML), pp. 6105–6114, 2019.

[8] S. McCloskey and M. Albright, "Detecting gan-generated imagery using color cues," 2018.

[9] N. Yu, L. Davis, and M. Fritz, "Attributing fake images to gans: Analyzing fingerprints in generated images," arXiv preprint arXiv:1811.08180, vol. 2, 2018.

[10] R. Wang, F. Juefei-Xu, L. Ma, X. Xie, Y. Huang, J. Wang, and Y. Liu, "Fakespotter: A simple yet robust baseline for spotting ai-synthesized fake faces," 2019.

[11] H. Dang, F. Liu, J. Stehouwer, X. Liu, and A. K. Jain, "On the detection of digital face manipulation," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2020.

[12] P. Zhou, X. Han, V. I. Morariu, and L. S. Davis, "Two-stream neural networks for tampered face detection," in 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2017.

[13] X. Yang, Y. Li, and S. Lyu, "Exposing deep fakes using inconsistent head poses," in ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 8261–8265, IEEE, 2019.

[14] A. Bharati, R. Singh, M. Vatsa, and K. W. Bowyer, "Detecting facial retouching using supervised deep learning," IEEE Transactions on Information Forensics and Security, vol. 11, no. 9, pp. 1903–1913, 2016.

[15] A. Jain, R. Singh, and M. Vatsa, "On detecting gans and retouching based synthetic alterations," in 2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS), pp. 1–7, 2018.

[16] F. Matern, C. Riess, and M. Stamminger, "Exploiting visual artifacts to expose deepfakes and face manipulations," in 2019 IEEE Winter Applications of Computer Vision Workshops (WACVW), 2019.

[17] X. Zhang, S. Karaman, and S. Chang, "Detecting and simulating artifacts in gan fake images," in 2019 IEEE International Workshop on Information Forensics and Security (WIFS), 2019.

[18] L. Nataraj, T. Manhar Mohammed, S. Chand rasekaran, A. Flenner, J. H. Bappy, A. K. Roy-Chowdhury, and B. S. Manjunath, "Detecting GAN generated Fake Images using Co-occurrence Matrices," arXiv e-prints, p. arXiv:1903.06836, Mar. 2019.

[19] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "Mesonet: a compact facial video forgery detection network," in 2018 IEEE International Workshop on Information Forensics and Security (WIFS), pp. 1–7, IEEE, 2018.

[20] H. H. Nguyen, F. Fang, J. Yamagishi, and I. Echizen, "Multi-task learning for detecting and segmenting manipulated facial images and videos," 2019.

[21] S. Tariq, S. Lee, H. Kim, Y. Shin, and S. S. Woo, "Detecting both machine and human created fake face images in the wild," in Proceedings of the 2nd international workshop on multimedia privacy and security, 2018.

[22] Y. Li, M.-C. Chang, and S. Lyu, "In ictu oculi: Exposing ai generated fake face videos by detecting eye blinking," arXiv preprint arXiv:1806.02877, 2018.

[23] A. Khodabakhsh, R. Ramachandra, K. Raja, P. Wasnik, and C. Busch, "Fake face detection methods: Can they be generalized?," in 2018 International Conference of the Biometrics Special Interest Group (BIOSIG), pp. 1–6, IEEE, 2018.

[24] P. Korshunov and S. Marcel, "Deepfakes: a new threat to face recognition? assessment and detection," arXiv preprint arXiv:1812.08685, 2018.

[25] C. Sanderson and B. C. Lovell, "Multi-region probabilistic histograms for robust and scalable identity inference," in International conference on biometrics, pp. 199–208, Springer, 2009.

[26] D. Güera and E. J. Delp, "Deepfake video detection using recurrent neural networks," in 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), pp. 1–6, IEEE, 2018.

[27] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2625–2634, 2015.

[28] E. Sabir, J. Cheng, A. Jaiswal, W. AbdAlmageed, I. Masi, and P. Natarajan, "Recurrent convolutional strategies for face manipulation detection in videos," Interfaces (GUI), vol. 3, no. 1, 2019.

[29] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778, 2016.

[30] Y. Li and S. Lyu, "Exposing deepfake videos by detecting face warping artifacts," arXiv preprint arXiv:1811.00656, 2018.

[31] M. Koopman, A. M. Rodriguez, and Z. Geradts, "Detection of deepfake video manipulation," in Conference: IMVIP, 2018.

[32] J. Lukas, J. Fridrich, and M. Goljan, "Digital camera identification from sensor pattern noise," IEEE Transactions on Information Forensics and Security, vol. 1, no. 2, pp. 205–214, 2006.

[33] K. Rosenfeld and H. T. Sencar, "A study of the robustness of prnu-based camera identification," in Media Forensics and Security, vol. 7254, p. 72540M, International Society for Optics and Photonics, 2009.

[34] S. Agarwal, H. Farid, Y. Gu, M. He, K. Nagano, and H. Li, "Protecting world leaders against deep fakes.," in CVPR Workshops, pp. 38–45, 2019.