



北京大学

# 数字媒体软件与系统开发期末作业

题目：深度伪造与检测期末报告

姓 名：干皓丞

学 号：2101212850

院 系：信息工程学院

专 业：计算机应用技术

研究方向：通信及信息安全技术

导 师：王荣刚 教授

二〇二二年五月



## 摘要

在近年大数据、人工智能等计算机学科的蓬勃发展下，其机器学习领域中的深度学习已经大范围的成功应用于许多从大数据分析到计算机视觉等各种复杂问题。同样的深度学习等演算法的新兴领域也有可能会被使用在造成隐私、民主和严重的国家安全造成威胁的用途上。近期出现的基于深度学习影响最大的应用之一是"Deepfake"，而所谓的 Deepfake 演算法可以创造出人类用肉眼也不容易辨别出真假的影像与照片。因此，在面对个困境来说能够进行自动检测和评估影像、图片、语音等媒体完整性的技术的研究与讨论是必不可少的过程。本作业先说明介绍了人工智能、深度学习与用于深度伪造与检测的背景，第二章再来说明其研究跟工具的分类、第三章则说明目前当下可用的资料集与素材、第四章则将该领域的研究进行归纳整理、第五章则对近来的研究进行说明，最后将这些调查工作进行总结。

该作业所进行得调研工作於此 GitHub 项目：

<https://github.com/kancheng/kan-cs-report-in-2022/tree/main/DMSASD/final>

关键词：深度伪造、深度伪造的检测



## 目录

<b>第一章 深度伪造与检测的背景</b>	1
1.1 深度伪造技术	1
1.1.1 针对人类脸部的伪造技术	2
1.1.2 人类脸部的表情伪造	9
1.1.3 人类语音的伪造	13
1.1.4 现有开源工具	15
<b>第二章 深度伪造的资料集</b>	17
2.1 深度伪造的影像资料集	17
2.2 深度伪造的语音资料集	18
<b>第三章 深度伪造的研究工具与检测手段的分类</b>	19
3.1 深度伪造的图像检测与人脸视频检测整理	19
3.2 过往传统的检测手段	20
3.3 人类的生理特征的检测手段	27
3.4 图像伪造后留下痕迹的检测手段	29
3.5 GAN 模型所产生的检测手段	30
3.6 数据驱动的深度学习之检测手段	34
3.7 图片的深度学习之检测手段	35
3.8 影像的深度学习之检测手段	41
3.9 人类语音伪造检测手段	42
3.10 代表性检测技术整理与比较	45
<b>第四章 深度伪造的对抗性研究與近期研究發展</b>	49
4.1 深度伪造生成的对抗性	49
4.2 深度伪造检测的对抗性	51
4.3 GAN	54
4.4 Transformers 與增量学习	54
<b>第五章 结论</b>	55
<b>参考文献</b>	57
<b>致谢</b>	67



## 主要符号对照表

$x, y, m, n, t$	标量, 通常为变量
$K, L, D, M, N, T$	标量, 通常为超参数
$x \in \mathbb{R}^D$	D 维列向量
$(x_1, \dots, x_D)$	D 维行向量
$(x_1, \dots, x_D)^T$ or $(x_1; \dots; x_D)^T$	D 维行向量
$x \in \mathbb{R}^{KD}$	( $KD$ ) 维的向量
$\mathbb{M}_i$ or $\mathbb{M}_i(x)$	第 $i$ 列为 $\mathbf{1}$ (或者 $x$ ), 其余为 $\mathbf{0}$ 的矩阵
$diag(\mathbf{x})$	对角矩阵, 其对角元素为 $x$
$I_N$ or $I$	( $N \times N$ ) 的单位阵
$A \in \mathbb{R}^{D_1 \times D_2 \times \dots \times D_K}$	大小为 $D_1 \times D_2 \times \dots \times D_K$ 的张量
$\{x^{(n)}\}_{n=1}^N$	集合
$\{(x^{(n)}, y^{(n)})\}_{n=1}^N$	数据集
$\mathcal{N}(x; \mu, \Sigma)$	变量 $x$ 服从均值为 $\mu$ , 方差为 $\Sigma$ 的高斯分布

① 本符号对照表内容选自邱锡鹏老师的《神经网络与深度学习》<sup>[1]</sup>一书。



## 第一章 深度伪造与检测的背景

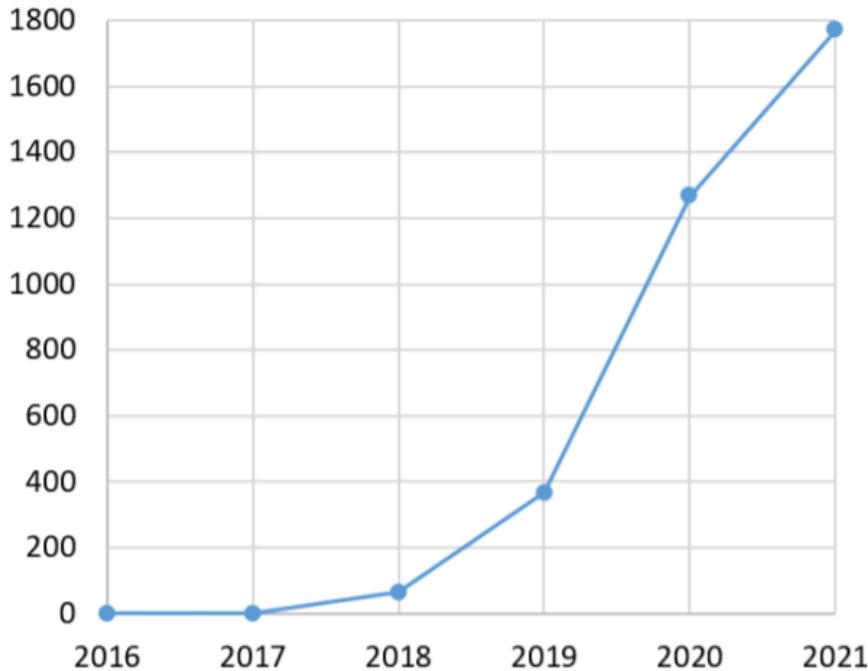
近年来人工智能与深度学习等计算机学科的蓬勃发展，连带也造成的不同研究领域与多样的议题需要进行研讨，比如讨论在以人工智慧应有的社会治理架构下，讨论其机器学习领域等演算法等对于法学所造成的挑战<sup>[2]</sup>，同时另一篇研究也讨论其人工智能与演算法在法律上应用的可能<sup>[3]</sup>。而在机器学习领域下的深度伪造技术则发展有越来越广泛的趋势，而其深度伪造属于机器学习下的深度学习的一部分，而深度学习已经广泛应用于各个领域，其领域包含了计算机视觉与自然语言处理，而本作业则关注当中快速兴起的则是深度伪造的领域，其深度伪造技术<sup>[4]</sup>虽造成了风险，但同时在这些年也有需多研究去分析深度伪造的工作原理，并且引入了许多基于深度学习的方法来检测深度伪造的影像或图像。

综上所述这些技术好的部分则是应用于将古老的照片变成动态的影像，或者是用于一些艺术与网路次文化的创作，又或者是 Reface APP<sup>[5]</sup> 等服务带给大众娱乐，但同时负面的因素也有造成社会动荡的可能，包含知名人士被伪造影像、进而被广泛散布不实资讯与谣言的危险，造成当事人的声誉、社会地位与事业严重打击，还有近期大量知名女性人士被用于成人色情网站，而受害当事人却因执法基层人员不理解相对应技术亦或是没有完善的检测工具，因而在此方面无法给予有效的协助，而受害人在描述其受害过程时受到再次的心理创伤，同时其深度伪造的假影片在网路散布时，对当事人的伤害就已经造成。再者此技术近期被应用在战争宣传战，将敌对方的政治人物伪造出用于其不正确的政治发言影片，进而造成某方的士气遭到打击。

另外需要注意的在于这些工具非常容易取得，与之相应的是一些相继机构发现这些问题后，进而举办相应的竞赛<sup>[6]</sup>，来推广该技术在此领域的热度。所以本作业即目标即探讨在人工智能下的深度学习领域中深度伪造与其检测的研究整理，同时调研过去 Girish N 等人所汇整的早期图像篡改工作<sup>[7]</sup>与 Nguyen TT 等人对该领域工作地早期总结<sup>[8]</sup>与 Li XR 等人近期来的汇整工作<sup>[9]</sup>与研读，同时对使用深度学习方法下深度伪造与相对应的前沿检测技术进行调查，并对目前最新的研究进行补充。

### 1.1 深度伪造技术

目前对深度伪造技术在视觉上所修改后的影像与图片，其大多是针对人类脸部的替换。而在此大致分为两大部分，其一为对人类的人脸表情进行伪造，让指定窜改者所改造的对象做到窜改者想要的脸部表情与动作，但不对该人脸进行目标人脸的替换，另一类则是根据两个不同影像与影片的人脸进行替换，经过将另一个完全不同身份的



**Fig. 1.** Number of papers related to deepfakes in years from 2016 to 2021, obtained from <https://app.dimensions.ai> at the end of 2021 with the search keyword “deepfake” applied to full text of scholarly papers.

图 1.1 从 Nguyen TT 等人<sup>[8]</sup>的工作中可以看到进来深度伪造的发展

人脸替换过去，从而达到该内容目标人物是窜改者所要之人。该技术从过往运用的三维重建技术等方法来修改之外，一路发展到运用深度学习的方法至今则用生成对抗网路为基础进行仿造，比如 Almarsi, A. M 等人在该领域工作汇整之一的 CycleGAN，此方法为无监督方法，它提取一张图像的特征，并通过 GAN 架构生成另一张具有相同特征的图像。该方法应用循环损失函数，使他们能够学习潜在特征，且该方法应用循环损失函数，使他们能够学习潜在特征，可以在不使用配对示例的情况下执行图像到图像的转换。换句话说，该模型从源和目标中学习不需要相互关联的图像集合的特征。而更重要的是目前的仿造技术还运用人类语音的修改，从而导致伪造出来的影像结果会更逼真。

### 1.1.1 针对人类脸部的伪造技术

#### 1.1.1.1 过往根据图形学所进行的脸部伪造技术

在过往几年来使用图形学来对人类的脸部进行替换和仿造的技术，一直被很多研究者持续的关注，而在 Zollhöfer M 等人对其领域进行调研总结的工作则说明地当下几个主要根据三维模型重建与追踪再该领域技术上的应用。该研究将讨论重点放在中心

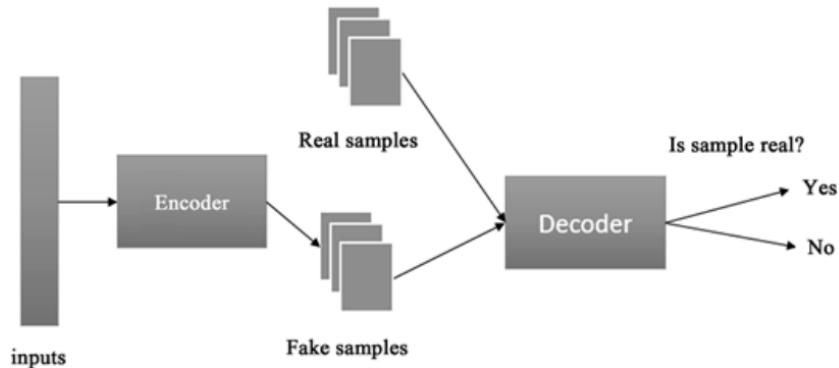


图 1.2 Almars, A. M 等人整理的 GAN 深度伪造示意

任务是使用基于优化的重建算法来恢复和跟踪人脸的三维模型的方法上，同时对现实世界图像形成的基本概念进行了深入的概述，并讨论了使这些算法实用的常见假设和简化。此外，该研究广泛涵盖了用于更好地约束欠约束单目重建问题的先验，并讨论了用于从单目 2D 数据中恢复密集的照片几何 3D 人脸模型的优化技术。最后，在动作捕捉、面部动画以及图像和视频编辑的背景下讨论了所审查算法的各种用例。

而 FaceSwap<sup>[10]</sup> 是一个根据图形学的人脸替换方法，该应用是 Marek Kowalski 于华沙理工大学就读多媒体数学时，所做的练习成果，其应用程序是用 Python 编写的，并使用人脸对齐、高斯牛顿优化和图像混合来将相机看到的人脸与提供的图像中的人脸交换。同时该应用的新版本则基于深度对齐网络方法，如果在 GPU 上运行，它比当前使用的方法更快，并且提供更稳定和更精确的面部标志。另外 Dale K 等人<sup>[11]</sup>提出了一种替换视频中人类脸部的方法，该研究的方法考虑了源视频和目标视频之间在身份、视觉外观、语音和时间方面的差异。该研究与以前的工作不同，它不需要大量的手动操作或复杂的采集硬件，只需要单机视频，研究者使用 3D 多线性模型来跟踪两个视频中的面部表现，使用相应的 3D 几何，最后将源扭曲到目标面并重新定时源以匹配目标性能。然后，研究者通过视频体积计算最佳接缝，以保持最终合成中的时间一致性。

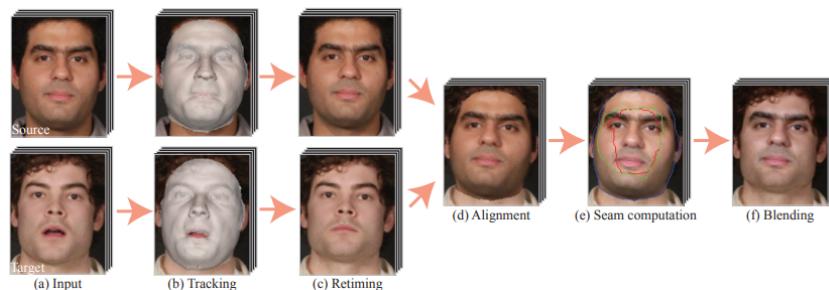


图 1.3 Dale K 等人一种基于图像的面部重建系统

Garrido P<sup>[12]</sup> 等人则研究提出了一种基于图像的面部重建系统，该系统将现有目标

视频中的演员面部替换为源视频中用户的面部，同时保留原始目标表现，其系统是全自动的，不需要源表达式数据库。相反，它能够从使用现成相机（例如网络摄像头）捕获的短源视频中产生令人信服的重演结果，用户在其中执行任意的面部表情，研究者的重演流程被设想为部分图像检索和部分面部转移：图像检索基于目标帧的时间聚类和一种新颖的图像匹配度量，该度量结合了外观和运动以从源视频中选择候选帧，而面部转移使用保留用户身份的 2D 变形策略。其系统在简单性方面表现出色，因为它不依赖于 3D 人脸模型，它在头部运动下很稳健，并且不需要源和目标性能相似。

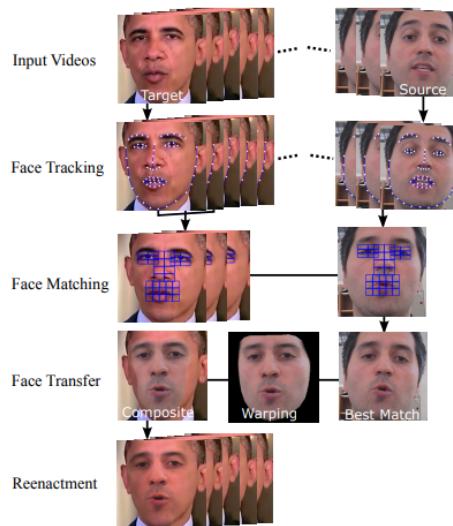


Figure 1. Overview of the proposed system.



Figure 3. Comparison of warping approaches. Left: Selected user frame. Right: Target pose. Middle left to right: non-rigid warping (Eq. (5)), affine warping (Eq. (6)), and our approach (Eq. (7)).

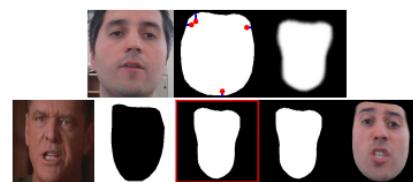


Figure 4. Seam generation. Top: User at rest, source mask with landmarks closest to the boundary in red, and eroded mask. Bottom left: Target frame and mask. Bottom Right: Transferred source frame and mask. Bottom middle: Final blending seam.

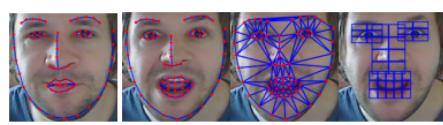
Figure 2. (a) Annotated reference frame. (b) Expressive face aligned to the reference. Left to right: estimated landmarks, triangulation, and detected regions of interest. The mouth, eyes and nose regions are split into  $3 \times 5$ ,  $3 \times 2$  and  $4 \times 2$  tiles, respectively.

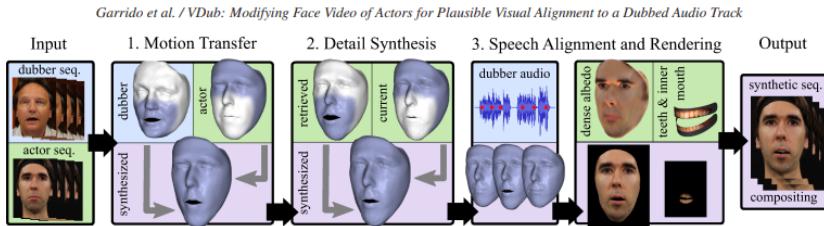
图 1.4 Garrido P 等人一种基于图像的面部重建系统

同样也是 Garrido P 等人<sup>[13]</sup>，考虑到在许多国家，外国电影和电视作品被配音，即演员的原声被配音演员用该国自己的语言所说的翻译代替，配音是一个复杂的过程，需要特定的翻译和准确定时的朗诵，以使新音频至少粗略地贴合视频中的嘴巴动作。然而，由于原作和配音语言中的音素和视位序列不同，视频与音频的匹配永远不会完美，这是视觉不适的主要来源，在本文中，研究者提出了一种系统来改变视频中演员的嘴巴动作，使其与新的音轨相匹配。其研究建立在对配音和目标演员的 3D 面部表演、照明和反照率的高质量单目捕捉的基础上，并结合使用音频分析和时空检索方法来合成一个新的照片般逼真的渲染和高度详细的 3D 形状嘴区域模型来替换目标性能。

而 Nirkin Y 等人<sup>[14]</sup>的研究让我们知道即使人脸图像不受约束且任意配对，它们之



**Figure 1:** We modify the lip motion of an actor in a target video (a) so that it aligns with a new audio track. Our set-up consists of a single video camera that films a dubber in a recording studio (b + c). Our system transfers the mouth motion of the voice actor (d) to the target actor and creates a new plausible video of the target actor speaking in the dubbed language (e).



**Figure 2:** Overview of our method

图 1.5 Garrido P 等人提出了一种系统来改变视频中演员的嘴部动作，使其与新的音轨相匹配

间的人脸交换实际上也非常简单。为此，该研究做出以下贡献。(a) 没有像其他人之前提出的那样为人脸分割定制系统，而是展示了标准的全卷积网络 (FCN) 可以实现非常快速和准确的分割，前提是它在足够丰富的示例集上进行训练。为此，描述了新的数据收集和生成例程，这些例程提供了具有挑战性的分割人脸示例。(b) 使用该研究的分割在前所未有的条件下实现强大的面部交换。(c) 与以前的工作不同，该研究的交换足够强大，可以进行广泛的定量测试。为此，研究者使用野外标记人脸 (LFW) 基准测试并测量对象内和对象间人脸交换对识别的影响。研究表明，其受试者内部交换的面孔仍然与其来源一样可识别，证明了我们方法的有效性。与众所周知的感知研究一致，而更好的面部交换会产生不太可识别的主体间结果。这是第一次在机器视觉系统中定量证明这种效果。

### 1.1.1.2 现在根据深度学习所进行的脸部伪造技术

由于过往图形学在面对伪造人类脸部技术有着极大的成本等诸多因素，从而导致该技术很难普遍的进行应用。然而自进入人工智能与机器学习所带动的深度学习热潮下，深度伪造技术在此之后有着非常快速的进展，此时许多研究者们开始关心其深度学习在人类脸部进行替换等应用技术。比如 Lu Z 等人<sup>[15]</sup>则该研究领域所涉及传统方法和高级深度学习方法的典型人脸合成工作进行了全面回顾。特别是，Generative Adversarial Net (GAN) 被突出显示以生成照片般逼真和身份保持的结果。此外，还详细介绍了公开可用的数据库和评估指标。

当中 FaceSwap<sup>[4]</sup> 是较早的一种利用深度学习来识别和交换图片和视频中的人脸的



Figure 1: *Inter-subject swapping*. LFW G.W. Bush photos swapped using our method onto very different subjects and images. Unlike previous work [4, 19], we do not select convenient targets for swapping. Is Bush hard to recognize? We offer quantitative evidence supporting Sinha and Poggio [40] showing that faces and context are both crucial for recognition.

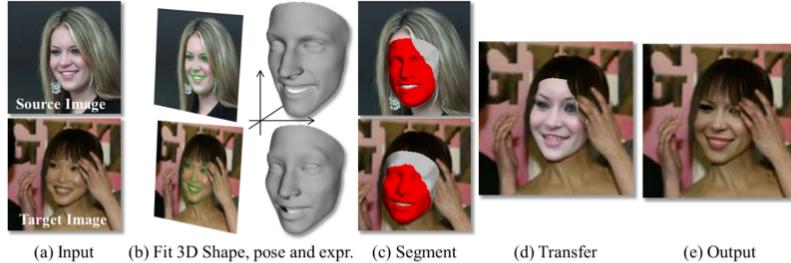


Figure 2: *Overview of our method*. (a) Source (top) and target (bottom) input images. (b) Detected facial landmarks used to establish 3D pose and facial expression for a 3D face shape (Sec. 3.1). We show the 3DMM regressed by [42] but our tests demonstrate that a generic shape often works equally well. (c) Our face segmentation of Sec. 3.2 (red) overlaid on the projected 3D face (gray). (d) Source transferred onto target without blending, and the final results (e) after blending (Sec. 3.3).

图 1.6 Nirkin Y 用分割的思路促进换脸



图 1.7 FaceSwap 的 Jennifer Lawrence/Steve Buscemi FaceSwap using the Villain model

工具的 GitHub 开源项目，为具有多平台 Deepfakes 软件，其技术由 Tensorflow、Keras 和 Python 提供支持，并在 Windows、macOS 和 Linux 上运行。

该原理为在训练模型期间，去训练两个有共享权重参数的自动编码器，并让其两编码器获得有人类脸部重建效果的能力，其后在所谓换脸阶段，交换两解码器，使之有人类脸部交换的效果。重点在于此行为仅仅需要被篡改目标人物的脸部图像与原人物脸部图像即可进行模型的训练，与过往的成本相比，已大大降低其交换人类脸部成本，但此过程也需要一定的训练模型的基础知识与技术，不然其生成器所产生的品质是很可能达不到理想的成果，此 Deepfakes 生成概念可以从 Li XR 等人<sup>[9]</sup>所整理的工作总结看到。由于上述诸多原因，研究者们开始注意在生成对抗模型上的混合应用，

而所谓的生成对抗模型是源于 Goodfellow, I 等人<sup>[16]</sup>的工作，其研究者提出了一个通过对抗过程估计生成模型的新框架，该研究同时训练两个模型：一个生成模型 G 捕获数据分布，一个判别模型 D 估计样本来自训练数据而不是 G 的概率。G 的训练过程是最大化 D 出错的概率。这个框架对应于一个极小极大的两人游戏，在任意函数 G 和 D 的空间中，存在唯一解，G 恢复训练数据分布，D 处处等于 1/2。在 G 和 D 由多层

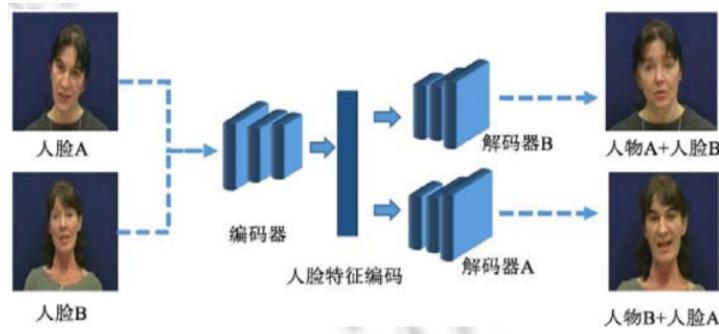

 Fig.1 Framework for Deepfakes generation<sup>[1]</sup>

 图 1 Deepfakes 生成框架<sup>[1]</sup>

图 1.8 Deepfakes 框架

感知器定义的情况下，整个系统可以通过反向传播进行训练。在训练或生成样本期间，不需要任何马尔可夫链或展开的近似推理网络，实验通过对生成的样本进行定性和定量评估，证明了框架的潜力。而所谓的 Faceswap-GAN 混合 GAN 的深部伪造技术，其模型加入了判别器的对抗损失函数，另外再生成的部分对其生成的图像与原本的图像进行相似程度判断，该过程的好处在于会让生成的图像品质提高，且此模型也为了眼珠转动的部分加入了对应的感知损失函数。综上所述，我们可以知道因为 GAN 模型加入了该领域的发展后，使深度伪造技术的成本降低，仿造人类的换脸技术因此变得更加逼真，同时也使此技术流行起来。



图 1.9 Faceswap-GAN 展示了 v2 GAN 生成的 160 个随机结果，带有自注意力机制

而 Korshunova 等人<sup>[17]</sup> 将所谓的深度伪造的换脸问题视为风格迁移问题，他们训练一个卷积神经网络，运用对非结构化的图像进行学习，其研究考虑图像中的人脸交换问

题，其中输入身份被转换为目标身份，同时保留姿势、面部表情和照明，为了执行这种映射，研究者使用经过训练的卷积神经网络从他/她的照片的非结构化集合中捕获目标身份的外观，此方法是通过在风格转移方面构建人脸交换问题来实现的，其目标是用另一个风格来渲染图像。基于该领域的最新进展，研究者设计了一种新的损失函数，使网络能够产生高度逼真的结果，通过将神经网络与简单的预处理和后处理步骤相结合，研究者的目标是让面部交换实时工作，无需用户输入。另外 Nirkin Y 等人<sup>[18]</sup>则提出了人脸交换 GAN (FSGAN) 用于人脸交换和重演，与以前的工作不同，FSGAN 与主题无关，可以应用于成对的面孔，而无需对这些面孔进行培训。为此，研究者描述了一些技术贡献，其推导出了一种新颖的基于循环神经网络 (RNN) 的面部重演方法，该方法可针对姿势和表情变化进行调整，并可应用于单个图像或视频序列，对于视频序列，我们引入了基于重演、Delaunay 三角剖分和重心坐标的人脸视图的连续插值，而被遮挡的人脸区域由人脸补全网络处理。最后，研究者使用人脸融合网络无缝融合两张脸，同时保留目标肤色和光照条件。该网络使用一种新颖的泊松混合损失，它将泊松优化与感知损失相结合。该研究将其方法与现有的最先进系统进行比较，并显示研究的结果在质量和数量上都优越。

同时也可以看到单纯窜改人类脸部属性与虚拟的人类脸部的 GAN 技术应用。比如 Choi Y 等研究者提出了 StarGAN<sup>[19]</sup>，这是一种新颖且可扩展的方法，可以仅使用单个模型为多个域执行图像到图像的转换，StarGAN 的这种统一模型架构允许在单个网络中同时训练具有不同域的多个数据集。与现有模型相比，这导致 StarGAN 具有卓越的翻译图像质量，以及将输入图像灵活翻译到任何所需目标域的新颖能力，研究者凭经验证明了其方法在面部属性转移和脸部表情合成任务上的有效性。Zhang H 等人提出的 Stackgan<sup>[20]</sup>，其研究者提出了一种用于文本到图像合成的两阶段生成对抗网络架构 StackGAN-v1，Stage-I GAN 根据给定的文本描述勾勒出对象的原始形状和颜色，产生低分辨率图像，Stage-II GAN 将 Stage-I 结果和文本描述作为输入，并生成具有照片般逼真细节的高分辨率图像，其次，针对有条件和无条件的生成任务提出了一种先进的多阶段生成对抗网络架构 StackGAN-v2。研究者的 StackGAN-v2 由树状结构中的多个生成器和判别器组成；从树的不同分支生成对应于同一场景的多个尺度的图像，StackGAN-v2 通过联合逼近多个分布，显示出比 StackGAN-v1 更稳定的训练行为。大量实验表明，所提出的堆叠生成对抗网络在生成照片般逼真的图像方面明显优于其他最先进的方法。Karras T 等人所提出 PGAN<sup>[21]</sup> 的描述了一种用于生成对抗网络的新训练方法。关键思想是逐步增长生成器和判别器：从低分辨率开始，研究者添加了新的层，随着训练的进行，对越来越精细的细节进行建模。这既加快了训练的速度，又极大地稳定了训练，使研究者能够生成质量前所未有的图像，例如 1024<sup>2</sup> 的 CelebA 图像，

其研究者还提出了一种简单的方法来增加生成图像的变化，并在无监督 CIFAR10 中实现 8.80 的创纪录初始分数。此外，研究者描述了几个实现细节，这些细节对于阻止生成器和判别器之间的不健康竞争很重要。最后，该研究建议在图像质量和变化方面评估 GAN 结果的新指标，作为额外的贡献，研究者构建了更高品质的 CelebA 资料集版本。这些一系列的 GAN 技术也可以生成虚拟人物的人类面部图像。

另外则是 Grigory 等人<sup>[22]</sup>运用 Mirza M 等人的 conditional-GAN<sup>[23]</sup>，其在这项工作中，Mirza M 等研究者介绍了生成对抗网络的条件版本，它可以通过简单地输入数据  $y$  来构建，我们希望同时对生成器和判别器进行条件处理。该研究展示了该模型可以生成以类标签为条件的 MNIST 数字。其研究还说明了该模型如何用于学习多模态模型，并提供了图像标记应用的初步示例，其中研究者演示了该方法如何生成不属于训练标签的描述性标签。Grigory 等人则在此基础上提出了基于 GAN 的自动人脸老化方法。与以前使用 GAN 来改变面部属性的工作相反，研究者特别强调在他/她的面部老化版本中保留原始人的身份。为此，该研究引入了一种新的方法来优化 GAN 的潜在向量的“身份保持”。通过最先进的人脸识别和年龄估计解决方案对产生的老化和恢复活力的人脸图像进行客观评估，证明了所提出方法的巨大潜力。还有 Huang R 等人<sup>[24]</sup>提出了一种双路径生成对抗网络 (TP-GAN)，用于通过同时感知全局结构和局部细节来进行逼真的正面视图合成。除了常用的全局编码器-解码器网络之外，还提出了四个具有里程碑意义的补丁网络来处理局部纹理。除了新颖的架构外，研究者通过引入对抗性损失、对称性损失和身份保持损失的组合来很好地约束这个不适定问题，组合的损失函数利用正面人脸分布和预训练的判别式深度人脸模型来指导从侧面看正面视图的身份保持推断。与之前主要依赖中间特征进行识别的深度学习方法不同，该研究的方法直接利用合成的身份保持图像来完成人脸识别和属性估计等下游任务，其实验结果表明，该研究的方法不仅提供了令人信服的感知结果，而且在大姿势人脸识别方面也优于最先进的结果。综上所述因为 GAN 技术的应用于深度伪造领域使其成果越来越真实，从而引来该技术滥用在负面用途开端。

### 1.1.2 人类脸部的表情伪造

所謂的人类脸部的表情伪造在於不去對該目標人物去進行替換，讓其他不同臉部表情的特徵來改變目標人物的表情。

比如 Thies J 等人<sup>[25]</sup> 所提出的了一种将面部表情从源视频中的演员实时传输到目标视频中的演员的方法，从而能够对目标演员的面部表情进行临时控制。其方法的新颖之处在于将面部变形和细节的转移和逼真的重新渲染到目标视频中，新合成的表情与真实视频几乎没有区别。为了实现这一点，该研究使用商品 RGB-D 传感器实时准确

地捕捉源对象和目标对象的面部表现，对于每一帧，研究者将身份、表情和皮肤反射率的参数模型与输入颜色和深度数据联合拟合，并重建场景照明。对于表达式转移，该研究计算参数空间中源表达式和目标表达式之间的差异，并修改目标参数以匹配源表达式。一个主要挑战是将合成的目标人脸重新渲染到相应的视频流中，这需要仔细考虑照明和阴影设计，两者都必须与现实世界环境相对应。研究者在现场设置中演示了其方法，并修改了视频会议来源，以便实时匹配不同人（例如翻译）的面部表情。另外也是 Thies J 等人<sup>[26]</sup>提出了 Face2Face，这是一种用于对单目目标视频序列（例如 Youtube 视频）进行实时面部重演的新颖方法，源序列也是单目视频流，使用商品网络摄像头实时捕获。其目标是通过源演员为目标视频的面部表情制作动画，并以照片般逼真的方式重新渲染经过处理的输出视频。为此，研究者首先通过基于非刚性模型的捆绑解决了从单目视频中恢复面部身份的约束不足问题。同时运行时使用密集的光度一致性测量来跟踪源视频和目标视频的面部表情。然后通过源和目标之间的快速有效的变形传递来实现重演，从目标序列中检索与重新定位的表情最匹配的嘴巴内部，并对其进行扭曲以产生准确的拟合。最后，研究者令人信服地在相应的视频流之上重新渲染合成的目标人脸，使其与现实世界的照明无缝融合。另外还提出 HeadOn 这是第一个实时的源到目标重演方法，用于完整的人类肖像视频，可以传输躯干和头部运动、面部表情和眼睛注视。给定目标演员的简短 RGB-D 视频，其研究者自动构建个性化几何代理，该代理嵌入参数化头部、眼睛和运动学躯干模型。而一种新颖的实时重演算法使用此代理将捕获的运动从源演员逼真地映射到目标演员。在粗略的几何代理之上，研究者提出了一种基于视频的渲染技术，该技术通过与视图和姿势相关的纹理合成修改后的目标肖像视频，并在新颖的躯干和头部姿势下创建目标演员的照片般逼真的图像，面部表情和注视方向。为此，研究者建议对源演员的面部和躯干进行稳健的跟踪。

Kim H 等人<sup>[27]</sup>提出了一种新颖的方法，该方法仅使用输入视频就可以对肖像视频进行逼真的重新动画处理。与仅限于面部表情操作的现有方法相比，研究者率先将完整的 3D 头部位置、头部旋转、面部表情、眼睛注视和眨眼从源演员转移到目标的肖像视频演员，其方法的核心是具有新颖时空架构的生成神经网络。该网络将参数化人脸模型的合成渲染作为输入，并据此预测给定目标演员的照片般逼真的视频帧，这种从渲染到视频的传输的真实性是通过仔细的对抗训练来实现的，因此，研究者可以创建修改后的目标视频，以模仿合成创建的输入的行为。为了实现源到目标视频的重新动画，研究者使用从源视频中重建的头部动画参数渲染合成目标视频，并将其输入到训练好的网络中，从而完全控制目标，凭借自由重组源参数和目标参数的能力，研究者能够演示各种视频重写应用程序，而无需明确建模头发、身体或背景。例如，该研究可以使用交互式用户控制的编辑来重新制作完整的头部，并实现高保真视觉配音。为

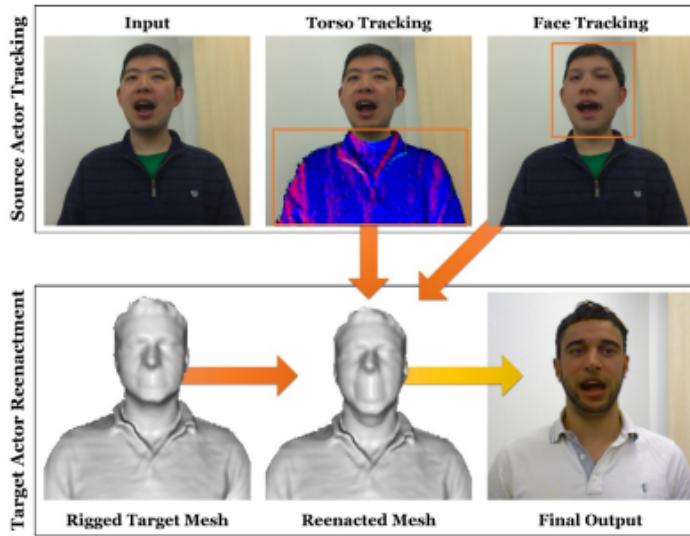


Fig. 2. Overview of our proposed *HeadOn* technique. Based on the tracking of the torso and the face of the source actor, we deform the target actor mesh. Using this deformed proxy of the target actor's body, we use our novel view-dependent texturing to generate a photo-realistic output.

图 1.10 HeadOn

了证明其成果输出的高质量，研究者进行了一系列广泛的实验和评估，例如，用户研究表明该研究的视频编辑很难检测到。

Thies J 等人<sup>[28]</sup>探索了不完美的 3D 内容的使用，例如，从具有噪声和不完整表面几何的光度重建中获得，同时仍然旨在产生照片般逼真的（重新）渲染。为了解决这个具有挑战性的问题，研究者引入了延迟神经渲染，这是一种将传统图形管道与可学习组件相结合的图像合成新范例。具体来说，该研究提出了神经纹理，它是作为场景捕获过程的一部分进行训练的学习特征图。与传统纹理类似，神经纹理作为贴图存储在 3D 网格代理之上；然而，高维特征图包含更多信息，可以通过我们新的延迟神经渲染管道进行解释，神经纹理和延迟神经渲染器都是端到端训练的，即使原始 3D 内容不完美，我们也能够合成照片般逼真的图像。与传统的黑盒 2D 生成神经网络相比，研究者的 3D 表示使我们能够明确控制生成的输出，并允许广泛的应用领域。例如，该研究可以合成时间一致的视频重新渲染记录的 3D 场景，因为研究者表示固有地嵌入在 3D 空间中。通过这种方式，可以利用神经纹理在静态和动态环境中以实时速率连贯地重新渲染或操作现有视频内容。该研究在几个关于新视图合成、场景编辑和面部重演的实验中展示了其方法的有效性，并与利用标准图形管道和传统生成神经网络的最先进方法进行了比较。

Suwajanakorn S 等人<sup>[29]</sup>提出了根据循环神经网络来建立人类面部嘴型与人类语音资料的对应，并利用美国总统巴拉克奥巴马的音频，研究者合成了一段高质量的他讲

话的视频，并具有准确的口型同步，并合成到目标视频剪辑中，一个循环神经网络在他每周的演讲视频中经过数小时的训练，可以学习从原始音频特征到嘴形的映射。给定每个时刻的嘴巴形状，研究者合成高质量的嘴巴纹理，并将其与适当的 3D 姿势匹配合成，以改变他在目标视频中所说的内容，以匹配输入音轨。其方法产生逼真的结果。

Zakharov E 等人<sup>[30]</sup>研展示了一个具有人物特写镜头画面的图像合成。它在大型视频数据集上执行冗长的元学习，然后能够将以前看不见的人的神经说话头模型的少量和一次性学习构建为具有高容量生成器和鉴别器的对抗性训练问题。至关重要的是，该系统能够以特定于人的方式初始化生成器和判别器的参数，因此尽管需要调整数千万个参数，但训练可以仅基于几张图像并快速完成。其研究表明，这种方法能够学习高度逼真和个性化的新人说话头部模型，甚至是肖像画。

类似的还有 Fried O 等人<sup>[31]</sup> 提出了一种新颖的方法来编辑说话头视频的脚本，以生成逼真的输出视频，其中说话者的对话已被修改，同时保持无缝的视听流（即没有跳转）。其方法使用音素、视位、3D 面部姿势和几何形状、反射率、表情和每帧场景照明自动注释输入的说话头视频，要编辑视频，用户只需编辑脚本，然后优化策略选择输入语料库的片段作为基础材料。与所选片段对应的注释参数无缝拼接在一起，并用于生成中间视频表示，其中脸部的下半部分用参数化脸部模型渲染，最后该循环视频生成网络将此表示转换为与编辑后的文字记录相匹配的逼真视频。研究者展示了各种各样的编辑，例如单词的添加、删除和更改，以及令人信服的语言翻译和完整的句子合成。

Averbuch-Elor H 等人<sup>[32]</sup>提出了一种自动为静止肖像制作动画的技术，使照片中的主体能够栩栩如生并表达各种情感，该研究使用（不同主题的）驾驶视频，并开发了将驾驶视频中主题的表现力转移到目标肖像的方法。与之前需要目标面部的输入视频来重现面部表现的工作相比，我们的技术仅使用单个目标图像，研究者通过模仿驾驶视频中的面部变换的 2D 扭曲对目标图像进行动画处理。由于单独的经线不能承载面部的全部表现力，因此研究者添加了通常与面部表情相关的精细动态细节，例如折痕和皱纹。此外，该研究对隐藏在输入目标面部中的区域产生幻觉，尤其是在嘴巴内。其技术产生了反应式配置文件，静止图像中的人可以自动与他们的观众互动，最后展示了该研究在互联网上的大量静态肖像上运行的技术。

Lample G 等人<sup>[33]</sup> 介绍了一种新的编码器-解码器架构，该架构经过训练，可通过直接在潜在空间中解开图像的显著信息和属性值来重建图像。因此，经过训练，其模型可以通过改变属性值来生成输入图像的不同真实版本。通过使用连续的属性值，研究者可以选择在生成的图像中可以感知多少特定属性。此属性可以允许用户使用滑动旋钮修改图像的应用程序，例如混合控制台上的推子，以更改肖像的面部表情或更新

某些对象的颜色。与主要依赖于通过在训练时更改属性值来训练像素空间中的对抗性网络的最先进技术相比，该研究的方法产生了更简单的训练方案并且可以很好地扩展到多个属性。其提供的证据表明，该模型可以显著改变属性的感知价值，同时保持图像的自然性。

### 1.1.3 人类语音的伪造

所谓的语音伪造则是根据人工智能的发展下利用相关的技术对人类语音进行合成，其技术分为两大部分，其一为所谓的语音变换 (Voice Conversion)，其二为文本对语音资料的合成 (Text-to-speech)，前者意义为将人的说话的音调改变为目标人物的音调，后者则是搜集大量的语音与对应文本资料来完成指定语音资讯的输出。这些成果除了可以让一些人无法辨识出是否为当事人，跟甚者能够欺骗过一些语音辨识系统。过往语音伪造都为高斯混合模型 (Gaussian Mixture Model, GMM) 与隐马尔可夫模型 (Hidden Markov Model, HMM)，但随人工智能等领域带动深度学习的技术改进，在这块语音伪造技术则有着明显的突破。

其 Van Den Oord A 等人提出第一个端到端的语音合成器 WaveNet，一种用于生成原始音频波形的深度神经网络。该模型是完全概率和自回归的，每个音频样本的预测分布都以所有先前的样本为条件；尽管如此，该研究证明它可以在每秒数万个音频样本的数据上进行有效训练。当应用于文本到语音时，它产生了最先进的性能，人类听众认为它比英语和普通话的最佳参数和连接系统听起来更自然。单个 WaveNet 可以以相同的保真度捕获许多不同说话者的特征，并且可以通过调节说话者身份在它们之间切换。当训练为音乐建模时，我们发现它会生成新颖且通常高度逼真的音乐片段。其研究者还表明它可以用作判别模型，为音素识别返回有希望的结果。

Arik S 等人<sup>[34]</sup>展示了 Deep Voice，这是一个完全由深度神经网络构建的生产质量的文本到语音系统，Deep Voice 为真正的端到端神经语音合成奠定了基础。该系统包括五个主要构建块：定位音素边界的分割模型、字素到音素的转换模型、音素持续时间预测模型、基频预测模型和音频合成模型。对于分割模型，该研究提出了一种使用连接主义时间分类 (CTC) 损失的深度神经网络执行音素边界检测的新方法，对于音频合成模型，研究者实现了 WaveNet 的变体，它需要的参数更少，训练速度比原始模型快。通过为每个组件使用神经网络，研究者的系统比传统的文本到语音系统更简单、更灵活，传统的文本到语音系统的每个组件都需要费力的特征工程和广泛的领域专业知识。最后，研究者展示其系统的推理可以比实时更快地执行，并描述了在 CPU 和 GPU 上优化的 WaveNet 推理内核，与现有实现相比可实现高达 400 倍的加速。

Wang Y 等人<sup>[35]</sup>文本到语音合成系统通常由多个阶段组成，例如文本分析前端、声

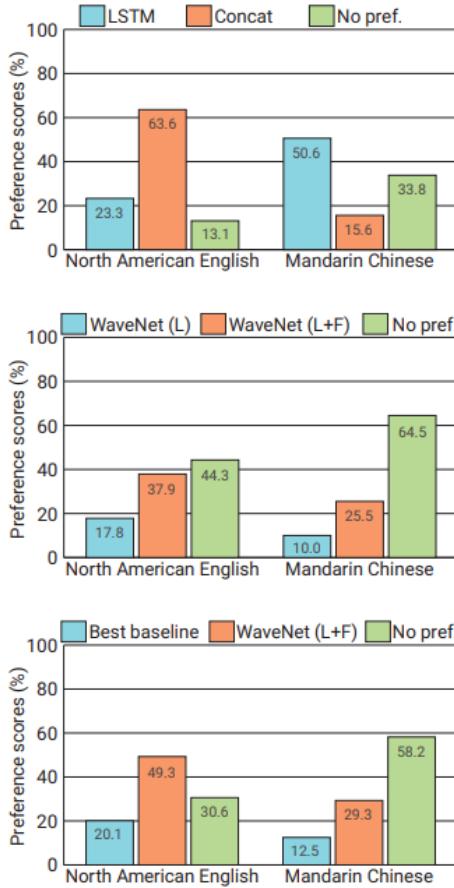


Figure 5: Subjective preference scores (%) of speech samples between (top) two baselines, (middle) two WaveNets, and (bottom) the best baseline and WaveNet. Note that LSTM and Concat correspond to LSTM-RNN-based statistical parametric and HMM-driven unit selection concatenative baseline synthesizers, and WaveNet (L) and WaveNet (L+F) correspond to the WaveNet conditioned on linguistic features only and that conditioned on both linguistic features and log  $F_0$  values.

图 1.11 Van Den Oord A 等人提出 WaveNet

学模型和音频合成模块，构建这些组件通常需要广泛的领域专业知识，并且可能包含脆弱的设计选择。该研究介绍了 Tacotron，这是一种直接从字符合成语音的端到端生成文本到语音模型。给定 text and audio 对，可以通过随机初始化完全从头开始训练模型。研究者提出了几种关键技术，以使序列到序列框架在这项具有挑战性的任务中表现良好。Tacotron 在美国英语上的主观 5 级平均意见得分为 3.82，在自然度方面优于生产参数系统。此外，由于 Tacotron 在帧级别生成语音，因此它比样本级别的自回归方法要快得多。

同时 Arik S 等人<sup>[36]</sup>介绍了一种使用低维可训练说话人嵌入来增强神经文本到语音 (TTS) 的技术，以从单个模型生成不同的声音。作为起点，研究者展示了对单说话人神经 TTS 的两种最先进方法的改进：Deep Voice 1 和 Tacotron。该研究介绍了 Deep Voice 2，它基于与 Deep Voice 1 类似的管道，但使用更高性能的构建块构建，并展示了与 Deep Voice 1 相比显著的音频质量改进，其研究通过引入后处理神经声码器来改

进 Tacotron，并展示了显著的音频质量改进。然后，研究者在两个多说话人 TTS 数据集上演示了该研究的 Deep Voice 2 和 Tacotron 多说话人语音合成技术。并展示了单个神经 TTS 系统可以从每个说话者不到半小时的数据中学习数百个独特的声音，同时实现高音频质量合成并几乎完美地保留说话者的身份。

Ping W 等人<sup>[37]</sup>展示了 Deep Voice 3，这是一种基于全卷积注意力的神经文本到语音(TTS)系统，Deep Voice 3 在自然度上与最先进的神经语音合成系统相匹配，同时训练速度提高了十倍。该研究将 Deep Voice 3 扩展到 TTS 前所未有的数据集大小，对来自 2000 多个扬声器的超过 800 小时的音频进行训练。此外，研究确定了基于注意力的语音合成网络的常见错误模式，演示了如何减轻它们，并比较了几种不同的波形合成方法，最后还描述了如何在一台单 GPU 服务器上将推理扩展到每天一千万个查询。

Pascual S 等人<sup>[38]</sup>当前的语音增强技术在频谱域上运行和/或利用一些更高级别的特征，它们中的大多数处理有限数量的噪声条件并依赖一阶统计数据。为了规避这些问题，由于深度网络能够从大型示例集中学习复杂功能，因此越来越多地使用它们。在这项工作中，研究建议使用生成对抗网络进行语音增强。与当前技术相比，该研究在波形级别上操作，端到端训练模型，并将 28 个扬声器和 40 种不同的噪声条件合并到同一模型中，以便在它们之间共享模型参数。研究者使用具有两个扬声器和 20 种替代噪声条件的独立、看不见的测试集来评估所提出的模型，增强的样本证实了所提出模型的可行性，客观和主观评价都证实了它的有效性。有了这个，研究者开始探索用于语音增强的生成架构，这可能会逐渐结合进一步的以语音为中心的设计选择，以提高它们的性能。

Donahue C 等人<sup>[39]</sup>音频信号以高时间分辨率进行采样，学习合成音频需要在一系列时间尺度上捕获结构。生成对抗网络(GAN)在生成本地和全局连贯的图像方面取得了广泛成功，但它们在音频生成方面的应用却很少。该研究的研究者介绍了 WaveGAN，这是将 GAN 应用于原始波形音频的无监督合成的首次尝试，WaveGAN 能够合成一秒钟的具有全局相干性的音频波形切片，适用于音效生成。其实验表明，在没有标签的情况下，WaveGAN 在小词汇量语音数据集上训练时学会了生成可理解的单词，并且还可以合成来自其他领域的音频，例如鼓、鸟的发声和钢琴。研究者将 WaveGAN 与将设计用于图像生成的 GAN 应用于类似图像的音频特征表示的方法进行比较，发现这两种方法都有前景。

#### 1.1.4 现有开源工具

由 Thanh Thi Nguyen 等人<sup>[40]</sup>的工作可以看到深度学习已成功应用于解决从大数据分析到计算机视觉和人类水平控制的各种复杂问题。然而，深度学习的进步也被用于

创建可能对隐私、民主和国家安全造成威胁的软件。最近出现的深度学习驱动的应用之一是 deepfake。Deepfake 算法可以创建人类无法区分真实图像和视频的虚假图像和视频。因此，能够自动检测和评估数字视觉媒体完整性的技术的提议是必不可少的。本文介绍了用于创建深度伪造的算法的调查，更重要的是，介绍了迄今为止文献中提出的用于检测深度伪造的方法。我们对与深度伪造技术相关的挑战、研究趋势和方向进行了广泛的讨论。通过回顾 deepfake 的背景和最先进的 deepfake 检测方法，本研究提供了 deepfake 技术的全面概述，并有助于开发新的、更强大的方法来处理日益具有挑战性的 deepfake。当中他们为其为深度伪造的工具进行现有的总结条列如下：

- Faceswap : <https://github.com/deepfakes/faceswap>
- Faceswap-GAN : <https://github.com/shaoanlu/faceswap-GAN>
- Few-Shot Face Translation : <https://github.com/shaoanlu/fewshot-facetranslation-GAN>
- DeepFaceLab : <https://github.com/iperov/DeepFaceLab>
- DFaker : <https://github.com/dfaker/df>
- DeepFake tf : <https://github.com/StromWine/DeepFake>
- AvatarMe : <https://github.com/lattas/AvatarMe>
- MarioNETte : <https://hyperconnect.github.io/MarioNETte>
- DiscoFaceGAN : <https://github.com/microsoft/DiscoFaceGAN>
- StyleRig : <https://gvv.mpi-inf.mpg.de/projects/StyleRig>
- FaceShifter : <https://lingzhili.com/FaceShifterPage>
- FSGAN : <https://github.com/YuvalNirkin/fsgan>
- StyleGAN : <https://github.com/NVlabs/stylegan>
- Face2Face : <https://justusthies.github.io/posts/face2face/>
- Neural Textures : <https://github.com/SSRSGJYD/NeuralTexture>
- Transformable Bottleneck Networks : <https://github.com/kyleolsz/TB-Networks>
- "Do as I Do" Motion Transfer : <https://github.com/carolineec/EverybodyDanceNow>
- Neural Voice Puppetry : <https://justusthies.github.io/posts/neural-voicepuppetry>

## 第二章 深度伪造的资料集

由于深度伪造的技术蓬勃发展，同时为了因应深度伪造技术的往负面应用的过程中，与其类似于攻守关系而逐渐发展起来的深度伪造检测手段的需求，间接带动了各个机构与研究者对深度伪造与检测领域上的数据集的迫切需要，因此根据 Li XR 等人工作总结中的所描述的开源数据集的各概况资讯等成果，本作业则将其整理而后条列，同时分为影像资料集与语音资料集如下：

### 2.1 深度伪造的影像资料集

- UADFV<sup>[41]</sup>: 早期的研究数据来源，使用 FakeAPP<sup>[42]</sup> 工具进行合成，同时在 Youtube 平台搜集素材，其资料分别有 49 真实未修改的影像与 49 个已经修改过的伪造影像。
- FaceForensics(FF)<sup>[43]</sup>: 从 Youtube8M<sup>[44]</sup> 的来源中将与人类脸部有关联的目标中取出 1004 的影像，并用 Face2Face 进行改造 1004 个资料集。
- FaceForensics++(FF++)<sup>[45]</sup>: 与 FaceForensics(FF) 类似，该来源从 Youtube 平台取得 1000 个影像，同时使用 4 种方式进行伪造，而当中四种方法包含了 Deepfakes、Face2Face、FaceSwap、Neural Textures。
- Deepfake-TIMIT<sup>[46]</sup>: 根据 Faceswap-GAN 方法进行伪造，同时该资料集也是第一个使用 GAN 所产生的伪造资料集。其资料是根据 VidTIMIT 来源去选 32 人，然后进行两两替换产生。
- Mesonet data<sup>[47]</sup>: 从 Youtube 所产生的数据集。
- Celeb-DF<sup>[48]</sup>: 来源从 Youtube 进行搜集，同时考量 UADFV、FaceForensics++(FF++)、Deepfake-TIMIT 等缺陷后，对伪造的方法进行改良。
- DeepfakeDetection(DFD)<sup>[49]</sup>: 由 Google 公司所制作的资料集，当中请 28 个演员来做出 363 个原始影像资料。
- DFDC preview Dataset<sup>[50]</sup>: 由脸书在 The Deepfake Detection Challenge 所举办的比赛中所开放的测试资料集，当中有 5214 个影像。
- DFDC<sup>[51]</sup>: 由脸书在 The Deepfake Detection Challenge 比赛所提供的正式资料集。
- DeeperForensics-1.0<sup>[52]</sup>: 由南洋理工和商汤科技从 26 个国家收集 100 名演员的脸部数据，过程中将 FaceForensics++ 资料集中的 1000 笔原始影像作为目标影像进行训练，产生 50000 笔未修改影像跟 10000 笔修改影像。

## 2.2 深度伪造的语音资料集

- ASVspoof 2015 database<sup>[53]</sup>: 为因应语音伪造欺骗的问题，而在 2015 年所举办的 synthetic and converted speech 竞赛，当中开放第一个大规模的语音伪造资料集，该资料集根据 106 位不同的人的语音纪录，其分别为 45 名男性，61 名女性组成的训练资料集有 3750 笔原始语音资料片段与 12625 组成的欺骗片段，另外验证资料集则有 3497 笔原始语音资料片段与 49875 组成的欺骗片段，而最后测试资料集则为 9404 笔原始资料集与 184000 笔欺骗语音资料所组成。
- ASVspoof 2019 database<sup>[54]</sup>: 2019 年所举办的 synthetic and converted speech 竞赛则根据 107 名不同的人士所撷取的原始资料，所建立起的资料集。

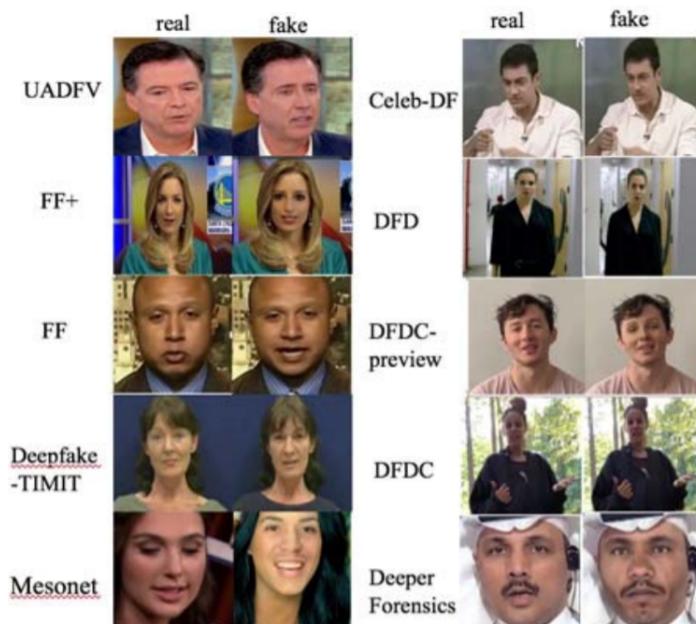
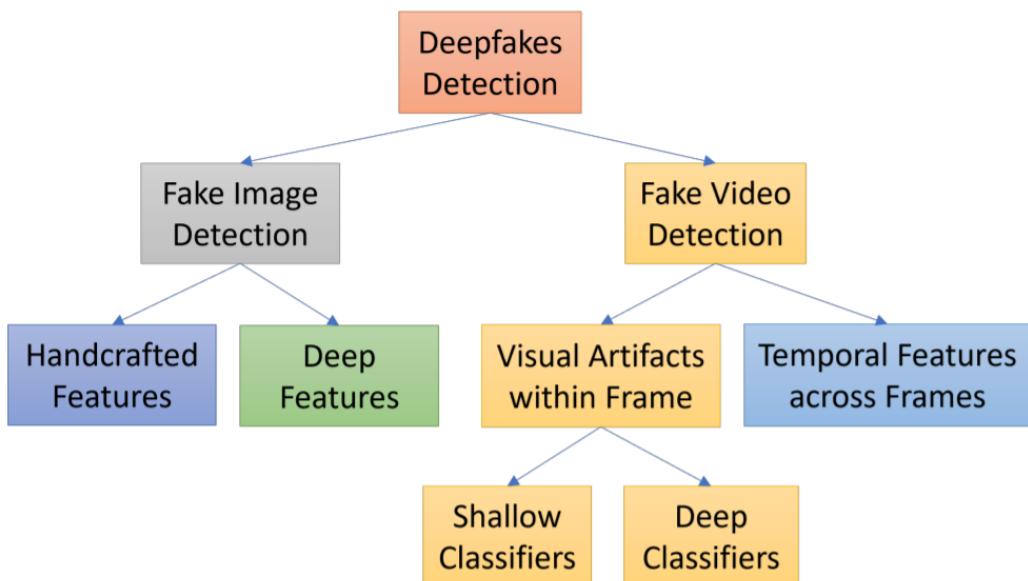


图 2.1 Li XR 等人<sup>[9]</sup>的工作总结的资料对比

## 第三章 深度伪造的研究工具与检测手段的分类

本章节探讨深度伪造的研究工具与检测手段的分类，首先会根据 Thanh Thi Nguyen 等人<sup>[40]</sup> 所进行的研究工作的简述，在他们的工作总结中将其工具分为两大类，也就是图像检测与人脸视频检测，并有着详细的整理，此外根据 Li XR 等人近期对深度伪造与检测的汇整工作<sup>[9]</sup> 也有着详尽的说明，本作业除了针对究工具与检测手段等研究者们的近来工作将其汇整于此章。



**Fig. 6.** Categories of reviewed papers relevant to deepfake detection methods where we divide papers into two major groups, i.e., fake image detection and face video detection.

图 3.1 从由 Thanh Thi Nguyen 等人<sup>[40]</sup> 将深度伪造的伪造分为两大类

### 3.1 深度伪造的图像检测与人脸视频检测整理

根据由 Thanh Thi Nguyen 等人<sup>[40]</sup>在深度伪造的工作中所整理的分类整理的过程中将其分为了图像检测与人脸视频检测。同时该工作也有将其深度伪造的检测手段进行总结，本作业根据原总结工作再汇整为深度伪造检测手段整理两表，两表之中总共有 24 种检测手段，这当中又因该工作所进行的分类，而分为三部分，其一为针对图像的检测，列于表二，共有 9 种手段，其二为影像的检测手段，其列于表一，共有 13 种手段，其三为图像检测与人脸视频检测皆可的检测，只有两种手段，并与表二图像检

测并列。

表 3.1 深度伪造检测手段整理

方法	技术	检测种类	备注
Eye blinking	LRCN	Videos	Yuezun Li 等人
Intra-frame and temporal inconsistencies	CNN and LSTM	Videos	David Guera 等人
Using face warping artifacts	VGG16, ResNet models	Videos	Yuezun Li 等人
MesoNet	CNN	Videos	Darius Afchar 等人
Eye, teach and facial texture	Logistic regression and neural network(NN)	Videos	Falko Matern 等人
Spatio-temporal features with RCN	RCN	Videos	Ekraam Sabir 等人
Spatio-temporal features with LSTM	Convolutional bidirectional recurrent LSTM network	Videos	Akash Chintha 等人
Analysis of PRNU	PRNU	Videos	Marissa Koopman 等人
Phoneme-viseme mismatches	CNN	Videos	Shruti Agarwal 等人
Using attribution-based confidence (ABC) metric	ResNet50 model, pre-trained on VGGFace2	Videos	Steven Fernandes 等人
Using appearance and behaviour	Rules based on facial and behavioural features	Videos	Shruti Agarwal 等人
FakeCatcher	CNN	Videos	Umur Aybars Ciftci 等人
Emotion audiovisual affective cues	Siamese network	Videos	Trisha Mittal 等人

## 3.2 过往传统的检测手段

过往的方式多为根据图像的特征来进行，其手段大多为信号的处理方式，且多数依赖某些特有的窜改证据，同时根据图像本身自有的频域特征与统计特征来进行分类，这些方式如噪音分析、设备指纹、光照、图像品质、处理复制移动、拼接、移除等问题，

其实际工作比如 De Carvalho TJ 等人<sup>[55]</sup>，提出了一种伪造检测方法，该方法利用图像照明颜色的细微不一致，其方法是基于机器学习的，并且需要最少的用户交互。该技术适用于包含两个或更多人的图像，并且不需要专家交互来做出篡改决定，为了实现这一点，研究者将来自基于物理和统计的光源估计器的信息整合到类似材料的图像区域上。从这些光源估计中，该研究者提取基于纹理和边缘的特征，然后将其提供给机器学习方法以进行自动决策。使用 SVM 元融合分类器的分类性能是有希望的。它在由 200 张图像组成的新基准数据集上产生 86% 的检测率，在从 Internet 收集的 50 张图像上产生 83% 的检测率。

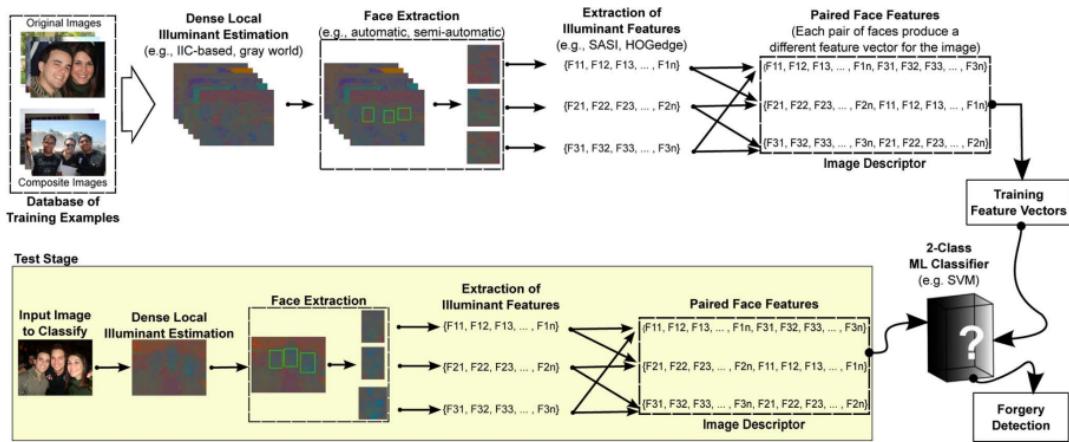


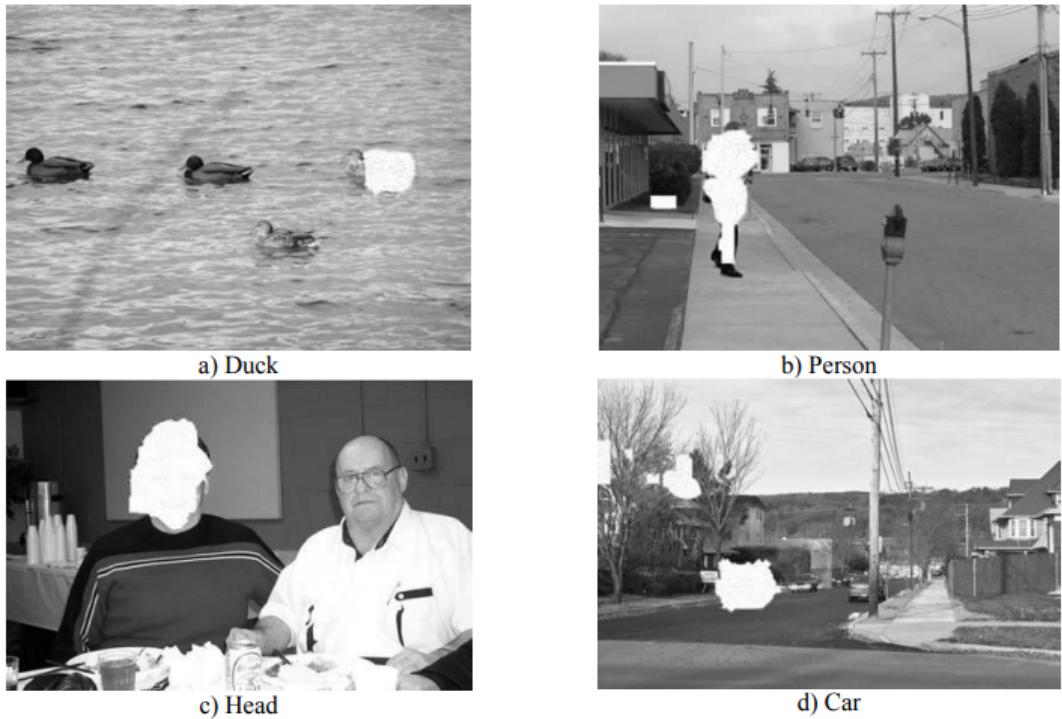
Fig. 4. Overview of the proposed method.

图 3.2 De Carvalho TJ 等人<sup>[55]</sup>

另外 Amerini I 等人<sup>[56]</sup>则探究图像是否被伪造的检测问题；特别是，图像的一个区域被复制然后粘贴到另一个区域用以创建复制或取消一些尴尬的东西的情况。通常来说为了使图像补丁适应新的上下文，需要进行几何变换。为了检测这种修改，提出了一种基于尺度不变特征变换 (SIFT) 的新方法。这种方法使研究者既可以了解是否发生了复制移动攻击，还可以恢复用于执行克隆的几何变换。广泛的实验结果证实了该技术能够精确地个体化改变区域，此外，还能够以高可靠性估计几何变换参数。该方法还处理多重复制。

同理可推，其深度伪造的影像的本质，实质上也是一连串的图像伪造的工作合成后所完成的结果，也因为如此，此类检测方式与手段就能使用到深度伪造的检测工作上。而 Luká J<sup>[57]</sup> 等人提出了一种检测数字图像中的伪造品的新方法，其方法基于检测图像中各个区域中相机模式噪声的存在，这是成像传感器的独特随机特性。而伪造区域被确定为缺少图案噪声的区域。噪声的存在是使用相关性确定的，如在扩频水印的检测中。研究者们提出了两种方法在第一个中，用户选择一个区域进行完整性验证。第二种方法试图在不假设任何先验知识的情况下自动确定伪造区域。这些方法在真实伪

造的例子和非伪造图像上都进行了测试，其研究者还研究了应用于伪造图像的进一步图像处理，例如有损压缩或过滤，如何影响验证图像完整性的能力。



**Figure 6:** Examples of automatically detected ROIs shown in white.

图 3.3 Luká J 等人<sup>[57]</sup>

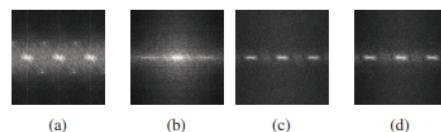
Chierchia G 等人<sup>[58]</sup>引入了光响应非均匀性 (PRNU) 作为检测图像伪造的强大工具，尽管它在许多情况下都有效，但所提出的方法无法检测到小的操作。在这项工作中，研究者基于图像的初步分割提出了中描述的检测算法的修改版本，这保证了对小尺寸加法伪造品的更好检测性能。而 Fridrich J 等人<sup>[59]</sup>则提出一种新的通用策略，用于为数字图像构建隐写检测器，该过程首先组装一个丰富的噪声分量模型，作为由使用线性和非线性高通滤波器获得的量化图像噪声残差的相邻样本的联合分布形成的许多不同子模型的联合。与以前的方法相比，研究者使模型组装成为训练过程的一部分，该过程由从相应的覆盖和隐秘源中抽取的样本驱动。集成分类器用于组装模型以及最终的隐写分析器，因为它们的计算复杂度低并且能够有效地处理高维特征空间和大型训练集。该研究在三种隐写算法上演示了所提出的框架，这些演算法旨在隐藏空间域中表示的图像中的消息：HUGO、Luo 的边缘自适应算法和优化编码的三元  $\$1$  嵌入。对于每种算法，研究者应用一种简单的子模型选择技术来提高每个模型维度的检测精度，并展示检测如何随着丰富模型的复杂性增加而饱和。通过观察不同子模型如何参与检测之间的差异，揭示了嵌入和检测之间有趣的相互作用。围绕丰富的图像模型构建的隐写分析与集成分类器相结合，是为广泛的隐写方案自动化隐写分析的一个有前途的

方向。另外 Wang W 等人<sup>[60]</sup>则专注于局部图像篡改检测。对于 JPEG 图像，其 DCT 系数的概率分布会受到篡改操作的干扰，篡改区域和未改动区域分布不同，是定位篡改的重要线索。基于未量化的 ac DCT 系数的拉普拉斯分布假设，可以估计这两个分布以及被篡改区域的大小，从而得到每个 DCT 块被篡改的概率，当研究者考虑常见篡改区域的先验知识时，可以获得更准确的定位结果，其研究者还设计了三种可以区分真实篡改区域和虚假区域的特征，以减少误报。对于以无损压缩格式保存的篡改图像，而研究者还提出了一种专门的方法，该方法利用高频 DCT 系数的量化噪声来提高篡改定位性能。在大规模数据库上的大量实验证明了该研究提出的方法的有效性，并证明其方法适用于定位不同尺度的篡改区域。

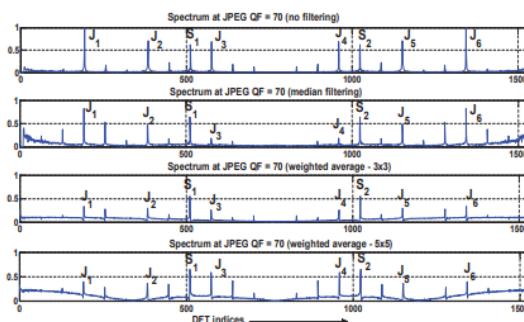
此外還對 JPEG 壓縮的圖像中添加噪聲來進行檢測的研究，比如 Nataraj L 等人<sup>[61]</sup>认为影响大多数图像大小调整检测算法的一个常见问题是它们容易受到 JPEG 压缩的影响，这是因为 JPEG 引入了周期性伪影，因为它适用于 8×8 块。其研究者提出了一种新颖但反直觉的技术，通过添加高斯噪声来“去噪”JPEG 图像。其研究者将适量的高斯噪声添加到调整大小和 JPEG 压缩的图像中，以便抑制由于 JPEG 压缩而导致的周期性，而由于调整大小而导致的周期性得以保留，受控的高斯噪声添加比中值滤波和基于加权平均的滤波更有效地抑制 JPEG 引起的周期性。

**Table 1.** Strength of AWGN to add for a given JPEG QF

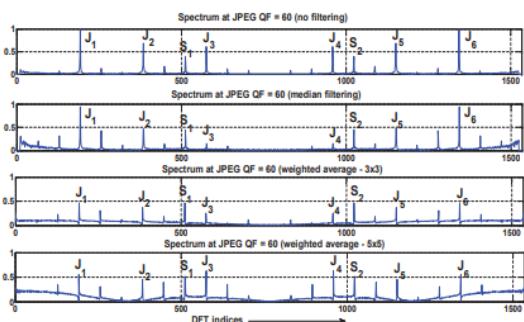
JPEG QF	SNR Range (dB)	JPEG QF	SNR Range (dB)
80	[20-50]	75	[20-50]
70	[20-40]	65	[20-30]
60	[20-30]	55	[20-25]
50	20	40	20



**Fig. 8.** DFT of the p-map after (a) resizing the image by a factor of 3 (b) JPEG on the resized image at a QF of 85 (c) adding AWGN on JPEG image at 35 dB SNR, (d) 40 dB SNR



**Fig. 6.** Filtering the JPEG compressed image, at QF of 70



**Fig. 7.** Filtering the JPEG compressed image, at QF of 60

图 3.4 Nataraj L 等人<sup>[61]</sup>

同时 Bianchi T 等人<sup>[62]</sup>提出了一种统计测试来区分 JPEG 图像中的原始区域和伪造区域，假设前者是双重压缩的，而后者是单次压缩的。提出了单压缩和双压缩区域 DCT 系数的新概率模型，以及双压缩情况下估计主量化因子的可靠方法，基于这样的模型，推导出每个 DCT 块被伪造的概率，其实验结果证明了相对于先前提出的方法更

好的区分为行为。

另外有些研究者则运用局部噪音方差分析的特性来拼接痕迹 Pan X 等人<sup>[63]</sup> 所提出的研究是基于来自不同来源的图像往往具有由传感器或后处理步骤引入的不同数量的噪声，研究者描述了一种通过检测局部噪声方差的不一致性来暴露图像拼接的有效方法，其方法基于以下观察估计局部噪声方差：带通滤波域中自然图像的峰度值倾向于集中在一个恒定值附近，并通过使用积分图像来加速。最后基于通过图像拼接生成的几组伪造图像证明了我们方法的有效性和鲁棒性。

而 Ferrara P<sup>[64]</sup> 等人则是运用色彩过滤矩阵 (Color Filter Array; CFA) 模型来找出修改过的地方，该研究提出了一种能够区分数码相机捕获的图像中的原始区域和伪造区域的取证工具，研究者假设图像是使用滤色器阵列获取的，并且由于去马赛克算法，篡改会消除伪影，其所提出的方法是基于一个新的特征来测量局部水平的去马赛克伪影的存在，以及一个新的统计模型，允许推导出每个  $2 \times 2$  图像块的篡改概率，而无需先验地知道伪造的位置。最后在配备不同去马赛克算法的不同相机上的实验结果证明了理论模型的有效性和方案的有效性。

由于近年来人工智能领域的进步，当中的机器学习下的深度学习方法逐渐导入检测手段中，其卷积神经网络、生成对抗网路与 Transformer 等模型，在该领域的应对皆超越了过往的纪录，而且成功率大大的提升，相较上述过往图像特征的检测方式，前者的鲁棒性更好。Cozzolino D 等人<sup>[65]</sup> 则认为数字图像的取证分析在很大程度上依赖于在获取的图像上留下的相机内和相机外过程的痕迹，这样的痕迹代表了一种相机指纹，若能够恢复它们，通过抑制高级场景内容和其他干扰，可以轻松完成多项取证任务。一个值得注意的例子是 PRNU 模式，它可以被视为设备指纹，在多媒体取证中受到了极大的关注。该研究提出了一种提取相机模型指纹的方法，称为噪声指纹，其中场景内容在很大程度上被抑制，与模型相关的伪影得到增强。这是通过连体网络获得的，该网络使用来自相同（标签 +1）或不同（标签 -1）相机的成对图像块进行训练。尽管噪声印记可用于多种取证任务，但这里我们专注于图像伪造定位，而在广泛使用的几个数据集上的实验表明，基于噪声印记的方法可以提供最先进的性能。Zhou P 等人<sup>[66]</sup> 认为图像操作检测不同于传统的语义对象检测，因为它更关注篡改伪影而不是图像内容，这表明需要学习更丰富的特征，研究者提出了一个双流 Faster R-CNN 网络并对其进行端到端训练，以检测给定操纵图像的篡改区域。两个流之一是 RGB 流，其目的是从 RGB 图像输入中提取特征，以发现篡改伪影，如强对比度差异、不自然的篡改边界等。另一种是噪声流，它利用从隐写分析丰富的模型过滤层中提取的噪声特征来发现真实区域和篡改区域之间的噪声不一致。然后，研究者通过双线性池化层融合来自两个流的特征，以进一步结合这两种模式的空间共现。对四个标准图像处理数据集

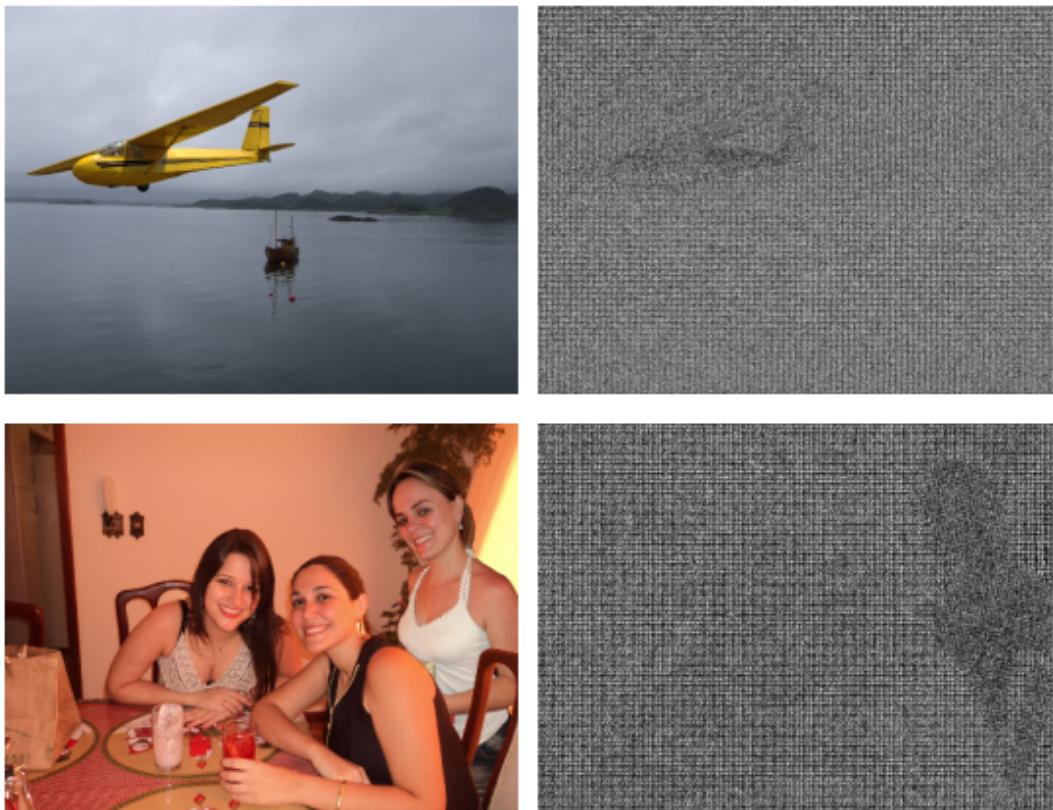


Fig. 1. Two forged images (left) with their noiseprints (right). The inconsistencies caused by the manipulation are visible in the extracted noiseprint.

图 3.5 Cozzolino D 等人<sup>[65]</sup>

的实验表明，其双流框架优于每个单独的流，并且与具有调整大小和压缩鲁棒性的替代方法相比，还实现了最先进的性能。

Rao Y 等人<sup>[67]</sup>提出了一种基于深度学习技术的新图像伪造检测方法，该方法利用卷积神经网络 (CNN) 从输入的 RGB 彩色图像中自动学习层次表示，所提出的 CNN 专为图像拼接和复制移动检测应用而设计，其网络第一层的权重不是随机策略，而是使用空间丰富模型 (SRM) 中残差图计算中使用的基本高通滤波器集进行初始化，作为正则化器，可以有效地抑制图像内容并捕获由篡改操作引入的细微伪影，将预训练的 CNN 作为补丁描述符从测试图像中提取密集特征，然后探索特征融合技术以获得 SVM 分类的最终判别特征。在几个公共数据集上的实验结果表明，所提出的基于 CNN 的模型优于一些最先进的方法。Liu B 等人<sup>[68]</sup>提出了一种新颖的深度融合网络，通过跟踪其边界来定位篡改区域，首先训练了一组称为 Base-Net 的深度卷积神经网络来分别响应某种类型的拼接伪造。然后，选择 Base-Net 的一些层并组合为深度融合神经网络 (Fusion-Net)，经过极少量的图片微调后，Fusion-Net 能够辨别图像块是否是从不同来

源合成的。在基准数据集上的实验表明，该研究方法在各种情况下都是有效的，并且优于最先进的方法。Huh M 等人<sup>[69]</sup>认为照片编辑和操作工具的进步使得创建假图像变得更加容易，突出了对更好的视觉取证算法的需求，然而，由于缺乏被操纵的视觉内容的良好数据集，学习从标记的训练数据中检测操纵是困难的，该研究介绍了一种自我监督方法，用于学习仅使用未标记数据来检测视觉操作。给定大量带有自动记录的 EXIF 元数据的真实照片，研究者训练一个模型来确定图像是否是自洽的——也就是说，它的内容是否可以由单个成像管道产生。研究者将这种自我监督学习方法应用于定位拼接图像内容的任务。其取证模型在许多基准上都取得了最先进的结果，尽管在没有实际操作示例的情况下进行了训练，也没有对特定的检测线索进行建模。除了手工制作的基准之外，研究者还展示了在 Reddit 和 The Onion 上发现假货以及检测计算机生成的拼接的有希望的结果。Cun X 等人<sup>[70]</sup>解决了图像拼接定位的问题：给定一张输入图像，定位从另一张图像中剪下的拼接区域，并将其制定为分类任务，但关键的是，我们不是通过局部补丁对拼接区域进行分类，而是利用整个图像和局部补丁的特征来对补丁进行分类。而研究者们称这种结构为 Semi-Global Network，其方法利用了拼接区域不仅应与局部特征（拼接边缘）高度相关，还应与整个图像的全局特征（语义信息、照明等）高度相关的观察结果。此外，该研究首先将全连接条件随机场作为图像拼接中的后处理技术，以提高输入图像和网络输出之间的一致性。

Cozzolino D 等人<sup>[71]</sup>基于图像噪声残差的局部描述符已被证明对于许多取证应用非常有效，例如伪造检测和定位，尽管如此，在计算机视觉领域取得可喜成果的推动下，研究界的重点现在正在转向深度学习。该研究展示了一类基于残差的描述符实际上可以被视为一个简单的约束卷积神经网络 (CNN)。通过放松约束，并在相对较小的训练集上微调网络，其成果相对于传统检测器获得了显著的性能提升。

Chen C 等人<sup>[73]</sup>认为随着通过社交媒体渠道传播的错误信息的兴起，以及图像处理工具的自动化和真实性的提高，图像取证成为一个越来越重要的问题，经典的图像取证方法利用低级线索，例如元数据、传感器噪声指纹等，当图像在上传到 Facebook 等时重新编码时很容易被愚弄。这需要使用更高级别的物理和语义线索，这些线索曾经在野外难以可靠地估计，但由于计算机视觉的能力越来越强，它们变得更加有效。特别是，该研究的检测由图像的人工模糊引入的操作，这会在图像强度和各种线索之间产生不一致的光度关系。在一个新的模糊操作数据集中，研究者在最具挑战性的情况下实现了 98% 的准确率，其中模糊在几何上是正确的并且与场景的一致物理安排。这种操作现在很容易生成，例如，通过具有硬件来测量深度的智能手机相机，例如。iPhone 7 Plus 的“人像模式”。最后还在挑战数据集上展示了良好的性能，该数据集评估了代表“野外”条件的图像中更广泛的操作。

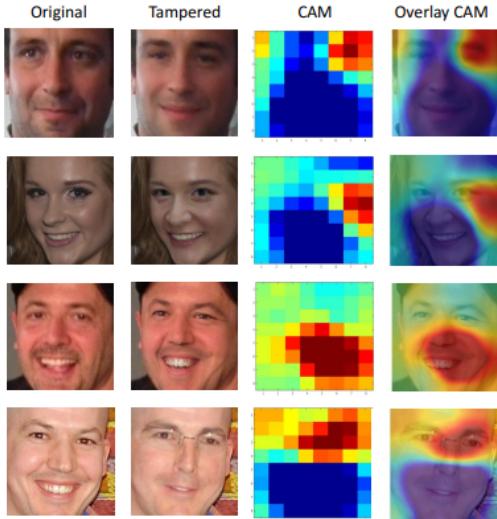


Figure 7. Class Activation Maps (CAMs) obtained from the face classification network. Each row shows the original image, the corresponding tampered face, the CAM, and a smoothing CAM overlaid with the tampered face for better visualization. In CAMs, red denotes high probability of tampering, and blue denotes low probability of tampering. We can observe that our face classification stream learns important artifacts created by the application during face tampering, such as stitching artifacts near face boundaries, strong edges around lips, and blurring effect when glasses are involved.

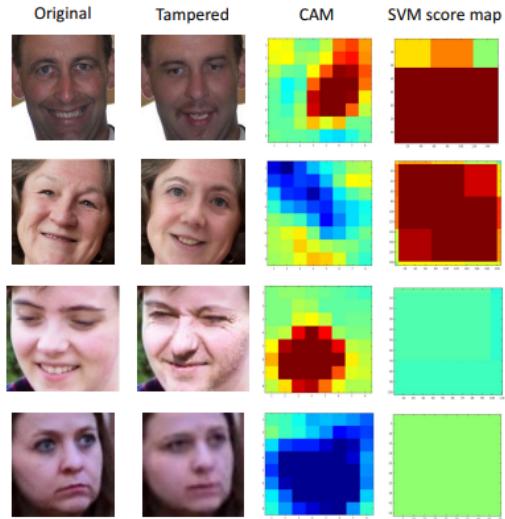


Figure 8. Heat map visualization of our two-stream network. Each row contains the original and tampered face, the corresponding CAM generated in the face classification stream and SVM score map derived from the SVMs in the patch triplet stream. In CAMs, red denotes high probability of tampering, and blue denotes low probability of tampering. In SVM score maps, red regions are more likely to be from different images other than the tampered images. In the first example, both streams can detect the tampered face. In the second and third examples, one stream fails while the other stream works, and fusing two streams successfully detects the tampered faces. Last row shows a failure case when the input face is small.

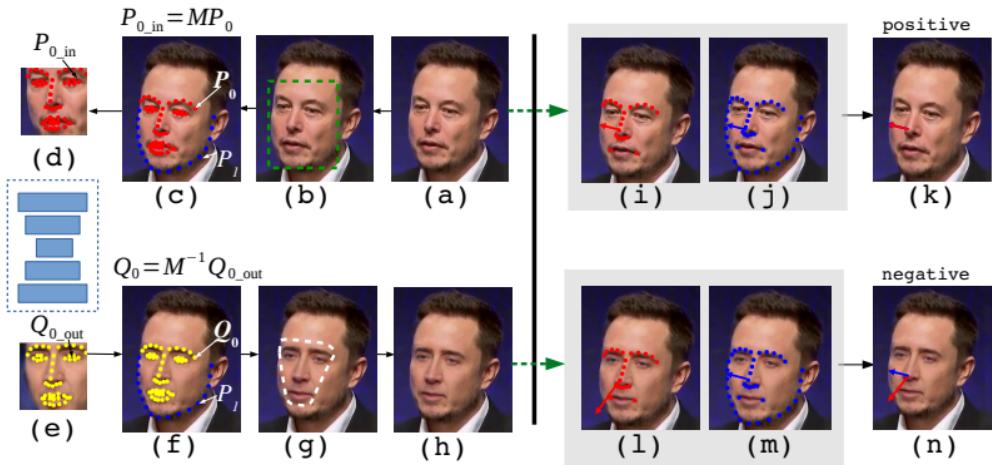
图 3.6 Zhou P 等人<sup>[72]</sup>

Zhou P 等人<sup>[72]</sup>则根据 Szegedy C 等人<sup>[74]</sup>所提出的基础上，做出了一种用于人脸篡改检测的双流网络，其训练 GoogLeNet 以检测人脸分类流中的篡改伪影，并训练基于补丁的三元组网络以利用捕获局部噪声残差和相机特征的特征作为第二个流。此外，研究者使用两个不同的在线人脸交换应用程序来创建一个新的数据集，该数据集由 2010 个篡改图像组成，每个图像都包含一个被篡改的人脸。最后在新收集的数据集上评估提出的双流网络。实验结果证明了其方法的有效性。

### 3.3 人类的生理特征的检测手段

深度伪造的影像往往会忽视人类正常活动的生理特征，因而无法在整体的状况下与真实人类的行动与反应一致，所以根据人类的生理信号特征也进入了研究者们关注的部分。Yang X 等人<sup>[75]</sup>所提出的方法是基于观察到 DeepFakes 是通过将合成的人脸区域拼接到原始图像中来创建的，并且在这样做的过程中，当从人脸图像估计 3D 头部姿势时可以发现错误。研究者进行实验来证明这种现象，并进一步开发基于这种线索的分类方法。使用基于此线索的特征，使用一组真实面部图像和 DeepFakes 评估 SVM 分类器。同样也是 Yang X 等人<sup>[75]</sup>发现生成对抗网络（GAN）最近导致了高度逼真的图

像合成结果。在这项工作中，研究者描述了一种使用面部标志点的位置来展示 GAN 合成图像的新方法。其方法是基于观察到，由于缺乏全局约束，GAN 模型生成的面部部件配置与真实面部不同。研究者进行了演示这种现象的实验，并表明使用面部标志点的位置训练的 SVM 分类器足以将 GAN 合成的面部实现良好的分类性能。



**Fig. 1.** Overview of Deep Fake work-flow (Left) and our method (Right). In (Deep Fake work-flow): (a) is the original image. (b) Detected face in the image. (c) Detected 2D facial landmarks. (d) Cropped face in (a) is warped to a standardized face using an affine transformation  $M$ . (e) Deep Fake face synthesized by the deep neural network. (f) Deep Fake face is transformed back using  $M^{-1}$ . (g) The mask of transformed face is refined based on landmarks. (h) The synthesized face is merged into the original image. (i) The final fake image. For (our method): The top row corresponds to a real image and the bottom corresponds to a Deep Fake. We compare head poses estimated using facial landmarks from the whole face (j),(m) or only the central face region (i),(l). The alignment error is revealed as differences in the head poses shown as their projections on the image plane. The difference of the head poses is then fed to an SVM classifier to differentiate the original image (k) from the Deep Fake (n).

图 3.7 Yang X 等人<sup>[75]</sup>

另外 Li Y 等人<sup>[76]</sup>认为深度生成网络的新发展显著提高了生成逼真的假人脸视频的质量和效率。在这项工作中，研究者描述了一种新方法来暴露由深度神经网络模型生成的假人脸视频，其方法基于检测视频中的眨眼，这是一种生理信号，在合成的假视频中没有很好地呈现。该方法在眨眼检测数据集的基准上进行了评估，并在检测使用基于 DNN 的软件 DeepFake 生成的视频方面表现出良好的性能。

而 Ciftci UA 等人<sup>[77]</sup>提出了一种检测肖像视频中合成内容的新方法，作为针对深度伪造威胁的预防性解决方案。换句话说，该研究引入了一个深度伪造检测器，研究观察到，盲目地利用深度学习的检测器无法有效捕捉虚假内容，因为生成模型会产生非常逼真的结果。其研究者的关键断言是，隐藏在肖像视频中的生物信号可以用作真实性的隐含描述符，因为它们既不会在空间上也不会在时间上保留在虚假内容中。为了证明和利用这一断言，该研究首先对成对分离问题进行了几次信号转换，达到了 99.39% 的准确率。其次，通过分析建议的信号转换和相应的特征集，利用这些发现为虚假内容制定通用分类器。第三，生成新的信号图并使用 CNN 来改进我们用于检测合成内容的传统分类器。最后，发布了一个“在野外”的假肖像视频数据集，而后在评估过程中收集了该数据集。研究在几个数据集上评估 FakeCatcher，分别在人脸取证、人脸取

证 ++、CelebDF 和研究本身新的 Deep Fakes 数据集上获得 96%、94.65%、91.50% 和 91.07% 的准确率。该研究还分析了来自不同面部区域的信号，在图像失真下，具有不同的片段持续时间，来自不同的生成器，针对看不见的数据集，以及在几种降维技术下。

同时 Fernandes S 等人<sup>[78]</sup> 获得了原始视频的心率并训练了最先进的神经常微分方程 (Neural-ODE) 模型。然后，研究者使用商业软件制作了 deepfake 视频，其十个原始视频获得的平均损失为 0.010927，十个捐赠视频为 0.010041，经过训练的 Neural-ODE 能够预测我们使用商业软件生成的 10 个 deepfake 视频和 deepfakeTIMI 数据库的 320 个 deepfake 视频的 heart rate，据该研究者所知，这是首次尝试在原始视频上训练 Neural-ODE 来预测假视频的 heart rate。

### 3.4 图像伪造后留下痕迹的检测手段

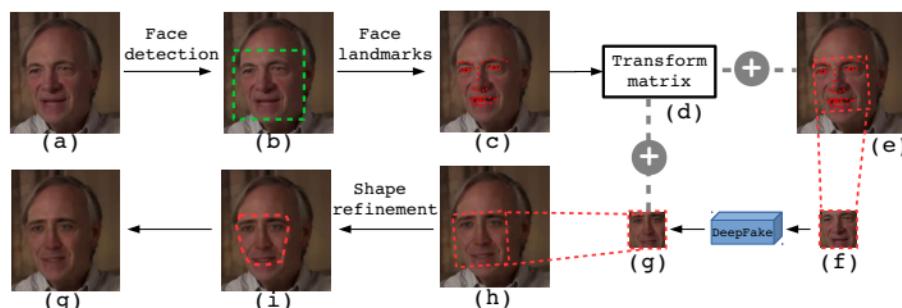


Figure 1. Overview of the DeepFake production pipeline. (a) An image of the source. (b) Green box is the detected face area. (c) Red points are face landmarks. (d) Transform matrix is computed to warp face area in (e) to the normalized region (f). (g) Synthesized face image from the neural network. (h) Synthesized face warped back using the same transform matrix. (i) Post-processing including boundary smoothing applied to the composite image. (g) The final synthesized image.

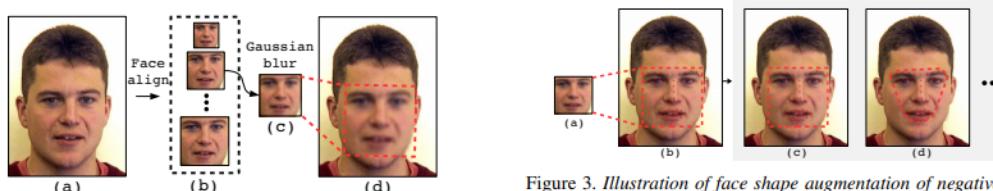


Figure 2. Overview of negative data generation. (a) is the original image. (b) are aligned faces with different scales. We randomly pick a scale of face in (b) and apply Gaussian blur as (c), which is then affine warped back to (d).



Figure 3. Illustration of face shape augmentation of negative examples. (a) is the aligned and blurred face, which then undergoes an affine warped back to (b). (c, d) are post-processing for refining the shape of face area. (c) denotes the whole warped face is retained and (d) denotes only face area inside the polygon is retained.

图 3.8 Li Y 等人<sup>[79]</sup>

其深度伪造的成果受限于过往深度学习的技术发展，用此技术所生成的人类脸部替换或者是伪造的细节皆没有很理想，因此不少研究者为此投入这个部分，而 Li Y 等人<sup>[79]</sup> 描述了一种新的基于深度学习的方法，可以有效地将 AI 生成的假视频（以下称为 DeepFake 视频）与真实视频区分开来。其方法基于当前 DeepFake 算法只能生成分辨率有限的图像的观察结果，这些图像需要进一步变形以匹配源视频中的原始人脸，这

种变换在生成的 DeepFake 视频中留下了独特的伪影，我们证明它们可以被卷积神经网络 (CNN) 有效地捕获，与以前使用大量真实和 DeepFake 生成的图像来训练 CNN 分类器的方法相比，该方法不需要 DeepFake 生成的图像作为负训练示例，因为研究将仿射面部扭曲中的伪影作为区分真假的显著特征图片。其研究方法的优点有两个：(1) 可以直接对图像使用简单的图像处理操作来模拟这种伪影，使其成为反例。由于训练 DeepFake 模型生成负样本既费时又需要资源，因此该研究的方法在训练数据收集方面节省了大量时间和资源；(2) 由于此类伪影普遍存在于来自不同来源的 DeepFake 视频中，因此该研究的方法与其他方法相比更加稳健，研究的方法在两组 DeepFake 视频数据集上进行了评估，以了解其在实践中的有效性。

最后 Li Y 等人<sup>[79]</sup> 则是运用了 He K 等人<sup>[80]</sup>所做的一个残差学习框架，以简化比以前使用的网络更深的网络的训练，该研究明确地将层重新定义为参考层输入的学习残差函数，而不是学习未参考的函数。其提供了全面的经验证据，表明这些残差网络更容易优化，并且可以从显著增加的深度中获得准确性。在 ImageNet 数据集上，研究者评估深度高达 152 层的残差网络——比 VGG 网络深 8 倍，但仍然具有较低的复杂度，这些残差网络的集合在 ImageNet 测试集上实现了 3.57% 的误差。该结果在 ILSVRC 2015 分类任务中获得第一名。我们还对具有 100 层和 1000 层的 CIFAR-10 进行了分析。表示的深度对于许多视觉识别任务至关重要，其研究在 COCO 对象检测数据集上获得了 28% 的相对改进，深度残差网络是该研究提交 ILSVRC 和 COCO 2015 比赛的基础，研究者还在 ImageNet 检测、ImageNet 定位、COCO 检测和 COCO 分割任务中获得了第一名，也就是所谓的 ResNet 框架。

另外 Matern F 等人<sup>[41]</sup> 回顾了当前的面部编辑方法和来自其处理管道的几个特征工件。其研究还表明，相对简单的视觉伪影在暴露此类操作方面已经非常有效，包括 Deepfakes 和 Face2Face 的成果。而该研究团队所用的辨别手段如下：

- 整体不一致性：其伪造手段所生成的人类脸部会有不协调这状况，比如左右眼珠、脸部、鼻子在颜色上不一致。
- 光影的不一致性：光线的照射往往都会被伪造的模型给忽略掉。
- 几何的不一致：细节上的牙齿、眼睛得缺失，又或者只有生成部分。

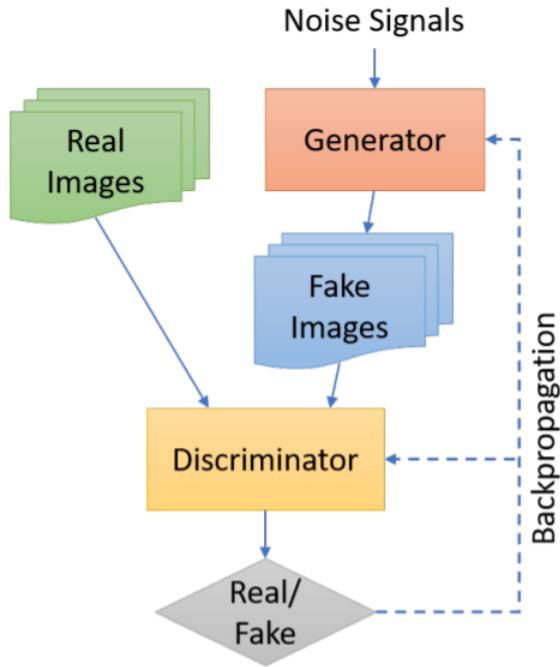
### 3.5 GAN 模型所产生的检测手段

目前深度伪造领域最广泛使用的是 Goodfellow I 等人<sup>[16]</sup> 所提出了一个通过对抗过程估计生成模型的框架，该研究同时训练两个模型：一个生成模型 G 捕获数据分布，一个判别模型 D 估计样本来自训练数据而不是 G 的概率。G 的训练过程是最大化 D 出错的概率。这个框架对应于一个极小极大的两人游戏，在任意函数 G 和 D 的空间中，存

在唯一解，G 恢复训练数据分布，D 处处等于 1/2。在 G 和 D 由多层感知器定义的情况下，整个系统可以通过反向传播进行训练。在训练或生成样本期间，不需要任何马尔可夫链或展开的近似推理网络，实验通过对生成的样本进行定性和定量评估，证明了框架的潜力。而从 Thanh Thi Nguyen 等人<sup>[40]</sup>对这领域所做的工作总结则是如图所见可以看到，GAN 架构由生成器和判别器组成，每个都可以通过神经网络实现，整个系统可以通过反向传播进行训练，从而使两个网络都能提高其能力。当中也描述了两个生成器之间的结构比较：一个 PGGAN (a) 和另一个 StyleGAN (b)。在 PGGAN 中，潜在代码仅被馈送到输入层。而在 StyleGAN 中，潜在代码首先被映射到中间潜在空间 W，然后通过每个卷积层的自适应实例归一化 (AdaIN) 将其注入生成器。在每个卷积之后，但在 AdaIN 操作之前添加高斯噪声。另外使用 StyleGAN 混合样式的示例：输出图像 是通过从源中复制指定的样式子集生成的 B 并从源 A 中获取其余部分。a) 从源 B 复制粗略样式（即对应于粗略空间分辨率  $4^2$ - $8^2$  的样式）将生成具有高级方面的图像，例如来自源 B 的姿势、一般发型、脸型和眼镜，并且具有来源 A 的所有颜色（眼睛、头发、灯光）和更精细的面部特征；b) 如果从 B 复制中等分辨率 ( $16^2$ - $32^2$ ) 的样式，则输出图像将具有较小比例的面部特征、发型、睁眼/闭眼从 B，而姿势、一般脸型和眼镜从 A 保存；c) 如果从源 B 复制精细的样式（对应空间分辨率  $64^2$  -  $1024^2$ ），生成的图像将主要具有源 B 的配色方案和微观结构。

另外 Nataraj L 等人<sup>[81]</sup>则是发现，生成对抗网络 (GAN) 的出现带来了全新的方式来转换和操纵数位图像中的像素，而基于 GAN 的技术，如图像到图像的转换、DeepFakes 和其他自动化方法在创建假图像方面变得越来越流行。在该研究中，研究者们提出了一种结合共现矩阵和深度学习来检测 GAN 生成的假图像的新方法，也就是在像素域中的三个颜色通道上提取共现矩阵，并使用深度卷积神经网络 (CNN) 框架训练模型。基于未配对的图像到图像转换 (cycleGAN) 和面部属性/表情 (StarGAN) 的两个多样化且具有挑战性的 GAN 数据集包含超过 56,000 张图像的实验结果表明，该研究的方法很有前景，并取得了超过两个数据集中的分类准确率为 99%。此外，当在一个数据集上训练并在另一个数据集上进行测试时，此研究的方法也可以很好地泛化并取得良好的结果。

同时 Li H 等人<sup>[82]</sup>，则借助强大的深度网络架构，例如生成对抗网络，人们可以轻松生成逼真的图像。尽管生成的图像并非专门用于欺骗人类或欺骗生物特征认证系统，但研究界和公共媒体对这些图像引起的安全问题表示了极大的关注。该研究解决了识别深度网络生成 (DNG) 图像的问题，考虑到相机成像和 DNG 图像生成之间的差异，研究者分析了不同颜色分量的 DNG 图像和真实图像之间的差异。其研究观察到 DNG 图像在色度分量中与真实图像更容易区分，尤其是在残差域中。基于这些观察，研究者

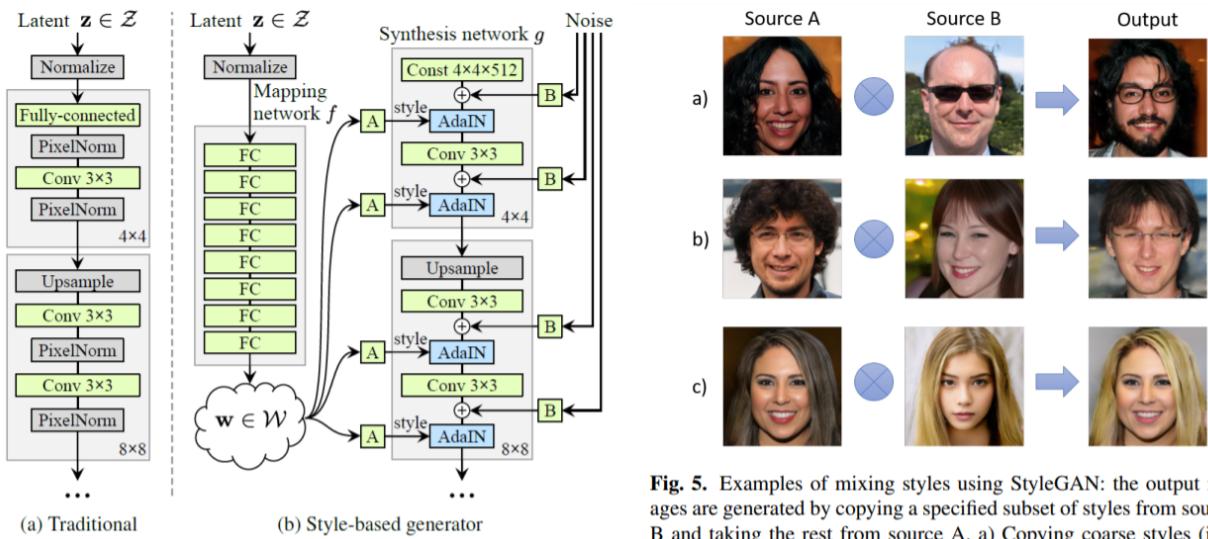


**Fig. 3.** The GAN architecture consisting of a generator and a discriminator, and each can be implemented by a neural network. The entire system can be trained with backpropagation that allows both networks to improve their capabilities.

图 3.9 Thanh Thi Nguyen 等人<sup>[40]</sup> GAN

们提出了一个特征集来捕获用于识别 DNG 图像的彩色图像统计信息。此外，研究者评估了几种检测情况，包括训练测试数据在图像源或生成模型中匹配或不匹配，以及仅使用真实图像进行检测。大量实验结果表明，该方法可以准确识别 DNG 图像，并且在训练和测试数据不匹配时优于现有方法。此外，当 GAN 模型未知时，研究者的方法也通过仅使用真实图像进行训练，实现了一类分类的良好性能。

而在图像的预处理上，Xuan X 等人<sup>[83]</sup>也发现最近 GAN 生成的人脸图像越来越逼真，质量越来越高，甚至人眼也很难检测到，另一方面，取证社区不断开发检测这些生成的虚假图像的方法，并试图保证视觉内容的可信度。尽管研究人员已经开发了一些检测生成图像的方法，但很少有人探索取证模型泛化能力的重要问题。随着新型 GAN 的快速涌现，取证模型检测新型 GAN 图像的泛化能力绝对是一个必不可少的研究课题。在该研究中，研究者们探讨了这个问题，并建议使用预处理图像来训练取证 CNN 模型。通过对真实和虚假的训练图像应用相似的图像级预处理，取证模型被迫学习更多的内在特征来对生成的和真实的人脸图像进行分类。而其的实验结果也证明了所提方法的有效性。



**Fig. 4.** A structure comparison between two generators: one of a PGGAN [61] (a) and another of a StyleGAN [51] (b). In PGGAN, the latent code is fed to the input layer only. In StyleGAN, the latent code is first mapped into an intermediate latent space  $W$ , which is then injected into the generator via the adaptive instance normalization (AdaIN) at each convolution layer. Gaussian noise is added after each convolution, but before the AdaIN operations [51].

**Fig. 5.** Examples of mixing styles using StyleGAN: the output images are generated by copying a specified subset of styles from source B and taking the rest from source A. a) Copying coarse styles (i.e., the styles corresponding to coarse spatial resolutions  $4^2 - 8^2$ ) from source B will generate images that have high-level aspects such as pose, general hair style, face shape, and eyeglasses from source B, and have all colors (eyes, hair, lighting) and finer facial features from source A; b) if copying the styles of middle resolutions ( $16^2 - 32^2$ ) from B, the output images will have smaller scale facial features, hair style, eyes open/closed from B, while the pose, general face shape, and eyeglasses from A are preserved; c) if copying the fine styles (corresponding to spatial resolutions  $64^2 - 1024^2$ ) from source B, the generated images will mainly have the color scheme and microstructure of source B [51].

图 3.10 Thanh Thi Nguyen 等人<sup>[40]</sup> PGGAN 和 StyleGAN

另外 McCloskey S 等人<sup>[84]</sup> 图像取证是一个越来越相关的问题，因为它可以潜在地解决在线虚假信息活动并减轻社交媒体的问题方面。鉴于其最近的成功，特别令人感兴趣的是由生成对抗网络 (GAN) 生成的图像的检测，例如‘深度伪造’，其利用大型训练集和广泛的计算资源，最近的工作表明，可以训练 GAN 生成合成图像，这（在某些方面）与真实图像无法区分，研究者们分析了一个流行的 GAN 实现的生成网络的结构，并表明该网络对颜色的处理在两个方面与真实相机明显不同。该研究进一步表明，这两个线索可用于区分 GAN 生成的图像和相机图像，证明了 GAN 图像和用于训练 GAN 的真实相机图像之间的有效区分。

另一方面在 GAN 的指纹上，Marra F 等人<sup>[85]</sup> 则认为在过去的几年里，生成对抗网络 (GAN) 在计算机视觉和相关领域的许多应用中显示出巨大的潜力，以目前的发展速度，可以肯定的是，GAN 很快就能生成与真实图像和视频几乎无法区分的高质量图像和视频。不幸的是，真实的 GAN 生成的图像对安全构成了严重威胁，首先可能出现大量虚假多媒体，因此迫切需要多媒体取证对策。在该项工作中，研究者们展示了每个 GAN 在其生成的图像中留下其特定的指纹，就像现实世界的相机用它们的照片响应非均匀模式的痕迹标记所获取的图像一样。几种流行的 GAN 的源识别实验表明，这种指纹代表了类似于法医分析的宝贵资产。与此同时 Yu N 等人<sup>[86]</sup> 也认为生成对抗网络

(GAN) 的最新进展表明，在生成逼真的图像方面取得了越来越大的成功，但 GAN 也对视觉取证和模型归因提出了挑战。该研究提出了学习 GAN 指纹对图像属性的第一项研究，并使用它们将图像分类为真实图像或 GAN 生成的图像，而对于 GAN 生成的图像，研究者进一步识别它们的来源，其研究的实验表明 (1) GAN 带有不同的模型指纹，并在其生成的图像中留下稳定的指纹，从而支持图像属性；(2) GAN 训练中即使是微小的差异也会导致不同的指纹，从而实现细粒度的模型认证；(3) 指纹在不同的图像频率和补丁上持续存在，并且不受 GAN 伪影的影响；(4) 指纹微调对五种对抗性图像扰动有效免疫；(5) 比较还表明，研究者的学习的指纹在各种设置中始终优于几个基线。

最后 Wang R 等人<sup>[87]</sup> 提出了一种名为 FakeSpotter 的新方法，该方法基于监视神经元行为来发现 AI 合成的假脸，对神经元覆盖和交互的研究已经成功地表明，它们可以作为深度学习系统的测试标准，尤其是在暴露于对抗性攻击的情况下。在这里，研究者们推测监控神经元行为也可以作为检测假脸的资产，因为逐层神经元激活模式可能会捕获对假人检测器很重要的更细微的特征。检测用最先进的 GAN 合成的四种假人脸并避免四种扰动攻击的实验结果表明了该研究方法的有效性和鲁棒性。

### 3.6 数据驱动的深度学习之检测手段

由 Thanh Thi Nguyen 等人<sup>[40]</sup> 与 Li XR 等人近期来的汇整工作<sup>[91]</sup>，图片的深度学习之检测手段与影像的深度学习之检测手段皆为数据驱动的深度学习，而所谓的数据驱动深度学习，其选择器都为深部结构，而特征是经过训练形成，而不是手工，大致上可将数据驱动的主动深度学习分成根据元学习下的主动深度学习、根据强化学习下的主动深度学习、根据不确定性学习下的主动深度学习、基于数据增强下的主动深度学习。

而在深度伪造的发展下，其算法的规模与对应的数据量都相较过去有显著的增加，而且有不断成长的趋势，在此则分为图片处理与影像处理两大分类，前者的原理是将影片的处理成一帧帧的图像，而后对每帧的图像进行伪造检测处理，另外后者则是影像类，根据循环神经网路去对整个影片来做伪造检测。由下 Thanh Thi Nguyen 等人<sup>[40]</sup> 所做的工作总结，可以看到第一种是图像用于人脸操作检测的两步过程，其中预处理步骤旨在检测、裁剪和对齐一系列帧上的人脸，第二步通过结合卷积神经网络 (CNN) 和循环神经网络 (RNN) 来区分经过操作的和真实的人脸图像）。另一种使用卷积神经网络 (CNN) 和长短期记忆 (LSTM) 提取给定视频序列的时间特征的深度伪造检测方法，这些时间特征通过序列描述符表示。由全连接层组成的检测网络用于将序列描述符作为输入，并计算帧序列属于真实类或 deepfake 类的概率。而后续图片的深度学习之检测手段与影像的深度学习之检测手段的细节则在后两节进行说明。

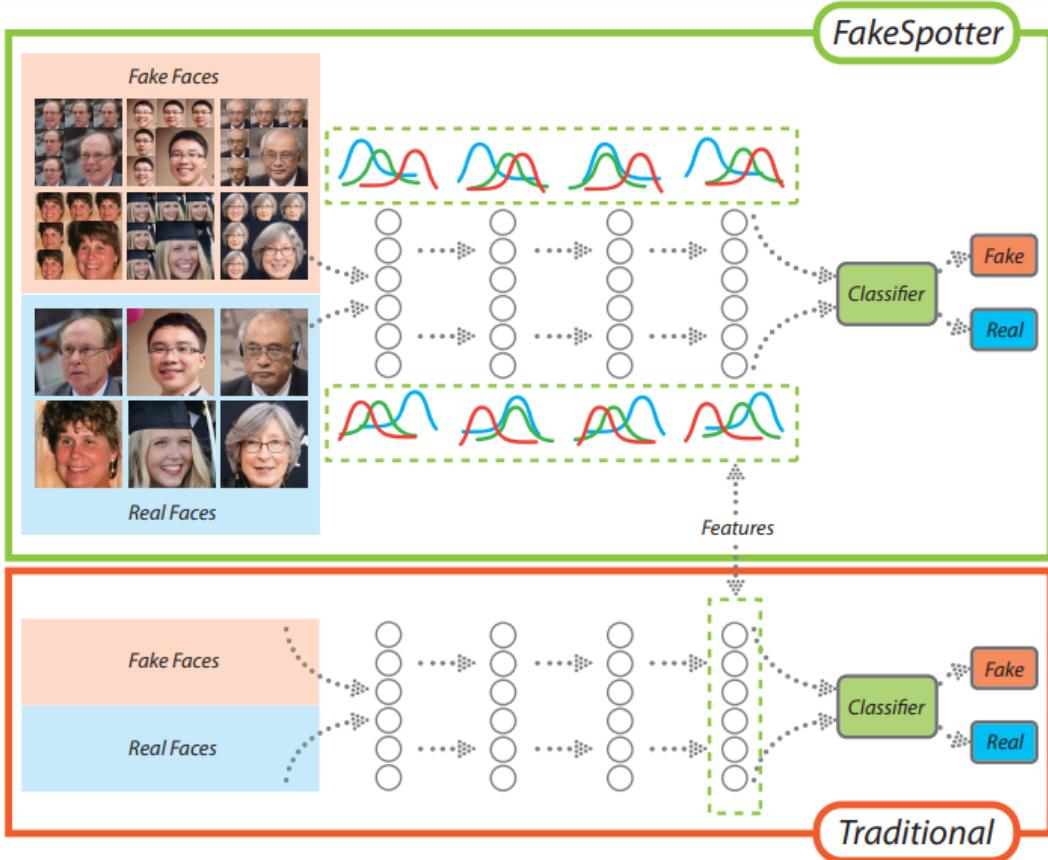


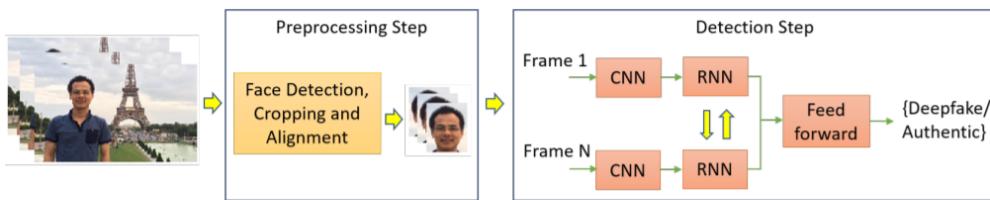
Figure 2: An overview of the proposed fake face detection method, FakeSpotter. Compared to the traditional learning-based method (shown at the bottom), the FakeSpotter uses layer-wise neuron behavior as features, as opposed to final-layer neuron output. Our approach uses a shallow neural network as the classifier while traditional methods rely on deep neural networks in classification.

图 3.11 Wang R 等人<sup>[87]</sup>运用神经元覆盖的方案来找出伪造出来的人脸

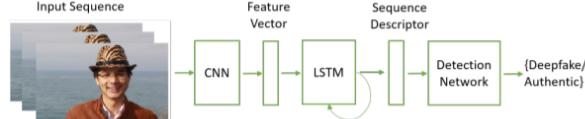
### 3.7 图片的深度学习之检测手段

此节讲述图片的深度学习之检测手段，Afchar D 等人<sup>[47]</sup>提出了一种自动有效地检测视频中的人脸篡改的方法，特地关注最近用于生成超逼真伪造视频的两种技术：Deepfake 和 Face2Face，由于压缩会严重降低数据质量，传统的图像取证技术通常不太适合视频。因此，该研究采用深度学习方法提出了两个网络，两者都具有较少的层数，以专注于图像的细观特性。该研究在现有数据集和研究者从在视频构成的数据集上评估这些快速网络。最后测试证明了非常成功的检测率，Deepfake 的检测率超过 98%，Face2Face 的检测率超过 95%。

Rossler A 等人<sup>[45]</sup>认为合成图像生成和处理的快速进展现在已经到了引发对社会影



**Fig. 7.** A two-step process for face manipulation detection where the preprocessing step aims to detect, crop and align faces on a sequence of frames and the second step distinguishes manipulated and authentic face images by combining convolutional neural network (CNN) and recurrent neural network (RNN) [103].



**Fig. 8.** A deepfake detection method using convolutional neural network (CNN) and long short term memory (LSTM) to extract temporal features of a given video sequence, which are represented via the sequence descriptor. The detection network consisting of fully-connected layers is employed to take the sequence descriptor as input and calculate probabilities of the frame sequence belonging to either authentic or deepfake class [112].

图 3.12 Thanh Thi Nguyen 等人<sup>[40]</sup> 影像和图片

响的重大担忧的地步。充其量，这会导致对数字内容失去信任，但可能会通过传播虚假信息或虚假新闻而造成进一步的伤害，该研究研究了最先进的图像处理的真实性，以及自动或人工检测它们的难度，而为了标准化检测方法的评估，研究者提出了面部操作检测的自动化基准。特别是，该基准基于 DeepFakes、Face2Face、FaceSwap 和 NeuralTextures 作为随机压缩级别和大小的面部操作的突出代表。该基准是公开的，包含一个隐藏的测试集以及一个包含超过 180 万张操纵图像的数据库，而且该数据集相对于可比较的、公开可用的伪造数据集大一个数量级。基于这些数据，研究者们对数据驱动的伪造检测器进行了彻底的分析。其研究表明，即使在存在强压缩的情况下，使用额外的特定领域知识也可以将伪造检测提高到前所未有的准确性，并且明显优于人类观察者，其前者运用 Chollet F 等人<sup>[88]</sup> 的 Xception 架构进行对影像每一帧跟人类脸部进行分别训练，其训练结果好于全帧的模型训练结果。当中所谓 Xception 架构则是将卷积神经网络中的 Inception 模块解释为介于常规卷积和深度可分离卷积操作（深度卷积后跟点卷积）之间的中间步骤，从这个角度来看，深度可分离卷积可以理解为具有最大数量的塔的 Inception 模块，这一观察使研究者提出了一种受 Inception 启发的新型深度卷积神经网络架构，其中 Inception 模块已被深度可分离卷积取代。其研究表明，这种被称为 Xception 的架构在 ImageNet 数据集 (Inception V3 的设计目标) 上略微优于 Inception V3，并且在包含 3.5 亿张图像和 17,000 个类别的更大图像分类数据集上显著优于 Inception V3。由于 Xception 架构与 Inception V3 具有相同数量的参数，因此性能提升不是由于容量增加，而是更有效地使用模型参数。

而运用找出人脸的关键部位来提升模型训练成果，则有 Songsri-in K 等人<sup>[89]</sup> 在其研

究进行验证，其研究是提出了第一个严格的人脸取证定位数据集，该数据集由真实的、生成的和经过处理的人脸图像组成。特别是，原始部分包含来自 CelebA 和 FFHQ 数据集的人脸图像。假图像是由各种 GANs 方法生成的，即 DCGANs、LSGANs、BEGANs、WGAN-GP、ProGANs 和 StyleGANs。最后，编辑的子集是基于自由形式掩码从 StarGAN 和 SEFCGAN 生成的。该数据集总共包含大约 130 万张用相应的二进制掩码标记的面部图像。基于所提出的数据集，研究者们证明了在输入图像之外显式添加面部标志信息可以提高性能。此外，该研究提出的方法由两个分支组成，可以连贯地预测人脸取证检测和定位，以优于以前在新提出的数据集以及 faceforecsic++ 数据集上的最新技术，尤其是在低质量视频上。而 Nguyen HH 等人<sup>[90]</sup> 认为媒体生成技术的最新进展使攻击者更容易创建伪造的图像和视频，其最先进的方法可以实时创建从社交网络获得的单个视频的伪造版本。尽管已经开发了许多用于检测伪造图像和视频的方法，但它们通常针对某些领域，并且随着新型攻击的出现很快就过时了。该研究介绍的方法使用胶囊网络来检测各种欺骗，从使用打印图像或录制视频的重放攻击到使用深度卷积神经网络的计算机生成视频，它将胶囊网络的应用扩展到解决逆图形问题的初衷之外，该研究更重要的一点在运用了 Simonyan K 等人<sup>[91]</sup> 所做的 VGG-19，其主要贡献是使用具有非常小的 (3x3) 卷积滤波器的架构对深度增加的网络进行了全面评估，这表明通过将深度提升到 16-19 个权重层可以实现对现有技术配置的显著改进，这些发现是研究者们提交 2014 年 ImageNet 挑战赛的基础，该研究团队分别获得了定位和分类轨道的第一和第二名。该研究还表明，其表示可以很好地推广到其他数据集，并在这些数据集上取得了最先进的结果。

Mo H 等人<sup>[92]</sup> 则发现生成对抗网络 (GAN) 是一种突出的生成模型，广泛用于各种应用，最近的研究表明，基于这种新颖的模型可以获得具有高视觉质量的假人脸图像。如果这些假脸被滥用于图像篡改，将导致一些潜在的道德、伦理和法律问题。因此，在该研究的研究者们首先提出了一种基于卷积神经网络 (CNN) 的方法来识别当前最佳方法生成的假人脸图像，并提供实验证据表明该方法可以达到令人满意的结果，平均准确率超过 99.4 %。此外，研究者们提供了对所提出的 CNN 架构的一些变体进行评估的比较结果，包括高通滤波器、层组数和激活函数，以进一步验证我们方法的合理性。

Durall R 等人<sup>[93]</sup> 运用离散傅立叶变换的方法来进行特征学习就此提出了一种检测此类假人脸图像的简单方法，也就是检测所谓的 DeepFakes，该研究的方法基于经典的频域分析，然后是基本分类器。与需要输入大量标记数据的先前系统相比，我们的方法仅使用少量带注释的训练样本就显示出非常好的结果，甚至在完全无监督的情况下也取得了很好的准确性。对于高分辨率人脸图像的评估，研究者将几个真实和虚假人脸的公共数据集组合成一个新的基准：Faces-HQ。鉴于如此高分辨率的图像，当该

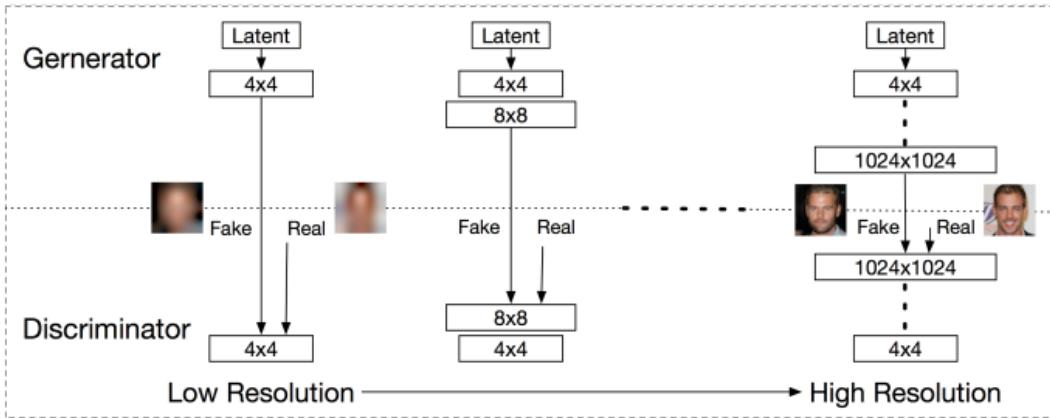


Figure 1: The progressive training strategy employed in [20]. Here  $N \times N$  refers to layers operating on images of  $N \times N$  resolution.

图 3.13 Mo H 等人<sup>[92]</sup>

研究的方法在少至 20 个带注释的样本上进行训练时，其方法达到了 100% 的完美分类准确率。在第二个实验中，在 CelebA 数据集的中等分辨率图像的评估中，其方法在有监督的情况下达到了 100% 的准确率，在无监督的情况下达到了 96%。最后，评估 FaceForensics++ 数据集的低分辨率视频序列，该研究的方法检测操纵影片的准确率达到 91%。Ding X 等人<sup>[94]</sup>在这项研究中，使用深度迁移学习进行人脸交换检测，也就是运用 Resnet18 进行改良，其结果显示大于 96% 的真阳性率，并且误报率非常低。与仅提供检测准确性的现有方法不同，该研究还为每个预测提供不确定性，这对于信任此类检测系统的部署至关重要。此外，研究者们提供了与人类受试者的比较。为了捕捉人类识别性能，此研究建立了一个网站来收集人类受试者图像的成对比较。基于这些比较，研究者们推断出从被认为最真实的图像到被认为最假的图像的共识排名，总体而言，结果显示了该研究的方法有效性。作为这项研究的一部分，研究者们创建了一个新的数据集。

而 Cozzolino D 等人<sup>[95]</sup>引入了取证转移 (FT)。研究者们设计了一种基于学习的取证检测器，它可以很好地适应新领域，即新颖的操作方法，并且可以处理在训练期间只有少数假样本可用的场景。为此，研究者学习了一种基于新型自动编码器架构的取证嵌入，该架构可用于区分真假图像，其学习嵌入充当异常检测器的一种形式；即，如果从不可见的方法处理的图像与真实图像集群足够远，则该图像将被检测为假图像。与之前的工作相比，FT 显示出可迁移性的显著改进，该研究在一系列关于尖端基准的实验中证明了这一点。例如，在未见过的例子上，研究者们的准确率高达 85%，而只有少数可见的例子，该研究性能已经达到了 95% 左右。

就 Cozzolino D 等人<sup>[95]</sup>的研究成果上，其 Nguyen HH 等人<sup>[96]</sup>认为检测被操纵的图像和视频是数字媒体取证中的一个重要课题，大多数检测方法使用二进制分类来确

定查询被操纵的概率，另一个重要主题是定位被操纵区域（即执行分割），这主要是由三种常用攻击创建的：删除、复制移动和拼接。研究者们设计了一个卷积神经网络，它使用多任务学习方法来同时检测被操纵的图像和视频，并为每个查询定位被操纵的区域，其通过执行一项任务获得的信息与另一项任务共享，从而提高两项任务的性能。另外使用半监督学习方法来提高网络的可生成性，该网络包括一个编码器和一个 Y 形解码器。编码特征的激活用于二进制分类，其解码器一个分支的输出用于分割操作区域，而另一个分支的输出用于重构输入，这有助于提高整体性能。使用 FaceForensics 和 FaceForensics++ 数据库的实验证明了该网络对面部重演攻击和面部交换攻击的有效性，以及它处理先前看到的攻击的不匹配条件的能力。此外，仅使用少量数据进行微调使网络能够处理看不见的攻击。

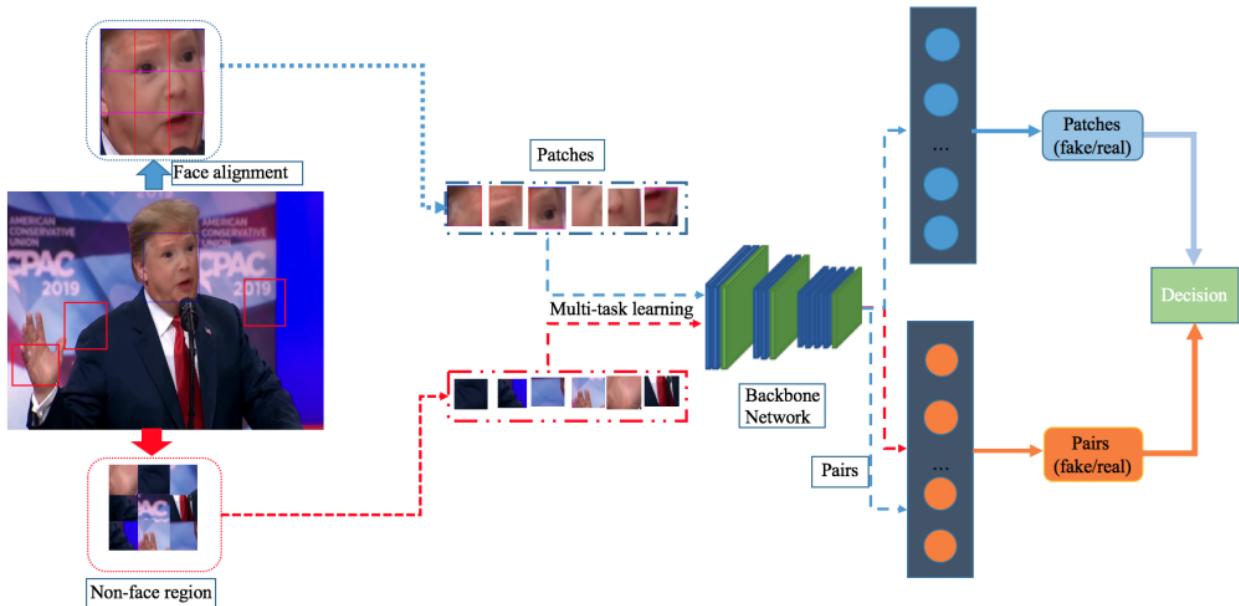
Hsu CC 等人<sup>[97]</sup> 开发一种深度伪造鉴别器（DeepFD）来有效地检测计算机生成的图像，其直接学习二元分类器相对比较棘手，因为很难找到共同的判别特征来判断不同 GAN 生成的假图像。为了解决这个缺点，研究者采用对比损失来寻找由不同 GAN 生成的合成图像的典型特征，然后连接一个分类器来检测这些计算机生成的图像，实验结果表明，所提出的 DeepFD 成功检测到由几个最先进的 GAN 生成的 94.7% 的假图像。同样也是 Hsu CC 等人<sup>[98]</sup> 提出了一种基于深度学习的方法，通过使用对比损失来检测假图像。首先，采用几种最先进的 GAN 来生成假-真图像对。接下来，将简化的 DenseNet 发展为双流网络结构，以允许成对信息作为输入。然后，使用成对学习来训练所提出的常见假特征网络，以区分假图像和真实图像之间的特征。最后，将分类层连接到所提出的常见假特征网络，以检测输入图像是假的还是真的，其实验结果表明，所提出的方法明显优于其他最先进的假图像检测器。

另外 Dang LM 等人<sup>[99]</sup> 提出了一种定制的卷积神经网络，即 CGFace，它是专门为计算机生成的人脸检测任务而设计的，通过自定义卷积层的数量，因此在检测计算机生成的人脸图像方面表现良好。之后，通过从 CGFace 层中提取特征并使用它们来训练 AdaBoost 和 eXtreme Gradient Boosting (XGB) 来改变 CGFace 的层结构以适应不平衡数据问题，从而创建了一个不平衡框架 (IF-CGFace)。接下来，研究者们将解释基于最先进的 PCGAN 和 BEGAN 模型生成大型计算机生成数据集的过程。而随后的进行了各种实验也表明所提出的具有增强输入的模型产生了 98% 的最高精度。最后，研究者们通过将所提出的 CNN 架构应用于其他 GAN 研究生成的图像来提供比较结果。

Bayar B 等人<sup>[100]</sup> 提出了一种通用的取证方法来使用深度学习执行操作检测。具体来说，研究者们提出了一种新的卷积网络架构，能够直接从训练数据中自动学习操作检测特征，而在目前的形式中，卷积神经网络将学习捕捉图像内容的特征，而不是操作检测特征。为了克服这个问题，该研究开发了一种新形式的卷积层，专门用于抑制

图像的内容并自适应地学习操作检测特征，通过一系列实验，其研究证明了研究者们提出的方法可以自动学习如何检测多个图像操作，而不依赖于预先选择的特征或任何预处理。这些实验的结果表明，该研究提出的方法可以自动检测几种不同的操作，平均准确率为 99.10%。

Li X 等人<sup>[101]</sup> 提出了一种新颖的 Patch&Pair 卷积神经网络 (PPCNN) 来区分 Deepfake 视频或图像与真实视频或图像，其通过对公共数据集的综合评估，研究者证明其研究的模型比现有的检测方法表现更好，并表现出更好的泛化性。



**Figure 1: The architecture of our PPCNN.**

图 3.14 Li X 等人<sup>[101]</sup> 提出了一种新颖的 Patch&Pair 卷积神经网络 (PPCNN)

Rahmouni N 等人<sup>[102]</sup> 提出了一种深度学习方法，用于将计算机生成的图形与真实的摄影图像区分开来。所提出的方法使用带有自定义池化层的卷积神经网络 (CNN) 来优化当前性能最佳的算法特征提取方案，来计算和聚合类概率的局部估计以预测整个图片的标签，研究者评估了我们在最近的照片般逼真的计算机图形方面的工作，并表明它在局部和完整图像分类方面都优于最先进的方法。

Dang H 等人<sup>[103]</sup> 建议利用注意力机制来处理和改进分类任务的特征图，而不是简单地使用多任务学习来同时检测操纵图像和预测操纵掩码（区域），其学到的注意力图突出显示信息区域以进一步改进二元分类（真人脸与假人脸），并可视化操作区域，为了使研究者能够研究操纵的面部检测和定位，该研究收集了一个包含多种类型的面部伪造的大型数据库。同时使用这个数据集的过程中，研究者对数据驱动的假人脸检

测进行了彻底的分析，成果展示了注意力机制的使用改进了面部伪造检测和操纵区域定位。

Brockschmidt J 等人<sup>[104]</sup> 研究了最先进的面部伪造检测架构的泛化能力，其研究者首先提出两个通用性标准：可靠地检测多种欺骗技术和可靠地检测看不见的欺骗技术，随后设计实验来衡量给定架构如何根据这些标准执行。研究者的分析侧重于两种最先进的面部伪造检测架构，MesoNet 和 XceptionNet，它们都是卷积神经网络 (CNN)，而实验使用来自六种最先进的面部伪造技术的样本：Deepfakes、Face2Face、FaceSwap、GANnotation、ICface 和 X2Face。研究者发现 MesoNet 和 XceptionNet 显示出泛化到多种欺骗技术的潜力，但在准确性上略有权衡，并且在很大程度上无法对抗看不见的技术。最后将这些结果松散地推断为类似的 CNN 架构，并强调需要更好的架构来应对普遍性的挑战。Sohrawardi SJ 等人<sup>[105]</sup> 研究者们提出了一个系统，该系统将强大而有效地使用户能够确定在线发布的影片是否是 deepfake，该研究从记者的角度处理问题，并努力开发一种工具以无缝融入他们的工作流程，结果表明对内部数据集和不匹配数据集的准确检测。综上所述，若是在影像中每一帧帧去检测伪造的方式，在一部影像伪造地方过少，很有可能没有办法检测的到。那图片的检测手段方式很可能会有严峻的挑战与不理想得结果。

### 3.8 影像的深度学习之检测手段

此节讲述影像的深度学习之检测手段，Agarwal S 等人<sup>[106]</sup> 发现从每个个体来看，运用针对人类脸部的追踪，跟面对人体的头部移动抽取特定动作，其脸部的肌肉变化的可以做为动作单元，同时运用皮尔森系数找出对每个特征间的关联，最后根据此来建立一个 SVM 分类伪造出来的影像。尽管在视觉上并不明显，但这些相关性经常被深度伪造视频的创建方式所破坏，因此可以用于身份验证。另外 Amerini I<sup>[107]</sup> 在这项工作中，给出了一种能够区分假视频序列和原始视频序列的新取证技术；与采用单个视频帧的其他最先进的方法不同，研究者们建议采用光流场来利用可能的帧间差异。然后将这样的线索用作 CNN 分类器要学习的特征，其分类器为 VGG16，最后在 FaceForensics++ 数据集上获得的初步结果突出了非常有前景的性能。

Güera D 等人<sup>[108]</sup> 提出了一种时间感知管道来自动检测 deepfake 视频。研究者的系统使用卷积神经网络 (CNN) 来提取帧级特征，然后使用这些特征来训练循环神经网络 (RNN)，该网络学习对视频是否受到操纵进行分类。同时研究者针对从多个视频网站收集的大量 deepfake 视频评估该研究的方法，最后展示了该研究的系统如何在使用简单架构的同时在这项任务中取得有竞争力的结果。

其框架分成了两个不同阶段的分析器，一个为 CNN 去找出每一个帧内的特征，然

后放进所需要的序列，最后交给 LSTM 进行分析，回传一个机率的判断结果。另外在 Li XR 等人近期来的汇整工作<sup>[9]</sup> 中也有详细说明其循环神经网路和卷积的判断机制。

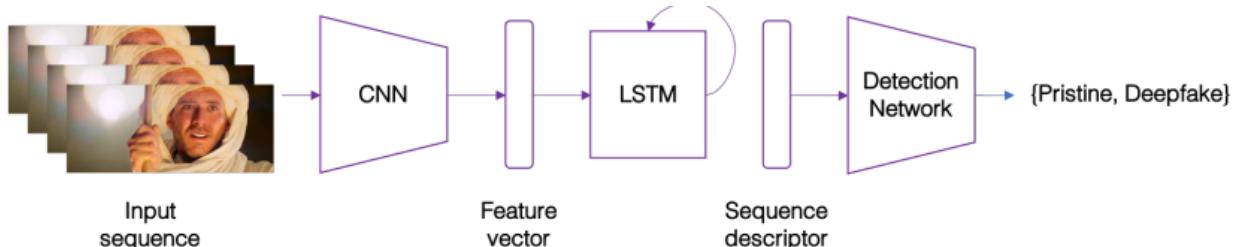


Figure 3. Overview of our detection system. The system learns and infers in an end-to-end manner and, given a video sequence, outputs a probability of it being a deepfake or a pristine video. It has a convolutional LSTM subnetwork, for processing the input temporal sequence.

图 3.15 Güera D<sup>[108]</sup> 架構

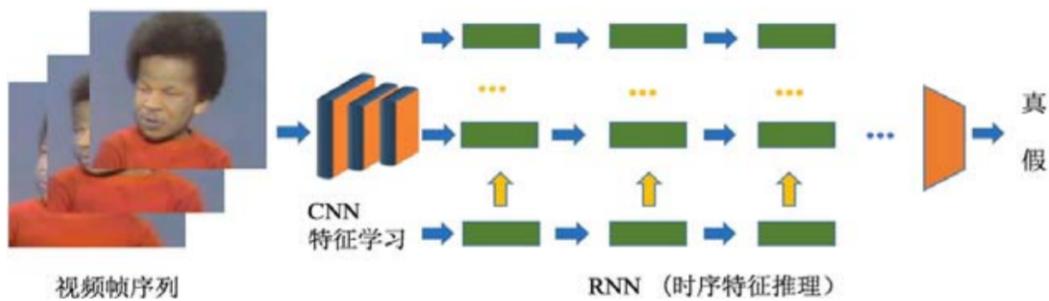


Fig.5 Frame sequences are learned by recurrent neural networks and convolutional neural networks  
图 5 循环神经网络和卷积神经网络学习帧序列

图 3.16 Li XR 汇整工作<sup>[9]</sup> 中详细说明其循环神经网路和卷积的判断机制

另外 Sabir E 等人<sup>[109]</sup> 通过广泛的实验提取了将这些模型的变化与特定领域的面部预处理技术相结合的最佳策略，以在公开的基于视频的面部操作基准上获得最先进的性能。具体来说，该研究尝试检测视频流中的 Deepfake、Face2Face 和 FaceSwap 篡改人脸，其评估是在最近推出的 FaceForensics++ 数据集上进行的，将之前最先进的准确率提高了 4.55%，其关键的地方在采用了双向时序网络和人脸对齐并用的方式去判断伪造出来的生成结果，当中最好的部分是的人脸对齐跟 Bidirectional-recurrent-dense 的伪造检测上。由上述可知，虽然影像的深度学习等之检测手段，泛化性高，同时可以找到影片中少量的篡改，但是在面对压缩过后的影像、光影变化时的伪造检测却很无力。

### 3.9 人类语音伪造检测手段

随着深度伪造在人类语音技术合成的发展，此类工作也受到研究者的关注。Todisco M 等人<sup>[110]</sup> 发现近年来，开发新的对策以保护自动说话人验证免受欺骗的努力已经加

强，其 ASVspoof 2015 计划表明，检测欺骗攻击的潜力很大，但检测以前无法预见的欺骗攻击仍然具有挑战性，该研究认为，从研究特征而不是分类器中可以获得更多收益，并介绍了一种基于恒定 Q 变换的欺骗检测新特征，这是一种在音乐研究中流行的受感知启发的时频分析工具。使用标准 ASVspoof 2015 数据库获得的实验结果表明，当与基于标准高斯混合模型的分类器结合使用时，所提出的恒定 Q 倒谱系数 (CQCC) 的性能明显优于所有先前报告的结果，特别是，对于未知欺骗攻击子集（未使用匹配的训练数据）的攻击为 0.46%，相对于先前报告的最佳结果提高了 72%。

同时 Wu Z 等人<sup>[111]</sup> 首先展示了评估当前最先进的说话人验证系统的脆弱性的新结果：具有联合因子分析 (GMM-JFA) 和概率线性判别分析 (PLDA) 系统的高斯混合模型，以防止欺骗攻击。而所谓的欺骗攻击则是通过两种语音转换技术模拟：基于高斯混合模型的转换和基于单元选择的转换。为了降低由欺骗攻击引起的错误接受率，研究者提出了一种用于说话人验证系统的通用反欺骗攻击框架，其中采用转换后的语音检测器作为说话人验证系统接受决策的后处理模块，其检测器决定接受的声明是人类语音还是转换后的语音。NIST SRE 2006 语料库中的核心任务子集用于评估说话人验证系统的脆弱性和转换后的语音检测器的性能。其研究结果表明，两种转换技术都可以提高 GMM-JFA 和 PLDA 系统的误认率，而转换后的语音检测器可以将 GMM-JFA 和 PLDA 的误认率从 31.54% 和 41.25% 降低到 1.64% 和 1.71% 基于单元选择的转换语音系统。同样地，Wu Z 等人<sup>[112]</sup> 建议使用从相位谱中获得的特征来检测转换后的语音。这些特征在转换后的语音检测器的三种不同训练情况下进行测试：a) 只有基于高斯混合模型 (GMM) 的转换后的语音数据可用；b) 只有基于单元选择的转换语音数据可用；c) 没有转换后的语音数据可用于训练转换后的语音模型。而在美国国家标准与技术研究院 (NIST) 2006 说话人识别评估 (SRE) 语料库上进行的实验表明，从相位谱派生的特征的性能大大优于梅尔频率倒谱系数 (MFCC)：即使没有经过转换的语音进行训练，等错误率 (EER) 从 MFCC 的 20.20% 降低到 2.35%。

Das RK 等人<sup>[113]</sup> 考虑了基于远程声学特征的新对策，这些对策在许多方面都是独一无二的，因为它们是使用倍频程功率谱和子带得出的，而不是常用的线性功率谱，在挑战后研究中，研究者进一步研究了使用深度特征来增强真实和欺骗性语音之间的区分能力。而该研究在从远程声学和深度特征的角度总结了欺骗检测的发现，并从不同类型的欺骗攻击的性质和系统开发进行了综合分析。

Zeinali H 等人<sup>[114]</sup> 的研究介绍了布尔诺理工大学 (BUT) 和 Omilia 共同努力的系统描述—ASVSpoof2019 Spoofing and Countermeasures Challenge 的对话智能，其物理访问 (PA) 的主要提交是两个 VGG 网络的融合，并在单通道和双通道特征上进行了训练。对于逻辑访问 (LA)，研究者的主要系统是 VGG 和最近引入的 SincNet 架构的融

合，其 PA 上的结果表明，所提出的网络在所有条件下都产生了非常有竞争力的性能，并且与官方基线相比实现了 86% 的相对改进。另一方面，LA 上的结果表明，尽管所提出的架构和训练策略在某些欺骗攻击上表现得非常好，但它无法推广到在训练期间看不见的某些攻击下。另外前者运用了 CQT 特征和功率谱图进行学习，而所谓的 CQT 则是 Schörkhuber C 等人<sup>[115]</sup> 所提出了一种计算时域信号恒定 Q 变换 (CQT) 的高效计算方法。CQT 指的是一种时频表示，其中频率区间是几何间隔的，并且所有区间的 Q 因子（中心频率与带宽的比率）相等，该研究提出了一种逆变换，它能够根据其 CQT 系数对原始信号进行合理质量（大约 55dB 信噪比）的重构。在这里，具有高 Q 因子的 CQT（相当于每倍频程 12-96 个 bin）特别令人感兴趣。所提出的方法在每倍频程的 bin 数量、应用的窗口函数和 Q 因子方面是灵活的，并且特别适用于音乐信号的分析。而提出方法的参考实现作为 Matlab 工具箱发布，该工具箱包括用户界面工具，可促进光谱数据可视化以及索引和使用 CQT 生成的数据结构。

Gomez-Alanis A 等人<sup>[116]</sup> 这项工作的目的是开发一个单一的反欺骗系统，该系统可用于有效检测 ASVspoof 2019 挑战赛中考虑的所有类型的欺骗攻击：从文本到语音、语音转换和基于重放的攻击，而该研究为了实现这一点，研究者们建议使用轻卷积门控循环神经网络 (LC-GRNN) 作为深度特征提取器，以稳健地将语音信号表示为话语级嵌入，稍后由后端识别器使用，该识别器执行最终的真实/欺骗分类，这种新颖的架构结合了轻卷积层在帧级别提取判别特征的能力与基于门控循环单元的 RNN 学习后续深度特征的长期依赖关系的能力。所提出的系统已作为对 ASVspoof 2019 挑战赛的贡献而提出，与基线系统相比，结果显示出显著改进。此外，还在 ASVspoof 2015 和 2017 语料库上进行了实验，结果表明我们的提议明显优于最近提出的其他流行方法和其他类似的基于深度特征的系统。另外 Chen T 等人<sup>[117]</sup> 则由于最近语音合成和语音转换技术的突破，Audio Deepfakes，在技术上被称为逻辑访问语音欺骗技术，已经成为语音接口上越来越大的威胁。为了有效检测这些攻击对于包括自动说话人验证系统在内的许多语音应用程序至关重要，同时随着新型语音合成和语音转换技术的迅速出现，欺骗对策的泛化能力正成为越来越关键的挑战。该研究重点通过使用大余量余弦损失函数 (LMCL) 和在线频率掩蔽增强来强制神经网络学习更稳健的特征嵌入来克服这个问题，而研究者在 ASVspoof 2019 逻辑访问 (LA) 数据集上评估了拟议系统的性能。此外，研究者使用公开可用的噪声在 ASVspoof 2019 数据集的噪声版本上对其进行评估，以模拟更真实的场景，最后在通过电话通道逻辑重放的资料集副本上评估所提出的系统，以模拟呼叫中心场景中的欺骗攻击。研究者们的基线系统基于残差神经网络，并在 ASVspoof 2019 挑战期间的所有单系统提交中实现了 4.04% 的最低等错误率 (EER)，此外，该研究提出的额外改进将 EER 降低到 1.26%。

Table 3: Physical access detailed results based on min-tDCF for different conditions. The first section shows the baseline results and the second section shows the primary and single best results of the best-performing systems, both from team T28.

System	Development set									Evaluation set								
	AA	AB	AC	BA	BB	BC	CA	CB	CC	AA	AB	AC	BA	BB	BC	CA	CB	CC
CQCC-GMM	0.4928	0.0539	0.0213	0.3999	0.0360	0.0197	0.4338	0.0414	0.0149	0.4975	0.1751	0.0529	0.4658	0.1483	0.0433	0.5025	0.1360	0.0461
T28 Primary	0.0132	0.0030	0.0009	0.0073	0.0017	0.0009	0.0065	0.0023	0.0008	0.0190	0.0079	0.0034	0.0113	0.0083	0.0022	0.0127	0.0075	0.0024
T28 Single	0.0185	0.0044	0.0013	0.0146	0.0043	0.0014	0.0146	0.0081	0.0024	0.0251	0.0107	0.0055	0.0152	0.0114	0.0058	0.0183	0.0111	0.0063
Primary	0.0389	0.0062	0.0039	0.0243	0.0049	0.0048	0.0233	0.0073	0.0028	0.0776	0.0217	0.0091	0.0586	0.0223	0.0088	0.0557	0.0256	0.0110
Single best	0.0611	0.0046	0.0040	0.0404	0.0052	0.0053	0.0402	0.0085	0.0039	0.1061	0.0267	0.0117	0.0901	0.0277	0.0115	0.0843	0.0330	0.0128
Contrastive1	0.0523	0.0245	0.0151	0.0256	0.0156	0.0130	0.0280	0.0229	0.0135	0.0695	0.0383	0.0148	0.0493	0.0383	0.0141	0.0437	0.0394	0.0192
Contrastive2	0.0726	0.0323	0.0170	0.0562	0.0283	0.0153	0.0633	0.0353	0.0167	0.0969	0.0547	0.0187	0.0843	0.0519	0.0193	0.0842	0.0532	0.0229

Table 4: Logical access detailed results based on min-tDCF for different conditions. The first section shows the baseline results and the second section shows the primary system results of the best performing team (T05) as well as the overall best single system results (team T45). The bold numbers show conditions where our single system performs better or the same as the best single system.

System	Development set						Evaluation set												
	A01	A02	A03	A04	A05	A06	A07	A08	A09	A10	A11	A12	A13	A14	A15	A16	A17	A18	A19
CQCC-GMM	0.0000	0.0000	0.0020	0.0000	0.0261	0.0011	0.0000	0.0007	0.0060	0.4149	0.0020	0.1160	0.6729	0.2629	0.0344	0.0000	0.9820	0.2818	0.0014
T05 Primary	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0009	0.0014	0.0000	0.0077	0.0055	0.0045	0.0028	0.0035	0.0050	0.0015	0.0341	0.0276	0.0020
T45 Single	0.0027	0.0000	0.0000	0.0036	0.0068	0.0085	0.0034	0.0308	0.0000	0.0130	0.0017	0.0058	0.0034	0.0042	0.0065	0.0071	0.9833	0.1171	0.0895
Primary	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0029	0.5672	0.0425	0.0425	0.1098	0.0005	0.5525	0.0000	0.3775	0.6473	0.0000
Single best	<b>0.0000</b>	<b>0.0000</b>	<b>0.0000</b>	<b>0.0000</b>	<b>0.0000</b>	<b>0.0000</b>	<b>0.0002</b>	<b>0.0004</b>	0.1393	0.9423	0.0426	1.0000	0.3693	<b>0.0000</b>	<b>1.0000</b>	<b>0.0004</b>	<b>0.4764</b>	0.6731	<b>0.0000</b>
Contrastive1	0.0000	0.0000	0.0000	0.0000	0.0003	0.0000	0.0654	0.2004	0.1663	0.5031	0.0002	0.9297	0.8583	0.0000	0.0002	0.0007	0.0263	0.5749	0.3217
Contrastive2	0.0000	0.0000	0.0000	0.0010	0.0000	0.0000	0.0017	0.0026	0.1505	0.9992	0.0253	1.0000	0.4737	0.0000	1.0000	0.0022	0.4131	0.9420	0.0009

Table 5: Physical access results of different submissions

System	Development set		Evaluation set	
	EER[%]	min-tDCF	EER [%]	min-tDCF
CQCC-GMM	9.87	0.1953	11.04	0.2454
Primary	0.66	0.0170	1.51	0.0372
Single best	1.02	0.0254	2.11	0.0527
Contrastive1	1.07	0.0253	1.49	0.0401
Contrastive2	1.59	0.0401	2.31	0.0591

Table 6: Logical access results of different submissions

System	Development set		Evaluation set	
	EER[%]	min-tDCF	EER [%]	min-tDCF
CQCC-GMM	0.43	0.0123	9.57	0.2366
Primary	0.00	0.0000	8.01	0.2080
Single best	0.00	0.0000	20.11	0.3563
Contrastive1	0.00	0.0000	10.52	0.2790
Contrastive2	0.03	0.0003	22.99	0.3811

图 3.17 Zeinali H 等人<sup>[114]</sup>

### 3.10 代表性检测技术整理与比较

Li XR 等人近期来的汇整工作<sup>[9]</sup>，其针对各个检测的代表性的主流深度伪造的手段皆有整理，在此本作业根据其技术总结条列其特性、使用模型与相关研究者，使对应由 Thanh Thi Nguyen 等人<sup>[40]</sup>所整理的深度伪造检测工具表列，做一个互相呼应，而更重要的在于 Li XR 等人总结出五大深度伪造检测方法的局限性跟在优势与劣势。

- Fridrich J 等人<sup>[59]</sup> 使用 SVM 模型，其特点为高通图像的隐写特征。
- Cozzolino D 等人<sup>[71]</sup> 使用 CNN 模型，特色为残差特征的学习。
- Afchar D 等人<sup>[47]</sup> 使用 CNN，特点是微观特征的学习。
- Rossler A 等人<sup>[45]</sup> 的 Xception 模型，其特色为对整帧的人脸区域学习。
- Nguyen HH 等人<sup>[90]</sup> 使用模型为 CNN + 胶囊网络，而特色为胶囊网络分类。
- Cozzolino, D 等人<sup>[95]</sup> 使用模型为 Autoencoder，而特色是用于分类和分割双任务。
- Nguyen HH 等人<sup>[96]</sup> 使用模型为 Autoencoder，其特色为分类和分割重建融合。
- Agarwal S 等人<sup>[106]</sup>，其模型为 SVM，而特色则为动作单元编码。

**Table 4** Advantages and disadvantages of various detection methods are summarized  
**表 4** 各类检测方法优劣总结

方法	特点	缺陷
基于图像取证的方法	技术成熟,特征可解释	主要面向图像,压缩等预处理会加大提取难度
基于生理信号的方法	捕获特定的生理特征, 关注图像的局部信息	在压缩的视频里特征提取误差大 一些特征在新技术中被隐藏,准确度不高
基于图像篡改 痕迹的方法	学习局部信息,针对 粗糙的 Deepfakes 有效	通用性不强, 精准度不高
基于 GAN 图像 特征的方法	聚焦 GAN 指纹信息	数据依赖性强,依赖 生成算法,通用性不好
基于数据驱动的方法	数据量大、可学习 信息多,准确度高	依赖同分布数据集,未知 类型以及压缩对性能影响大

图 3.18 Li XR 等人近期来的汇整工作<sup>[9]</sup>，其五大深度伪造检测方法的局限性跟在优势与劣势

- Güera D 等人<sup>[108]</sup> 使用了 CNN + RNN 的模型，而特点则是图片的时序信息。
- Sabir E 等人<sup>[109]</sup> 使用模型为 CNN + Bi-LSTM，而特色为图片的时序信息。
- Zhou P 等人<sup>[72]</sup> 使用 CNN + SVM 模型，并有人脸和隐写特征结合。
- Li Y 等人<sup>[79]</sup> 使用 CNN 模型，特点为学习人脸边框篡改遗留痕迹。
- Matern F 等人<sup>[41]</sup> 使用 Logistic Regression + MLP，其特色为学习篡改痕迹的细节缺失。
- Yang X 等人<sup>[75]</sup> 使用 SVM 模型，并针对头部姿态评估。
- Korshunova I 等人<sup>[17]</sup>，而模型方案则是 PCA + RNN 和 PCA + LDA，其特色为图像质量与声频校对
- Bayar B 等人<sup>[100]</sup>
- Dang H 等人<sup>[103]</sup> 其使用模型为 CNN + Attention，特色则是增加注意力机制。
- Chen T 等人<sup>[117]</sup> 使用模型则是 Deep Residual Network + Frequency Masking，而特点则是大边际距离损失函数。
- Gomez-Alanis A, 等人<sup>[116]</sup> 使用模型为 LightCNN + RNN，而特色为混合光卷积和门递归单元。
- Li R 等人<sup>[118]</sup> 模型则是 Butterfly Unit Multi-Task，特色为多特征融合与多任务学习。
- Zeinali H 等人<sup>[114]</sup> 而模型为 Light CNN 、VGG、SincNet，而特点则是多网络融合。

表 3.2 深度伪造检测手段整理 (续)

方法	技术	检测种类	备注
Head poses	SVM	Videos and Images	Xin Yang 等人
Capsule-forensics	Capsule networks	Videos and Images	Huy H Nguyen 等人
Preprocessing combined with deep network	DCGAN, WGAN-GP and PGGAN.	Images	Xinsheng Xuan 等人
Analyzing convolutional traces	KNN, SVM, and linear discriminant analysis (LDA)	Images	Luca Guarnera 等人
Bag of words and shallow classifiers	SVM, RF, MLP	Images	Ying Zhang 等人
Pairwise learning	CNN concatenated to CFFN	Images	Chih-Chung Hsu 等人
Defenses against adversarial perturbations in deepfakes	VGG and ResNet	Images	Apurva Gandhi 等人
Face X-ray	CNN	Images	Chih-Chung Hsu 等人
Using common artifacts of CNN-generated images	ResNet-50 pre-trained with ImageNet	Images	Sheng-Yu Wang 等人
Using convolutional traces on GAN-based images	KNN, SVM, and LDA	Images	Luca Guarnera 等人
Using deep features extracted by CNN	A new CNN model, namely SCnet	Images	Zhiqing Guo 等人



## 第四章 深度伪造的对抗性研究與近期研究發展

本作业在此章节分为三个部分，其一是针对于深度伪造的对抗性部分，这方面则可以从其深度伪造的生成面与检测面来看，其 Li XR 等人近期来的汇整工作<sup>[19]</sup> 有可以找到工作成果，另外还有 Yisroel Mirsky 等人<sup>[119]</sup> 对其 GAN 的工作总结与近来应用 Transformers 、增量学习在深度伪造领域的检测上的研究工作总结。

### 4.1 深度伪造生成的对抗性

由于近年来深度伪造技术所生成的人类脸部技术能够轻易修改人类的身分，甚至可以使目标人脸做出想要的脸部肌肉表情，使之在人脸身分辨识的领域遭遇到极大的挑战，同时因应人脸识别的对抗性攻击也不曾间断。其 Goswami G 等人<sup>[120]</sup>，发现基于深度神经网络 (DNN) 架构的模型具有很高的表达能力和学习能力，然而，它们本质上是一种黑盒方法，因为在其多层表示中学习到的函数在数学上表示不容易。意识到这一点后，许多研究人员已经开始设计方法来利用基于深度学习的演算法的缺点，质疑它们的鲁棒性并暴露它们的奇点。在此研究中，研究者们试图解开与 DNN 在人脸识别方面的鲁棒性相关的三个方面：(i) 评估深度架构对人脸识别的影响，以应对攻击的脆弱性，这些攻击受到现实世界中普遍观察到的扭曲的启发，浅层学习方法和基于学习的对手可以很好地处理这些扭曲；(ii) 通过表征深层网络隐藏层中的异常滤波器响应行为来检测奇异点；和 (iii) 对处理管道进行更正以缓解问题。该研究则使用多个开源的基于 DNN 的人脸识别网络（包括 OpenFace 和 VGG-Face）以及两个公开可用的资料库（MEDS 和 PaSC）的实验评估表明，基于深度学习的人脸识别算法的性能在存在这样的扭曲。同时该方法还与现有的检测算法进行了比较，结果表明，通过使用网络中隐藏层的响应适当地设计分类器，它能够以非常高的精度检测攻击。最后，研究提出了几种有效的对策来减轻对抗性攻击的影响并提高基于 DNN 的人脸识别的整体鲁棒性，其结果发现只要对图片与影像增加一定的遮挡或者在影像加入人类肉眼看不见得噪声，就能够有骗过机器的可能，该工作展示对 Parkhi OM 等人<sup>[121]</sup> 所以提出 VGGface ，与 Baltruaitis T 等人<sup>[122]</sup> 所做的 Openface 等模型的实验。

Song Q 等人<sup>[123]</sup> 专注于一种对人脸识别网络进行攻击的新颖方法，该方法会误导网络将某人识别为目标人，而不是不明显地错误分类，同时，因为此缘故，研究者引入了一个特定的注意力对抗攻击生成网络来生成假人脸图像。为了捕获目标人的语义信息，这项工作添加了条件变分自动编码器和注意模块来学习人脸之间的实例级对应关系。与传统的双人 GAN 不同，这项工作引入了人脸识别网络作为第三个参与者参与



Figure 1: We show that deep learning based OpenFace (OF) and VGG-Face can be deceived even by image processing operations that mimic real world distortions.

图 4.1 Goswami G 等人<sup>[120]</sup>

生成器和判别器之间的竞争，这使得攻击者可以更好地模仿目标人。生成的结果难以引起旁观者注意的人脸可以逃避最先进网络的识别，并且大多数人都被识别为目标人。

Majumdar P 等人<sup>[124]</sup>提出了部分面部篡改攻击，其中面部区域被替换或变形以生成篡改样本，而在 CMU-MultiPIE 数据集上使用两个最先进的人脸识别系统 VGG-Face 和 OpenFace 进行的人脸验证实验表明了这些系统对攻击的脆弱性。此外，该研究提出了一种部分人脸篡改检测（PFTD）网络来检测所提出的攻击，该网络通过结合输入图像的原始信息和高频信息来捕获原始图像和篡改图像之间的不一致性，以检测篡改图像。所提出的网络在篡改图像检测方面超越了现有基线深度神经网络的性能。

另外 Korshunov P 等人<sup>[46]</sup>，通过使用预先训练的生成对抗网络（GAN），将视频中的一个人的脸自动替换为另一个人的脸变得越来越容易。最近的公共丑闻，例如，名人的面孔被交换到色情视频上，需要自动检测这些 Deepfake 视频的方法，为了帮助开发此类方法，在本文中，我们展示了第一组从 VidTIMIT 数据库的视频中生成的公开可用的 Deepfake 视频。研究者使用基于 GAN 的开源软件来创建 Deepfakes，该研究强调训练和混合参数可以显著影响结果视频的质量。为了证明这种影响，研究者使用不同调整的参数集生成了具有低和高视觉质量的视频（每个 320 个视频），其研究展示了

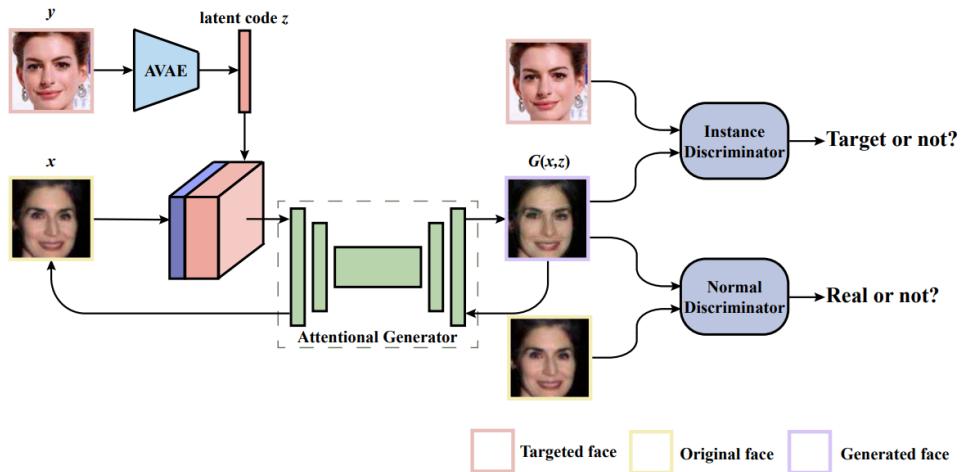


Figure 2. Overview of  $A^3GN$ . Attentional variational autoencoder (AVAE) captures the latent code  $z$  from target face  $y$ . And then the original face  $x$  is concatenated with  $z$  to generate  $\hat{x}$ ,  $G(x,z) \rightarrow \hat{x}$  in attentional generator.  $G(x,z)$  is sent into normal discriminator to determine whether it is a real image or not with  $x$  and sent into instance discriminator to determine whether it can be classified as the target person or not with  $y$ .

图 4.2 Song Q 等人<sup>[123]</sup>

基于 VGG 和 Facenet 神经网络的最先进的人脸识别系统容易受到 Deepfake 视频的攻击，错误接受率分别为 85.62% 和 95.00%，这意味着检测 Deepfake 视频的方法是必要的。通过考虑几种基线方法，我们发现基于口型同步不一致检测的视听方法无法区分 Deepfake 视频。性能最佳的方法，基于视觉质量指标，常用于演示攻击检测领域，在高质量 Deepfakes 上的错误率为 8.97%，最后的实验表明，GAN 生成的 Deepfake 视频对人脸识别系统和现有检测方法都具有挑战性，而人脸交换技术的进一步发展将使其变得更加困难。同样的也是 Korshunov P 等人<sup>[125]</sup>展示了 Deepfake 视频的公开数据集，其中的人脸使用基于 GAN 的算法变形，同时为了生成这些影像，研究者使用了基于 GAN 的开源软件，并且我们强调训练和混合参数可以显著影响生成影像的品质，该研究表明，基于 VGG 和 Facenet 神经网络的最先进的人脸识别系统容易受到深度变形视频的影响，错误接受率分别为 85.62 和 95.00，这意味着检测这些视频的方法是必要的。同时研究也考虑了几种检测深度变形的基线方法，并发现基于视觉品质指标的方法（通常用于演示攻击检测领域）导致最佳性能，错误率为 8.97。而最后研究者的实验表明，GAN 生成的深度变形视频对人脸识别系统和现有检测方法都具有挑战性，而深度变形技术的进一步发展将使其更加如此。

## 4.2 深度伪造检测的对抗性

其深度伪造领域的演算法多数都是运用类神经网路等技术，而其神经网路模型本身即有着对抗样本的攻击，从 Szegedy C 等人<sup>[126]</sup>报告了两个这样的属性。首先，根据单元分析的各种方法，该研究发现单个高级单元和高级单元的随机线性组合之间没

有区别。其研究表明，在神经网络的高层中，包含语义信息的是空间，而不是单个单元。其次，研究者们发现深度神经网络学习的输入-输出映射在很大程度上是不连续的。可以通过应用某种不可察觉的扰动来导致网络对图像进行错误分类，这种扰动是通过最大化网络的预测误差来发现的。此外，这些扰动的具体性质并不是学习的随机伪影：相同的扰动可能导致在数据集的不同子集上训练的不同网络对相同的输入进行错误分类。另 Goodfellow IJ 等人<sup>[127]</sup> 认为神经网络易受对抗性扰动影响的主要原因是它们的线性性质，这种解释得到了新的定量结果的支持，同时给出了关于它们最有趣的事实的第一个解释：它们在架构和训练集上的泛化。此外，这种观点产生了一种简单而快速的生成对抗样本的方法。使用这种方法为对抗性训练提供示例，该研究减少了 MNIST 数据集上 maxout 网络的测试集误差。

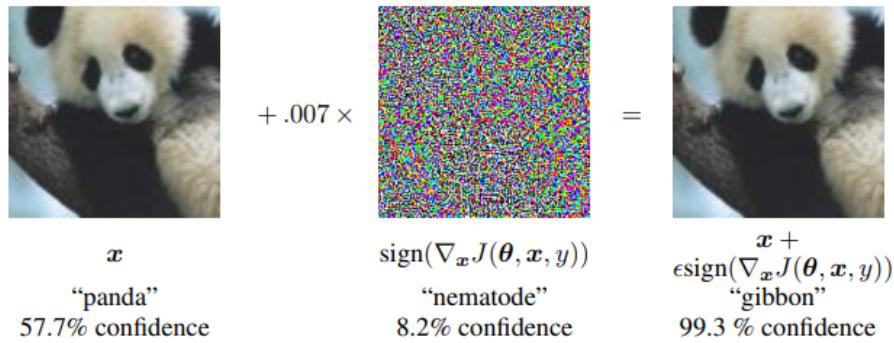


Figure 1: A demonstration of fast adversarial example generation applied to GoogLeNet (Szegedy et al., 2014a) on ImageNet. By adding an imperceptibly small vector whose elements are equal to the sign of the elements of the gradient of the cost function with respect to the input, we can change GoogLeNet’s classification of the image. Here our  $\epsilon$  of .007 corresponds to the magnitude of the smallest bit of an 8 bit image encoding after GoogLeNet’s conversion to real numbers.

图 4.3 Goodfellow IJ 等人<sup>[127]</sup>

另外 Kurakin A 等人<sup>[128]</sup> 发现大多数现有的机器学习分类器都非常容易受到对抗性示例的影响，对抗性示例是输入数据的样本，该样本经过了非常轻微的修改，旨在导致机器学习分类器对其进行错误分类。在许多情况下，这些修改可能非常微妙，以至于人类观察者甚至根本没有注意到修改，但分类器仍然会出错。对抗性示例会带来安全问题，因为它们可用于对机器学习系统进行攻击，即使对手无法访问底层模型。到目前为止，所有先前的工作都假设了一个威胁模型，其中攻击者可以将数据直接输入机器学习分类器。在物理世界中运行的系统并非总是如此，例如那些使用来自相机和其他传感器的信号作为输入的系统。该研究表明，即使在这样的物理世界场景中，机器学习系统也容易受到对抗性示例的影响。研究者们通过将从手机摄像头获得的对抗性图像馈送到 ImageNet Inception 分类器并测量系统的分类精度来证明这一点，最后研究者也发现，即使通过摄像头感知，大部分对抗性示例也被错误分类。

从上述几项研究的结果可以得知，这些由神经网络组成的模型在面对对抗性样本攻击时，会导致模型受到干扰，从而产生误判，由此也造成深度伪造技术在生成时可隐藏自身特征，从而绕过检测，由此有必要对当下的模型跟算法做对抗性的评估，另外 Wang SY 等人<sup>[128]</sup> 研究尝试是否有可能创造一个“通用”检测器，用于将真实图像与 CNN 生成的图像区分开来，无论使用何种架构或数据集。为了测试这一点，研究者们收集了一个数据集，该数据集由 11 种不同的基于 CNN 的图像生成器模型生成的假图像组成，这些模型跨越当今常用架构的空间（ProGAN、StyleGAN、BigGAN、CycleGAN、StarGAN、GauGAN、DeepFakes、级联细化网络，隐式最大似然估计，二阶注意力超分辨率，在黑暗中看到）。其研究者证明，通过仔细的预处理和后处理以及数据增强，仅在一个特定的 CNN 生成器 (ProGAN) 上训练的标准图像分类器能够很好地泛化到看不见的架构、数据集和训练方法（包括刚刚发布的 StyleGAN2）。而研究者的研究结果表明，当今 CNN 生成的图像存在一些常见的系统缺陷，从而阻止它们实现逼真的图像合成，这一有趣的可能性，同时该工作对训练资料进行类似于 JPEG 压缩、模糊等操作手法，可以提高模型的泛化性能。

另外 Neves JC 等人<sup>[129]</sup> 专注于整个面部图像的合成，这是一种特定类型的面部操作。该研究的主要贡献有四方面：i) 描述了一种从基于自动编码器的合成假图像中去除 GAN “指纹”的新策略，以欺骗面部操作检测系统，同时保持结果图像的视觉质量；ii) 对近期面部操作检测文献的深入分析；iii) 对这种类型的面部操作进行完整的实验评估，考虑到最先进的假检测系统（基于整体深度网络、隐写分析和局部伪影），并指出这项任务在不受约束的场景中的挑战性；最后 iv) 研究者们宣布了一个名为 iFakeFaceDB 的新型公共数据库，该数据库将该研究所提出的 GAN 指纹去除方法 (GANprintR) 应用于已经非常逼真的合成假图像。在该研究的实证评估中获得的结果表明，需要额外的努力来开发针对看不见的条件和欺骗技术的强大的面部操作检测系统，例如本研究中提出的技术。

Brockschmidt J 等人<sup>[104]</sup> 对都属于卷积神经网络 (CNN) 的 Rossler A 等人<sup>[45]</sup> 所做的 Xception 与 Afchar D 等人<sup>[47]</sup> 所做的 Mesonet 进行对抗性评估，其实验使用来自六种最先进的面部伪造技术的样本：Deepfakes、Face2Face、FaceSwap、GANnotation、ICface 和 X2Face。研究者发现 MesoNet 和 XceptionNet 显示出泛化到多种欺骗技术的潜力，但在准确性上略有权衡，并且在很大程度上无法对抗看不见的技术。最后将这些结果松散地推断为类似的 CNN 架构，并强调需要更好的架构来应对普遍性的挑战。

Marra F 等人<sup>[130]</sup> 研究了几种图像伪造检测器对图像到图像转换的性能，无论是在理想条件下，还是在存在压缩的情况下，通常在上传到社交网络时执行。该研究在 36302 张图像的数据集上进行，表明传统和深度学习检测器都可以实现高达 95% 的检

测精度，但只有后者在压缩数据上保持高达 89% 的高精度，换言之现有的检测器若面对未知的压缩与类型，其表现并不理想。Zhang X 等人<sup>[131]</sup> 检测 GAN 生成的图像，传统的监督机器学习算法需要从目标 GAN 模型中收集大量真实和虚假图像。但是，攻击者使用的特定模型通常是不可用的。为了解决这个问题，该研究提出了一个 GAN 模拟器 AutoGAN，它可以模拟由几个流行的 GAN 模型共享的公共管道产生的伪影，此外，研究者确定了由通用 GAN 管道中包含的上采样组件引起的独特伪影。该研究从理论上表明，这种伪影表现为频域中光谱的复制，因此提出了一种基于光谱输入而不是像素输入的分类器模型。通过使用模拟图像来训练基于频谱的分类器，即使在训练期间没有看到目标 GAN 模型产生的假图像，此研究的方法在检测由流行 GAN 模型（如 CycleGAN）生成的假图像方面也取得了最先进的性能。

Du M 等人<sup>[132]</sup> 提出了 Locality-Aware AutoEncoder (LAE) 来弥补泛化差距。在训练过程中，研究者使用像素级掩码来规范 LAE 的局部解释，以强制模型从伪造区域学习内在表示，而不是在训练集中捕获伪影并学习表面相关性来执行检测。而该研究进一步提出了一个主动学习框架来选择具有挑战性的标签候选者，该框架需要不到 3% 的训练数据使用人工蒙版，从而大大减少了规范化解释的注释工作。三个 deepfake 检测任务的实验结果表明，LAE 可以专注于伪造区域来做出决策。分析进一步表明，就先前未见过的操作的泛化精度而言，LAE 在三个深度伪造检测任务上的性能分别优于现有技术 6.52%、12.03% 和 3.08%。

Huang, R 等人<sup>[133]</sup> 通过实验证明了个体对抗性扰动 (IAP) 和普遍对抗性扰动 (UAP) 的存在，它们可能导致表现良好的 FFM 行为不端。基于迭代过程，梯度信息用于生成两种可用于制造分类和分割输出的 IAP。相比之下，UAP 是在过火的基础上生成的。研究者们设计了一个新的目标函数，鼓励神经元过度激发，即使不使用训练数据也可以生成 UAP，实验证明了 UAP 在未见数据集和未见 FFM 之间的可转移性。此外，研究者对对抗性扰动的不可察觉性进行了主观评估，表明精心制作的 UAP 在视觉上可以忽略不计。这些发现为评估 FFM 的对抗性安全性提供了基准。

### 4.3 GAN

### 4.4 Transformers 與增量学习

## 第五章 结论



## 参考文献

- [1] 邱锡鹏. 神经网络与深度学习[M/OL]. 北京: 机械工业出版社, 2020. <https://nnndl.github.io/>.
- [2] 邱文聪. 第二波人工智能知识学习与生产对法学的挑战—资讯、科技与社会研究及法学的对话[J], 2021.
- [3] 陈弘儒. 初探目的解释在法律人工智能系统之运用可能[J], 2021.
- [4] Deepfakes.[EB/OL]. <https://github.com/deepfakes/faceswap>.
- [5] Reface app.[EB/OL]. <https://hey.reface.ai/>.
- [6] Deepfake detection challenge.[EB/OL]. <https://www.kaggle.com/c/deepfake-detection-challenge>.
- [7] GIRISH N, NANDINI C. A review on digital video forgery detection techniques in cyber forensics[J]. Science, Technology and Development, 2019, 3(6): 235-239.
- [8] NGUYEN T T, NGUYEN Q V H, NGUYEN C M, et al. Deep learning for deepfakes creation and detection: A survey[J]. ArXiv preprint arXiv:1909.11573, 2019.
- [9] 李旭嵘纪守领吴春明刘振广邓水光程鹏杨珉孔祥维. 深度伪造与检测技术综述[J]. 软件学报, 2021, 32(2): 496. DOI: 10.13328/j.cnki.jos.006140.
- [10] FaceSwap.[EB/OL]. <https://github.com/MarekKowalski/FaceSwap/>.
- [11] DALE K, SUNKAVALLI K, JOHNSON M K, et al. Video face replacement[C]//Proceedings of the 2011 SIGGRAPH Asia conference. [S.l. : s.n.], 2011: 1-10.
- [12] GARRIDO P, VALGAERTS L, REHMSEN O, et al. Automatic face reenactment[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. [S.l. : s.n.], 2014: 4217-4224.
- [13] GARRIDO P, VALGAERTS L, SARMADI H, et al. Vdub: Modifying face video of actors for plausible visual alignment to a dubbed audio track[C]//Computer graphics forum: vol. 34: 2. [S.l. : s.n.], 2015: 193-204.
- [14] NIRKIN Y, MASII, 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). [S.l. : s.n.], 2018: 98-105.
- [15] LU Z, LI Z, CAO J, et al. Recent progress of face image synthesis[C]//2017 4th IAPR Asian Conference on Pattern Recognition (ACPR). [S.l. : s.n.], 2017: 7-12.
- [16] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial nets[J]. Advances in neural information processing systems, 2014, 27.
- [17] KORSHUNOVA I, SHI W, DAMBRE J, et al. Fast face-swap using convolutional neural networks[C] //Proceedings of the IEEE international conference on computer vision. [S.l. : s.n.], 2017: 3677-3685.
- [18] NIRKIN Y, KELLER Y, HASSNER T. Fsgan: Subject agnostic face swapping and reenactment[C] //Proceedings of the IEEE/CVF international conference on computer vision. [S.l. : s.n.], 2019: 7184-7193.

- [19] CHOI Y, CHOI M, KIM M, et al. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. [S.l. : s.n.], 2018: 8789-8797.
- [20] ZHANG H, XU T, LI H, et al. Stackgan++: Realistic image synthesis with stacked generative adversarial networks[J]. IEEE transactions on pattern analysis and machine intelligence, 2018, 41(8): 1947-1962.
- [21] KARRAS T, AILA T, LAINE S, et al. Progressive growing of gans for improved quality, stability, and variation[J]. ArXiv preprint arXiv:1710.10196, 2017.
- [22] ANTIPOV G, BACCOUCHE M, DUGELAY JL. Face aging with conditional generative adversarial networks[C]//2017 IEEE international conference on image processing (ICIP). [S.l. : s.n.], 2017: 2089-2093.
- [23] MIRZA M, OSINDERO S. Conditional generative adversarial nets[J]. ArXiv preprint arXiv:1411.1784, 2014.
- [24] HUANG R, ZHANG S, LI T, et al. Beyond face rotation: Global and local perception gan for photo-realistic and identity preserving frontal view synthesis[C]//Proceedings of the IEEE international conference on computer vision. [S.l. : s.n.], 2017: 2439-2448.
- [25] THIES J, ZOLLHÖFER M, NIESSNER M, et al. Real-time expression transfer for facial reenactment.[J]. ACM Trans. Graph., 2015, 34(6): 183-1.
- [26] THIES J, ZOLLHOFER M, STAMMINGER M, et al. Face2face: Real-time face capture and reenactment of rgb videos[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. [S.l. : s.n.], 2016: 2387-2395.
- [27] KIM H, GARRIDO P, TEWARI A, et al. Deep video portraits[J]. ACM Transactions on Graphics (TOG), 2018, 37(4): 1-14.
- [28] THIES J, ZOLLHÖFER M, NIESSNER M. Deferred neural rendering: Image synthesis using neural textures[J]. ACM Transactions on Graphics (TOG), 2019, 38(4): 1-12.
- [29] SUWAJANAKORN S, SEITZ S M, KEMELMACHER-SHLIZERMAN I. Synthesizing Obama: Learning Lip Sync from Audio[J/OL]. ACM Trans. Graph., 2017, 36(4). <https://doi.org/10.1145/3072959.3073640>. DOI: 10.1145/3072959.3073640.
- [30] ZAKHAROV E, SHYSHEY A, BURKOV E, et al. Few-Shot Adversarial Learning of Realistic Neural Talking Head Models[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). [S.l. : s.n.], 2019.
- [31] FRIED O, TEWARI A, ZOLLHÖFER M, et al. Text-Based Editing of Talking-Head Video[J/OL]. ACM Trans. Graph., 2019, 38(4). <https://doi.org/10.1145/3306346.3323028>. DOI: 10.1145/3306346.3323028.
- [32] AVERBUCH-ELOR H, COHEN-OR D, KOPF J, et al. Bringing Portraits to Life[J/OL]. ACM Trans. Graph., 2017, 36(6). <https://doi.org/10.1145/3130800.3130818>. DOI: 10.1145/3130800.3130818.
- [33] LAMPLE G, ZEGHIDOUR N, USUNIER N, et al. Fader Networks:Manipulating Images by Sliding Attributes[C/OL]//GUYON I, LUXBURG U V, BENGIO S, et al. Advances in Neural Information

- Processing Systems: vol. 30. [S.I.]: Curran Associates, Inc., 2017. <https://proceedings.neurips.cc/paper/2017/file/3fd60983292458bf7dee75f12d5e9e05-Paper.pdf>.
- [34] ARIK S Ö, CHRZANOWSKI M, COATES A, et al. Deep voice: Real-time neural text-to-speech[C] //International Conference on Machine Learning. [S.I. : s.n.], 2017: 195-204.
- [35] WANG Y, SKERRY-RYAN R, STANTON D, et al. Tacotron: Towards end-to-end speech synthesis[J]. ArXiv preprint arXiv:1703.10135, 2017.
- [36] GIBIANSKY A, ARIK S, DIAMOS G, et al. Deep voice 2: Multi-speaker neural text-to-speech[J]. Advances in neural information processing systems, 2017, 30.
- [37] PING W, PENG K, GIBIANSKY A, et al. Deep Voice 3: 2000-Speaker Neural Text-to-Speech.[J]., 2017.
- [38] PASCUAL S, BONAFONTE A, SERRA J. SEGAN: Speech enhancement generative adversarial network[J]. ArXiv preprint arXiv:1703.09452, 2017.
- [39] DONAHUE C, MCAULEY J, PUCKETTE M. Adversarial audio synthesis[J]. ArXiv preprint arXiv:1802.04208, 2018.
- [40] NGUYEN T T, NGUYEN Q V H, NGUYEN D T, et al. Deep Learning for Deepfakes Creation and Detection: A Survey[EB/OL]. arXiv. 2019. <https://arxiv.org/abs/1909.11573>.
- [41] 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). [S.I. : s.n.], 2019: 83-92.
- [42] FakeApp.[EB/OL]. <https://www.deepfakescn.com>.
- [43] RÖSSLER A, COZZOLINO D, VERDOLIVA L, et al. Faceforensics: A large-scale video dataset for forgery detection in human faces[J]. ArXiv preprint arXiv:1803.09179, 2018.
- [44] ABU-EL-HAIJA S, KOTHARI N, LEE J, et al. Youtube-8m: A large-scale video classification benchmark[J]. ArXiv preprint arXiv:1609.08675, 2016.
- [45] ROSSLER A, COZZOLINO D, VERDOLIVA L, et al. Faceforensics++: Learning to detect manipulated facial images[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. [S.I. : s.n.], 2019: 1-11.
- [46] KORSHUNOV P, MARCEL S. Deepfakes: a new threat to face recognition? assessment and detection[J]. ArXiv preprint arXiv:1812.08685, 2018.
- [47] 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). [S.I. : s.n.], 2018: 1-7.
- [48] LI Y, YANG X, SUN P, et al. Celeb-df: A new dataset for deepfake forensics[J]., 2019.
- [49] DeepfakeDetection[EB/OL]. <https://github.com/ondyari/FaceForensics>.
- [50] DOLHANSKY B, HOWES R, PFLAUM B, et al. The deepfake detection challenge (dfdc) preview dataset[J]. ArXiv preprint arXiv:1910.08854, 2019.
- [51] DFDC.[EB/OL]. <https://www.kaggle.com/c/deepfake-detection-challenge/data>.

- [52] JIANG L, LI R, WU W, et al. Deeperforensics-1.0: A large-scale dataset for real-world face forgery detection[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. [S.l. : s.n.], 2020: 2889-2898.
- [53] ASVspoof 2015 database.[EB/OL]. <https://datashare.is.ed.ac.uk/handle/10283/853>.
- [54] ASVspoof 2015 database.[EB/OL]. <https://datashare.is.ed.ac.uk/handle/10283/853>.
- [55] DE CARVALHO T J, RIESS C, ANGELOPOULOU E, et al. Exposing digital image forgeries by illumination color classification[J]. IEEE Transactions on Information Forensics and Security, 2013, 8(7): 1182-1194.
- [56] AMERINI I, BALLAN L, CALDELLI R, et al. A sift-based forensic method for copy-move attack detection and transformation recovery[J]. IEEE transactions on information forensics and security, 2011, 6(3): 1099-1110.
- [57] LUKÁ J, FRIDRICH J, GOLJAN M. Detecting digital image forgeries using sensor pattern noise[C] //Security, Steganography, and Watermarking of Multimedia Contents VIII: vol. 6072. [S.l. : s.n.], 2006: 362-372.
- [58] CHERCHIA G, PARRILLI S, POGGI G, et al. PRNU-based detection of small-size image forgeries[C]//2011 17th International Conference on Digital Signal Processing (DSP). [S.l. : s.n.], 2011: 1-6.
- [59] FRIDRICH J, KODOVSKY J. Rich models for steganalysis of digital images[J]. IEEE Transactions on information Forensics and Security, 2012, 7(3): 868-882.
- [60] WANG W, DONG J, TAN T. Exploring DCT coefficient quantization effects for local tampering detection[J]. IEEE Transactions on Information Forensics and Security, 2014, 9(10): 1653-1666.
- [61] NATARAJ L, SARKAR A, MANJUNATH B S. Adding gaussian noise to “denoise” JPEG for detecting image resizing[C]//2009 16th IEEE International Conference on Image Processing (ICIP). [S.l. : s.n.], 2009: 1493-1496.
- [62] BIANCHI T, DE ROSA A, PIVA A. Improved DCT coefficient analysis for forgery localization in JPEG images[C]//2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). [S.l. : s.n.], 2011: 2444-2447.
- [63] PAN X, ZHANG X, LYU S. Exposing image splicing with inconsistent local noise variances[C] //2012 IEEE International Conference on Computational Photography (ICCP). [S.l. : s.n.], 2012: 1-10.
- [64] FERRARA P, BIANCHI T, DE ROSA A, et al. Image Forgery Localization via Fine-Grained Analysis of CFA Artifacts[J]. IEEE Transactions on Information Forensics and Security, 2012, 7(5): 1566-1577. DOI: 10.1109/TIFS.2012.2202227.
- [65] COZZOLINO D, VERDOLIVA L. Noiseprint: a CNN-based camera model fingerprint[J]. IEEE Transactions on Information Forensics and Security, 2019, 15: 144-159.
- [66] ZHOU P, HAN X, MORARIU V I, et al. Learning rich features for image manipulation detection[C] //Proceedings of the IEEE conference on computer vision and pattern recognition. [S.l. : s.n.], 2018: 1053-1061.

- [67] RAO Y, NI J. A deep learning approach to detection of splicing and copy-move forgeries in images[C]//2016 IEEE International Workshop on Information Forensics and Security (WIFS). [S.l. : s.n.], 2016: 1-6.
- [68] LIU B, PUN C M. Deep fusion network for splicing forgery localization[C]//Proceedings of the European Conference on Computer Vision (ECCV) Workshops. [S.l. : s.n.], 2018: 0–0.
- [69] HUH M, LIU A, OWENS A, et al. Fighting fake news: Image splice detection via learned self-consistency[C]//Proceedings of the European conference on computer vision (ECCV). [S.l. : s.n.], 2018: 101-117.
- [70] CUN X, PUN C M. Image splicing localization via semi-global network and fully connected conditional random fields[C]//Proceedings of the European Conference on Computer Vision (ECCV) Workshops. [S.l. : s.n.], 2018: 0–0.
- [71] COZZOLINO D, POGGI G, VERDOLIVA L. Recasting residual-based local descriptors as convolutional neural networks: an application to image forgery detection[C]//Proceedings of the 5th ACM Workshop on Information Hiding and Multimedia Security. [S.l. : s.n.], 2017: 159-164.
- [72] ZHOUP, HAN X, MORARIU V I, et al. Two-stream neural networks for tampered face detection[C] //2017 IEEE conference on computer vision and pattern recognition workshops (CVPRW). [S.l. : s.n.], 2017: 1831-1839.
- [73] CHEN C, MCCLOSKEY S, YU J. Focus manipulation detection via photometric histogram analysis[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. [S.l. : s.n.], 2018: 1674-1682.
- [74] SZEGEDY C, LIU W, JIA Y, et al. Going deeper with convolutions[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. [S.l. : s.n.], 2015: 1-9.
- [75] YANG X, LI Y, LYU S. Exposing deep fakes using inconsistent head poses[C]//ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). [S.l. : s.n.], 2019: 8261-8265.
- [76] LIY C M, INICTUOCULIL. Exposing AICreated FakeVideosby DetectingEyeBlinking[C]//2018 IEEE InterG national Workshop on Information Forensics and Security (WIFS). IEEE. [S.l. : s.n.], 2018.
- [77] CIFTCI U A, DEMIR I, YIN L. Fakecatcher: Detection of synthetic portrait videos using biological signals[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020.
- [78] FERNANDES S, RAJ S, ORTIZ E, et al. Predicting heart rate variations of deepfake videos using neural ode[C]//Proceedings of the IEEE/CVF international conference on computer vision workshops. [S.l. : s.n.], 2019: 0–0.
- [79] LI Y, LYU S. Exposing deepfake videos by detecting face warping artifacts[J]. ArXiv preprint arXiv:1811.00656, 2018.
- [80] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. [S.l. : s.n.], 2016: 770-778.
- [81] NATARAJ L, MOHAMMED T M, MANJUNATH B, et al. Detecting GAN generated fake images using co-occurrence matrices[J]. Electronic Imaging, 2019, 2019(5): 532-1.

- [82] LI H, LI B, TAN S, et al. Identification of deep network generated images using disparities in color components[J]. *Signal Processing*, 2020, 174: 107616.
- [83] XUAN X, PENG B, WANG W, et al. On the generalization of GAN image forensics[C]//Chinese conference on biometric recognition. [S.l. : s.n.], 2019: 134-141.
- [84] MCCLOSKEY S, ALBRIGHT M. Detecting gan-generated imagery using color cues[J]. ArXiv preprint arXiv:1812.08247, 2018.
- [85] MARRA F, GRAGNIELLO D, VERDOLIVA L, et al. Do gans leave artificial fingerprints?[C] //2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR). [S.l. : s.n.], 2019: 506-511.
- [86] YU N, DAVIS L S, FRITZ M. Attributing fake images to gans: Learning and analyzing gan fingerprints[C]//Proceedings of the IEEE/CVF international conference on computer vision. [S.l. : s.n.], 2019: 7556-7566.
- [87] WANG R, MA L, JUEFEI-XU F, et al. Fakespotter: A simple baseline for spotting ai-synthesized fake faces. arXiv 2019[J]. ArXiv preprint arXiv:1909.06122,
- [88] CHOLLET F. Xception: Deep learning with depthwise separable convolutions[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. [S.l. : s.n.], 2017: 1251-1258.
- [89] SONGSRI-IN K, ZAFEIRIOU S. Complement face forensic detection and localization with facial-landmarks[J]. ArXiv preprint arXiv:1910.05455, 2019.
- [90] NGUYEN H H, YAMAGISHI J, ECHIZEN I. Capsule-forensics: Using capsule networks to detect forged images and videos[C]//ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). [S.l. : s.n.], 2019: 2307-2311.
- [91] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[J]. ArXiv preprint arXiv:1409.1556, 2014.
- [92] MO H, CHEN B, LUO W. Fake faces identification via convolutional neural network[C]// Proceedings of the 6th ACM workshop on information hiding and multimedia security. [S.l. : s.n.], 2018: 43-47.
- [93] DURALL R, KEUPER M, PFREUNDT F J, et al. Unmasking deepfakes with simple features[J]. ArXiv preprint arXiv:1911.00686, 2019.
- [94] DING X, RAZIEI Z, LARSON E C, et al. Swapped face detection using deep learning and subjective assessment[J]. EURASIP Journal on Information Security, 2020, 2020(1): 1-12.
- [95] COZZOLINO D, THIES J, RÖSSLER A, et al. Forensictransfer: Weakly-supervised domain adaptation for forgery detection[J]. ArXiv preprint arXiv:1812.02510, 2018.
- [96] NGUYEN H H, FANG F, YAMAGISHI J, et al. Multi-task learning for detecting and segmenting manipulated facial images and videos[J]. ArXiv preprint arXiv:1906.06876, 2019.
- [97] HSU C C, LEE C Y, ZHUANG Y X. Learning to detect fake face images in the wild[C]//2018 International Symposium on Computer, Consumer and Control (IS3C). [S.l. : s.n.], 2018: 388-391.
- [98] HSU C C, ZHUANG Y X, LEE C Y. Deep fake image detection based on pairwise learning[J]. Applied Sciences, 2020, 10(1): 370.

- [99] DANG L M, HASSAN S I, IM S, et al. Deep learning based computer generated face identification using convolutional neural network[J]. *Applied Sciences*, 2018, 8(12): 2610.
- [100] BAYAR B, STAMM M C. A deep learning approach to universal image manipulation detection using a new convolutional layer[C]//Proceedings of the 4th ACM workshop on information hiding and multimedia security. [S.l. : s.n.], 2016: 5-10.
- [101] LI X, YU K, JI S, et al. Fighting against deepfake: Patch&pair convolutional neural networks (PPCNN)[C]//Companion Proceedings of the Web Conference 2020. [S.l. : s.n.], 2020: 88-89.
- [102] RAHMOUNI N, NOZICK V, YAMAGISHI J, et al. Distinguishing computer graphics from natural images using convolution neural networks[C]//2017 IEEE Workshop on Information Forensics and Security (WIFS). [S.l. : s.n.], 2017: 1-6.
- [103] DANG H, LIU F, STEHOUWER J, et al. On the detection of digital face manipulation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern recognition. [S.l. : s.n.], 2020: 5781-5790.
- [104] BROCKSCHMIDT J, SHANG J, WU J. On the generality of facial forgery detection[C]//2019 IEEE 16th International Conference on Mobile Ad Hoc and Sensor Systems Workshops (MASSW). [S.l. : s.n.], 2019: 43-47.
- [105] SOHRAWARDI S J, CHINTHA A, THAI B, et al. Poster: Towards robust open-world detection of deepfakes[C]//Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security. [S.l. : s.n.], 2019: 2613-2615.
- [106] AGARWAL S, FARID H, GU Y, et al. Protecting World Leaders Against Deep Fakes.[C]//CVPR workshops: vol. 1. [S.l. : s.n.], 2019.
- [107] AMERINI I, GALTERI L, CALDELLI R, et al. Deepfake video detection through optical flow based cnn[C]//Proceedings of the IEEE/CVF international conference on computer vision workshops. [S.l. : s.n.], 2019: 0-0.
- [108] GÜERA 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). [S.l. : s.n.], 2018: 1-6.
- [109] SABIR E, CHENG J, JAISWAL A, et al. Recurrent convolutional strategies for face manipulation detection in videos[J]. *Interfaces (GUI)*, 2019, 3(1): 80-87.
- [110] TODISCO M, DELGADO H, EVANS N W. A New Feature for Automatic Speaker Verification Anti-Spoofing: Constant Q Cepstral Coefficients.[C]//Odyssey: vol. 2016. [S.l. : s.n.], 2016: 283-290.
- [111] WU Z, KINNUNEN T, CHNG E S, et al. A study on spoofing attack in state-of-the-art speaker verification: the telephone speech case[C]//Proceedings of The 2012 Asia Pacific Signal and Information Processing Association Annual Summit and Conference. [S.l. : s.n.], 2012: 1-5.
- [112] WU Z, CHNG E S, LI H. Detecting converted speech and natural speech for anti-spoofing attack in speaker recognition[C]//Thirteenth Annual Conference of the International Speech Communication Association. [S.l. : s.n.], 2012.

- [113] DAS R K, YANG J, LI H. Long range acoustic and deep features perspective on ASVspoof 2019[C] //2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). [S.l. : s.n.], 2019: 1018-1025.
- [114] ZEINALI H, STAFYLAKIS T, ATHANASOPOULOU G, et al. Detecting spoofing attacks using vgg and sincnet: but-omilia submission to asvspoof 2019 challenge[J]. ArXiv preprint arXiv:1907.12908, 2019.
- [115] SCHÖRKHUBER C, KLAPURI A. Constant-Q transform toolbox for music processing[C]//7th sound and music computing conference, Barcelona, Spain. [S.l. : s.n.], 2010: 3-64.
- [116] GOMEZ-ALANIS A, PEINADO A M, GONZALEZ J A, et al. A light convolutional GRU-RNN deep feature extractor for ASV spoofing detection[C]//Proc. Interspeech: vol. 2019. [S.l. : s.n.], 2019: 1068-1072.
- [117] CHEN T, KUMAR A, NAGARSHETH P, et al. Generalization of audio deepfake detection[C]// Proc. Odyssey 2020 The Speaker and Language Recognition Workshop. [S.l. : s.n.], 2020: 132-137.
- [118] LI R, ZHAO M, LI Z, et al. Anti-Spoofing Speaker Verification System with Multi-Feature Integration and Multi-Task Learning.[C]//Interspeech. [S.l. : s.n.], 2019: 1048-1052.
- [119] MIRSKY Y, LEE W. The Creation and Detection of Deepfakes: A Survey[J/OL]. CoRR, 2020, abs/2004.11138. arXiv: 2004.11138. <https://arxiv.org/abs/2004.11138>.
- [120] GOSWAMI G, RATHA N, AGARWAL A, et al. Unravelling robustness of deep learning based face recognition against adversarial attacks[C]//Proceedings of the AAAI Conference on Artificial Intelligence: vol. 32: 1. [S.l. : s.n.], 2018.
- [121] PARKHI O M, VEDALDI A, ZISSERMAN A. Deep face recognition[J], 2015.
- [122] BALTRUAITIS T, ROBINSON P, MORENCY L P. Openface: an open source facial behavior analysis toolkit[C]//2016 IEEE Winter Conference on Applications of Computer Vision (WACV). [S.l. : s.n.], 2016: 1-10.
- [123] YANG L, SONG Q, WU Y. Attacks on state-of-the-art face recognition using attentional adversarial attack generative network[J]. Multimedia Tools and Applications, 2021, 80(1): 855-875.
- [124] MAJUMDAR P, AGARWAL A, SINGH R, et al. Evading face recognition via partial tampering of faces[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. [S.l. : s.n.], 2019: 0–0.
- [125] KORSHUNOV P, MARCEL S. Vulnerability of face recognition to deep morphing[J]. ArXiv preprint arXiv:1910.01933, 2019.
- [126] SZEGEDY C, ZAREMBA W, SUTSKEVER I, et al. Intriguing properties of neural networks[J]. ArXiv preprint arXiv:1312.6199, 2013.
- [127] GOODFELLOW I J, SHLENS J, SZEGEDY C. Explaining and harnessing adversarial examples[J]. ArXiv preprint arXiv:1412.6572, 2014.
- [128] KURAKIN A, GOODFELLOW I J, BENGIO S. Adversarial examples in the physical world[G]// Artificial intelligence safety and security. [S.l.]: Chapman, 2018: 99-112.

- [129] NEVES J C, TOLOSANA R, VERA-RODRIGUEZ R, et al. Ganprintr: Improved fakes and evaluation of the state of the art in face manipulation detection[J]. IEEE Journal of Selected Topics in Signal Processing, 2020, 14(5): 1038-1048.
- [130] MARRA F, GRAGNIELLO D, COZZOLINO D, et al. Detection of gan-generated fake images over social networks[C]//2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR). [S.l. : s.n.], 2018: 384-389.
- [131] ZHANG X, KARAMAN S, CHANG S F. Detecting and simulating artifacts in gan fake images[C] //2019 IEEE International Workshop on Information Forensics and Security (WIFS). [S.l. : s.n.], 2019: 1-6.
- [132] DU M, PENTYALA S, LI Y, et al. Towards generalizable forgery detection with locality-aware autoencoder[J], 2019.
- [133] HUANG R, FANG F, NGUYEN H H, et al. Security of facial forensics models against adversarial attacks[C]//2020 IEEE International Conference on Image Processing (ICIP). [S.l. : s.n.], 2020: 2236-2240.



## 致谢

非常感谢王荣刚教授，在数字媒体软件与系统开发课让学生上进行了充分搜集了深度伪造与检测的文献搜集，该工作也帮助到学生目前的开发与研究工作进度，同时也对目前深度伪造的进展有所调研，同时也将此流程在其他课程的作业上进行测试获得良好的回馈。最后感谢在这一年来一起寒窗苦读得同学与所有老师，还有默默在开源社群与前沿研究奉献的技术人员跟研究者们。