華中科技大學

本科生毕业设计[论文]

人脸伪造视频检测技术研究

院　　系　软件学院

专业班级　软件工程 **201702** 班

姓　　名　林浩坤

学　　号　**U201717002**

指导教师　黄立群

2021 年 5 月 27 日

学位论文原创性声明

本人郑重声明：所呈交的论文是本人在导师的指导下独立进行研究所取得的研究成果。除了文中特别加以标注引用的内容外，本论文不包括任何其他个人或集体已经发表或撰写的成果作品。本人完全意识到本声明的法律后果由本人承担。

作者签名：　　2021 年　5 月 27 日

学位论文版权使用授权书

本学位论文作者完全了解学校有关保障、使用学位论文的规定，同意学校保留并向有关学位论文管理部门或机构送交论文的复印件和电子版，允许论文被查阅和借阅。本人授权省级优秀学士论文评选机构将本学位论文的全部或部分内容编入有关数据进行检索，可以采用影印、缩印或扫描等复制手段保存和汇编本学位论文。

本学位论文属于 1、保密□，在　　年解密后适用本授权书

2、不保密 ☑　。

（请在以上相应方框内打“ √ ”）

作者签名：　　2021 年 5 月 27 日

作者签名：　　2021 年 5 月 27 日

摘 要

近年来深度学习技术被广泛应用到图像生成领域，大量虚假视频被制作并发布在互联网中，这对网络信息安全带来了巨大的风险和隐患。以 Deepfakes 为代表的深度伪造技术可以篡改视频中的人脸，如果这些伪造技术被滥用，会对网络信息的真实性和可信度造成巨大影响。因此如何准确检测伪造视频成为近年来诸多学者关注的热门研究方向，数字媒体取证领域也迎来了新的发展机遇。

由于学术界对于深度伪造生成技术和检测技术的研究缺乏统一的共识，所以本文首先对伪造视频生成与检测领域进行了梳理。本文将深度伪造生成技术划分为四类，包括人脸交换、表情操纵、全脸合成、属性编辑，分别介绍了不同类别内的代表性生成方法。对于数字媒体取证领域，本文对现有的伪造视频数据集进行系统的总结，分别从图像级别和视频级别回顾了深度伪造检测技术的发展并对检测方法的性能进行了评估。

考虑到当今伪造视频数据集中的数据分布单一，导致检测模型的泛化性较差，本文针对通用人脸检测问题进行了深入的探索和研究。我们实现了一个基于学习权重的通用人脸检测模型。该模型使用元学习的框架将一个基础的二分类检测模型和一个权重感知网络融合起来，可以学习到不同数据集之间的共有特征，以此来有效提升检测方法的泛化性。通过在主流人脸深度伪造数据集上进行的大量实验，我们证明了该模型检测人脸伪造视频的有效性，并在不可见数据集上保持较好的检测能力。

关键词：深度学习、伪造、数据集、数字媒体取证、元学习

Abstract

In recent years, deep learning technology has been applied to the field of image generation. Many fake videos are produced and published on the Internet, which bring huge risks to network security. The deep forgery technology represented by Deepfakes can tamper with the face of the original video. If these forgery technologies are abused, it will have a huge impact on the authenticity and credibility of network information. Therefore, how to accurately detect forged videos has become a hot research direction that many scholars have paid attention to in recent years. The field of digital media forensics has also ushered in new development opportunities.

Since the academic community lacks a unified consensus on the research of deep forgery generation and detection technology, this article first sorts out the field of video forgery generation and detection. This article divides the forgery generation technology into four categories, including face exchange, expression swap, entire face synthesis, and attribute manipulation. Representative forgery methods in different categories are introduced respectively. For the field of digital media forensics, this article summarizes the existing Deepfakes datasets, and reviews the development of face forgery detection technology from the image level and the video level.

Considering that the data distribution in the fake video datasets is single, resulting in insufficient generalization of the detection model. This article pays attention to the general face forgery scenario. We implement a general face detection model based on learning weights (LTW), which configures different weights for face images from different domains. The model uses a meta-learning framework to integrate a basic two-class detection model and a weight-aware network, and it can balance the model's generalizability across multiple domains. At the same time, extensive experiments are designed to prove that the model can effectively detect synthetic faces and can maintain a good detection ability on unseen datasets.

Key Words: Deep Learning; Forgery; Datasets; Digital media forensics; Meta Learning

目 录

1 绪论

1.1 研究背景

图像与视频是计算视觉应用的载体，人们可以通过图像与视频来认知事物、了解世界，在 IT 时代，互联网上的图像数量早已超过千亿级别，图像和视频早已成为当今最具有影响力的传播利器和营销手段。因此，图像和视频的处理一直是计算机视觉领域的热门研究方向，在深度学习发展火热之前，传统的图像处理方法主要通过人为设计的方法和技巧对图像进行分析（如直方图分析、统计分析等）和加工处理（如数字滤波、图像增强等），其本质上是对一维信号处理的二维扩展，所以传统图像处理几乎只能做一些直观的表层图像处理任务。

近年来，深度学习的快速发展，尤其是其在图像视觉领域的突破和进步，使得计算机对于图像的理解、处理和控制能力大幅提升。随着生成对抗网络 GAN[1] 以及其他如 VAE[2]这样的生成模型出现，图像生成逐渐发展为计算机视觉中一个火热的研究领域。生成技术的发展使得人们可以使用深度学习的方法轻松生成逼真的虚假图像，达到以假乱真的效果，大量开源数据集和开源方法使得伪造图像变得愈发容易，由此网上产生了大量的虚假图像和视频，这无疑会增加公众分辨信息真实性的难度。

在图像生成技术中，特别值得注意的是对于人脸的伪造技术，图 1-1 列举了一些伪造出的人脸视频。由于面部和人身份的强关联性，虚假人脸可能会导致现实中的人物关系混乱，进而导致严重的政治、社会、经济问题。

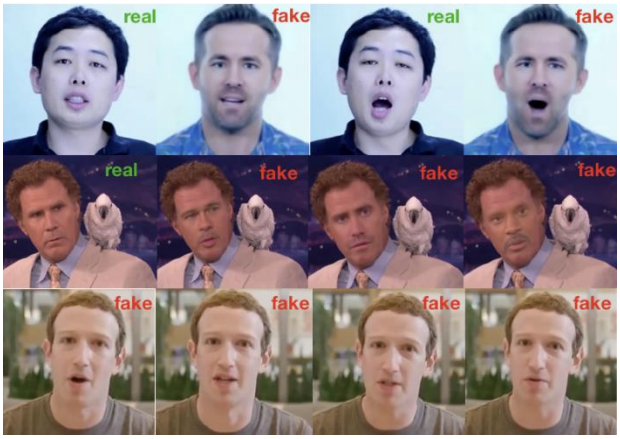

图 1-1 人脸伪造视频举例

深度伪造技术的代表即为 DeepFakes，DeepFake 这一单词由 DeepLearning 和 Fake 组合而成，表示使用深度学习技术将一个人的脸与另一个视频中的人脸进行交换来制造虚假视频。该技术起源于一名为"deepfake"的 Reddit 用户在 2017 年使用一种机器学习的算法将许多名人的面孔生成为色情视频，并发布在互联网上，这些视频迅速传遍全网，之后类似的伪造视频大量出现，而伪造生成算法的成熟也使伪造人脸图像难辨真伪。除了伪造色情内容外，这类伪造方法可以用于伪造新闻、虚假演讲、金融诈骗等活动，对国家和社会稳定造成严重影响。因此，开发一种区分真假面孔的方法至关重要，我们需要高效的 DeepFake 检测算法，这一问题近年来备受研究界的关注。

1.2 国内外研究概况

深度伪造技术和伪造检测技术是一种攻防关系，他们之间相辅相成，生成技术的发展会增加伪造检测的难度，而检测技术的发展会促进生成技术的提高，两者在博弈对抗中不断完善。

先前由于缺乏简单的编辑工具、需要专业的领域专业知识、生成过程复杂且耗时的原因，传统人脸交换方法的发展受到了限制。近年来，在图像和视频中合成不存在的面孔或操纵一个人的真实面孔变得越来越容易，这要归功于：i）大规模公共数据的可访问性，以及 ii）深度学习的发展减少了许多需要手动编辑的步骤。现今已经发布了许多开源软件和移动应用程序（例如 ZAO 和 FaceApp)，这使得任何人都可以方便快速地创建伪造的图像和视频。

具体而言，当今的伪造生成技术主要分为四类，包括全脸合成、属性编辑、人脸交换和表情操纵。

伪造人脸的图像或视频在网上的传播会造成许多负面的影响，为了检测那些复杂的操纵方式，研究界正在做出巨大努力来设计更方便快捷的检测方法，主要方向分为 DeepFake 开源数据集的建立和伪造检测算法的设计两大部分。

在伪造检测领域，我们需要大量的测试数据来对检测方法进行测试，同时基于深度学习的检测方法也需要众多训练数据，因此对大规模 DeepFake 视频数据集的需求不断增长。近些年来也有一些公开的数据集发布，根据发布时间和综合算法，我们将 UADFV[3], DF-TIMIT[4]和 FF++[5]归为 DeepFake 数据集的第一代，

而 DFFD[6], DFDC[7], Celeb-DF[8]数据集被分为第二代。通常，第二代数据集在视频数量和质量上都比第一代有所提高。

检测人脸伪造视频的算法主要为三类。第一类方法是基于 DeepFake 视频在物理特性或生物特征表现出的不一致之处。DeepFake 检测算法的第二类使用在生成过程中引入的信号级别伪像。DeepFake 检测方法的第三类是数据驱动的，它直接采用在真实视频和伪造视频上训练的各种 DNN，而不依赖任何特定的伪像。

虽然已经有非常多的研究工作开展对伪造图像或视频的检测，但是依然缺乏完美的解决方案，检测方法依然存在很多的问题，检测领域也有很多研究难点。

(1) DeepFake 数据集的质量不高，如果仔细查看当今存在的伪造视频数据集，我们会很容易找到肉眼可见的漏洞，而当今网络上传播的伪造视频质量较高，难以分辨真假。这就导致在当今数据集上训练得到的检测模型，在实际运用中效果较差。

(2) 生成算法复杂多变，由于不同生成方法会注重不同的篡改点，所以生成视频的特征各不一样，所以对某一特定数据集训练得到的模型，通常会过拟合已有的生成器特征，在未知生成方法的数据集上会表现出较差的性能，模型的泛化性和鲁棒性较差。

(3) 视频的分辨率不同，由于互联网上存在视频的分辨率各不一样，如果对其做统一的缩放处理，会导致视频中的部分特征丢失，自然会影响检测模型的特征提取，类似的情况会要求模型去考虑特征融合。

1.3 研究内容及问题

深度学习技术的快速发展使得当今生成虚假图像变得方便快捷，互联网上出现了大量的人脸伪造图像和视频，伪造方法的不断改善使得这些图像已经达到了肉眼无法区分的程度，这给传统的图像取证领域带来了更大的困难，对伪造视频检测方法的需求日益增大，风险增大的同时，深度学习技术同样为伪造检测领域带来了机遇。本文将深入分析人脸伪造视频生成与检测领域的发展现状，总结归纳伪造生成方法和伪造检测领域的各类方法及数据集，并针对通用人脸检测领域这一问题进行更深入的探索，复现Ke Sun等人提出的基于学习权重的通用人脸检

测模型（LTW模型）[9]，同时设计实验评测该模型的性能以及证明模型具有更好的泛化性。

本文分为以下部分。第二章介绍深度伪造生成技术，对生成对抗网络以及四类主流的人脸伪造生成方法进行介绍。第三章对伪造检测领域进行梳理和总结，人脸伪造数据集是检测领域发展的重要基石，本章中将分析现有的十大主流数据集并列表进行对比，同时从图片级别和视频级别对伪造检测方法进行归纳，并列表对涉及的检测方法进行了对比。第四章介绍基于学习权重的通用人脸检测模型，对模型结构、优化过程以及实现过程进行说明。第五章详细介绍了针对LTW模型的实验过程。第六章总结了本文完成的工作并讨论了未来检测领域的潜在方向。

2 深度伪造生成技术概述

深度学习在计算机视觉领域的应用是较早开展也是较为成熟的，其在图像物体识别、目标检测、目标跟踪等等传统的应用场景中都取得了十分成功的结果，随着生成对抗网络 GAN 的出现，生成真实逼真的图像成为可能，目前的人脸深度伪造方式也主要是基于 GAN。本章中我们将对生成对抗网络和当今主流的伪造视频生成方法进行介绍。

2.1 生成对抗网络

2014 年，Ian Goodfellow 提出了生成对抗网络 GAN[1]，该模型是利用一种对抗学习的过程来同时学习两个模型，一个是用于学习数据样本分布的生成器 G，另一个是用于学习判断样本是否来自真实数据分布的判别器 D。生成器 G 的目的是最大化判别器 D 出错的概率，即“欺骗”判别器，而判别器 D 的目的是最小化自身的判别损失，即提升鉴别能力。这两个模型类似于警察和盗贼的角色，相互对抗，具体的生成对抗网络的目标函数如下：

$$\min_G \max_D V(D,G) = \mathbb{E}_{\boldsymbol{x}\sim p_{\text{date}}(\boldsymbol{x})}[\log D(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{z}}(\boldsymbol{z})}[\log(1 - D(G(\boldsymbol{z})))].$$

这是原始生成对抗网络中的目标函数公式，后续有非常多的研究人员对生成对抗网络进行了改进和提升，使其在理论层面更加完善，像模型崩溃等问题都得到了极大的改善，生成图像的质量也越来越好。除了图像处理领域中的应用，生成对抗网络在文本、音频等各式各样的数据形式上都取得了不错的结果，生成对抗网络俨然已经成为一个被广泛使用的深度学习模型。

2.2 人脸交换图像伪造

人脸交换是一种将某一人物的面部交换到另一人物面部的技术。2017 年 Reddit 用户使用的 DeepFakes 技术便是基于深度学习的人脸交换项目。其原理如图 2-1 所示，该方法首先训练两个共享权值的编码器，利用编码器提取输入人脸的主要特征，从而学会如何重建人脸。解码器 A 利用人脸 A 的信息重建人脸 A，解码器 B 利用人脸 B 的信息重建出人脸 B。训练结束后，交换解码器 A 和

B，从而达到换脸效果。这个过程不需要传统图像处理技术，只需要原始人物和目标人物的图片，极大的降低了换脸技术的应用难度。

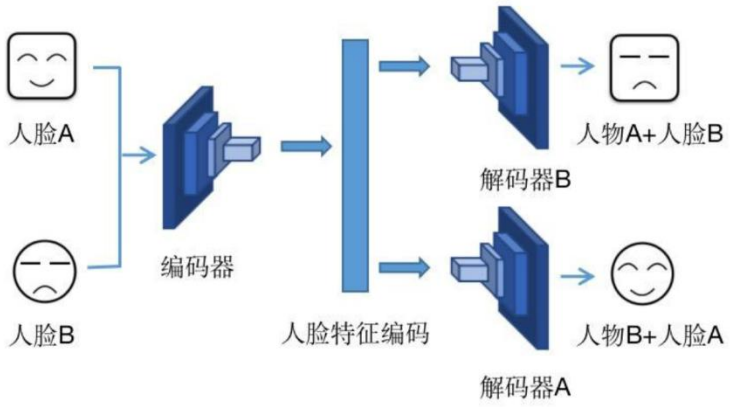

图 2-1 deepfakes 技术原理图

为了获得更好的生成效果，Goodfellow 等人将 GAN 技术融合进换脸领域，其提出的 Faceswap-GAN 引入判别器作为对抗损失函数，损失函数会在图像生成的过程中判别原始图像和生成图像的相似度，进而极大的提升了生成效果。

Nirkin 等人利用 2D 人脸的特征点来适应 3D 人脸的面部模型[10]，并使用全卷积神经网络来从背景和遮挡物中分割出人脸，并且投影到三维空间，建立三维模型，使得纹理迁移后得到的面部与背景更加协调。

Nrikin 等人还提出了 FSGAN[11]，使用循环神经网络来调整人脸的姿态和表情变化，其损失函数也与传统方法不同，使用优化和感知损失结合的混合损失函数，使替换后的人脸更加自然。

2.3 表情操纵图像伪造

表情操纵即实时控制目标者的表情和嘴唇，操纵者利用此技术可以生成虚假的讲话视频。Face2Face[12]是一种非常火热的表情操纵技术，该技术是一种传统的图形学方法，借助传统的 OpenCV 库和 dlib 工具包来将原始人物的表情转移到目标任务身上，并且可以保持目标人物的身份不变。具体而言，该方法先从原始人物的图像中利用人脸检测器来标记出人脸的关键点，再使用转换模型把关键点转换为目标人物的图像。

Thies 等人提出一种新的方法，名为 Neural Texture[13]，其是利用 Neural Textures 的神经纹理来进行渲染，从视频数据中学习目标人物的神经纹理，对目

标人物的身份信息和原始人物的表情信息进行渲染和追踪，通过编码解码网络来生成更加细致真实的人脸图像。图 2-2 显示了表情操纵图像伪造技术的效果。

图 2-2 表情操纵效果图

2.4 全脸合成图像伪造

全脸合成主要利用 GAN，主要思想是给生成器输入一个随机的信号噪声，使生成器生成相应的人脸图像，训练生成器制造虚假数据，来对真实数据分布进行拟合。判别器对生成器生成的人脸图像和真实人脸图像进行比较判断，尽可能的判断数据是否由生成器生成。生成器和判别器在相互博弈中不断提升，逐渐使生成器生成的图片接近真实样本。

大量 GAN 的衍生网络在该领域都取得了良好的效果，可以生成逼真的人脸图像，如 DCGAN[14]、BEGAN[15]、PGGAN[16]等，图 2-3 列举了通过全脸合成技术生成的图像。

图 2-3 全脸合成效果图

GAN 网络的生成器参数更新来自于判别器的反向传播，而不是真实数据样本，正是因为这一原因，生成对抗网络可以创造真实世界中不存在的数据样本。但是，训练复杂和难于收敛也是非常明显的缺点，而且生成的数据分布多样性也不足。

2.5 属性编辑图像伪造

人脸属性编辑主要指对人脸属性的改变，可以改变人脸的发型、面部表情、皱纹、肤色、发色等属性以达到伪造人脸的目的。这一方法在很多娱乐领域都得到了应用，比如通过人脸年龄编辑算法，模拟人的一生面部变化，可以帮助影视剧演员拍摄跨年龄的戏份。

Zhang 等人提出了 CAAE（Conditional Adversarial Autoencoder）[17]。该模型可以进行年龄编辑，其假定人脸图像处在高维流形中，首先通过一个条件编码器将人脸映射到一个隐向量上，再用一个解码器映射到基于年龄的人脸流形上。如果我们引导图像在流形中沿着指定的方向移动，这个隐向量便保留了个人的人脸特征，且该向量具有年龄条件约束，因此人在图像中的年龄会随之变化。

Li 等人设计了 BeautyGAN[18]，用于人脸自动上妆，使用了 CycleGAN[19]的基本结构，输入人脸素颜照片，以另一张带妆人脸照片作为参考，根据带妆照片，模型对素颜人脸的眼部、唇部和面部整体进行化妆。

2.6 本章小结

当下主要流行的伪造生成技术都是对人面部进行篡改，本章首先介绍了大部分技术中都需要使用的生成对抗网络 GAN，之后分别介绍了四大主要的生成领域以及其对应的主流方法。从技术分类来看，现有的生成方法中有部分是利用图形学的方法，如 FaceSwap 和 Neural Texture，基于图形学的方法整体上时间开销较大、成本代价大且使用门槛高，导致这些技术较难被普及。深度学习的普及大大降低了技术使用门槛，在本章中提到了大量使用 GAN 技术的生成方法，这些方法使得人脸的生成和篡改都越发容易，技术的两面性导致人脸伪造的滥用会带来极大的危害，因此我们在利用强大生成技术带来的种种好处之时，也必须研究开发更好的伪造检测技术，这样才能更好的保障人脸隐私安全。

3 深度伪造数据集及检测技术

3.1 深度伪造数据集

3.1.1 UADFV

UADFV[3]是最早的公共数据库之一，由美国的纽约州立大学奥法罗分校 Siwei Lyu 团队在 2018 年提出。这个数据库包括来自 Youtube 的 49 个真实视频，团队成员通过 FakeApp 移动应用程序创建了 49 个假视频，所有这些视频都与尼古拉斯·凯奇的原貌进行了交换。因此，所有的假视频中都只有一个身份。每个视频代表一个人，典型分辨率为 294×500 像素，平均 11.14 秒。数据集目前需要以邮件的方式向论文作者申请。

这一数据集是最早期深度伪造研究的数据集，存在较多的缺陷，如视频质量差，分辨率低下，数据量较小，换脸痕迹明显等等。

3.1.2 FaceForensics（FF）

FF 数据集[20]规模较大，是早期大规模深度伪造数据集，视频从 Youtube8M 数据集和 YouTube 上选取了标签为人脸、新闻播报员的 1004 个视频，视频分辨率必须大于 480p。研究人员采用 Face2Face 的方法构造了对应的 1004 个虚假视频，并过滤掉人脸遮挡较多的视频。

FF 数据的规模较大，初始视频的人脸质量较高，但是仅使用了单一的伪造方法，并且可以看到明显的篡改痕迹。

3.1.3 FaceForensics++（FF++）

FaceForensics++数据集[5]由德国的慕尼黑工业大学在 2019 年提出，是当今规模较大且伪造方法较多的深度伪造数据集之一。该数据集共有三种不同的视频质量（Raw，Crf 23 和 Crf40），不同的视频质量都包含 1000 段真实视频，4000 段对应的伪造视频，真实视频取自 YouTube，采用了 Deepfakes、FaceSwap、NeuralTextures 和 Face2Face 这四种伪造方法，每类生成方法各有 1000 段视频。此外，数据集还提供了对应人脸篡改区域的 mask。数据集的下载网址为：https://github.com/ondyari/FaceForensics。

Deepfakes 方法主要利用自动编码模型，将原始人脸替换为目标视频中的人

脸。Faceswap 方法和 Face2Face 方法都是图形学的方法，对重建的 3D 模型进行伪造编辑，从而实现人脸造假。NeuralTextures 方法主要基于神经纹理进行渲染，利用原始视频数据来学习得到目标对象的神经纹理。

该数据集的缺点是伪造视频质量不高，肉眼可看见明显的伪造痕迹。

3.1.4 Deepfake-TIMIT

Deepfake-TIMIT[4]是由瑞士的 Idiap 研究所于 2018 年提出，该数据集由 Faceswap-GAN（Faceswap 的一种改进方法）生成。该数据集在公开可用的 VidTIMIT 数据库(http://conradsanderson.id.au/vidtimit/)中人工选择了 16 组肉眼看起来相似的人。对于 32 名受试者，生成了两种质量不同的视频：较高质量(HQ)的视频（128 x 128）和较低质量(LQ)的视频（64 x 64）。VidTIMIT 数据库中每个对象都有 10 个视频，因此数据集制作者生成了每个人对应的 320 个视频，所以一共有 640 个视频交换了人脸。数据集中的音频部分保留了每个视频的原始音频轨道，没有对音频通道进行编辑或修改。数据集的申请网址为：https://www.idiap.ch/dataset/deepfaketimit。

该数据集的视频质量比 FF++数据集高，但是视频面部边界仍有篡改痕迹且视频分辨率不高。

3.1.5 DeepFakeDetection（DFD）

Deepfake Detection 由美国的 Google 公司于 2019 年提供，该数据集由 DeepFakes 方法生成，包含来自 28 个演员在不同场景中的 3000 多个被操纵的视频，该数据集也提供了三种视频质量（Raw、Crf23、Crf40）。DFD 数据集比 FF++数据集质量更高，数据集的网址为：https://github.com/ondyari/FaceForensics。

3.1.6 Celeb-DF

Celeb-DF 数据集[8]由美国的纽约州立大学奥法罗分校 Siwei Lyu 团队在 2018 年提出。这个数据库旨在提供视觉质量更好的伪造视频，类似于在互联网上共享的流行视频。Celeb DF 包括从 Youtube 提取的 408 个真实视频和 795 个假视频，这些视频是通过 Deepfakes 生成算法的改进版本创建的，改善了合成人脸的低分辨率和颜色不一致等问题，视频的帧率为 30，平均长度 13s。数据集的网址为：http://www.cs.albany.edu/~lsw/celeb-deepfakeforensics.html。

3.1.7 Diverse Fake Face Dataset(DFFD)

DFFD 数据集[6]由美国的密歇根州立大学于 2019 年发布，该数据集由多个公共可用的子数据集组成，这些子数据集是使用开源的代码生成的。利用多种真实图像的来源，能够得到不同分辨率和图像质量的真实图片及伪造图片。数据集的网址为：http://cvlab.cse.msu.edu/dffd-dataset.html。

3.1.8 DFDC Preview

DFFD Preview 数据集[21]由美国的 FaceBook 公司于 2019 年提出，由 66 个付费演员的 1131 个真实视频和 4119 个假视频组成，数据集使用了两种人脸修改算法。该数据集公布于 DFDC 竞赛前夕，相当于一个预赛数据集，假视频主要在相似人脸之间进行篡改，如皮肤颜色、头发等，每个视频的长度约为 15s。数据集的获取网址为：https://ai.facebook.com/datasets/dfdc/

3.1.9 DeeperForensics-1.0

该数据集由新加坡的南洋理工大学和中国的商汤公司[22]于 2020 年提出。作者邀请来自 26 个国家/地区的 100 名付费演员录制源视频。收集的高质量数据在身份，姿势，表情，情绪，光照条件和 3DMM 混合形状等方面各不相同。使用基于 Deepfakes 的自动编码器（DF-VAE）进行篡改。DF-VAE 改善了可伸缩性，样式匹配和时间连续性，以确保人脸交换质量。作者在 5 个强度级别上应用了 7 种类型的后处理。某些视频会受到混合的多种后处理方法处理，从而更好地模拟现实情况。数据集网址：https://github.com/EndlessSora/DeeperForensics-1.0。

3.1.9 DFDC

DFDC 数据集[7]由美国的 FaceBook 公司于 2020 年提出，是 The Deepfake Detection Challenge 比赛中使用的数据集。该数据集中共有来自 3426 名付费演员的超过 10 万段数据，使用多种深度伪造、基于 GAN 的方法以及非深度学习方法制作。创建 DFDC 数据集的过程中使用的换脸方法包括：DFAE，MM/NN face swap，NTH，FSGAN，StyleGAN 和 Refinement 和 audio swapping 等。该数据集还使用多种数据增强技术提升图像质量，如几何变换或干扰等。数据集获取网址为：https://ai.facebook.com/datasets/dfdc/。

3.1.10 ForgeryNet

ForgeryNet 数据集[23]由中国的商汤公司于 2021 年提出，是目前公开的最大规模深度人脸伪造数据集，该数据集从数据规模（290 万张图像，221247 个视频）、manipulations 操纵（7 种图像级方法，8 种视频级方法）、perturbations 扰动（36 种独立的和更多的混合扰动）和标注（630 万个分类标签，290 万个操纵区域标注和 221247 个时空伪造段标签）这些方面来看都是规模最大的。数据集的获取网址：https://yinanhe.github.io/projects/forgerynet.html。

3.1.11 对比归纳

对上述的 10 余种当今公开的深度伪造检测数据集，我对其进行了统计和比较，结果如下表所示。

图 3-1 Deepfake 数据集比较

数据集	视频		图片		方法	人物	后处理	标注
	真实	伪造	真实	伪造				
UADFV	49	49	241	252	1	49	-	591
DF-TIMIT	320	640	-	-	2	43	-	1600
Deep Fake Detection	363	3068	-	-	5	28	-	3431
Celeb-DF	590	5639	-	-	1	59	-	6229
DFFD	1000	3000	58703	240336	7	-	-	8000
FaceForensics++	1000	4000	-	-	5	-	2	11000
DeeperForensics-1.0	50000	10000	-	-	1	100	7	60000
DFDC Preview	1171	4073	-	-	2	66	3	5244
DFDC	23564	104500	-	-	8	960	19	128064
ForgeryNet	99630	121617	1438201	1457861	15	5400+	36	9393574

3.2 深度伪造图像检测

视觉上的深度伪造生成技术一般分为全脸合成、属性编辑、人脸交换和表情操纵四个方面，但是对于图像级的伪造检测方法，大多数并不是针对特定检测方法提出的，图像级的深度伪造检测方法大多分为如下三种。

3.2.1 通用的图像检测分类网络

此类方法中较有代表性的有 2017 年 Chollet 等人发表在 CVPR 上的 Xception[24]网络，和 Tan 等人 2019 年发表在 ICML 上的 EfficientNet[25]网络，这两个分类网络也成为了现今很多检测方法的 backbone。

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

3.2.2 针对特定线索的检测方法

这类方法主要是通过寻找某些特定的伪造线索来进行伪造检测，方法的可解释性较强。

S. McCloskey 和 M. Albright 采用了颜色这一特征[26]，并使用 SVM 为分类器完成伪造检测。文章中通过研究表明，GAN 网络生成的图像和真实相机拍摄的图像在颜色处理上有所不同。因此可以用颜色特征作为区分 GAN 生成图像和真实图像的判别线索。

J. Stehouwer 等人[6]提出，针对不同的伪造样本，每一个样本造假痕迹明显的地方不一定是同一区域，如果更加关注于造假的区域，会更易对伪造样本进行检测，由此提出了一种基于注意力机制的方法，在原有的 CNN 模型基础上，增加注意力机制，自适应关注篡改区域，提高检测准确率。

P. Zhou 等人提出了一种双流网络[27]，他们考虑图像的细节伪影可以很好的帮助原有分类网络提升鉴伪性能，因此使用双流网络，一个流依靠分类网络从全局角度检测是否为篡改过的人脸，另一流基于隐写特征捕捉噪声残留，最后将两个流进行融合，从而判断是否为伪造的人脸。

除了时域特征，X. Zhang 等提出了一种基于频域的方法[28]，使用 AutoGAN 来生成图像，再转换至频域上进而训练分类器，对于 StarGAN 生成的面部属性编辑图片，该模型能够有很好的检测效果，准确率较高。

3.2.3 基于 CNN 的检测方法

H. Nguyen 等人提出了一种基于多任务学习的 CNN 方法[29]，同时检测伪造样本并定位篡改区域，基于自编码器来构建检测系统。

S. Tariq 等以 Adobe Photoshop CS6 生成面部属性编辑的伪造样本[30]，比如上妆，眼镜，太阳镜，头发和帽子等，用 VGG16、VGG19、ResNet 和 XceptionNet 作为检测网络，进行伪造样本检测。

总而言之，当今在图像级的深度伪造检测方法的种类繁多，基于各种线索以及不同的分类模型等等，当前不同方法之间更多注重在性能进行提升，检测方法的鲁棒性和可解释性较差，这也是领域内更需关注和研究的问题。

3.3 深度伪造视频检测

3.3.1 基于生理特征的检测方法

对于单张图像，无论图像真假都无法反映出人脸的呼吸、心跳、眨眼等人体生理特征，对于视频数据来讲，根据人脸面部说话的口型变化以及眨眼频率可以判断出该视频的真假，因此我们可以利用视频中人体生理特征来鉴定伪造视频。

基于生理特征的第一篇工作是 Li 等人提出的基于眨眼来鉴别深度伪造视频的方法[31]。首先检测每一帧的人脸，定位人脸关键点信息，然后利用人脸对齐算法将人脸关键点定位到统一的空间，降低人脸头部转向和移动带来的干扰。做完上述操作后，再定位提取并缩放眼睛区域的关键点，形成一段帧序列，送入长期循环卷积网络 (LRCN) [32]中，先由 CNN 网络提取人眼特征，然后在 LSTM-RNN 网络中学习序列级的特征，最后输出到全连接层中来检测人在视频中眨眼的频率。论文显示在真实视频中可以检测到每分钟 34.1 次的眨眼频率，在虚假视频中却只有每分钟 3.4 次眨眼频率，设定一个正常人眨眼频率阈值为每分钟 10 次，通过这一差别区分出真假视频。该方法在 EBV[33]等数据集上取得了较好的结果。但是倘若伪造的视频考虑到眨眼频率的真实性，特意去训练具备眨眼能力的模型，则该种方法无效。

3.3.2 基于视频帧间时序特征的检测方法

无论伪造的视频是何种类型，伪造视频的生成过程基本都是一帧一帧操作的，

这些操作必然会带来视频帧之间的时间不连续或者抖动。所以将视频逐帧输入到时间序列网络中，再让分类器给出真假分类结果也是可行方案。

基于时间序列的第一篇工作是 Güera 等人提出的一种端到端的时间感知的方法[33]。作者在论文中首先证明了深度伪造视频的帧间具有不一致的特性，进而基于 CNN 和 LSTM 来检测深度伪造视频。给一段视频序列，CNN 网络首先提取特征，每一帧得到一个 2048 维的向量，然后将特征向量输入到 LSTM 网络中，再将输出送入全连接层和 softmax 层，得到最终真假视频分类结果。作者从网站上收集看 300 个虚假视频，测试了视频帧长度不同的视频，对于以每秒 24 帧的速度采样了 40 帧的片段，可以精准判断这一片段是否来源于一个深度伪造视频，准确率达到百分之九十七。该方法的缺点是鲁棒性不足，容易受到对抗样本的攻击，而且需要真实和伪造数据作为训 练数据，比较低效。

Sabir 等人利用行为识别领域中时间信息处理视频的方法[34]，利用递归卷积网络 (RCN) 对视频流时空特征进行检测。首先对视频序列进行预处理，包括人脸检测、人脸裁剪、人脸对齐，然后把每一帧输入到 RNN 网络中，进行端到端的学习。该方法在数据集 FaceForensics++ 上比之前的最好结果提升了 4.55 左右的准确率。

3.3.3 基于视频内视觉伪像的检测方法

基于视频帧内视觉伪像的检测方法流程主要是先提取视频帧内的判别特征，随后将提取到的特征送入深层或浅层分类器中进行训练，从而实现真假数据分类。所以此类方法与深度伪造图像的检测技术是相通的，使得有些方法既可以检测虚假伪造图像，也可以检测虚假伪造视频。其中，深层分类器主要基于神经网络模型实现，而浅层分类器主要结合传统机器学习模型实现。

1. 深层分类器

Afchar 等人[35]经过分析，认为由于视频压缩带来的图像噪声强烈退化，基于图像噪声的微观分析并不会起作用。同时，在高层语义层面，特别是当图像描绘的是人脸，人眼会很难分辨出伪造的图像。所以作者进行了折中，在介于高层语义信息和低层微观信息之间，提出了少量层的神经网络模型 MesoNet, 包括 Meso-4 和 MesoInception-4。Meso-4 网络由四个连续的卷积神经网络构成，每层

后面加上批归一化和最大池化层，最后连接两个全连接层和 sigmoid 层进行分类。而 MesoInception-4 网络把 Meso-4 前面两个卷积层用 Inception 模块进行替代，相当于堆叠了几个不同卷积核大小的卷积层的输出，以此来增加函数优化空间。这种方法在保证了高性能的基础上构建了一个轻量级的检测网络，其参数数量明显比 XceptionNet、ResNet-50 等神经网络结构少了很多。

Afchar 等人在研究中也证明了眼睛和嘴巴部位的特征对于伪造视频检测是非常重要的。由于目前的深度伪造生成算法只能生成有限分辨率的图像，而且需要将目标人脸通过仿射变换 (如缩放、旋转和剪切等) 匹配到原始视频中，这就会造成合成区域和原始区域之间的分辨率不一致的问题，并在伪造的视频中留下视觉伪像。

Li 等人[36]利用这一发现，提出了一种基于 CNN 模型的深度伪造视频检测方法。作者直接模拟此类仿射变换，可以简化负样本的生成过程，同时对原始人脸图像的面部区域以及关键点坐标进行提取，之后从多个尺度进行对齐处理, 再将高斯模糊操作用于随机选取的缩放图像，并将其形变回原始图像，这样可以减少时间消耗和资源消耗，并且具有较好的泛化性能。但是该模型未在大量压缩视频上进行性能评估，并且可能对特定分布的伪造视频过拟合，因此训练数据的多样性需要提高。

2. 浅层分类器

浅层分类器将目标人脸拼接到原始人脸的面部区域过程中，会在从二维面部图像估计三维头部姿态（比如头的方向和位置）时引入误差。Yang 等人[3]基于这一观察，进行实验证明了这一现象，并且将这种特征送入 SVM 分类器进行分类。作者通过两种方法来估计图像或视频中的头部姿态，一种是用检测得到的 68 个关键点来估计，另一种是只用中心区域的关键点来估计，将两种估计方法得到的头部三维单位向量比较余弦距离，实验证明在真实人脸中两种方法估计得到的余弦距离较为接近 (0-0.02)，但是虚假人脸中两种方法估计得到的余弦距离较远 (0.02-0.08)，这是因为中间的人脸区域和外部的人脸轮廓关键点来自不同的域，所以两者的误差会比较大。因此可以通过这种方法将两种分布区分开，进而区分出真假数据。

最后介绍一种为国家领导人和世界名人（POIs）制定的深度伪造视频检测技

术。世界上没有一片树叶是一样的，同样，Agarwal 等人认为每一个人在说话时都会展现出不一样的面部表情和头部运动[37]，这称之为软生物特征模型，但是深度伪造的人物相关视频，则不会存在这样的软生物特征。因此基于这一观察，可以用此特征进行特定人物的真实虚假视频判断。给定一段视频片段，用 OpenFace2 开源工具追踪人脸和头部运动，面部肌肉的运动可以被编码成特定的运动单元（AU），利用 OpenFace2 提供的 AU，生成 20 个特征向量，然后用 Pearson 相关性系数测量向量之间的相似度，进而得到 190 维的特征向量，然后用 SVM 来进行真假数据的区分。

3.4 伪造检测技术比较

前述研究工作在提出的同时，大多数在开源数据集上进行了评测，我对前述主流的深度伪造检测算法在公开数据集上的检测表现总结如表 3-2，主要评估指标包括准确率（Acc），ROC 曲线面积（AUC），平均错误概率（EER），Raw、HQ、LQ 分别代表原生态、高清和低清，DF/F2F/FS/NT 分别是 FF++中四种生成方法的缩写。

图 3-2 代表性方法在主要测试集上的性能评估

研究工作	模型	特点	数据集	性能
Stehouwer 等人[6]	CNN +Attention	增加注意力机制	DFFD	AUC=99.4%, ERR =3.1%
Zhou 等人[27]	CNN+SVM	人脸和隐写特征结合	UADFV DeepfakeTIMIT-HQ DeepfakeTIMIT-LQ FF++/DF Celeb-DF	AUC =85.1% AUC=73.5% AUC =83.5% AUC=70.1% AUC =55.7%
Nguyen 等人[29]	Autoencoder	分类和分割重建融合	UADFV DeepfakeTIMIT-HQ DeepfakeTIMIT-LQ FF++/DF Celeb-DF	AUC =65.8% AUC=55.3% AUC =62.1% AUC=77.3% AUC =36.5%
Guera 等人[33]	CNN +RNN	图片的时序信息	Own	AUC=97.1%
Sabir 等人[34]	CNN +Bi-LSTM	图片的时序信息	FF++/LQ DF/F2F/FS	AUC 96.9%, 94.4%, 96.3%

Afchar 等人[35]	CNN	微观特征的学习	FF++	Acc%
			Raw(DF/F2F/FS/NT)	99.59 99.61 99.14 99.36
			HQ(DF/F2F/FS/NT)	98.85 98.36 98.23 94.5
			LQ(DF/F2F/FS/NT)	94.28 91.56 93.7 82.11
			Mesonet Data	Acc=98.4%
Li 等人[36]	CNN	学习人脸边框篡改遗留痕迹	UADFV	AUC=84.3%
			DeepfakeTIMIT-HQ	AUC=87.8%
			DeepfakeTIMIT-LQ	AUC=68.4%
			Celeb-DF	AUC=53.6%
			UADFV	AUC=97.4%
			DeepfakeTIMIT-HQ	AUC=93.2%
			DeepfakeTIMIT-LQ	AUC =99.9%
			FF++/DF	AUC=79.2%
			Celeb-DF	AUC =53.8%
Yang 等人[3]	SVM	头部姿态估计	UADFV	AUC=89.0%
			DeepfakeTIMIT-HQ	AUC=53.2%
			DeepfakeTIMIT-LQ	AUC =55.1%
			FF++/DF	AUC=47.3%
			Celeb-DF	AUC =54.8%
Agarwal 等人[37]	SVM	动作单元编码	Own (FaceSwap,HQ)	AUC=96.3%

3.5 本章小结

本章首先对深度伪造领域开源的数据集进行了整理，覆盖了领域内基本上所有数据集，对数据集的来源、生成方式等都进行了介绍，列表对比了不同数据集之间的差异。随后，我在本章中分别从图片级别和视频级别对伪造检测方法进行了归纳和介绍，领域内主流的检测方法基本都有涉及，列表对检测方法在主要测试集上的性能进行了比较。总体来看，本章对当今伪造检测领域的发展现状进行了梳理。

4 基于学习权重的通用人脸检测模型

4.1 引言

前文已经介绍了深度伪造生成方法，同时也总结对比了伪造检测领域的数据集和部分检测方法，这些研究工作极大推动了人脸伪造领域的进步和发展。

但是当今主流数据集中的伪造面孔的生成方式相对单一，导致训练集和测试集的数据分布和面部特征大致相同。在实际应用中，给定训练集（源域）上训练的模型始终用于不同的测试集（目标域），目标域的不可见会导致检测方法的效果大幅度下降。因此，在不可见目标域检测人脸是相对更有挑战性的工作，即我们必须提高检测模型的泛化能力，使其在未知数据集上依然能保持良好的检测效果。在 Guo 等人发表在 CVPR2020 的论文中[38]，不可见目标域的人脸检测这一场景被称为通用人脸检测，接下来的论述中也采用这一说法。

在本次毕业设计中，我尝试对通用人脸检测这一问题进行调研和探索，经过多篇论文阅读，我发现 Ke Sun 等人提出基于学习权重的通用人脸检测模型，即 learning-to-weight (LTW)模型[9]可以提高检测模型的泛化性，在数据集上训练得到的模型可以在未知数据域上表现出较好的效果，因此在毕业设计中我学习并实现了 LTW 模型，并设计了多个不同的 benchmark 来进行实验，证明该方法在检测伪造人脸方面的有效性。

当今非常多的伪造图片是由 GAN 模型生成的，由于生成器内部的固有偏差，所以伪造图片中也会有多种偏差，这种虚拟的偏差使得源数据间分布更加复杂，数据之间的语义差距很大。除此之外，生成图片的质量差距较大，会导致不同样本之间可以学习的特征不同，如果使用一样的权重在样本中训练会破坏模型的泛化能力。

LTW 模型针对上述问题进行了改进，其基本思想很简单但很有效。第一分支是基本的二分类检测模型，该模型提取每个图像的特征并确定输入图像是真实的还是伪造的。第二个分支是一个权重感知网络，该网络预测训练每个 batch 中每个图像的域-自适应权重得分。随后我们使用元学习框架将这两个分支在一起，元学习过程可以更新权重感知网络的参数，还可以优化检测模型梯度下降的方向。

4.2 元学习

元学习（Meta Learning），即“学习如何学习”，其目标是针对各种学习任务训练模型，以便仅使用少量训练样本即可解决新任务，即让模型利用已学习过的信息，快速学习到新的概念或技能，可以适应新的任务。

对于元学习的定义尚未统一，一般认为一个元学习系统需满足以下三点：1.系统必须包含一个学习子系统；2.系统利用学习过程中的元知识来获得通用性的经验，元知识源于不同领域或者单个的数据集；3.系统必须动态的选择学习偏差。

元学习领域的研究现在主要从以下几个方法的角度出发：

基于度量的方法主要是希望最大程度上抽取样本中的内部特征，使用特征比对的方法判定样本种类。基于优化器的方法则针对传统优化算法梯度不稳定的问题提出改进，在优化过程中使用神经网络对梯度进行更新，以此实现快速适应新任务的目标。基于数据增强的方法希望为小样本任务增加额外的样本来提供更多数据支持，解决样本数量不足的问题。基于泛化性较强的初始化方法研究进展较多，主要是模型无关元学习算法（MAML）[39]以及其后续改进的诸多方法。

元学习框架的结构与其他模型类似，分为特征提取和分类两部分。其一般流程是：在训练数据集和验证集上得到泛化性强的初始化网络参数，在测试任务上检验优化后模型的效果。

4.3 模型结构

4.3.1 模型的整体框架

Learning to weight（LTW）模型希望可以提高检测方法的泛化性，使得其在未知数据集上依然表现出好的检测效果。

LTW 模型定义基础检测模型为$f(\theta)$，θ表示一个神经网络的参数。在训练阶段，使用 N 个源域进行训练，源域表示为$D^s_{train} = \{d^s_{d1}, d^s_{d2}, \dots, d^s_{dN}\}$，其中$d^s_{di}$表示根据生成方法不同划分出的子集。随后，我们会将训练好的模型$f(\theta)$在 M 个未知的目标域上进行测试，测试域表示为$D^s_{test} = \{d^s_{d1}, d^s_{d2}, \dots, d^s_{dM}\}$，我们的目标在于希望在$D^s_{train}$上训练得到的检测模型$f(\theta)$，可以在测试域$D^s_{test}$中表现出良好的检测效果。

具体而言，LTW 模型有两个分支。第一个分支是一个二分类神经网络$f(\theta)$，其目标是提取每张人脸的特征并判断真实性。另一个分支是一个权重网络$p(\omega)$，其依赖于$f(\theta)$输出的特征图f_i，这一分支可以将域适应的权重分配给每个图片，帮助基础检测模型$f(\theta)$挖掘更多用于泛化性强的特征，进而在遇到未知数据域时可以获得更好的结果。

与 Shu[40]等人不同，LTW 模型的权重网络以 $f(\theta)$ 上一个卷积层输出的特征图 f_i 作为输入，f_i 被定义成从输入图片x_i 提取出的隐藏特征。为了避免太多参数进而影响参数传递效率，$p(\omega)$网络会被设计的非常小，仅仅包含 1.024M 个参数。具体而言，我们使用了两个 depth-wise 卷积层去压缩特征图的通道数量，紧接着使用一个全连接层去给出输入图片的预测分数。为了获得更好的效果，我们使用激活函数将分数正则化为[0,1]，我们使用$p(f_i;\omega)$去描述权重感知网络对输入图片x_i的影响。

为了更新网络$p(\omega)$的参数同时矫正检测模型$f(\theta)$的下降梯度，我们为 LTW 框架引入一个元学习的策略[9]。元学习策略可以带来两个好处:

(1) 基础检测模型$f(\theta)$可以避免在某一特定的数据集上过拟合。

(2) 权重感知网络$p(\omega)$利用元学习的梯度来更新参数。

LTW 模型的结构如下图：

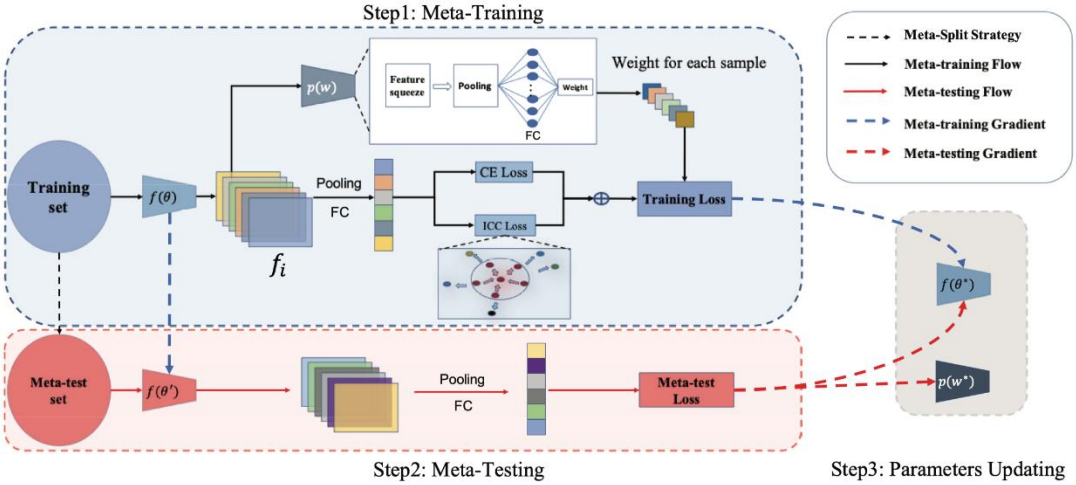

图 4-1 LTW 模型

4.3.2 元分割策略

在训练的一个 epoch 中将源域D^s_{train}划分为元训练域D^s_{nt}和元数据域D^s_{meta}。为了避免特定数据分布导致的过拟合，我们使用一种随机选取的策略去划分源域。

具体而言，我们随机将 N 个源域D_{train}^{s} 划分为两个子集D_{nt}^{s}和D_{meta}^{s}，其中D_{nt}^{s}包含了$N/2$个源域去训练检测器，D_{meta}^{s}则包含剩下的源域用于元测试来辅助模型训练。值得注意的是我们这种分割策略保证了不会存在太大的语义差距。

在训练过程中，一个 batch 的样本从对应 epoch 的元训练子集D_{nt}^{s}中选取，而其对应的元测试子集选取方法如下图。

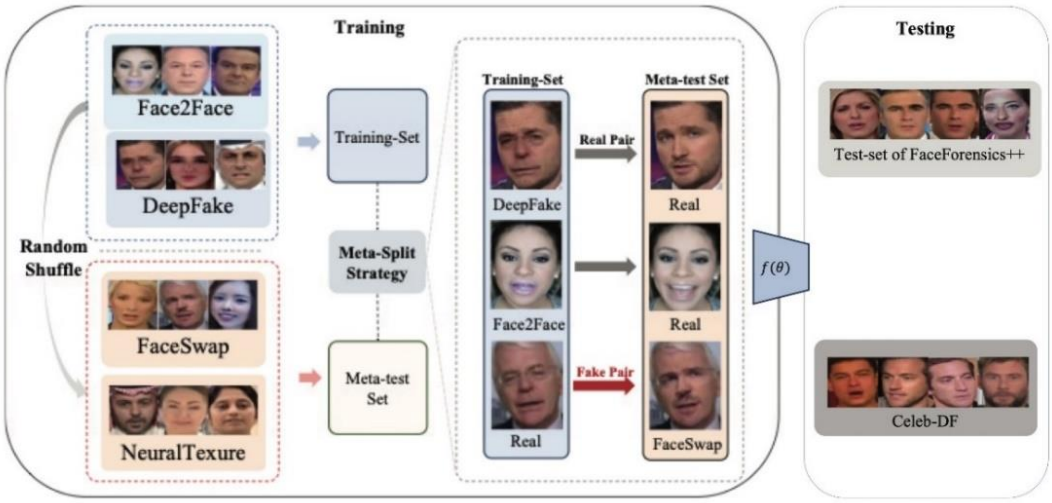

图 4-2 训练过程中元训练域和元测试域的划分

(1) 对于每一张真实图片，我们从D_{meta}^{s}选择其对应的虚假图片。

(2) 对于每一个虚假图片，我们从D_{meta}^{s}选择去对应的真实图片。

该策略保证了模型可以学习到不同域（即不同生成方法）上的 pair 信息。

4.3.3 学习过程

将整个模型的学习过程划分为三个步骤：元训练、元测试和参数更新，在图二中显示了这一过程。

步骤 1：元训练

这一步骤是为了计算带着权重感知网络的二分类器$f(\theta)$在元训练子集上的 loss，具体而言我们从D_{nt}^{s}中选择 K 个训练数据并标记为$X_s = \{x_i; y_i\}_{i=1}^{K}$。通常来讲，参数$\theta$的更新通过最小化损失函数 L 来实现，但避免过拟合，我们在计算最终损失函数时添加一个域适应的权重$p(f_i; \omega)$，因此，元训练时的损失$T(\theta, \omega)$表示如下，其中 L 会在 4.3 节中进行说明。

$$T(\theta, \omega) = \frac{1}{K}\sum_{i=1}^{K} L\left((x_i, y_i); \theta\right) * p(f_i; \omega) \quad (1)$$

步骤 2：元测试

在元训练之后，我们获得了在元训练子集上的损失函数 T，接下来我们应该考虑如何更新权重感知网络的参数ω，并更好的利用对于模型未知的源数据D_{meta}^{S}来提升模型的泛化性。

受到 Li[41]的启发，我们利用模型参数的二阶微分。首先我们利用梯度下降来模拟一次对参数θ的更新，参数α表示元训练阶段的学习率

$$\theta' = \theta - \alpha \nabla_{\theta} T(\theta, \omega) \qquad (2)$$

接下来使用虚拟参数θ'的检测模型会在元测试子集$X_m = \{x_i', y_i'\}_{i=1}^{K}$上进行测试，这部分数据在每个 epoch 中对模型是不可见的，所以模型$f(\theta)$可以学习到不同数据域上通用的特性。

元测试阶段的 loss 被定义为：

$$M(\theta', \omega) = \frac{1}{K} \sum_{i=1}^{K} L\left((x_i', y_i'); \theta', \omega\right) \qquad (3)$$

步骤 3：参数更新

元训练阶段在元训练子集上得到 loss T，元测试阶段在元测试子集得到了 loss M，下面是训练阶段对参数进行更新的方法。定义每次更新的目标为：

$$argmin_{\theta,\omega} T(\theta, \omega) + \beta M(\theta', \omega) \qquad (4)$$

β表示元测试 loss 的重要性，可以平衡元训练和元测试阶段，最小化公式(4)，我们可以使用梯度下降的方法来更新模型参数θ和参数ω。

$$\theta^* = \underbrace{\theta - \alpha \nabla_{\theta} T(\theta, w)}_{meta-train\ update} + \underbrace{\beta(\theta' - \gamma \nabla_{\theta'} M(\theta', w))}_{meta-test\ update} \qquad (5)$$

$$w^* = w - \phi \nabla_{w} M(\theta', w) \qquad (6)$$

参数 γ 表示元测试阶段的学习率，参数 Φ 控制更新 ω 的速率，即权重感知网络的学习率。

4.4 模型优化

考虑到存在许多种不同的攻击方式，模型可能无法覆盖到所有伪造人脸的种类，但是真实人脸的分布却相对稳定。为了更好的利用这一点，我们首先将这个问题归类为一个二分类问题。受到 Perera and Patel[42]的影响，文章提出了一个新

的损失函数[9]，命名为 Intra-Class-Compact(ICC) Loss。

优化的思想是为了使真实样本聚集，将虚假样本推离真实样本的中心。具体而言，定义 $O = \{o_1, o_2 \ldots o_n\} \in R^{n*1}$ 为检测模型$f(\theta)$经过池化层和一个全连接层的输出，样本的输出会带有标签$Y = \{y_1, y_2, \ldots, y_n\}, y_i \in \{(0,1)\}$，0 表示一个预测为真的样本，1 表示一个预测为假的样本，根据标签 Y 我们将样本集划分为真实样本集O^{real} 和虚假样本集 O^{fake}。

真实样本和真实样本中心的距离定义为

$$L_{positive} = \frac{1}{|O^{real}|} \sum_{j=1}^{|O^{real}|} (o_j^{real} - C_{real})^2, \qquad (7)$$

where

$$C_{real} = \frac{1}{N} \sum_{j=1}^{N} o_j^{real}. \qquad (8)$$

虚假样本和真实样本中心的距离被定义为

$$L_{negative} = \frac{1}{|O^{fake}|} \sum_{j=1}^{|O^{fake}|} (o_j^{fake} - C_{real})^2. \qquad (9)$$

所以我们定义 ICC-Loss 为

$$L_{icc} = L_{positive} - L_{negative} \qquad (10)$$

值得注意的是在每个 mini-batch 中C_{real}都会被更新，最终在模型中使用的损失函数为

$$L = L_{ce} + \lambda L_{icc} \qquad (11)$$

L_{ce} 是二分类损失函数，λ 用于平衡 CE loss 和 ICC loss。

4.5 模型实现

LTW 框架的实现主要包括了数据的预处理、网络搭建、训练函数编写等部分。LTW 模型主要使用 FF++数据集来进行训练，我用 MTCNN 框架[43]对视频中

每一帧的人脸进行了提取，根据模型的输出取一个正方形人脸框并放大至 1.3 倍。

MTCNN 主要有以下四个步骤：

(1) 首先需要制作图像金字塔（将图像按照从大到小的尺寸堆叠在一起类似金字塔形状），将输入图像 resize 为不同的尺寸。

(2) 将图像金字塔输入 P-Net（Proposal Network），获取人脸的边界框（Proposal boundding boxes），并通过非极大值抑制（NMS）算法去除冗余框，这样便初步得到一些人脸检测候选框。

(3) 将 P-Net 输出的人脸图像输入到 R-Net（Refinement Network），对人脸检测框进一步细化，并用 NMS 算法去除冗余框

(4) 将 R-Net 输出的人脸图像输入 O-Net（Output Network），进一步细化人脸检测框坐标，同时输出人脸的 5 个关键点（左嘴角、右嘴角、鼻子、左眼、右眼）

我选用在 ImageNet 上预训练的 EfficientNet-b0 作为基础检测模型，使用等步长（stepLR）的梯度更新策略（其中 step-size 设定为 5，gamma 设定 0.1），将元训练的学习率α和元测试的学习率γ都设定为 0.001，并使用 Adam 优化算法。权重感知网络的学习率Φ被设定为 0.01，平衡元训练和元测试的参数β设定为 1，而平衡 CE loss 和 ICC loss 的参数λ设定为 0.01。

训练使用的 FF++数据集依照官方对数据集的划分，720 个视频用于训练，140 个视频用于验证，140 个视频用于测试。训练前我们会将图片大小 resize 到 224 * 224，训练过程中我们等间距提取每个视频中的 10 帧，batch 大小设定为 25，一次训练跑 15 个 epoch。

4.6 本章小结

在本章中，首先介绍了当今通用人脸检测这一场景，即检测方法在不可见目标域上的性能，随后我对元学习的思想进行了简要介绍。本章重点介绍了基于学习权重的通用人脸检测模型，即 learning-to-weight(LTW)模型，该模型基于元学习的思想，在基础检测模型之上考虑不同数据域之间的权重，以此来学习不同数据集之间的共同特征，提高了检测模型的泛化性，本章详细介绍了模型的框架，其优化方法，并对该模型的实现过程进行了说明。

5 实验比较与分析

第四章中对基于学习权重的通用人脸检测模型进行了详细介绍，本章将通过多组实验来验证 LTW 模型检测伪造视频的有效性，使用多种评价指标评测模型的性能，并证明其可以获得较好的泛化效果。

本文的实验在 Linux 服务器下进行，环境如下表 6-1 所示。

表 6-1 实验硬件配置

名称	配置
操作系统	CentOS
CPU	Intel E5-2650 v4
GPU	Nvidia RTX 2080 * 4

5.1 Benchmark 划分

为了验证我们模型的泛化性，我们利用 FF++数据集和 Celeb-DF 数据集划分了多组不同的 benchmarks

由于 FF++数据集中运用了四种不同的生成方法，包括两种计算机图形学的方法和两种深度学习的方法，可以任务不同生产方法的数据集相当于不同的数据分布，所以我们基于攻击方法划分出源域和目标域，来验证在源域上训练的模型可以在目标域上取得较好的检测结果。我们同样考虑到视频质量的影响，分别会针对 c23 压缩率和 c40 压缩率的视频进行测试。除此之外，为了更好的测试模型的泛化性，我们设置 GCD benchmark，使用 FF++中的全部数据作为源域，目标域会使用 FF++和 Celeb-DF，这样可以更有效得测试检测模型的跨库性能。

表 6-2 实验 Benchmark 划分

Benchmark 名称	**压缩率**	**源域**	**目标域**
GID-DF23/40	c23/c40	Face2Face FaceSwap NeuralTextures	DeepFake
GID-F2F23/40	c23/c40	DeepFake FaceSwap NeuralTextures	Face2Face

GID-FS23/40	c23/c40	DeepFake Face2Face NeuralTextures	FaceSwap
GID-NT23/40	c23/c40	DeepFake Face2Face FaceSwap	NeuralTextures
GCD	c23	DeepFake Face2Face FaceSwap NeuralTextures	DeepFake Face2Face FaceSwap NeuralTextures Celeb-DF

5.2 性能评价指标

为了对检测模型的性能进行更好的评估，我们使用了四种常见的评价指标，包括检测准确率（ACC）、回归误差（LogLoss）、ROC 曲线面积（AUC）以及平均错误概率（EER）。其中回归误差我们选用 DFDC 竞赛中使用的二分类损失函数，函数表达式如下：

$$LogLoss = -\frac{1}{n}\sum_{i=1}^{n} y_i log(\hat{y}_i) + (1-y_i) log(1-\hat{y}_i), \quad (12)$$

其中有 n 张图片用于测试，y_i表示第 i 张图片的真假标签，$\hat{y}i$表示第 i 张图片的预测结果,log()函数以 e 为基底。

5.3 对照组设置

为了测试 LTW 模型的有效性和性能，我设置了三个不同的对照组作为 baseline。

(1) 基本模型（Basemodel），在 ImageNet 上训练得到的 backbone 直接用于目标域进行测试，提供了最简单的 baseline。

(2) 全训练模型（Alltrain-Basemodel），将所有的数据都用来当作源域训练，相当于模型训练阶段已经完整学习到了全部数据集的数据分布，提供了最权威的 baseline。

(3) 基于 FocalLoss 的基本模型（FocalLoss-Basemodel），在所有源域上进行训练，但是训练过程中的损失函数使用 Focal Loss。

5.4 实验结果分析

5.4.1 在 GID 上的结果

表 6-3 在 GID-DF23 和 GID-F2F23 上的测试结果

Benchmarks	**GID-DF23**				**GID-F2F23**			
Metric	**ACC**	**LOSS**	**AUC**	**EER**	**ACC**	**LOSS**	**AUC**	**EER**
Basemodel	0.482	0.721	0.479	0.612	0.499	0.781	0.533	0.527
Alltrain-Basemodel	0.793	0.911	0.890	0.165	0.621	1.821	0.711	0.341
FocalLoss-Basemodel	0.781	0.795	0.834	0.213	0.594	**0.890**	0.689	0.355
LTW	**0.826**	**0.584**	**0.986**	**0.176**	**0.677**	1.492	**0.782**	**0.293**

表 6-4 在 GID-FS23 和 GID-NT23 上的测试结果

Benchmarks	**GID-FS23**				**GID-NT23**			
Metric	**ACC**	**LOSS**	**AUC**	**EER**	**ACC**	**LOSS**	**AUC**	**EER**
Basemodel	0.492	0.778	0.486	0.563	0.488	0.732	0.412	0.556
Alltrain-Basemodel	0.513	3.211	0.499	0.527	0.601	2.012	0.685	0.387
FocalLoss-Basemodel	0.477	1.992	0.478	0.531	0.574	**1.431**	0.636	0.419
LTW	**0.564**	**1.140**	**0.585**	**0.434**	**0.623**	1.982	**0.745**	**0.321**

表 6-5 在 GID-DF40 和 GID-F2F40 上的测试结果

Benchmarks	**GID-DF40**				**GID-F2F40**			
Metric	**ACC**	**LOSS**	**AUC**	**EER**	**ACC**	**LOSS**	**AUC**	**EER**
Basemodel	0.480	0.722	0.490	0.621	0.487	0.773	0.526	0.533
Alltrain-Basemodel	0.621	1.929	0.689	0.334	0.599	1.832	0.668	0.397
FocalLoss-Basemodel	0.615	**0.874**	0.672	0.319	0.581	0.789	0.642	0.398
LTW	**0.686**	0.934	**0.757**	**0.315**	**0.638**	**0.721**	**0.682**	**0.364**

表 6-6 在 GID-FS40 和 GID-NT40 上的测试结果

Benchmarks	**GID-FS40**				**GID-NT40**			
Metric	**ACC**	**LOSS**	**AUC**	**EER**	**ACC**	**LOSS**	**AUC**	**EER**
Basemodel	0.473	0.754	0.483	0.577	0.462	0.755	0.404	0.540
Alltrain-Basemodel	0.536	3.983	0.503	0.523	0.548	2.338	0.564	0.402
FocalLoss-Basemodel	0.529	1.721	0.516	0.499	0.546	**1.131**	0.569	0.414
LTW	**0.552**	**1.555**	**0.590**	**0.444**	**0.563**	1.689	**0.592**	**0.425**

从表 6-2 到表 6-6 显示了不同检测模型在我们划分的 benchmark 上的实验结果，从四个评价指标综合来看，显然 LTW 模型的效果更佳，特别是在 GID-FS23、GID-F2F23、GID-DF40、GID-F2F40 这几组实验中，LTW 模型的准确率相比于全训练网络提供的 baseline 增长了 5%左右，在某些实验中，FocalLoss-Basemodel 可以获得比 LTW 模型更小的回归误差，但其检测的准确性不如 LTW 模型。

5.4.2 在 GCD 上的结果

表 6-7 在 GCD 上的测试结果

Benchmarks	GCD-CeleDF				GCD-Others			
Metric	**ACC**	**LOSS**	**AUC**	**EER**	**ACC**	**LOSS**	**AUC**	**EER**
Basemodel	0.554	0.883	0.491	0.517	0.510	0.663	0.499	0.517
Alltrain-Basemodel	0.603	2.681	0.588	0.466	0.901	0.460	0.977	0.101
FocalLoss-Basemodel	0.596	1.326	0.544	0.474	0.894	0.333	0.980	0.089
LTW	**0.646**	**1.031**	**0.629**	**0.404**	**0.920**	**0.299**	**0.975**	**0.067**

在 GCD benchmark 上的实验可以更有效得测试 LTW 模型的泛化能力，Celeb-DF 数据集和训练用的 FF++数据集关联性更小，由此能验证 LTW 模型在跨数据集上的表现。GCD-Others 显示 LTW 在 FF++数据集上的测试域的表现。我们可以看到在 Celeb-DF 数据集中，LTW 模型检测的准确率为 64.6%，相比于全训练模型提升了 4.3%，由此可以看出 LTW 模型具有较强的泛化能力，面对未知数据集依然保持了良好的检测效果。

5.5 本章小结

本章主要是针对 LTW 模型进行的实验过程，为了充分测试模型的性能，我们选用了准确率、ROC 曲线面积、平均错误概率四个评价指标。根据 FF++数据集中生成方法的不同，我划分了四组不同的 benchmark，同时和 Celeb-DF 数据集结合设置第五组 benchmark，以此来验证模型的泛化性，经过实验发现，基于学习权重的通用检测模型可以在不可见目标域上取得良好的效果，模型准确率对比实验的 baseline 均有所增加，部分实验中可以提升 5%左右，以此证明了模型的泛化能力得到实质提升。

6 总结与展望

人工智能技术的发展使得深度伪造技术生成的人脸愈发逼真，而大量开源的换脸软件，让没有计算机专业基础的人们也可以快速进行人脸伪造，虚假信息的泛滥毫无疑问会威胁国家安全以及个人隐私，因此在利用伪造生成技术的同时，我们必须防范生成技术滥用带来的危害，应该加强对可靠的深度伪造检测技术的探索。

人脸伪造生成和深度伪造检测技术之间相辅相成，相互博弈，本文的一大工作便是对这两个研究领域的发展现状进行了梳理，对生成技术、研究数据集、检测方法三个方面进行了总结，整理了许多领域内影响力较大的研究成果，并对这些技术和方法进行了科学的分类。

现今的检测方法基本上依赖于特定的数据集，而这些数据集中的生成算法较为单一，导致训练数据的分布较为相近，所以检测模型的泛化性较弱，本文针对这一问题，对通用人脸伪造检测的算法进行研究，复现了一个基于学习权重的通用人脸伪造检测模型，通过对共性特征的学习，使得检测模型的泛化性得到提高，可以在不可见目标域上保持较好的效果。通过充分的实验，我们有效验证了 LTW 模型可以提升检测方法的泛化性。

虽然本文实现的模型取得了一定的效果，但是对模型泛化性的研究依然是一个非常困难的任务，未来可以继续对这一方向进行更深入的探索，可以尝试从模型的原理出发，分析生成技术的固有缺陷，如生成器的指纹等等，以此使得现有的检测方法可以更好的应用于真实场景。

除了模型的泛化性，检测算法的鲁棒性也至关重要，由于深度伪造图像经过互联网传播，所以图像压缩等图像预处理操作很可能对检测模型造成干扰。因此我们可以进一步探索不同预处理方法对算法鲁棒性的影响。

综上所述，伪造检测算法有非常大的研究价值和实际意义，目前该领域的研究仍然有诸多问题需要解决，也存在很多的研究难点，因此未来我们可以从多个角度多个层次来深入探索伪造检测领域，促进领域的发展和进步。

致　谢

行文至此，毕业论文完稿了，我的大学时光也即将结束。回首在华科的生活，喻家山下的日子里忙碌充实也充满乐趣，不能说这四年非常圆满，但我也体验过了独属于自己的精彩。

感谢我未来的导师孙哲南老师和王伟老师，感谢两位老师帮助我选定毕业论文的题目，引领我初探科研工作的奥秘，未来的博士求学路还要在您们的指导下砥砺前行。感谢软件学院的黄立群老师，黄老师在毕业设计期间对我的工作提出了很多的建议，感谢黄老师不厌其烦的帮我修改论文，为我的大学画上一个完美的句号。感谢四年里华科和软院所有为我上过课的老师们，传道，受业，解惑，谆谆教诲，谨记于心。感谢自动化所的管伟楠师兄、张时润师兄，两位师兄在我实验遇到困难时给予了巨大的帮助，未来还要多向师兄师姐们学习和请教。

四年生活里，我认识了诸多志同道合的朋友，非常幸运能和你们在 1037 号森林公园里共度青春，难忘我们无数次在公用房熬夜复习时的相互帮助，难忘我们在每年工程实训时加班写代码调程序的光阴，难忘我们一起看江城美景品武汉美食的惬意。我的大学因为你们而变得更加灿烂，在此不一一列举大家的名字了，但衷心祝愿我们都有一个美好的未来。

父母之恩永不能忘，父母总是无条件支持我的想法，用无私的爱呵护我的成长。在此，我非常想对家人们说一声抱歉，大学生活里很多次冲动带来的困境都是你们的陪伴让我度过难关，很抱歉自己的不懂事让你们时刻操心，希望你们身体健康、快乐幸福，我也会不断成长，终将成为家里的顶梁柱。

最后我想感谢一下自己，四年里我面临过很多的选择，也遭受过巨大的打击，但我很庆幸自己始终在努力奔跑，我坚信天道酬勤，四年的刻苦求学让自己获得去中科院自动化所深造的机会，不枉费那些奋斗拼搏的日子。希望未来的自己能始终咬着牙微笑，迎着风坚定奔跑，在此也预祝自己可以顺利拿到博士学位。

逝者如斯乎，不舍昼夜。我们的青春留在了喻家山下，我们的回忆终将成过往，但我想得到了喻家山哺育的我们终将在未来成为母校的骄傲。感谢华科，我爱在这里遇到的一切，我爱 18-22 岁这属于我的黄金时代。

参考文献

[1] Goodfellow I J, Pouget-Abadie J, Mirza M, et al. Generative adversarial networks[J]. arXiv preprint arXiv:1406.2661, 2014.

[2] Kingma D P, Welling M. Auto-encoding variational bayes[J]. arXiv preprint arXiv:1312.6114, 2013.

[3] 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). IEEE, 2019: 8261-8265.

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

[5] 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. 2019: 1-11.

[6] 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. 2020: 5781-5790.

[7] Dolhansky B, Bitton J, Pflaum B, et al. The deepfake detection challenge dataset[J]. arXiv preprint arXiv:2006.07397, 2020.

[8] Li Y, Yang X, Sun P, et al. Celeb-df: A large-scale challenging dataset for deepfake forensics[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 3207-3216.

[9] Sun K, Liu H, Ye Q, et al. Domain General Face Forgery Detection by Learning to Weight[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2021, 35(3): 2638-2646.

[10] Nirkin Y, Masi I, 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). IEEE, 2018: 98-105.

[11] Nirkin Y, Keller Y, Hassner T. Fsgan: Subject agnostic face swapping and reenactment[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 7184-7193..

[12] 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. 2016: 2387-2395.

[13] 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.

[14] Radford A, Metz L, Chintala S. Unsupervised representation learning with deep convolutional generative adversarial networks[J]. arXiv preprint arXiv:1511.06434, 2015.

[15] Berthelot D, Schumm T, Metz L. Began: Boundary equilibrium generative adversarial networks[J]. arXiv preprint arXiv:1703.10717, 2017.

[16] 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.
[17] Zhang Z, Song Y, Qi H. Age progression/regression by conditional adversarial autoencoder[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 5810-5818.
[18] Li T, Qian R, Dong C, et al. Beautygan: Instance-level facial makeup transfer with deep generative adversarial network[C]//Proceedings of the 26th ACM international conference on Multimedia. 2018: 645-653.
[19] Zhu J Y, Park T, Isola P, et al. Unpaired image-to-image translation using cycle-consistent adversarial networks[C]//Proceedings of the IEEE international conference on computer vision. 2017: 2223-2232.
[20] 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.
[21] Dolhansky B, Howes R, Pflaum B, et al. The deepfake detection challenge (dfdc) preview dataset[J]. arXiv preprint arXiv:1910.08854, 2019.
[22] 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. 2020: 2889-2898.
[23] He Y, Gan B, Chen S, et al. ForgeryNet: A Versatile Benchmark for Comprehensive Forgery Analysis[J]. arXiv preprint arXiv:2103.05630, 2021.
[24] Chollet F. Xception: Deep learning with depthwise separable convolutions[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 1251-1258.
[25] Tan M, Le Q. Efficientnet: Rethinking model scaling for convolutional neural networks[C]//International Conference on Machine Learning. PMLR, 2019: 6105-6114.
[26] McCloskey S, Albright M. Detecting gan-generated imagery using color cues[J]. arXiv preprint arXiv:1812.08247, 2018.
[27] Zhou P, 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). IEEE, 2017: 1831-1839.
[28] LeCun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.
[29] Nguyen H H, Fang F, Yamagishi J, et al. Multi-task learning for detecting and segmenting manipulated facial images and videos[C]//2019 IEEE 10th International Conference on Biometrics Theory, Applications and Systems (BTAS). IEEE, 2019: 1-8.
[30] Tariq S, Lee S, Kim H, et al. Detecting both machine and human created fake face images in the wild[C]//Proceedings of the 2nd international workshop on multimedia privacy and security. 2018: 81-87.
[31] Li Y, Chang M C, Lyu S. In ictu oculi: Exposing ai created fake videos by detecting eye blinking[C]//2018 IEEE International Workshop on Information Forensics

and Security (WIFS). IEEE, 2018: 1-7.
[32] Donahue J, Anne Hendricks L, Guadarrama S, et al. Long-term recurrent convolutional networks for visual recognition and description[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 2625-2634.
[33] 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). IEEE, 2018: 1-6.
[34] Sabir E, Cheng J, Jaiswal A, et al. Recurrent convolutional strategies for face manipulation detection in videos[J]. Interfaces (GUI), 2019, 3(1).
[35] 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). IEEE, 2018: 1-7.
[36] Li Y, Lyu S. Exposing deepfake videos by detecting face warping artifacts[J]. arXiv preprint arXiv:1811.00656, 2018.
[37] Agarwal S, Farid H, Gu Y, et al. Protecting World Leaders Against Deep Fakes[C]//CVPR Workshops. 2019: 38-45.
[38] Li L, Bao J, Zhang T, et al. Face x-ray for more general face forgery detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 5001-5010.
[39] Finn C, Abbeel P, Levine S. Model-agnostic meta-learning for fast adaptation of deep networks[C]//International Conference on Machine Learning. PMLR, 2017: 1126-1135.
[40] Shu J, Xie Q, Yi L, et al. Meta-weight-net: Learning an explicit mapping for sample weighting[J]. arXiv preprint arXiv:1902.07379, 2019.
[41] Li D, Yang Y, Song Y Z, et al. Learning to generalize: Meta-learning for domain generalization[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2018, 32(1).
[42] Perera P, Patel V M. Learning deep features for one-class classification[J]. IEEE Transactions on Image Processing, 2019, 28(11): 5450-5463.
[43] Zhang K, Zhang Z, Li Z, et al. Joint face detection and alignment using multitask cascaded convolutional networks[J]. IEEE Signal Processing Letters, 2016, 23(10): 1499-1503.