上海大学

SHANGHAIUNIVERSITY

毕业论文（设计）

UNDERGRADUATE THESIS (PROJECT)

题目：Deepfake(深度伪造)检测研究与开发

学院	计算机工程与科学学院
专业	智能科学与技术
学号	18123158
学生姓名	沈纪元
指导教师	武星
起讫日期	2022.02.21– 2022.06.03

目 录

Deepfake（深度伪造）检测研究与开发

摘 要

近年来，由 Deepfake 伪造技术生成的视频和图像在网络上广泛流传，以其逼真性和广泛性在社会造成了极大的消极作用。而目前的检测技术还处于萌芽阶段，在检测速度和准确性、模型泛化性和可解释性、实时性上依然存在很大的提升空间。因此，本文摒弃了使用单一的原始图像，对基于频域分析的 Deepfake 检测作了更深入的研究，提出两个 Deepfake 检测算法。第一种是基于频域分析和双流网络的算法，主要改善了以往单纯的频域变换方式。具体而言，通过使用可学习的频率滤波器进行频域特征提取，运用双流网络、特征融合和多任务学习模式大大提升了模型的泛化性和可解释性。第二种算法在前者的基础上进行了改进，在大大减少模型参数量的同时，也保证了在大规模弱监督数据集上进行预训练的可行性。特别地，本文所提出的多模特图像融合注意力机制是第一个将嫁接思想运用到 Deepfake 检测领域的，同时所提出的模块也模拟了人类观测事物的直觉本能。通过结合多流网络，实现了从全局到局部细节再回归全局的思想，换言之，该思想遵从从时域出发，查询局部频域统计信息，最后再回到经过频域滤波的时域信息的过程。实验证明，使用此方法大约减少了十倍的模型参数量，同时收敛速度和模型的精度也都有显著的提升。

关键词：深度伪造检测，频域分析，融合注意力，多任务学习，多流网络

Deepfake Detection Research and Development

ABSTRACT

In recent years, videos and images generated by Deepfake have been widely spread on the Internet, which cause great negative effects in society due to their reality and widespread. However, the current detection technology is still in its infancy, and there is still a lot of room for improvement in terms of detection speed and accuracy, model generalization and interpretability, and real-time performance. Therefore, this thesis abandons the widely use of a single original image modality, dig deeper into a more advanced study on Deepfake detection based on frequency domain analysis, and proposes two Deepfake detection algorithms. The first is an algorithm based on frequency domain analysis and two-stream network, which mainly improves the previous simple frequency domain transformation method. Specifically, by using learnable frequency filters for frequency-domain feature extraction, applying two-stream networks, feature fusion and multi-task learning modes, the generalization and interpretability of the model have been improved greatly. The second algorithm then makes some other improvements on the basis of the former, which not only reduces the amount of model parameters, but also ensures the feasibility of pre-training on large-scale datasets under weakly supervision. In particular, the fusion attention mechanism of multi-model images proposed in this thesis is the first to apply the graft idea to the field of Deepfake detection, and the proposed module also well simulates the human instinction to observe things. By combining the multi-stream network, the idea of travelling from the global to the local information and then back to the global is realized. In other words, the idea follows the process of starting from the time domain, querying the local frequency domain statistical information, and finally returning to the time domain information filtered in the frequency domain. Experiments show that using this method reduces the number of model parameters by about ten times, and at the same time, the convergence speed and the accuracy of the model are also significantly improved.

Keywords: Deepfake Detection, Frequency Domain Analysis, Graft Attention, Multi-task Learning, Multi-stream Network

第 1 章　绪论

本章主要描述了 Deepfake（深度伪造）检测的研究背景及意义，分析了目前的研究问题以及难点，最后提出了本文所研究的内容及主要贡献。

§1.1　Deepfake 检测的研究背景及意义

近年来，得益于深度生成模型的发展，人脸的操控技术取得了巨大突破，以 Deepfake 为代表的人脸视频深度伪造技术在互联网快速流行，受到了学术界和工业界的广泛重视。Deepfake 定义为“以某种方式使合理的观察者错误地将其视为个人真实言语或行为的真实记录的方式创建或更改的视听记录”[1]，其中“视听记录”即指图像、视频和语音等数字内容。这种深度伪造技术通过交换原始人脸和目标人脸的身份信息、编辑目标人脸的属性信息或者直接虚拟合成的方式来生成虚假的人脸图片或者视频。而其中，视频伪造是 Deepfake 技术最为主要的代表，制作假视频的技术也被业界形象化地称为人工智能换脸技术（AI face swap），当然 Deepfake 不只包括换脸，换脸只是为了方便理解的一种简称。其核心原理是利用生成对抗网络[2] 或者卷积神经网络[3] 等算法将目标对象的面部“嫁接”到被模仿对象上。由于视频是连续的图片组成，因此只需要把每一张图片中的脸替换，就能得到变脸的新视频。具体而言，首先将模仿对象的视频逐帧转化成大量图片，然后将目标模仿对象面部替换成目标对象面部。最后，将替换完成的图片重新合成为假视频，而深度学习技术可以使这一过程实现自动化。

人脸深度伪造技术激发了很多相关的娱乐应用，如使用面部替换技术将使用者的人脸替换到某段电影片段中，或使用表情重演技术来驱动某个著名人物的静态肖像等[1]。目前市面中也可以发掘许多相关的应用产品[2]，但当前人脸深度伪造技术仍处于快速发展阶段，其生成的真实感和自然度仍有待进一步提升。然而另一方面，由于许多应用产品的诞生，降低了此类技术的实现和应用门槛，导致很多不法分子利用深度伪造技术来做违法的事情，比如用来制作色情电影、虚假新闻，甚至被用于政要人物来制造政治谣言等[3]，见图1.1，这对国家安全与社会稳定都带来了极大的潜在威胁，因此深度伪造的防御技术至关重要。

为了降低深度伪造人脸视频所带来的负面影响，众多学者对伪造人脸视频的检

1　https://derivative.ca/community-post/deepfake-salvador-dal%C3%AD-interacts-museum-visitors-takes-selfies/60968

2　https://apps.apple.com/us/app/reface-face-swap-videos/id1488782587

3　https://ars.electronica.art/center/en/obama-deep-fake/

图 1.1 Deepfake 的应用场景。(a) 驱动奥巴马发表演说；(b) 驱动静态图像画像；(c) 更换目标的脸为任意他人的脸

测鉴别技术进行了深入研究，并从不同视角提出了一系列防御方法。然而由于数据集分布形式单一、评价标准不一、针对性攻击、对抗性建模等因素[1]，使得防御检测技术在走向实用的道路上仍有很长一段距离。事实上，人脸深度伪造与防御技术的研究仍旧处在高速的发展期之中，其技术的内涵与外延正在快速的更新与迭代。

§1.2 Deepfake 检测的研究问题与难点

Deepfake 的检测研究面临着以下的几个问题和难点。事实上，由于目前多采用深度学习的生成模型来进行深度伪造，因此不可避免的出现对抗性的攻击，这是本研究的最大问题之一。而研究的难点就在于探寻一条普适、性能与精度较高的检测方法。

§1.2.1 研究存在的问题

Deepfake 的生成和检测之间的关系很像一种“竞争与寄生”关系，两者相互促进，又相互攻击。由于本文着重于检测相关研究，因此以下简述两点检测方面存在的问题：

1. 针对伪像提取的攻击：为了躲避基于伪像的检测器，生成器（G）只需要减轻单个缺陷就可以躲避检测。例如，G 可以通过添加监控这些信号的鉴别器来产生由[4,5] 监控的生物信号。为了避免大范围神经元激活的异常[6]，G 可以增加一个最小化神经元覆盖的损失。检测异常姿势和姿势的方法[7] 可以通过重现整个头部和从相同的数据库中学习姿势来避免。
2. 针对深度学习分类器的攻击：有许多检测方法将深度学习直接应用于 Deepfake 检测的任务[8–11]。然而，G 可以通过向目标对象添加小扰动来使用对抗性机器学习来逃避检测。对抗性机器学习的进展表明，不管使用的训练数据如何，这

些攻击都会跨多个模型而转移[12]。最近的工作表明，这些攻击不仅对 Deepfake 分类器有效[13]，而且在没有分类器或其训练集先验知识的情况下也有效[14]。

以上的问题，可以说是 Deepfake 检测自始自终都存在的问题，是该任务的特性以及 GAN 网络的结构决定的。但即使如此，研究者们还是努力向前，不断的提出高效的方法来解决 Deepfake 检测这一问题。

§1.2.2 研究的难点

Deepfake 的生成和检测是一对互相并行的“双胞胎”，了解了一方的难点就可以知道如何攻击或者防御另外一方，因此本小结会把生成和检测的难点都进行简要概述。

针对 Deepfake 生成伪造图像主要难点在于：

1. 数据量与质量方面[15,16]：现在绝大多数的 AI 模型都是极其依赖数据的，换言之是以数据驱动的。而众所周知，GAN 模型的训练是一个极其困难的事情，尽管到目前为止有很多的论文在阐述一些实用的方面，比如训练策略、优化方式和损失函数的改进等，但究其本源数据依然非常重要。这就导致一般的普通人相较于公众人物，拥有比较少的图片和视频数据，使得 Deepfake 应用于普通人会比较受限。

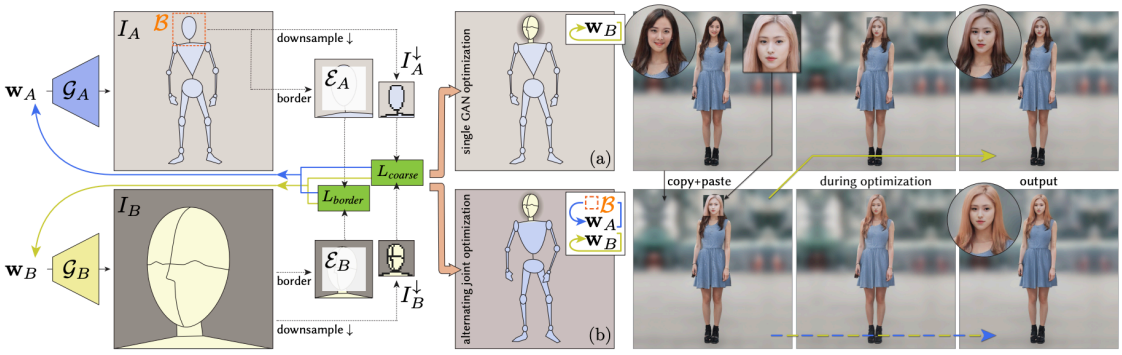


图 1.2 InsetGAN 流程图[17]。

2. 速度与质量方面[18,19]：要生成精细的图片或者视频需要高分辨率模型，这类模型往往具有极大的参数量，这会导致如果要生成深度换脸视频时，很难进行实时的换脸。即使是离线的操作也会导致大量计算资源的占用，尤其是显卡资源并非对于每个人都是易得的。尤其考虑到针对一些模型，会特意添加一个边缘修正网络[17]，见图1.2，运用双网络协同处理换脸的视频或图像，这导致虽然质量提升了，但推理速度和训练速度缺下降了。

除了以上有关质量方面的难点，目前的 Deepfake 还存在一些难点[1]。第一，参考图1.3换脸和再现技术的流程图所列出的技术组成可知：对于再现，内容总是由正

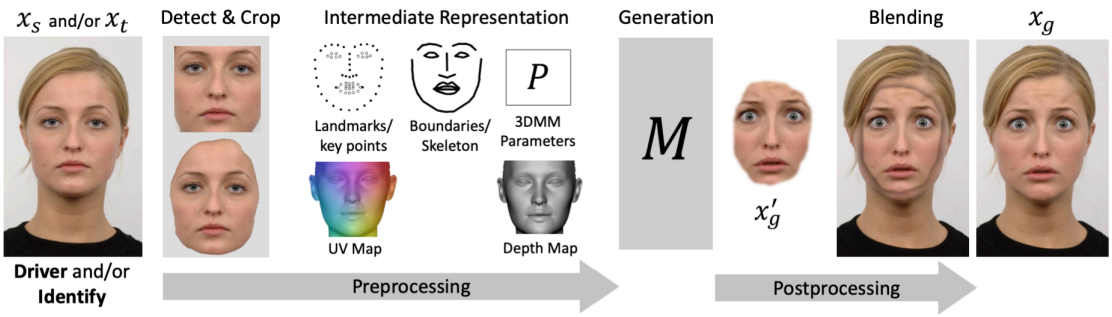
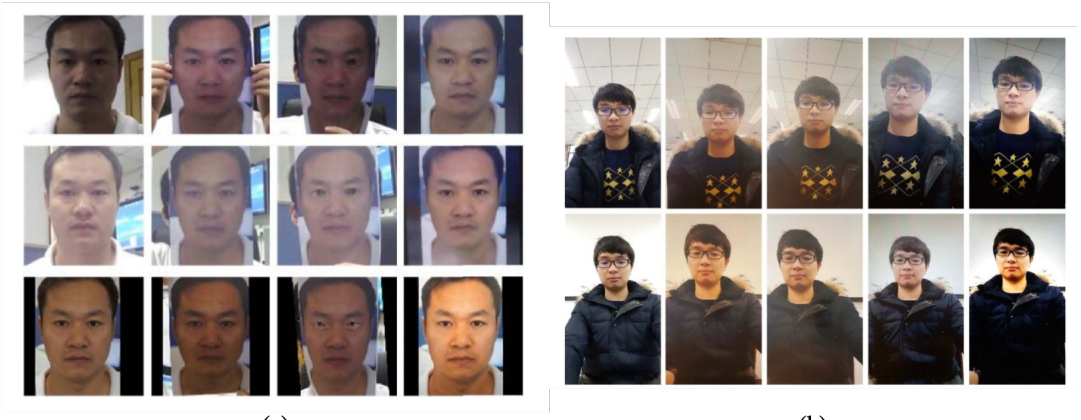


图 1.3　深度伪造中的换脸和再现技术的流程图[1]。一般而言的算法会采用图中所列的部分方法。

面姿态驱动和生成，这就把再现局限在了静止的模式。虽然这种情况可以通过在相似者的身体上进行面部交换来避免，但是很好的匹配并不总是可能的，而且这种方法的灵活性有限。第二，实时 Deepfake。目前有一些作品已经实现了 30fps 的实时 Deepfake，但其具有明显局限性：他的头发、牙齿、舌头、阴影很难实现连贯渲染 (尤其是触摸面部时)。

而正因为有这些生成的难点，在检测 Deepfake 的时候可以充分利用这些因素，做到针对性的防御。因此，检测方面的重难点主要应该集中在以下四点：

1. 基于最新的生成伪造技术提出检测方法。对基于伪影和不一致性的视觉深度伪造检测方法而言，如何保证其所针对的特征的新鲜度是一个亟待解决的难题，越来越多的新型生成伪造技术在伪造方面对原始伪造进行了修复，所以在伪造检测方面需要关注新的伪造生成技术，从生成方面入手，针对新的生成伪造技术进行分析以针对最新的生成伪造技术和真实的人脸，采用对比的方式，如孪生网络等，获取更鲁棒的特征提取，从而获得更好的深度伪造检测性能。
2. 利用高效的基于数据驱动的方法。由于基于数据驱动的方法往往需要大量的数据集和机器学习训练，随着计算显卡资源的大力发展，该伪造检测方法在近几年才逐渐发展起来，并且在未来也将拥有良好的发展前景。但是其提取特征的效率较低、耗费资源较广、耗时较长也是一个需要注意的问题。因此，对基于数据驱动的伪造检测方法而言，可以采用知识蒸馏的方式进行模型压缩，也可以采用教师-学生网络，通过复杂但性能完好的教师网络训练浅层的学生网络模型，从而达到压缩模型，提高模型工作效率的目的。
3. 利用泛化性更强的检测方法。一些检测方法在原理上有创新，但是通常在使用上局限于方法特定的生成技术和数据集上，欠缺泛化能力，容易出现对某个特定数据集检测准确率高、对其他常用的数据集准确率低的情况。因此，鉴别伪造图像方法的泛化能力也是一个非常重要的问题。为了增强检测方法的泛化能力，主要可以采取的方式有数据扩增、尝试进行单分类模型、引入专家先验知



(a) (b)

图 1.4 对抗样本示例图，来源于 H Chan[20]。(a) 为人为噪声，(b) 为光照变化噪声。

识、更多的考虑模型的可解释性等。

4. 利用对抗样本学习防御技术。由于深度伪造检测方法往往是基于机器学习的大量工作提取特征，而这类模型很少有经过鲁棒性测试，通过引入对抗样本学习防御的方法可以一定程度解决这一问题。具体而言，可以采用的方法有：一方面是对对抗样本的识别，另一方面可以在训练集当中加入对抗样本，见图1.4，包括自然噪声（如光照变化）、人为噪声，进而进行鲁棒的模型训练[20]。

以上，大致就是目前 Deepfake 检测存在的一些研究难点。本论文力求在运用最新的基于大数据驱动检测技术的同时，提升模型的泛化性以及可解释性，能够让结果不只是一个单纯的概率，还能体现出究竟是哪一些部分让模型认为他是深度伪造的，让深度学习部分的脱离“黑盒”的诟病，逐渐成为“白盒”。

§1.3 研究内容与主要贡献

为了解决以上问题以及改进存在的一些难点，本文着重研究了如何高效地将一系列专家先验知识和原始图像、传统方法和深度学习进行有机的融合，设计了基于多模态图像融合思想的注意力机制模块。据作者有限的学识所知，本文提出的方法是第一个将嫁接融合思想与注意力机制统一运用到 Deepfake 领域的。具体而言，本文认为使用频域和时域结合的方式能有助于网络更好的学习 Deepfake 残留的一些伪迹，同时区别于一般的注意力机制，本文提出的嫁接融合思想，将原始图像得到的特征表示与局部频域特征统计信息进行结合，生成注意力矩阵，然后与经过频域变换滤波后的原始图像进行相乘，不仅融合了时域和频域的特征，同时也由于局部频域信息统计的信息，辅助了模型的收敛，易于快速提取伪迹特征。本文认为这样子的注意力操作模式是具有可解释性的，代表着从时域出发，查询频域的局部信息，最后再回到时域的过程。

本文基于此模块设计了相关的基于卷积结构的网络框架，该框架运用从全局到局部再回到全局、时域到频域再回到时域的思想提取特征。同时，本文也提出了基于此框架的多任务结构，运用定位任务（Localization）辅助模型更快更高效准确的收敛。实验表明，本结构具有良好的性能与结果表现，不仅收敛速度快，同时占用内存小。总而言之，这项工作的贡献可以概括为以下几点:

- 本文充分分析了目前的主流研究方向以及运用的方法，认为其在先验知识和频域分析的运用上存在一定的局限性与缺陷。
- 本文提出了基于频域分析的双流网络和多流网络，将传统和深度学习进行结合，同时为可学习的滤波器设定一定约束，从而根据不同图像自适应分离出频率信息。
- 本文提出了基于多模态图像融合思想的注意力机制模块（Multi-Model Graft Attention），通过使用频域和时域结合的方式有助于网络更好的学习 Deepfake 残留的伪迹。
- 本文提出了基于此模块的多流网络结构，在收集的数据集上表现出了优秀的效果，同时收敛速度更快，占用内存也更小。

相关代码已开源，可登陆此链接进行查看：https://github.com/shenjiyuan123/Graduate_thesis。

§1.4 文章组织结构

本文总共划分为 5 章。

第 1 章主要描述了 Deepfake 检测的研究背景及意义，分析了目前的研究问题以及难点，提出了本文所要研究的内容及主要贡献。

第 2 章首先对 Deepfake 检测问题进行了描述，介绍了目前的 Deepfake 检测研究方向，大致分为四个研究方向（基于特定伪影的视觉深度伪造检测、基于数据驱动的视觉深度伪造检测、基于信息不一致的视觉深度伪造检测、其他类型视觉深度伪造检测），明确了本文的研究方向。之后阐明了和本文工作最相关的一些研究内容，包括视觉骨干网络（Backbone）、频域分析、双流网络、注意力机制的介绍。

第 3 章主要描述基于频域分析和双流网络的 Deepfake 检测，分别从骨干网络选择、频域特征提取、双流网络以其特征融合等方面介绍模型的构成部分。之后介绍了本工作所用的数据集和本文实验所用的评价指标，阐述了实验的参数设置和训练过程，最后给出了对应的消融实验和结果分析。

第 4 章主要介绍基于多模态图像融合注意力机制的 Deepfake 检测算法的各个模块以及技术细节。具体而言，分别从多流网络、多模态图像融合注意力机制、半监督学习、损失函数等方面介绍了本检测算法的各个模块实现。之后介绍了网络的实验

参数设置和训练过程，最后给出了相应的实验和结果分析。

第 5 章对全文进行了总结，归纳了基于频域分析和双流网络的 Deepfake 检测以及基于多模态图像融合注意力算法的主要工作与创新点。最后给出了 Deepfake 检测领域需要进一步研究的问题。

第 2 章　Deepfake 检测的相关工作

本章节主要描述了 Deepfake 检测的理论基础，首先明确地定义了 Deepfake 检测问题，然后从四个研究方向简介了目前主流的研究方向，明确了本文的主要研究方向为基于数据驱动加之特定伪影提取的检测方法。之后针对本文在第三、四章所采用的方法介绍了相关的研究背景和之前的研究进展成果，分别是从视觉网络 Backbone、频域分析和注意力机制方面展开概述。最后介绍了 Deepfake 的数据集和本文实验部分所用的评价指标。

§2.1　Deepfake 检测问题描述

近年来，随着深度学习的普及与大规模应用，越来越多的 Deepfake 采取深度学习的数据驱动方式进行生成。这类图像由于其高像素、逼真的特性使得人难以通过肉眼进行鉴别，因此许多研究者通过借助计算机的能力来对此进行鉴别。相对应的 Deepfake 检测问题也通常被描述为通过算法或者深度学习模型对图像或者视频进行检测，判读其是否为真实的亦或是伪造的。

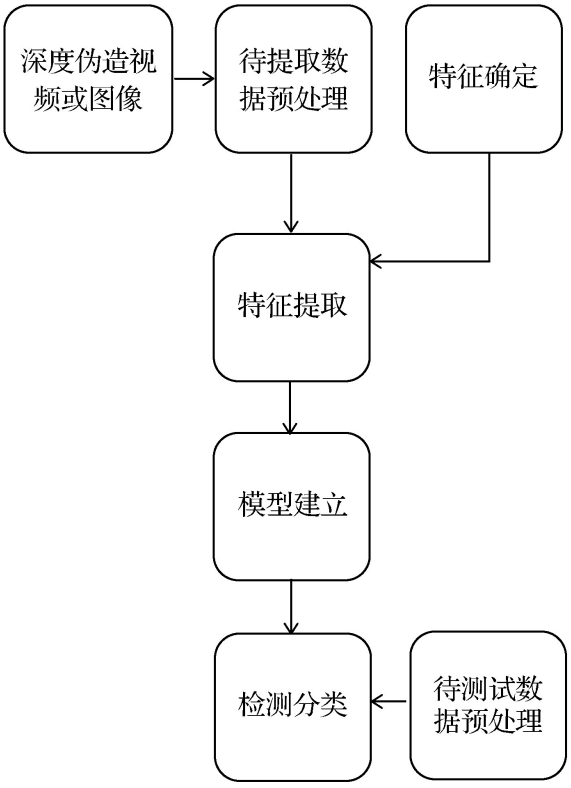


图 2.1　视觉深度伪造检测算法流程

具体而言，视觉深度伪造检测技术主要由特征提取、模型建立、检测分类等步骤进行。首先，需要将待检测的图像或视频数据进行预处理，并根据专家知识或图像处理的方式确定待检测特征。接着，设计相应算法提取特征信息，并建立与检测任务相匹配的网络模型。最后，使用待检测数据对检测算法的性能进行测试，从而验证所选取特征的科学性及分类模型的有效性。其中，决定检测性能的关键就在于如何选择可以有效区分真假人脸的相关特征，以及如何建立分类效果良好的模型。图2.1详细展示了检测算法的具体流程。

常见的 Deepfake 的伪造图像包含四种不同的领域，分别是重演、替代、编辑和合成，如图2.2所示。重演是指用某个 x_s 驱动 x_t 的动作[16,19,21,22]，驱动的内容包括表情，嘴部[15]，身体姿态等。替代是指一部分目标的内容被替换为来源的内容[19,23]。编辑指的是目标的部分属性被添加、修改[24]、或者去除，比如年龄、头发、衣服、体重、种族。合成指的是创造出清晰的高分辨率的但是世界上并不存在的人像图片或者视频[2,9]。

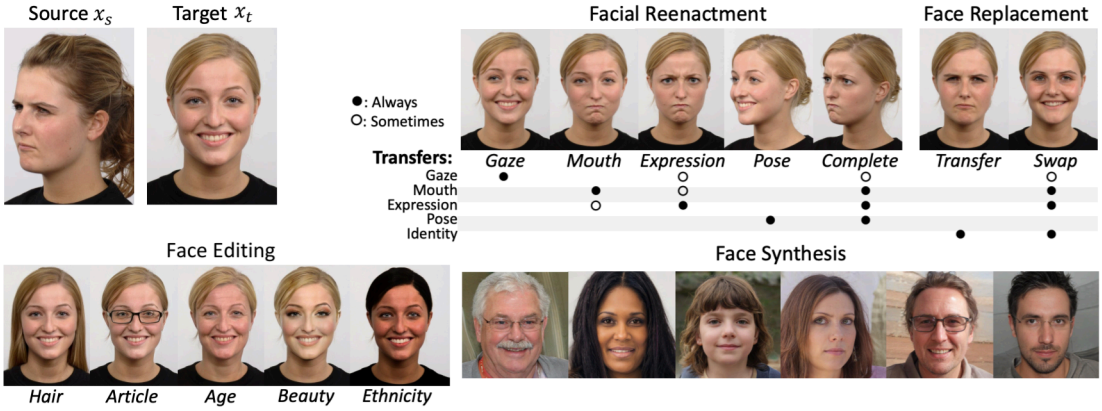


图 2.2 重演、替代、编辑和合成的 Deepfake 样例[1]。

而针对深度伪造的图像，常见的检测大方向有两种：基于异常检测的思想和基于分类的思想。前者思路着重于学习正确的内容应该是什么样子的。异常检测模型在正常数据上训练，然后在部署期间检测异常值。后者则主要利用 CNN，通过给定正常数据和伪造数据，训练模型能够做出正确的分类。在此基础上，为了解决 CNN 只能检测特定训练的攻击，有些使用了诸如分层存储网络 (HMN) 架构[25]，3D-CNN[26]，频域分析[27]，双流网络[28] 等方法。

§2.2 Deepfake 检测主要研究方向

基于特定伪影的视觉深度伪造检测。这些方法专注于寻找假伪影和真实图像之间的像素级差异。在机器学习的帮助下，研究人员可以有效地提取人眼难以察觉的

伪影。无需复杂的网络模型，在实际实验中计算效率高，检测更简单方便。然而，随着基于各种伪影的 Deepfake 检测方法的提出，Deepfake 生成技术逐渐修复了其生成过程中可能产生的伪影。这一点可从数据集的演变中观测一二，自 2016 年以来，视觉 Deepfake 检测数据集，参见表2.1已从第一代 FaceForensics、DF-TIMIT、UADFV 和 FaceForensics++ 数据集发展到第二代数据集，包括 DFDC、Celeb-DF 和 Deeperforensics-1.0。GAN 生成的新数据集和更严格的手工检查在纠正一些已发现的伪影方面发挥了很好的作用。鉴于假冒技术在这方面的逐步发展，检测技术应更多地关注发现普遍存在的伪影和尚未发现的不一致性，同时也需要针对最新提出的 GAN 技术进行研究，从伪造生成技术方面着手，找到对应的检测方法。

表 2.1 视觉深度伪造数据集汇总

数据集	真实数据/幅	伪造	真实内容数据来源	伪造内容数据来源	数据类型	发表日期
CelebA	202 699	0	CelebFaces	/	图像	2015
FaceForensics	1 004	1 004	YouTube	Face2Face	视频	2018.3
UADFV	49	49	YouTube	FakeApp	视频	2018.11
FFW	0	50	/	FakeApp	视频	2018.12
DF-TIMIT	320	LQ:320HQ:320	VidTIMIT	FaceSwap	视频	2018.12
FaceForensics++	1 000	4 000	YouTube	DeepFake, Face2Face, FaceSwap,NeuralTextures	视频	2019.1
WildDeepfake	3 805	3 509	网络上搜集	网络上搜集	视频	2019.11
Celeb-DF	590	5 639	YouTube	DeepFake	视频	2019.11
Deeperforensics-1.0	50 000	10 000	受邀实验者	DeepFake-VAE	视频	2020.1
DFDC	19 154	99 992	受邀实验者	FaceSwap, NTH,FSGAN,StyleGAN	视频	2020.1

基于数据驱动的视觉深度伪造检测。该检测方法从 2018 年至今经历了技术的更迭，由最简单的 CNN 架构训练预处理后的伪造数据出发，随着深度学习和计算资源的普及和发展，逐渐在神经网络结构上加入了胶囊网络、注意力融合模块、双流网络等，丰富了模型，同时也扩大了模型的规模，使得检测方法的准确率进一步提升，受众面也更为广泛。然而，基于数据驱动的检测方法也具有一定的缺陷：

1. 模型的复杂会导致训练的时间长、耗费资源多，在实操方面没有基于伪影的方法那么便捷；
2. 由于这类方法着重于研究神经网络的结构，因此可以使用对抗性的机器学习来添加扰动以避免检测[20]。

基于数据驱动的检测方法目前仍然处于高速发展中，由于其准确率高以及广泛的可移植性，在视觉深度伪造检测方法上被广泛看好，因此具有广阔的发展空间。然而如何有效地提高模型训练的效率和如何解决对抗机器学习的攻击将成为日后该类检测方法一个重要的方向。

基于信息不一致的视觉深度伪造检测。此类方法着重于寻找伪造制品与真实图像或视频之间的信息级差异。目前主流分为 3 个方向：生物信号的不一致性[29,30]、时间序列的不一致性、与真人行为的不一致性。与基于具体伪影的视觉深度伪造检测

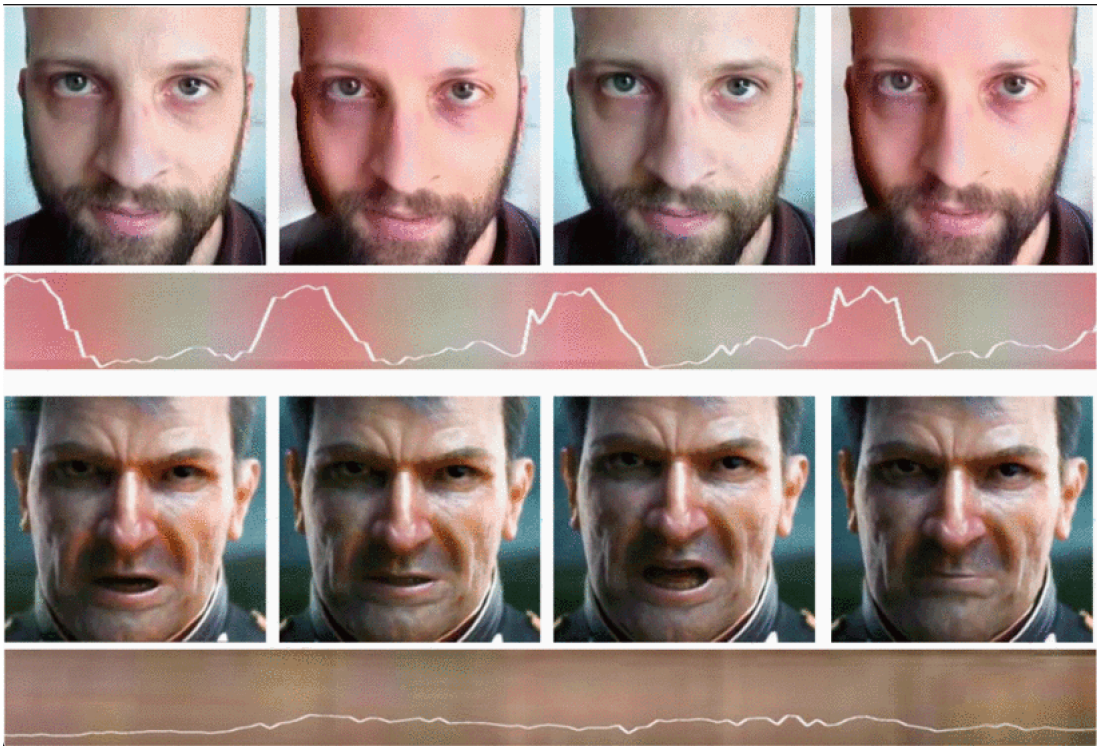


图 2.3 Conotter[29] 提出的由脉搏引起的可视变化

相似，一方在知道伪造制品的不一致性表现后，可以通过增加限制约束条件、优化模型或者加大模型训练数据集的方式减弱或消除此类表现。比如说针对伪造制品眨眼频率与真实人像不同的检测方法，在新的 GAN 被提出后，该不一致已经不在新的视觉深度伪造数据集中出现。因此，如何找寻到一种各种篡改模型共性且难以消除的不一致性特征，是此类研究方向未来需要研究的关键。因此许多一些研究工作聚焦于生物信号，比如图2.3，Conotter 提出了脉搏的不一致性检测方法；而图2.4，Ciftci 探究了 Deepfake 伪造视频中心跳不一致的信号。

其他类型视觉深度伪造检测。本文将不属于上述 3 种方法的检测方法归于此类，包含了许多思想各异的方案。比如运用统计思想，研究人员从海量的数据集中抽取出特定的统计特征，在无需大量训练计算量的情况下，使用该统计特征进行分类，最后的实验效果也能区分一定数量的真假图像与视频。当然，此类方法也在很大程度上受限于某种特定的数据集或篡改模型，缺乏较强的泛化能力是本方法的最大问题。

以上大致是目前 Deepfake 检测的几个方向，本文着重于探究数据驱动的视觉深度伪造检测，同时也借助于一定的特定伪迹半手工提取方式辅助模型的训练和泛化。因此，后文主要介绍这方面相关的研究工作。

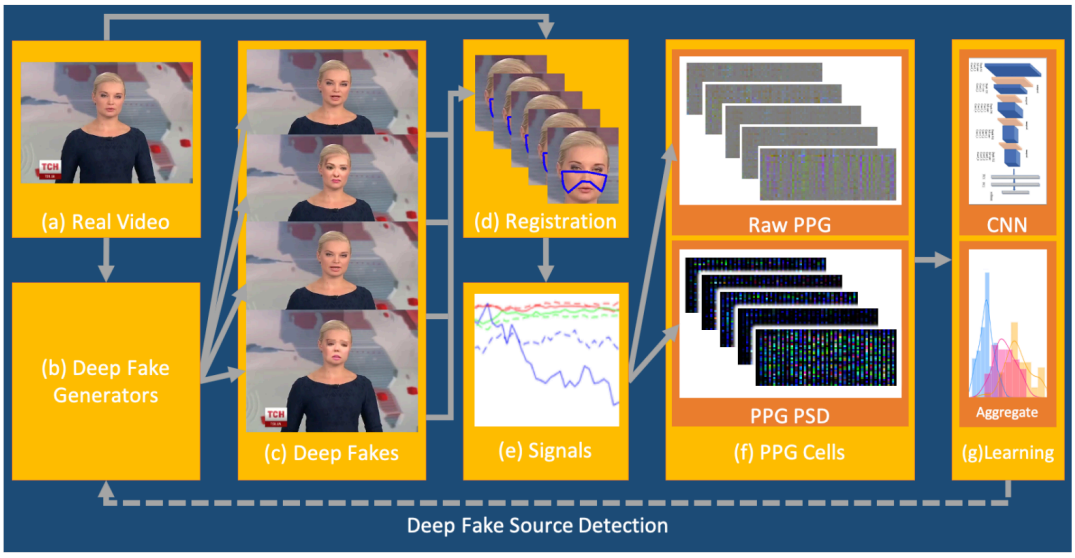


图 2.4 Ciftci[30] 提出的 PGG 检测架构：从真实视频 (a) 中，几个生成器 (b) 创建具有特定于每个模型 (c) 的残差的深度伪造。Ciftci 提出的系统提取面部 ROI (d) 和生物信号 (e)，以创建 PPG 单元 (f)，其中残差反映在空间和频域中。然后它通过对 PPG 单元的训练和聚合窗口预测 (g) 对任何视频 (c) 的真实性和来源进行分类。

§2.3 Deepfake 检测的关键技术

§2.3.1 Deepfake 检测中的视觉骨干网络

本工作的骨干网络（Backbone）主要采用基于卷积神经网络结构。

自 LeNet-5[3] 以来，卷积神经网络的使用愈发广泛，配合以一系列模块如批标准化、激活函数、池化操作等，在各种图像处理领域解决了很多问题。卷积在图像问题上有其天然的优势，他本身带有的归纳偏执，包括平移不变性和局部性，非常符合自然的经验。具体而言，局部性表示任何一个像素点总和周围的像素点具有最紧密的联系；而平移不变性可表示为 $f(g(x) = g(f(x))$，举例而言就是一个物体不管在图片的任何物体，他所计算得出的特征值都是一样的。

正是因为卷积神经网络具有以上这些优势，早期一些研究人员直接将卷积与一些池化等操作进行结合，仿造 LeNet-Style 的设计，对 Deepfake 进行检测，见图2.5。此类未经精心设计的网络结构在 GAN 生成伪造图片的初期，由于生成算法依然处于萌芽阶段，能较好的检测出伪造图像，在之后随着时间的发展，显然其检测能力是不足的。

另外一个值得注意的点在于，早期的卷积神经网络层数通常较浅，这是由于容易产生梯度消失的问题，所以限制了网络的层数。自 2016 年，He 提出残差连接后[31]，这一问题得到了显著的解决，越来越多深层的网络结构出现了，同时网络的表达能力也随着层数和参数的提升得到了显著的加强，卷积神经网络可以处理更加复杂的

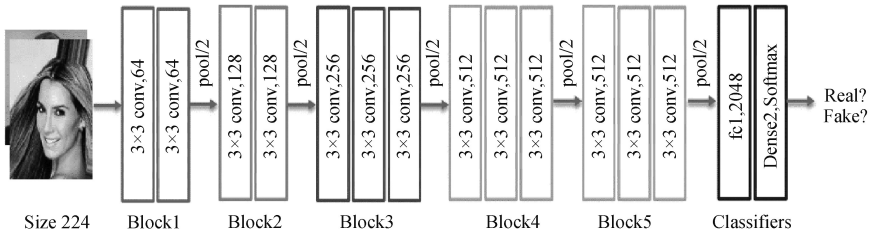


图 2.5 基于标准架构的简单卷积神经网络[3]。

问题了。具体而言，He 采用一种残差连接改进了梯度消失以及大型网络难以训练的问题。网络的前项与后项之前通过一条直连和正常路径，最大限度保留了梯度，其中最差的情况就退化为缺省这一模块，公式可表示为：$y = F(x, W_i) + W_s x$。F 表示任何一卷积模块，$W_s x$ 表示为输入的参数值。

ResNet[31] 的出现同时也标志着卷积神经网络向着模块化的进程，越来越多的网络通过精心设计一个模块，以此作为基础，然后通过不断的叠加该模块，组成了深层次的网络结构，主要的代表有 XceptionNet[32]，DenseNet，ResNext，EfficientNet 等。而与此同时，在 Deepfake 的检测上，研究人员发现运用 XceptionNet 的网络结构相对于其他结构具有一定程度的优越性，这可能是 XceptionNet 对于人脸信息的提取能力较好，因此绝大多数目前的 Deepfake 检测论文[33–36] 所用的 Backbone 均是基于 XceptionBlock 的原因。

§2.3.2 频域分析

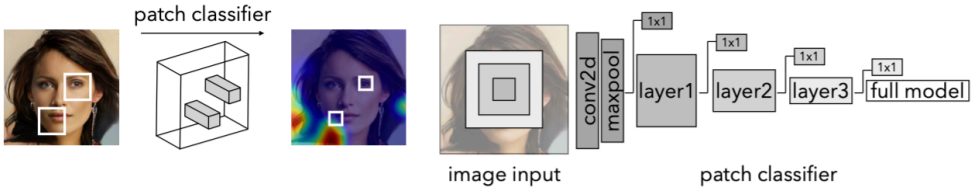


图 2.6 Patch-based CNN[36]

在运用频域进行图像分析之前，就有一系列工作关注于为什么数据驱动的深度学习建模方法能够判别检测深度伪造的图像。L Chai[36] 认为假脸检测和普通的人脸识别不一样，人脸识别需要很多内容上的信息，但是假脸检测更多需要细节信息。他们通过运用一种基于局部卷积的网络（Patch-based CNN）做了相对应的实验，简要而言是把完整网络中的某一层的像素接一个 1×1 卷积进行预测，最后的预测结果是所有 patch 的平均值，具体可参考图2.6。通过最后的实验可视化表明，见图2.7，对于 GAN 生成的图像来说，由于没见过人脸，只能从头发或者背景的细微处判断，而热力图（Heat-map）也主要集中在头发等细微处。由此可见，Deepfake 的检测模型会

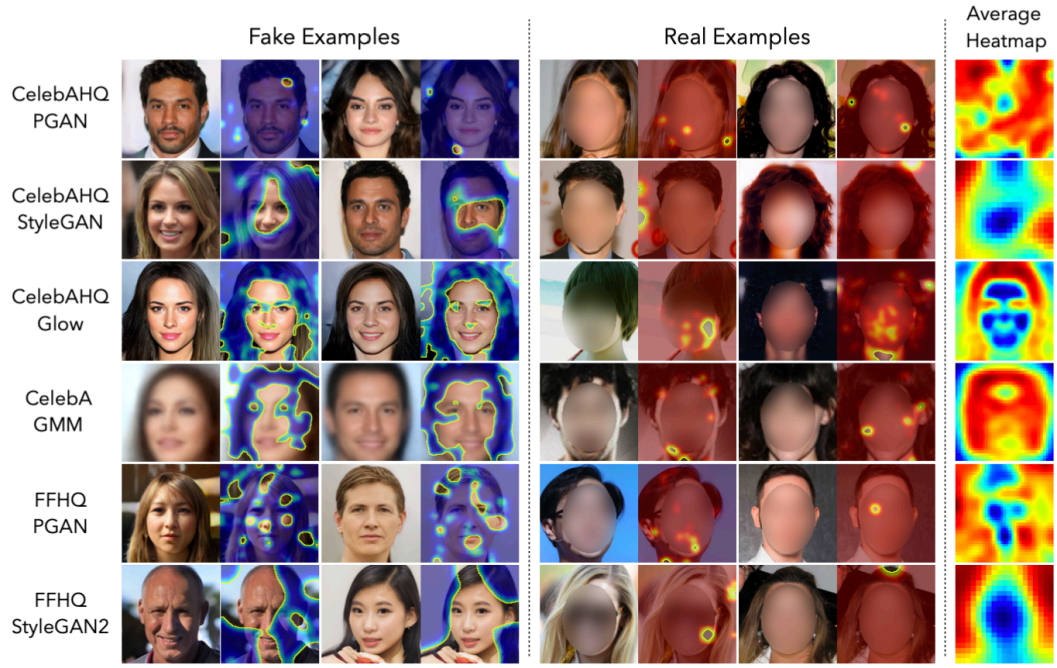


图 2.7 热图基于对来自每个数据集和假图像生成器的真实和虚假示例的 patch 预测，来源 L Chai[36]。对 0 和 1 之间的所有热图进行归一化，并以蓝色显示假值，以红色显示真实值。最右边一列表示 100 个最简单和虚假示例的平均热图，其中红色最能表示正确的类别。

主要集中关注于一些细微处。

但随着 GAN 模型的完善，以及一些网络结构会特意增加一个模块[17] 来修正一些细微处，如头发、耳朵、牙齿等，完全基于 CNN 的检测结构很难具有强的泛化能力和鉴别能力。因此，一些研究工作从 GAN 生成图像方式的角度进行了思考。R Durall[37] 提出要重点关注 GAN 生成阶段的上采样操作，如膨胀卷积和反卷积等，这造成了模型生成的图像在光谱分布上的不稳定，而这一现象的发生是独立于模型的架构层面的，完全由于其上采样方式的必然性决定的。如图2.8所示，通过下采样后再上采样的图像，在光谱分布上和原始图像具有很大的区别，在尾端有很明显的上扬。这也从实验角度证明了：如果通过某些变换，将原始图像进行一个映射或者变换，可以有效的对 Deepfake 图像进行检测。

而频域分析的使用也是从以上的研究历程中发展而来的，最终的实验结果也证明了 GAN 生成图像方式的确存在的这一问题。具体而言，J Frank[27] 提出了将 RGB 原始图像进行离散傅立叶变换（Discrete Fourier Transform, DFT）以及离散余弦变换（Discrete Cosine Transform, DCT），将时域信息通过映射，变换到频域空间之上。这类变换方式以往多见于图像分类[38,39]，纹理分类[40] 和超分辨率重建上[41]，这导致了很多学者忽视了 Deepfake 检测上频域分析的重要性，也可能由于之前鲜少有研究深度学习为什么能检测 Deepfake 的工作，缺乏这类工作的指导才导致直到最近慢慢开始重视起频域分析上的研究。

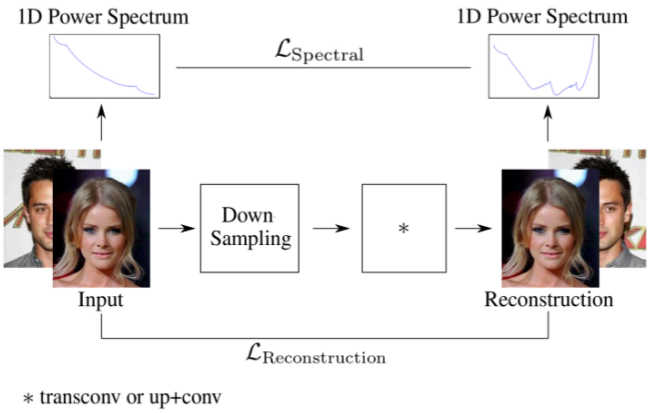


图 2.8 R Durall[37] 采用简单的 Encoder 和 Decoder 证实了光谱分布上上采样操作遗留的伪迹。

J Frank[27] 在文中认为也认可了[37] 的发现结果，并对此进行了一定程度的扩展和深入研究。具体实验可见图2.9，右半部分为通过 StyleGAN[42] 生成的图片中能观测到明显的伪影。然而上采样的伪影在时域上很难观测到，但一旦经过频域变换后就可以清晰的观测到，在图像中可以看到明显的耀斑，这表明了 GAN 从低维隐空间映射到高维空间过程中存在结构性的问题。作者之后也做了很多频谱实验，验证了这一结论的普适性，见图2.10，均可以观测到明显的栅格化问题（Grid-like pattern）。此项工作可以说是一个开创性的进展，后续很多工作也基于了频域分析的模式。

图 2.9 左半边为真实自然照片和它的 10000 次的频域平均热力图，右半边为 StyleGAN 生成的伪造图和它的 10000 次的频域平均热力图，来源于 J Frank[27]。

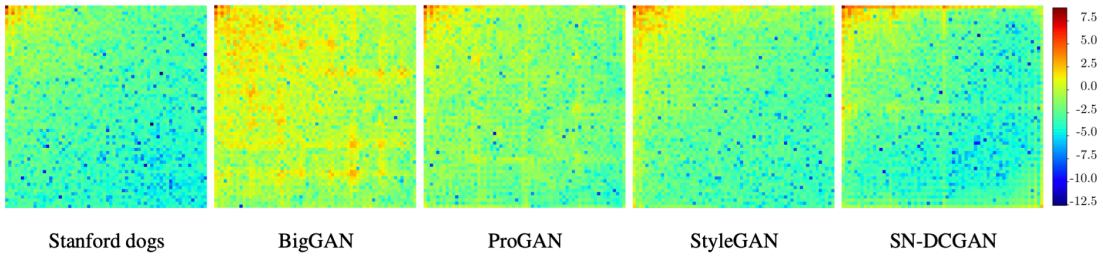


图 2.10 转换后的频域光谱图[27]。

之后的研究工作就不只局限于直接的变换，他们寻求使用一些过滤器，比如说一些研究者应用了高通滤波[43,44] 以及 Gabor[45] 滤波来提取感兴趣的元素。Phase Aware CNN[45] 使用手工制作的 Gabor 和高通滤波器以增强边缘和纹理特征。通用检测器[43] 发现经过高通滤波后的图像在光谱中可以获得显着差异。但是，这些中使用的过滤器研究通常是固定的和手工制作的，因此无法自适应地捕捉到伪造的图案。在本工作中，我利用了频域通道上的自适应图像分解以挖掘频率伪造线索。

§2.3.3 双流网络

双流网络的应用最早是出现在视频理解领域，是由 Simonyan 提出的基于原始视频任意一帧图像和光流信息作为双流的网络[46]。当时 Simonyan 的考量是如果单流网络只能应用一个模态的特征的话，给网络增加一支分流进行并行就能实现两个模态的融合。而在 Deepfake 检测上的使用是在 2018 年 P Zhou 提出的双流网络[28]。当时的朴素想法就是结合不同方面的特征信息，如图2.11所示，一支为原始图像的脸部特征表示；一支为局部的三元组流用来提取隐藏信息（Steganalysis features），用来确保两个部分是来源于一个隐空间。最后采用后融合的方式得到检测结果。

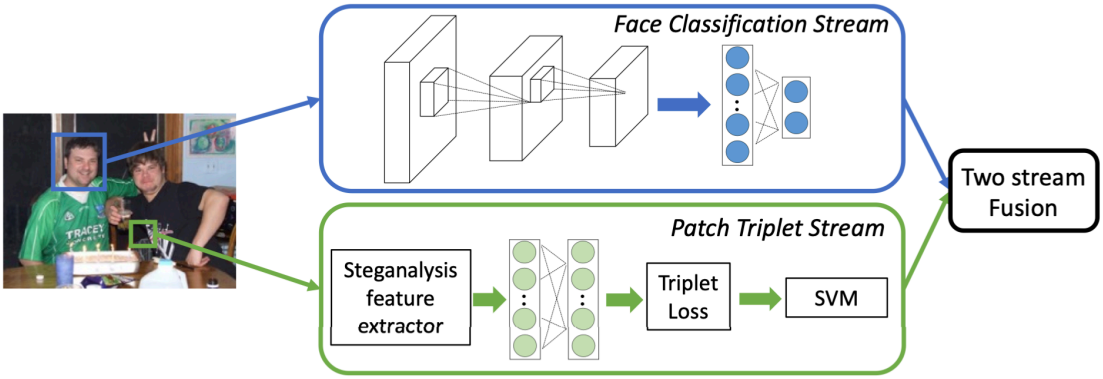


图 2.11 P Zhou 提出的双流网络[28]。面部分类流通过分类面部是否被篡改来模拟视觉外观。局部三元组流在隐写分析特征上进行训练，以确保来自同一图像的局部特征在隐空间中接近，并且基于学习特征训练的 SVM 对每个局部进行分类。最后，融合两个流的分数以识别被篡改的人脸。

后续的研究主要在于两个方面，一个是如何分配两条流所提取的信息来源，另一方面是如何较好的融合两条流的信息。针对前者，相关的研究工作包括利用人脸 24 个点位（Landmark）[34]，这类工作主要是利用人脸点位的区域性特性配合以原始图像，具体可见图2.12，Songsri-in K 的考量主要来源于人脸 24 点位能使得网络做出一个基于先验知识的定位，针对一些局部五官模糊的伪造图像具有优越性，它能显式地通过人脸点位约束网络的学习和泛化能力。除此以外，一些工作也将手工提取

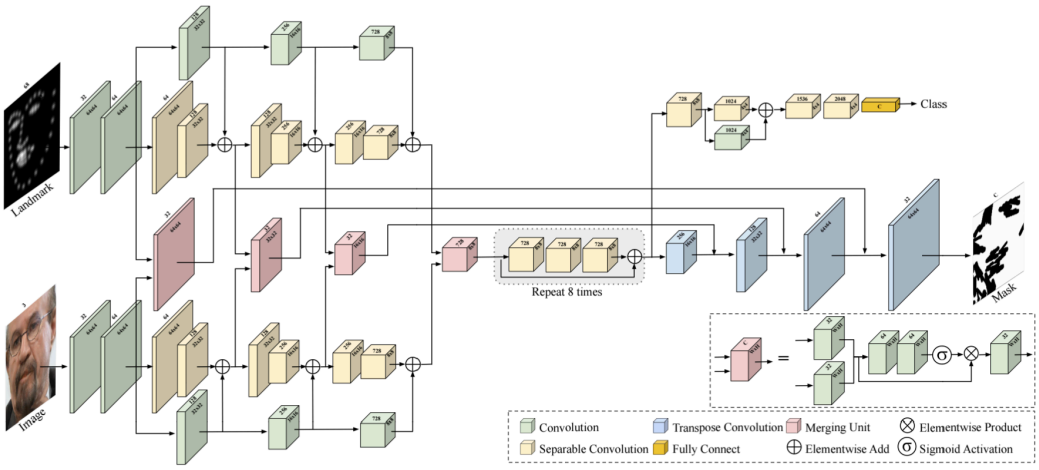


图 2.12　Songsri-in K[34] 提出的基于人脸 24 个点位和原始图像的双流网络结构

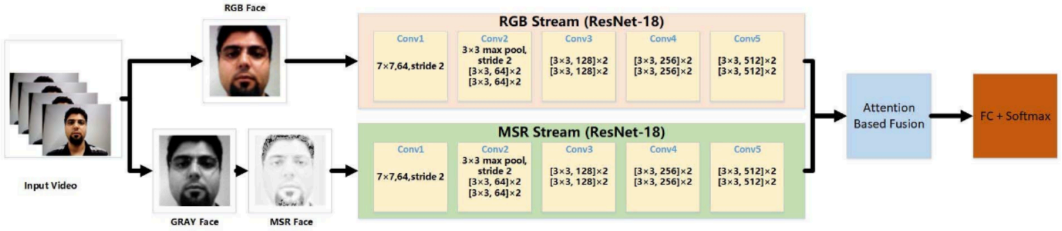


图 2.13　Y Qian[33] 提出的基于注意力机制的双流面部检测网络结构：包含 RGB Stream 和 MSR Stream。

特征的模式嵌入到了双流网络中。Y Qian[33] 通过将频域变换以及滤波器的使用，融合了频域信息和原始图像信息，取得了很好的效果。H Chen[20] 通过应用光照不变性，将原始图像转换为灰度图然后运用 MSR 流来提取特征，具体可见图2.13。

而针对后者如何融合特性信息的问题，方法主要包括直接分类取平均、双流特征聚合、特征平均或池化等。随着注意力机制的流行，一些研究者也运用了注意力机制。如图2.13，其核心在于生成注意力矩阵，生成的方式各种各样，H Chen[20] 主要是通过复制两支分流的特征信息，然后做一个映射得到注意力矩阵，最后将两支分流信息分别乘以注意力矩阵。除了这种方式，还包括可以先运用 1×1 的卷积进行融合，然后采用自注意力[47] 的机制进行融合。而本文所用的方式区别于以上所言，采用了一种嫁接模块的思想[48]，在更好的融合特征的同时，也充分考虑了算法的性能以及可解释性。

§2.3.4　注意力机制

注意力机制由来已久，而生成注意力矩阵的方式也是多种多样的，其核心在于生成一个概率矩阵，帮助网络学习应当着重注意的部分。由于本文所采用的为一种自注意力机制[47]（Self-attention）的变体，因此后文着重介绍一下自注意力机制的原

理。

自注意力机制中比较关键的几个元素分别是 Query（Q）、Key（K）和 Value（V），他们分别代表了查询、索引和值。那么通俗的举个例子，自注意力机制所做的事情就好比去搜索引擎查询内容，需要得到最想关注的、相关性最大的内容。具体而言，针对一个问题 Q，然后去搜索引擎里面搜，搜索引擎里面有许多文章，每个文章 V 有一个能代表其正文内容的标题 K，然后搜索引擎用问题 Q 和那些文章 V 的标题 K 进行匹配，计算相关度。$Q \times K^T$ 夹角小，正相关；$Q \times K^T$ 夹角大，负相关。然后利用这些检索到的不同相关度的文章 V 来表示该搜索的问题，于是使用这些相关度将检索的文章 V 进行加权和，就得到了一个新的 $Q^{'}$，此时的 $Q^{'}$ 融合了相关性强的文章 V 更多信息，而融合了相关性弱的文章 V 较少的信息。这就是自注意力机制，注意力度不同，重点关注（权值大）与想要的东西相关性强的部分，稍微关注（权值小）相关性弱的部分。

以上描述的是针对单一元素的情况，具体到实际情况，有许多的 Q，因此考虑使用矩阵加速计算，使其能够并行计算。而自注意力中的“自”体现在以上的 Q, K, V 均是用同一个矩阵映射而来的，比如通过一个前向网络映射得到，参考公式(2.1)。

$$\begin{aligned} Q &= W_q \times x_i \\ K &= W_k \times x_i \\ V &= W_v \times x_i \end{aligned} \tag{2.1}$$

通过使用三个可学习的参数矩阵 W_q, W_k, W_v 对于输入的矩阵进行映射，可以采用全连接完成该操作。之后，采用之前上文描述的搜索引擎描述的操作，对特征进行融合提取，公式如下：

$$Attention(Q, K, V) = softmax(\frac{Q \times K^T}{\sqrt{d_k}}) \times V \tag{2.2}$$

在此基础上，一般的注意力都采用多头注意力（Multi-head attention）。具体而言，就是将一个输入 token，通过多个可学习的映射参数 W_q, W_k, W_v，生成多个 Q, K, V，然后在进行注意力操作，最后将其 Concate 作为多头注意力的输出结果，公式表示如下：

$$\begin{aligned} Multihead(Q, K, V) &= Concate(head_1, \ldots, head_h)W^O \\ where \quad head_i &= Attention(QW_i^Q, KW_i^K, VW_i^V) \end{aligned} \tag{2.3}$$

实验表明，通过采用多头注意力可以有效地让网络学习针对一个输入 token 的不同侧面和不同的特征提取方式，对于最后的模型精度具有很好的提升能力。

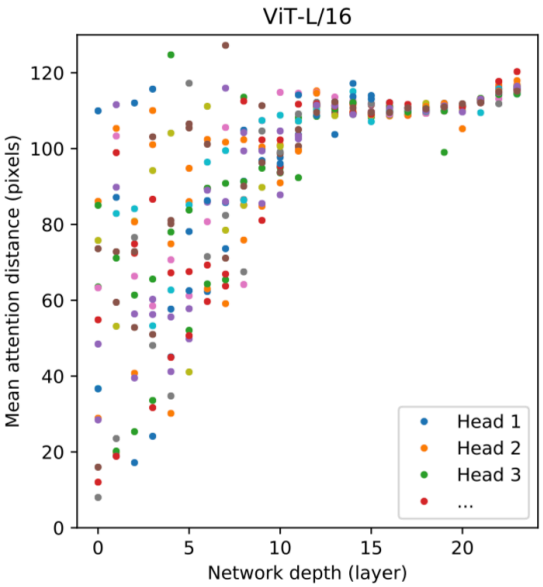
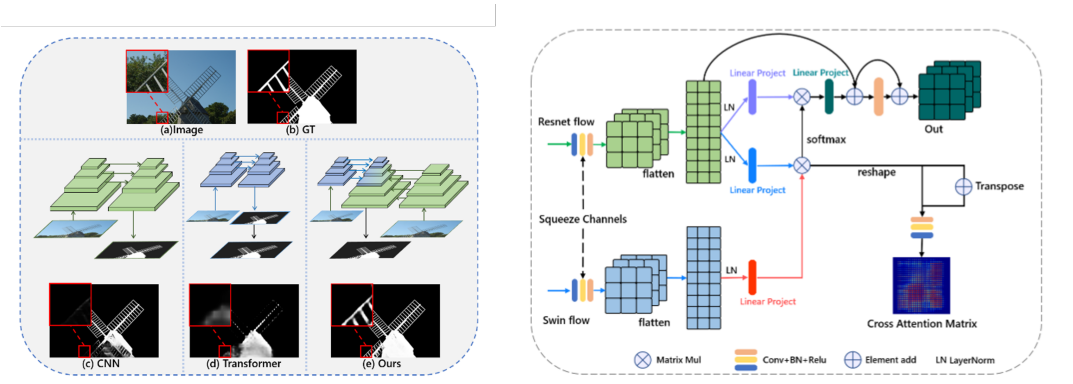


图 2.14 自注意力关注区域图示，来源于 ViT 中的实验[49]。

该论文的另一项实验则表明，通过采用注意力机制能使网络在初始阶段就更为注重全局，而非像卷积在初始阶段只能关注于局部周围的感受野，具体可见图2.14，从图中可以明显发现在浅层网络中，注意力模块也能很好的关注到很远的部分。然而，虽然理论上通过使用自注意力模块能在初始阶段就达到很长距离的感受野，但由于其完全抛弃了卷积所有的两点归纳偏执，因此需要极其大规模的训练数据集的支持，才能让运用自注意力的网络逐步学习到这一归纳偏置[49]，这可以算是一个“天下没有免费的午餐”（No Free Lunch Theorem）的常见场景。

因此一些研究工作就在思考如何在有限的数据量上更好的使用自注意力机制，更为具体而言就是如何将归纳偏执显式地添加到自注意力模块中。其中的一些工作就是将 CNN 与 Self-attention 进行有机的融合。比如说 F Yi[50] 的做法是：在 token 映射前先做一个卷积的操作，以此来获得一些的局部性归纳偏执。另外有些工作将卷积神经网络与 Transformer 网络进行结合，在中间的融合模块中，将 Q, K, V 的来源进行拆分[51]，将跨模型嫁接思想引入，如图2.15所示，将 Q, K 的来源设定于 CNN，而 V 设定为 Transformer，以此操作同时继承了 CNN 的局部特性以及 Attention 的全局特性。这种思想被命名为 Graft 模块，在显著性检测（Saliency Detection）中取得了良好的效果。但据作者所知，在 Deepfake 检测中还没有类似的操作，而本文所提出的基于嫁接注意力的模块在此基础上更近一步，运用多流网络的特性分别得到 Q, K, V，同时具有不像前文[51] 那么缺乏解释性（该文的融合理念可以理解，但是方式还是比较迷惑的）。



(a) (b)

图 2.15 C Xie[51] 所提出的 Cross-Model Grafting Module：(a) 单独的 CNN 和 Transformer 对于显著性检测存在的问题，以及该论文提出的方法的对照图；(b) 为 Graft Module 的算法细节展示。

§2.4 本章小节

本章节主要描述了 Deepfake 检测的理论基础，明确地定义了 Deepfake 检测问题，然后从四个研究方向简介了目前主流的研究方向，阐明本文的主要研究方向为基于数据驱动加之特定伪影提取的检测方法。之后针对本文在第三、四章所采用的方法介绍了相关的研究背景和之前的研究进展成果，分别从视觉骨干网络（Backbone）、频域分析、双流网络和注意力机制方面展开概述。

第 3 章　基于频域分析和双流网络的 Deepfake 检测

本章为本论文的重点章节，主要介绍了基于频域分析和双流网络的 Deepfake 检测的整体框架以及实现细节，分别从频域特征表示和融合、多任务学习模式等方面介绍了本检测算法的各个模块；之后介绍了本工作所用的数据集、本文实验所用的评价指标和训练参数设置。最后给出了基于此方法的消融实验并分析了实验结果。

§3.1　基于频域分析和双流网络的网络架构

§3.1.1　整体框架

对于网络架构的总览图可见图3.1。本文的网络输入格式是标准的 (N, C, H, W) 格式的，如果要开启多任务学习模式，还需要生成对应的 mask 图（在第4.1.3小节中，介绍了如何使用半监督信号在大规模缺乏 mask 的数据集中进行训练，当然也可以运用到本网络中）。

对于输入图片，首先需要通过频域特征表示和提取模块，这是一个双流网络的架构，在此过程中，两条支流的信息并不会进行融合和分享。之后对于提取到的特征信息，会进入 XceptionBlock 模块，由于两条支流得到的特征向量通道数 C 并不统一，因此在这一步会使其得以统一。在此过程中，需要重复 4 次 XceptionBlock，在

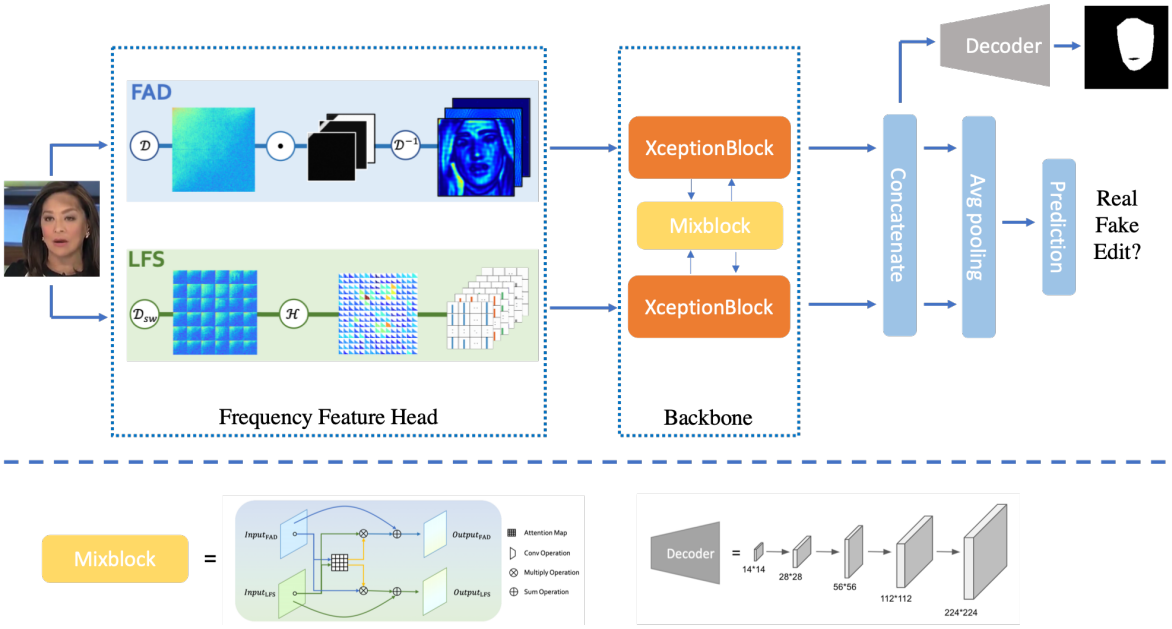


图 3.1　基于频域分析和双流网络的网络整体框架。对于输入图像，运用 FAD 和 LFS 结构对其分别提取特征，在骨干网络处使用融合模块（MixBlock）融合特征，最后使用多任务模式联合优化模型参数。

第二和第四次的时候加入了融合模块（MixBlock）。至此为骨干网络。

在最后的输出头上，本文使用了多任务学习的模式，在骨干网络的输出向量上进行共享，一只为分类头，一只为分割定位头。值得注意的是，分类头和分割头的权值是共享的，运用联合损失同步优化整体网络参数。

§3.1.2 骨干网络

本文延用以往的工作成果，继续使用 XceptionNet 作为骨干网络。XceptionNet 是基于 InceptionV3 所改进得来的，最主要的特点在于极限化了 Inception 里面的模块，如图3.2所示，图中 (a) 表示普通的 Inception 模块，对于输出通道数为 k_{out}，他首先运用 1×1 的卷积得到 k_{out} 组特征向量，然后将其平分为 3 组，之后对其进行 3×3 的卷积操作，最后进行 Concat。相较于普通的没有进行分组的卷积操作，如果输出特征图大小和通道保持一致，使用 Inception 模块可以减少约 3 倍的参数量。然而，在 XceptionNet 中，作者对其做了更大胆的假设，将空间与通道信息充分解耦，即分成两步：第一步做深度方面的卷积，考虑不同通道上的信息，第二步做局部像素点方面的卷积，考虑空间上的信息。具体而言，如果输入通道为 k_{in}，则用 1×1 的卷积核对输入特征图进行卷积，在这一过程中学习通道间的相关性；在第二步进行 3×3 的卷积核，而此时的卷积是只作用在一个通道上的，主要学习局部空间上的相关性。通过这样的两步操作，大大降低了参数量，使得推理训练的速度大大加快。

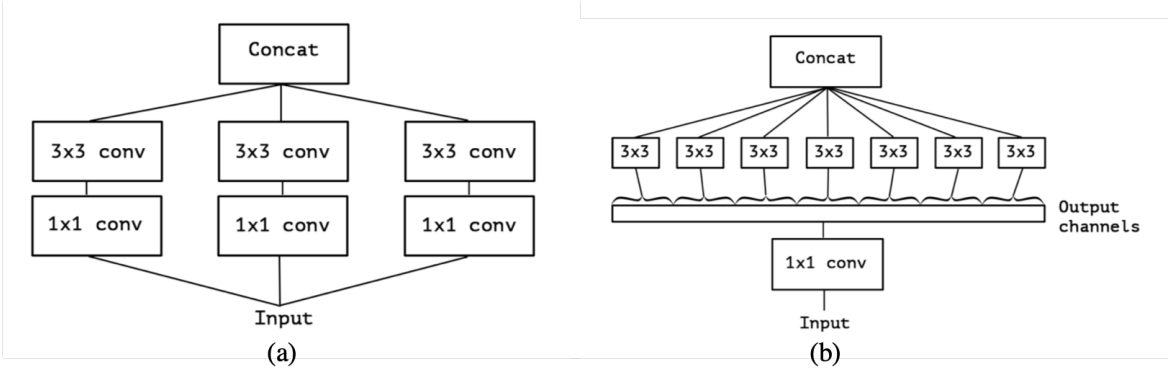


图 3.2 深度可分离卷积：(a)Inception 模块；(b)Xception 模块

§3.1.3 频域特征表示和提取

本小节主要提出了两种提取频率特征的方式，一个是频率图像分解（Frequency-aware Image Decomposition，FAD），另一个是局部频域特征统计（Local Frequency Statistics，LFS）。

§3.1.3.1 频率图像分解

对于频率图像分解，以前的研究通常在空间域应用手工制作的滤波器组[45]，因此未能覆盖整个频率域。同时，固定的过滤配置使得难以自适应地捕获伪造图案。为此，本文提出了一种新的频率感知分解 (FAD)，根据一组可学习的频率滤波器在频域中自适应地划分输入图像。分解的频率分量通过逆变换映射回到空间域，产生一系列频率图像分量。这些组件沿通道轴（Channel，C）堆叠，然后输入到卷积神经网络，以全面挖掘伪造模式。

具体而言，如图3.3所示：首先将原始图像通过 DCT（Discrete Cosine Transform）映射到频域空间，然后使用一组可学习的频率滤波器自适应地划分输入图像，通常会运用一个掩码器划分高、中、低三个频带，之后再将分解的频率分量通过 IDCT（Inversed Discrete Cosine Transform），可以用公式(3.1)阐述这一过程:

$$\mathrm{y}_i = \mathcal{D}^{-1}\left\{\mathcal{D}(\mathrm{x}) \odot \left[\mathrm{f}^i_{\text{base}} + \sigma\left(\mathrm{f}^i_w\right)\right]\right\}, \quad i = \{1, \ldots, N\} \tag{3.1}$$

1. 设计 N 个二分类滤波器（也就是所谓的掩码 mask）$\{f^i_{\text{base}}\}^N_{i=1}$，将图像的频率分为低，中，高三个频带。
2. 为了让其具备自适应能力，本文额外设计了三个可学习的滤波器 $\{f^i_{\text{w}}\}^N_{i=1}$。然后分别将这两种滤波器结合在一起:

$$f^i_{base} + \sigma\left(f^i_w\right) \tag{3.2}$$

其中，

$$\sigma = \frac{1 - e^{-x}}{1 + e^{-x}} \tag{3.3}$$

3. 对 $f^i_{base} + \sigma\left(f^i_w\right)$ 做 2D DCT 变换，之后再做 IDCT，即 x^{-1}。输出的通道数应为 12。

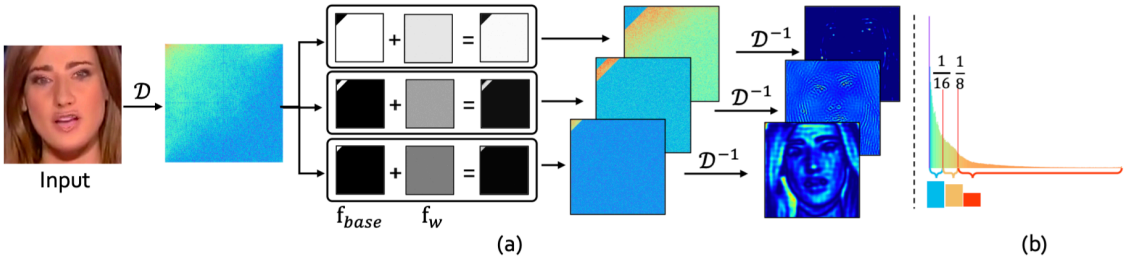


图 3.3 FAD 流程图: 首先将原始图像通过 DCT 映射到频域空间，然后使用一组可学习的频率滤波器自适应地划分输入图像，之后再将分解的频率分量通过 IDCT 映射，得到经过提取的特征元素。

§3.1.3.2 局部频域特征统计

FAD 尽管提取到了频域特征，但它最后是通过 IDCT 变换，转化到 RGB 空间上，再输入进 CNN。这些信息并不是直接的频域信息，事实上这只是一个滤波的过程，考虑到 CNN 在时域上直接提取伪造信息很困难（单纯的 CNN 很难提取长距离和大纹理信息，而 Deepfake 遗留的伪迹则多为长距离和大纹理信息[52]），因此需要通过一种方式能够直接的统计和评估频域上的信息特征，所以运用了局部频域特征统计（Local Frequency Statistics，LFS），它能满足 RGB 图片的平移不变性以及局部一致性，同时也能够输入卷积神经网络中以探索学习高层级的伪造模式特征。

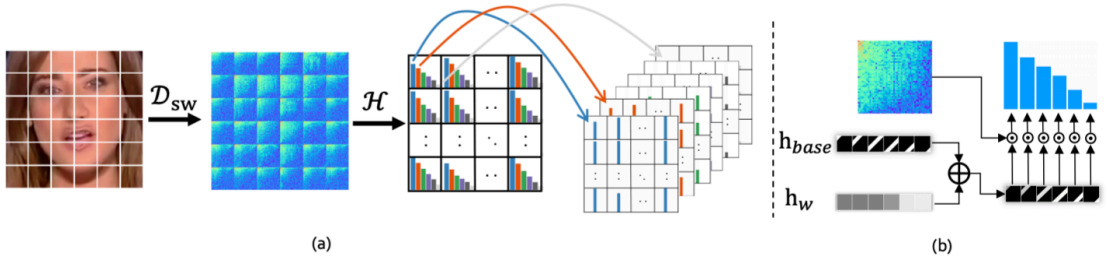


图 3.4 LFS 流程图：首先对于原始图像采用 SWDCT，提取基于局部的频域信息；随后通过可学习的频带计算频带响应均值；之后重新聚合成一个多通道的空间特征向量。

如图3.4所示，首先对于原始图像采用基于窗口的 DCT 变换（Sliding Window DCT, SWDCT），以提取基于局部的频域信息；随后通过一系列可学习的频带计算频带响应均值；然后这些频带响应重新聚合成一个多通道的空间特征向量，同时会进行一个映射操作以保证和输入图像具有相同的布局。至此，就完成了局部频域信息的统计。值得注意的是，本文所采用的频带是一系列可学习的自适应矩阵，因此能较好的提取局部频域信息。

$$\mathrm{q}_i = \log_{10} \left\| \mathcal{D}(\mathrm{p}) \odot \left[\mathrm{h}^i_{\mathrm{base}} + \sigma\left(\mathrm{h}^i_w\right) \right] \right\|_1, \quad i = \{1, \ldots, M\} \tag{3.4}$$

具体而言，如式3.4所示，大致的思路和表述形式和 FAD 中类似，本文设计了二分类滤波器 $\{h^i_{\mathrm{base}}\}_{i=1}^N$ 和可学习滤波器 $\{h^i_{\mathrm{w}}\}_{i=1}^N$；然后将二者结合起来，加入一个 σ 参数，得到：

$$h^i_{base} + \sigma\left(h^i_w\right) \tag{3.5}$$

其中，

$$\sigma = \frac{1 - e^{-x}}{1 + e^{-x}} \tag{3.6}$$

之后再点乘 D_p，代表滑窗 DCT 变换；最后的 $\log_{10}$ 是为了调整数值级别。经过上述的变换被转换为通道数为 6 的特征向量。

§3.1.4 双流网络与特征融合

正如前两小节所描述的，FAD 和 LFS 模块都是从原始图像中提取特征，都是运用了频域变换和自适应滤波器的方式提取有效特征，虽然二者在其中一些细节中有不同，一个是更关注了局部频域统计信息，因此添加了小窗口，一个则是运用 DCT 和 IDCT 恢复成了滤波后的时域图像，但二者是存在内在紧密联系的。因此我认为可以通过合适的方式融合两种不同的频域信息挖掘模块。这也就是图3.1展示的融合模块（MixBlock）。本文运用此模块完成了双流网络中 FAD 和 LFS 提取特征的融合。具体到实施细节，如图3.5在 Xception 为 FAD 和 LFS 提取完的特征进行进一步学习的过程中，添加了两个 MixBlock 模块，以此完成双流网络的特征融合。

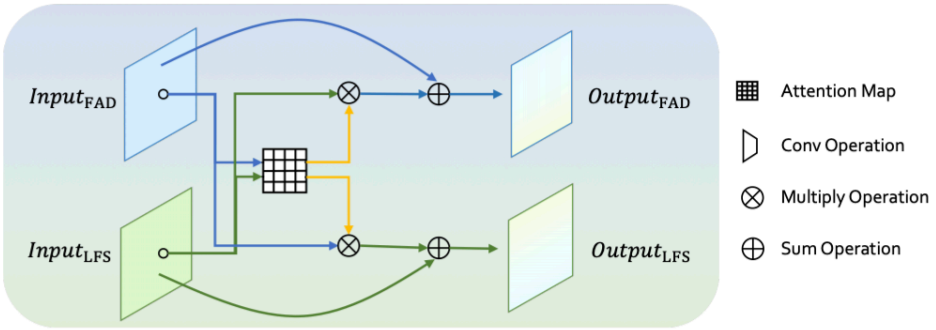


图 3.5 提出的融合模块（MixBlock）示意图。⊗ 表示矩阵点乘，⊕ 表示矩阵逐元素相加。

虽然称其为注意力矩阵，但这与大家广泛了解的自注意力机制2.1不同，这里的注意力生成是基于卷积的，而不是“自己”的。具体而言，首先聚合两条支流的特征矩阵，然后进行对其做 1×1 的卷积操作，融合二者信息成为一个注意力矩阵，然后让 FAD 和 LFS 两条支流分别与各自另外一条的支流的特征矩阵乘上注意力矩阵，最后再和原先的特征矩阵相加，完成特征融合。用数学公式表示为：

1. FAD 和 LFS 共同输入进卷积，得到一个 $Attention_{Map}$

$$Attention_{Map} = W \times (Concat(Input_{FAD}, Input_{LFS})) \tag{3.7}$$

2. LAD 和 LFS 分别与 $Attention_{Map}$ 相乘得到 $L_{Attention}$ 和 $F_{Attention}$

$$\begin{aligned} F_{Attention} &= Input_{LFS} \cdot Attention_{Map} \\ L_{Attention} &= Input_{FAD} \cdot Attention_{Map} \end{aligned} \tag{3.8}$$

3. 最后与 $Input_{FAD}$ 和 $Input_{LFS}$ 逐元素相加得到融合特征结果

$$\begin{aligned} Output_{FAD} &= F_{Attention} + Input_{FAD} \\ Output_{LFS} &= L_{Attention} + Input_{LFS} \end{aligned} \tag{3.9}$$

§3.1.5 多任务学习模式

多任务学习的模式近年来受到广泛关注，其核心在于将几个关系紧密的任务运用一个共同的骨干网络提取特征，再最后的输出头上配合以不同任务特定的输出网络结构进行输出。一般而言，在无人驾驶领域使用颇为常见，比如说 D Wu[53] 提出运用 YOLOv4 中的 CSPDarkNet 作为骨干网络进行提取共有特征，然后使用了三个不同的但又紧密联系的任务进行协同训练，见图3.6，分别是：车辆目标检测、车道线分割和汽车可行区域分割。文中任务虽然任务形式不同，但具有内在统一性和可解释性。具体而言，车道线一般都是包裹在汽车可行区域外沿的，而车辆一般都在汽车可行区域上开车。这三种任务的设置其实给了网络一种先验的归纳偏执，即：三种输出位置空间上的方位关系。这其实是非常有用的，卷积神经网络本身就擅长提取特征，再给他一些先验的可解释性的归纳偏执，网络能在较小的数据集上，运用较小的模型参数和计算资源就实现一个比较优秀的结果。

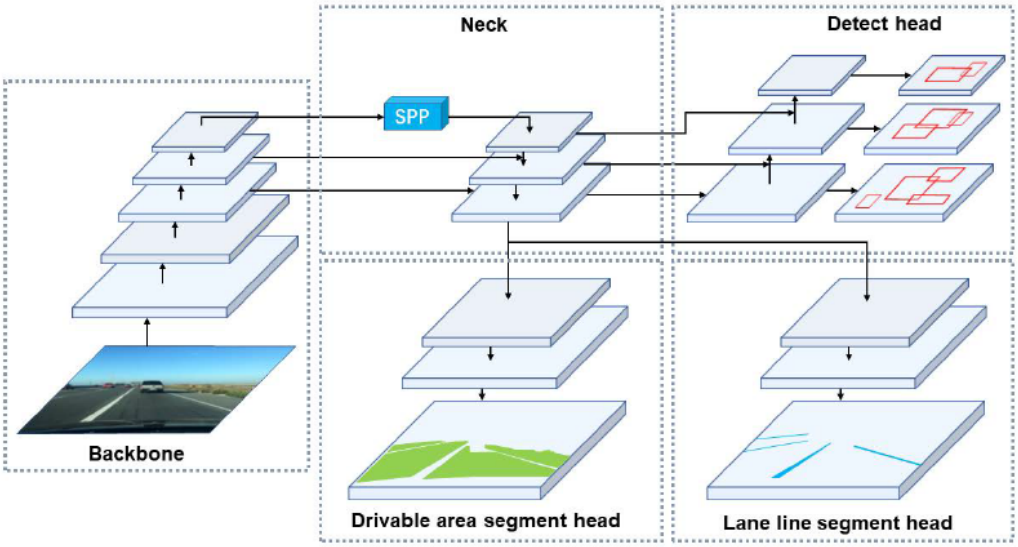


图 3.6 D Wu[53] 提出的 YOLOP：该网络共享一个编码器，同时运用三个解码器完成三个不同的任务。编码器由一个骨干网络和颈部网络（Neck）构成。

由此可见，多任务学习这一概念虽然不难理解，似乎每个大的任务总可以找到那么几个细分小任务进行多任务协同学习，但不可否认的是，任务的选择一定要合理且任务间要具备一定的内在联系，也就是说要具备一定的先验归纳偏执，这样子

才能真正达到多任务学习的优势。

回到 Deepfake 的检测上，本文寻找到了两个合理且关系紧密的任务进行协同训练，分别是 Deepfake 图像的分类检测任务和伪造区域的分割定位任务。这两者内在的关联是很明显的，事实上只要能定位到某张图像存在伪造区域，那么就可以认定该图像一定不是真实的图像，一定是通过某种方式伪造虚拟生成的或者是换脸编辑过的。这样一种强烈的先验假设给网络很强的归纳偏执，使其能够快速的训练拟合，并且最终的实验效果也是很不错的。值得注意的是，针对 Deepfake 图像，依然还需要区分是完全虚拟生成的还是换脸编辑的，此时本文设置的 mask 图也起到了正相关的作用。换脸编辑的 mask 图，编辑区域主要定位在脸部，可以是脸部的细节，也可以是整张脸；而相对的，完全虚拟的图像则是整张图像都是虚假的。通过使用分割定位任务和检测任务相结合，共同训练骨干网络的方式，能够加速训练过程，同时在增加不多的参数量的同时，提升模型的整体表现效果。

具体到网络结构实现的细节，参考图3.1可知，本文在输出头上的结构设置都比较简单，这是因为本文相信通过前面的双流网络与特征融合，已经足够网络提取图像的各方面的深层次的特征信息，因此在输出头上只需要做一些简单的操作即可。同时，由于多任务协同训练的方式，不需要在输出头部进行过于复杂的操作，关注的重点还是前面的双流网络和特征融合模块。因此，本文的检测分类头就进行了简单的全局平均池化和全连接层，而分割定位头则使用了反卷积进行上采样操作。

值得注意的一点是，有些时候 Deepfake 编辑过的伪造图很难获取其编辑区域的 mask 图，这是很多做 Deepfake 定位任务的难点，毕竟数据集是一切的来源。为了解决这个问题，采用了半监督的学习模式，具体在第4.1.3小节中，据本文作者有限学识所知，这是第一个将半监督学习与多任务学习模式在 Deepfake 检测上进行运用的工作。

§3.1.6 损失函数

针对损失函数的设置，也进行了针对多任务学习的优化，采用联合损失函数的方式，综合了两个不同任务的损失函数。

首先针对检测分类任务，本文采用基本的交叉熵损失函数（CELoss）。针对 $\{l_1, \ldots, l_N\}^\top$ 个类别，L_{CE} 的公式如下：

$$L_{CE} = -w_{y_n} \log \frac{\exp\left(x_{n,y_n}\right)}{\sum_{c=1}^{C} \exp\left(x_{n,c}\right)} \cdot 1 \tag{3.10}$$

而针对定位分割任务，由于是像素级别的分类，因此首先也使用了二分交叉熵

损失函数。L_{BCE} 的公式如下：

$$L_{BCE} = -w_n \left[y_n \cdot \log x_n + (1 - y_n) \cdot \log (1 - x_n) \right] \tag{3.11}$$

由于单独的二分交叉熵损失函数只度量每个像素的分类情况，可能会缺乏考虑分割定位任务的连续性，为了惩罚过于离散的定位情况，因此引入了交并比（IOU）的思想，加入了基于 Dice 系数的损失函数。L_{DICE} 的公式如下：

$$L_{DICE} = 1 - \frac{2 * \sum p_{\text{true}} * p_{\text{pred}}}{\sum p_{\text{true}}^2 + \sum p_{\text{pred}}^2 + \epsilon} \tag{3.12}$$

最后，将以上三个损失函数进行结合，所采用的方式是直接相加，当然也可以添加不同的系数以惩罚网络着重优化某一个方面。但这里通过实验认证，单纯的相加最后的效果也是不错的。定义总的损失函数 L_{Total} 公式如下：

$$L_{Total} = \alpha L_{CE} + \beta L_{BCE} + \gamma L_{DICE}, \quad where \quad \alpha, \beta, \gamma = 1 \tag{3.13}$$

§3.2 实验设计

§3.2.1 Deepfake 数据集

Deepfake 的数据集包含很多，主要分为视频和图像两大类，由于计算资源的限制以及现今的数据集规模都愈发庞大，因此选择了图像类作为研究，同时取了大规模数据集的一个子集作为本论文的 benchmark。虽然数据量规模较之主流的减少了一大部分（约为 5%），但依然是有借鉴意义和研究价值的，一方面是可以探究 Deepfake 检测在小规模数据集上的训练和泛化能力，另外一方面由于视频就是由每一帧图像组成的，能够很好的鉴别图像也说明对于视频也是能够检测的。

目前主流的数据集已经在表2.1中作了详尽的阐述。本论文的 Deepfake 数据集主要由三部分构成，分别是：真实人脸图像、GAN 生成的完全虚拟人脸图像和经过换脸变换的伪造图。

如表3.1所示，真实人脸数据主要来源于 FFHQ[54]，这是由 NVLab 搜集的一个超大规模高清晰度、横跨多个年龄阶层、包含不同人种和面部特征的数据集，本论文随机选取了其中的 15000 张图片作为真实图像。

而对于完全伪造的 GAN 生成的图像，本论文仿照其他 Deepfake 检测工作，运用多种 GAN 网络进行直接生成。值得注意的是，各种 GAN 网络的预训练模型来源比较多元化，有同样通过 FFHQ 训练的，也有用 Celeb-A[55] 训练的，这是为了保证数据集来源的多样性，提升模型的泛化能力。具体而言，所采用的三种 GAN 生成网

表 3.1　本论文所用的数据集构成

	Real		Fake		Edit	
All	FFHQ	15000	Mspie StyleGANv2	10000	Face2Face	1123
			StyleGANv1	10000	FaceSwap	1163
			WGAN-GP	10000		
Train	FFHQ	10000	Mspie StyleGANv2	8000	Face2Face	794
			StyleGANv1	8000	FaceSwap	785
			WGAN-GP	8000		
Validation	FFHQ	5000	Mspie StyleGANv2	2000	Face2Face	328
			StyleGANv1	2000	FaceSwap	377
			WGAN-GP	2000		

络分别是年龄较老的 WGAN-GP[56]，两年前著名的 StyleGAN 系列[54]，最近的一篇优秀工作运用训练策略和显示条件约束无条件 StyleGANv2 的结构[57]。选用这三个网络的原因是考虑到 Deepfake 检测中很重要的一点，是否能保证对于低清晰度和高清晰度同样可行。直觉上而言，越新的网络结构运用了更多的计算资源，使用了先进的网络模块和结构，效果肯定会更好，但为了保证模型不会出现一些不良的归纳偏执，因此也加入了 WGAN-GP 进行训练和测试。针对三种 GAN 算法的介绍就不多言了，毕竟本工作着重于检测相关的研究，对于生成模型算法的涉及只限于此，不会妨碍后文对于本工作的阐述。

第三个部分是部分编辑的伪造图，这一块主要体现在 Deepfake 的换脸技术上。所采用的算法主要是两个，分别是 Face2Face 和 FaceSwap[58]。这两块算法的数据集，除了生成编辑过的伪造图外，还生成了伪造区域的 mask 图，这是为了后文模型的可解释性作准备工作。

总而言之，本数据集总共规模大约为 5w 左右，包含了三块检测的内容，分别是真实图像、完全虚拟生成的伪造图以及部分编辑的换脸图。除了图像外，每张图像还包括各自的 mask 分割图。图3.7展示了部分数据集的图像。

§3.2.2　训练参数设置

本文在 Pytorch1.10 的环境上进行开发，使用了两种不同的骨干网络，分别是 ResNet 和 XceptionNet。对于输入图像并未做过多的数据增强或者变换操作，只是进行了简单的放缩，将所有图像的大小统一为 224×224。设置全局的随机种子为 10。

图 3.7 数据集掠影，包含三个部分，五种不同的算法生成方式。

在优化器选择上均使用了 Adam 作为优化器，将学习率（Learning rate）设为 0.002（本文发现将学习率设置的偏小一点能更好的收敛，普通的 lr=0.02 以及过小的 lr=0.0002 都不是很合适），权重衰减（Weight decay）设为 0.0001，$\beta = (0.9, 0.999)$。在使用三张 GTX1080Ti 进行训练的时候，设置 batch size 为 32，训练轮数保持为 30 轮。本实验并未使用学习率优化策略（如 Warmup 或者线性增长学习率），而是至始至终保持学习率不变。

§3.2.3 Deepfake 检测常用评价指标

本工作实验部分主要采取以下几个指标评判，分别是 Acc、混淆矩阵、F1-score 和 Dice 系数，下面依次做一个简要介绍。

Acc 精度是分类任务重最常用的性能度量，既适用于而分类任务，也适用于多分类任务。精度代表着分类正确的样本数占样本总数的比例。对样例集 D，定义精度为：

$$\begin{aligned} acc(f;D) &= \frac{1}{m}\sum_{i=1}^{m}\Phi(f(x_i)=y_i) \\ &= 1 - E(f;D) \end{aligned} \tag{3.14}$$

准确率作为最常用的指标，当出现样本不均衡的情况时，并不能合理反映模型的预测能力。例如测试数据集有 90% 的正样本，10% 的负样本，假设模型预测结果全为正样本，这时准确率为 90%，然而模型对负样本没有识别能力，此时高准确率不能反映模型的预测能力。因此就需要混淆矩阵了。

此时需要先引入两个概念，分别是查准率（Precision）和查全率（Recall），公式

定义如下：

$$\begin{aligned} P &= \frac{TP}{TP+FP} \\ R &= \frac{TP}{TP+FN} \end{aligned} \tag{3.15}$$

其中，TP、FP、TN、FN 分别代表真正例、假正例、真反例、假反例，同时有 $TP+FP+TN+FN=$ 样例总数，基于此就能推演出混淆矩阵的形式了，如表3.2所示：

表 3.2　分类结果混淆矩阵

真实情况	预测结果	
	正例	反例
正例	TP（真正例）	FN（假反例）
反例	FP（假正例）	TN（真反例）

一般而言查准率与查全率是一组相互对立的关系，因此通常采用 F1-score 来度量两者的关系。F1-score 是基于查准率与查全率的调和平均（Harmonic mean），模型的学习目标是在两者对立的关系中找到一个好的平衡点，具体定义为：

$$\frac{1}{F_1} = \frac{1}{2} \cdot (\frac{1}{P} + \frac{1}{R}) \tag{3.16}$$

Dice 系数（Dice similarity coefficient）是图像分割网络中最常采用的一个评价指标，主要用来度量网络分割结果和 mask 之间的相似性，具体的计算方式如下：

$$\begin{aligned} \mathrm{dice}(A,B) &= \frac{2|A \cap B|}{|A|+|B|} \\ &= \frac{2 \cdot TP}{(TP+FP)+(TN+FN)} \end{aligned} \tag{3.17}$$

§3.3　实验与结果分析

本实验分别选用了两种骨干网络进行测试，为了使其具有可比性，将二者参数量设置的大致相等（将 Xception 去掉了一部分的 Block）。如表3.3可见，Xception 作为骨干网络的表现明显好过 ResNet。这可能是由于深度可分离卷积的模式比较适合人脸的检测。因此，后文也都继续保持选用 Xception 作为骨干网络。

之后，首先实验了单流网络的表现效果，即单独使用频域图像分解（FAD）或者局部频域信息统计（LFS）进行实验。从表3.3第三四行中可以观察到，精度和 F1-score 都有明显的上升，其中 FAD 提升的最多。这可以清晰地表明，使用频域变换的方式对于 Deepfake 的检测具有帮助，能够很大程度的提升模型的判断能力。

表 3.3　针对不同模块的消融实验：在单流网络中，使用 FAD 效果最好；如果简单的合并两条支流，效果提升并不明显，需要使用 MixBlock 进行特征融合；多任务学习模式在不过多增加参数量的情况下，对于模型性能提升有积极作用。

骨干网络	频域特征提取方式	特征融合方式	多任务学习模式	ACC↑	F1-score↑	DICE 系数↑	参数量 (M)
ResNet	/	/	/	83.04	87.23	/	11.4
XceptionNet	/	/	/	90.56	88.89	/	14.35
	FAD	/	/	**96.28**	**97.23**	/	14.55
	LFS	/	/	94.03	95.87	/	14.35
	FAD+LFS	/	/	95.23	95.58	/	28.92
	FAD+LFS	MixBlock	/	96.76	97.86	/	33.17
	FAD+LFS	MixBlock	✓	**98.92**	**98.25**	**98.47**	35.17

然而虽然最后的结果表明单流的 FAD 相较于 LFS 的表现效果要很多，但在训练过程中 FAD 的收敛速度缺是最慢的，甚至比纯粹在时域空间上使用的 XceptionNet 都要慢。如图3.8可见，当都保持单流网络的时候，能明显发现 FAD 的收敛速度更缓慢，在大约 2500 步的时候才趋于平缓。而 LFS 的收敛速度最快，在 800 步的时候训练的损失就已经较小了。我认为其中的原因是基于局部信息统计的特征提取方式综合了局部信息，在网络较浅层次的时候相较于 FAD 能获得更大的感受野，因此使得网络能更快速的收敛。正是由于观测到二者收敛速度具有较大差异，在第四章中才提出了基于嫁接思想的注意力机制，希望融合二者的优势，做到取长补短。

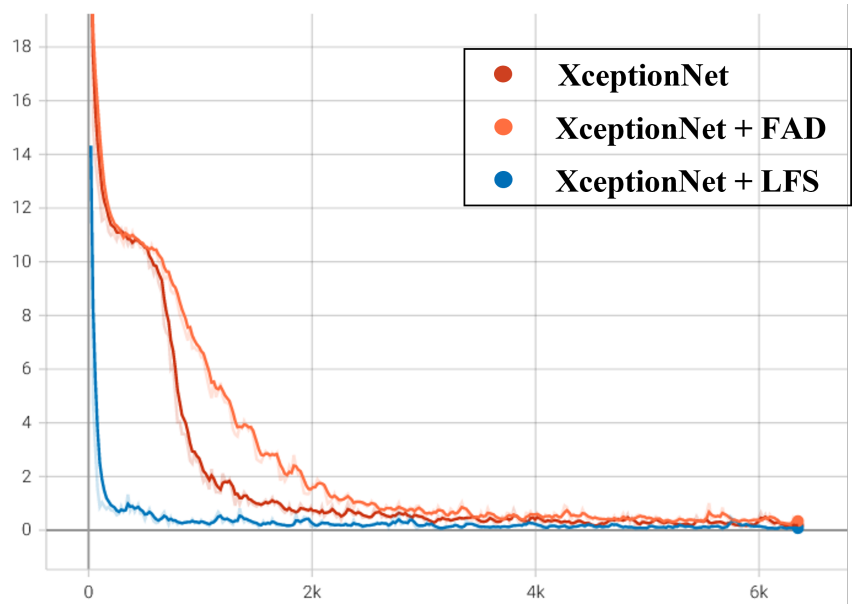


图 3.8　三种不同网络结构在训练时损失（Loss）的变化曲线。

当然除了训练的收敛速度和精度表现外，对于模型的可解释性也是比较好奇的，即为何通过 FAD 或者 LFS 提取的频域特征就能展示出比较明显的优势，因此我将

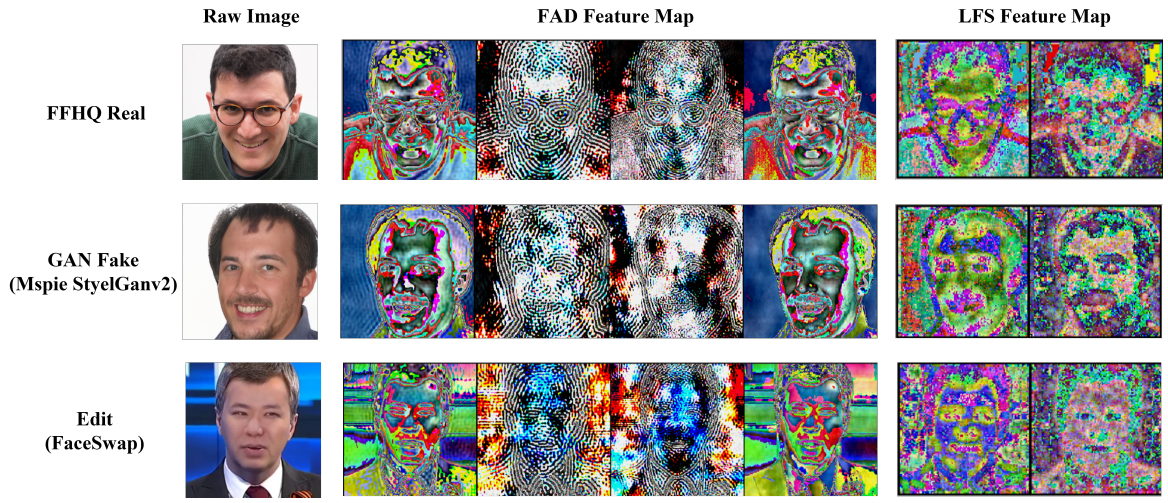


图 3.9　特征图可视化。FAD：将生成的包含 12 个通道的特征矩阵按照 3 个一组进行输出得到。LFS：将生成的包含 6 个通道的特征矩阵按照 3 个一组进行输出得到。GAN 生成的完全虚拟图选取了 Mepie StyleGANv2 算法，部分编辑伪造图选取了 FaceSwap 算法。

经过 FAD 和 LFS 的图像进行了可视化，希望能从中发现一些原因。本文的做法是将 FAD 的 12 个通道按照 3 个为一组进行输出，而 LFS 的 6 个通道则相应的划分为两组。由图3.9可发现，真实图像经过 FAD 支流后的特征图纹理要清晰的多。FAD 的二三列可以观测到真实图像脸部、耳朵处、眼睛处较少出现大面积的黑斑，而一四列的图像可以发现纹理信息和深度信息都更为复杂。作为对比，可以发现由 GAN 生成的虚拟伪造图脸部纹理信息较少，容易出现大面积的深色；而经过 FaceSwap 的编辑图像可以发现二三列的 FAD 图像容易把背景的图像颜色信息带入，且在一四列出现了“阴阳脸”的情况，而同样地，脸部的纹理感较差，均是大面积的色块。类似地，三类图片在经过 LFS 支流后的特征图在面部细节上的表现也是不同的，尤其是 FaceSwap 的图像在下巴处出现了明显的割痕。这类对于人眼比较显著的特征显然也能轻易被之后的骨干网络所学习到，因此通过频域变换的方式，使得一些原来 CNN 网络难以学习的特征变得显著，这也是本实验效果能有显著提升的一个依据。

之后测试了将 FAD 和 LFS 支流进行融合的实验效果，如表3.3中 5-7 行可以发现：如果单纯将二者进行并行，并且不做任何特征融合，只是在最后得到结果处采用相加平均的方式，其效果并不是特别好。但当添加了 MixBlock 融合模块的时候，可以发现效果能有明显提升，F1-score 上升了大约 0.6，ACC 精度也提升了 0.5 左右。将实验结果用混淆矩阵可视化，可以发现针对每个类别的准确率，从图3.10中可以发现对于各个类的准确率都在 95% 以上。容易出错的类别在于将编辑伪造图判断为真实图像，有 3% 的编辑图被检测错误了。

最后在使用一个简单的多层反卷积组成的分割定位头后，在增加一点参数量的同时，F1-score 能达到 98.25，ACC 精度也提上到了 98.92，与此同时衡量分割效果

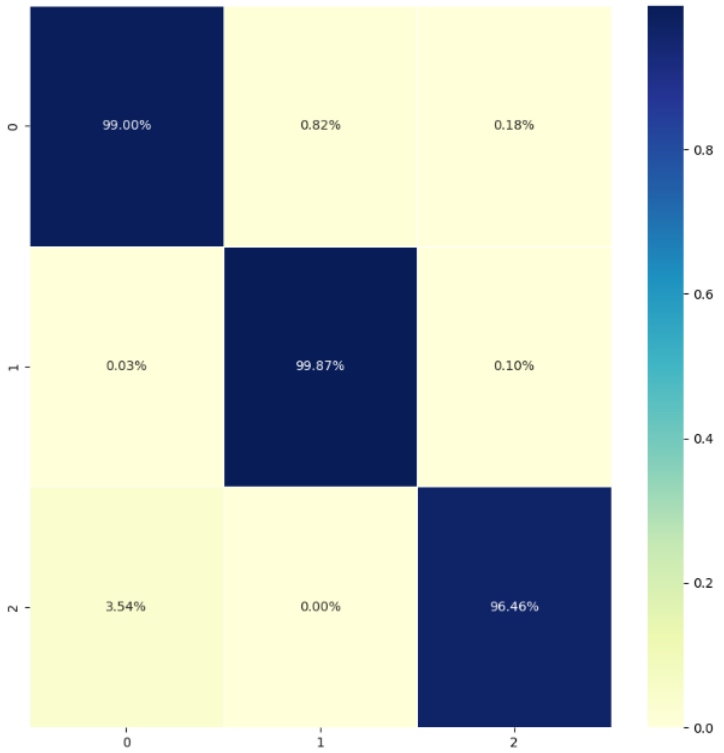


图 3.10 运用 MixBlock 后模型的混淆矩阵

的 DICE 系数也达到了 98.47。这表明使用多任务学习模式对于模型的收敛和表现是有积极作用的。这也启示我们在数据集具备模态多样性的时候，可以大胆地尝试多任务学习这一方法。

§3.4 本章小节

本章主要介绍了基于频域分析和双流网络的 Deepfake 检测算法，分别从骨干网络选择、频域特征提取、双流网络以其特征融合等方面介绍模型的各个组成元素。之后介绍了本工作所用的数据集和本文实验所用的评价指标，阐述了实验的参数设置和训练过程，最后给出了对应的实验和结果分析，包括特征提取后特征矩阵的可视化、训练收敛速度对比、消融实验等。当然本方法也存在一定的缺陷，比如模型参数量较大、融合方式可解释性不足以及缺乏 mask 图导致的难以做弱监督或无监督学习，因此在第四章中针对这些问题进行了一系列的优化和解决。

第 4 章　基于多模态图像融合注意力的 Deepfake 检测

本章为本论文的重点章节，主要介绍基于多模态图像融合注意力机制的 Deepfake 检测算法具体而言，分别从整体框架、多流网络、多模态图像融合注意力机制、半监督学习等方面介绍了本检测算法的各个模块实现。之后给出了相应的消融实验和结果分析。

§4.1　基于多模态图像融合注意力的网络架构

§4.1.1　整体框架

整体的网络框架可见图4.1。对于一张图像而言，会将该输入分别送入三个支流网络，分别是 FAD、LFS 和原始图像处理模块。然后分别送入两个 XceptionBlock 模块进行特征提取。之后会得到一个大小为 $(N, 32, 28, 28)$ 的特征矩阵，为了在之后进行类似自注意力的操作，需要将特征图像变为一个序列，因此采用直接折叠第三四维度的方式进行处理。

通过折叠变换后特征向量大小变化为 $(N, 32, 784)$，之后进行一个线性映射（Linear Projection of Flattened Feature Maps）。在这里保持了线性映射输入数和输出数的一致。之后进行多模态图像融合注意力操作，将原始图像看作是一个查询问题，需

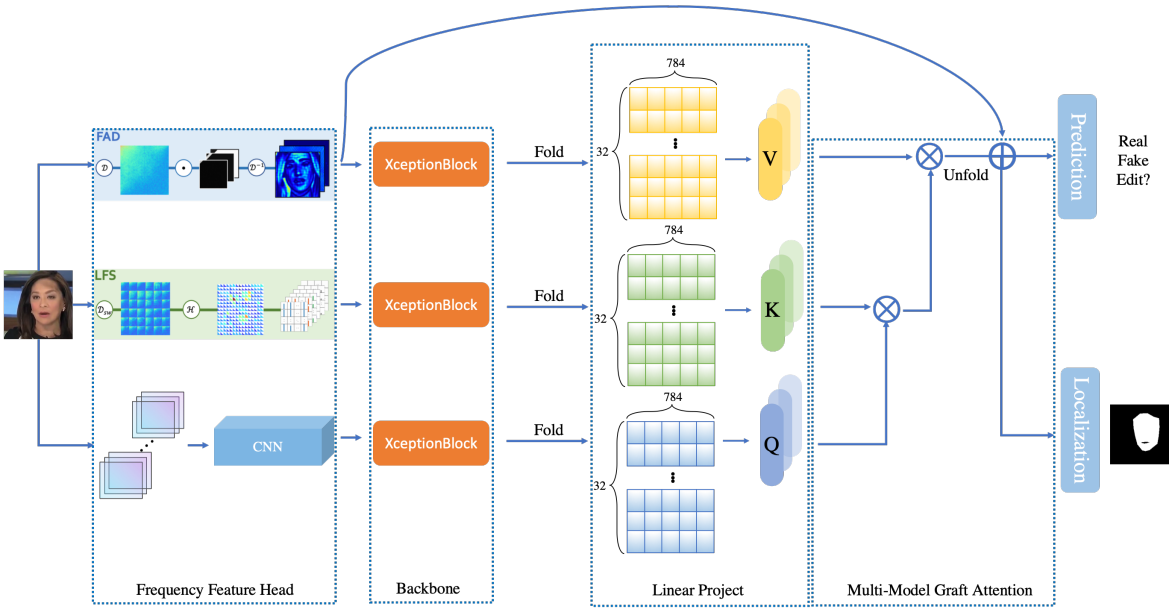


图 4.1　基于多模态图像融合注意力的网络整体框架。对于输入图像，运用多流网络结构提取特征，将图像折叠后进行线性映射，使用融合注意力（Multi-Model Graft Attention）提取强特征，最后使用多任务模式联合优化模型参数。

要通过 LFS 得到的局部特征进行索引匹配，$Q \times K$ 夹角越小，正相关；反之，负相关。最后得到的结果是基于 FAD 的经过频域变换和滤波的特征。之后将其第三维度重新展开为 H 和 W。如此的融合注意力机制一方面是出于实验发现的结论：单纯的 FAD 或者 LFS 支流提取信息，在收敛速度上明显 LFS 更快，但在最后表现效果却是 FAD 更好，如果能结合二者，运用 LFS 的快速收敛辅助 FAD 更好的拟合，同时也能发挥 FAD 的最大效果。另外一方面是出于融和方式的考量：在第三章中介绍的融合方式虽然也是基于注意力的，但却是由卷积得到的，或者说只是一个概率矩阵，缺乏一定的可解释性。在本章所用的模块是符合人类对于探究物体的直觉的：先从原始图像出发，然后探究每一个像素点和局部周围的关系，得到相关度矩阵，然后用它去查询 FAD 中哪些最相关。另外一点值得注意的是，这里将每个通道的 $28 * 28$ 矩阵展开，然后进行注意力融合，是为了更好的关注每一个值不同的表示方式，是一种在通道上做注意力的操作。

之后采用两个输出头，分别是检测分类头和定位分割头。在检测分类头上采用的模块和第三章保持一致，即简单的 CNN。而分割定位头上采用了一些跳跃连接，分别按照 FAD、LFS 和原始图像的顺序进行跳跃连接。这是借鉴了 UNet[59] 的结构，将编码器和解码器的特征进行融合，辅助网络收敛，尤其是对于图像分割这样比较底层级语义信息的任务，进行跳跃连接很有必要。

最后的损失函数设置和第3.1.6节保持一致，均采用了检测分类和定位分割的联合损失函数。

§4.1.2 多模态图像融合注意力机制

首先需要明确的一点是此处的多模态不是指的常说的图片、文本这类多模态，而是单纯指的图像上的多种模态。因为本文不是纯粹的从原始图像出发，而是运用了频域分析和频域变换等方法，因此对于融合的特征，它们的来源是多样的、多模态的，有来源于原始图像经过卷积提取信息的，也有来自于频域变换滤波又做了逆变换的，当然也包括频域空间上做局部信息统计的。本文的多模态指的是这些模态，而融合了这些模态的方法本文称其为多模态图像融合注意力机制或者嫁接注意力机制（Multi-Model Graft Attention）。

多模态图像融合注意力，或称嫁接注意力（Graft Attention）的来源在第二章中已经阐明了，思想是来源于 C Xie[51] 在图像显著性检测时的发现，其核心是将不同模态层级的特征信息进行注意力操作，目的是由于 CNN 关注局部，而 Transformer 更关注全局，为了综合二者，因此使用了融合注意力模块。事实上，这一思想和多流网络就能进行很好的结合。这也是本文的一个关键贡献。本文将多流网络与多模态融合注意力进行了结合，所提出的模块相比第三章的 MixBlock 更简单，同时性能和

表现效果也更优秀：占有的计算资源更少，收敛更快。第三章的实验已经表明，单纯的使用单流网络各有优劣，而使用双流网络和多任务学习的模式会增加较多的计算开销，同时收敛速度提升也有限。因此考虑结合一些人体观测物体的方式（直觉知识）来辅助模型更好的收敛，同时提升它的表现效果。

接下来描述下人类关注物体一般的顺序和方式：首先第一眼会从全局出发，大致摸清楚整张图像是个什么样子，整体的对象；此时大脑会联系到之前的一些学识和经验，做出一个判断：如果这副图像是表达的某个观点或对象的话，应该有个地方是它的重要特征，是需要去重点关注和复核的。因此，第二步人类会凑近了观测这副图像，运用脑中知识进行匹配，在局部区域进行复核和相关度的判断。这一步的重点就是和自己脑内固有的和对该对象局部的判断进行匹配。而在此基础上，人类就有了大致的自我判断，在第三步，他们会重新回顾整张图像，“站的稍微远一点”，再次从全局进行审核和观察。值得注意的是，第三步的全局观测和第一步是不同的，第三步已经有了自我的一个判断，而第一步则完完全全是空白的。这一系列的动作都是自发发生在人类观测任何物体的时候，我们会去调用一定的神经元，但对于这一任务并不涉及特别深层次的语义信息。Deepfake 的检测也是这样的，对于伪造抑或真实图像的感受很多时候都是直觉主导的，因此在本工作中对第三章的网络结构进行了修改，使得网络变得更为轻量化，本文认为通过模拟人类的一个行径，能对 Deepfake 的检测具有指导意义。

以上大致阐述了人类的直觉观测模式，具体到计算机的实现上则是改善了第三章的双流网络，变更为多流网络，同时运用多模态图像融合注意力机制实现的。如图4.1所示，本文使用原始 RGB 图像进行卷积操作生成的特征图模拟人类第一步中的操作，运用简单的卷积加上 XceptionBlock 模拟人类对图像第一印象的全局操作，并以此作为查询向量。在第二步上，采用了第3.1.3.2小节中提出的局部频域统计信息（LFS），并以此作为被查询的索引向量。这一步就相当于人类认知的第二部：利用局部特性。最后，重新回到全局信息，利用频域图像分解（FAD）得到的滤波后的时域信息，相当于是用 $Q \times K$ 的相关性矩阵，叠加上检索的具体内容信息。至此，得到的特征向量就是融合了全局以及局部信息的强特征。

值得注意的一点是，在使用多模态图像融合注意力的时候，也依然保持了 Vaswani A[47] 使用的多头注意力结构（Multi-head attention），如图4.2，目的是为了更好的融合不同通道层面的信息。

§4.1.3 半监督学习及其损失函数

目前 Deepfake 的定位分割问题存在一个难点，即很难获取局部编辑的 mask 图。因此本文也提出一个相应的解决方案，运用半监督的方式能够较好的在大规模数据

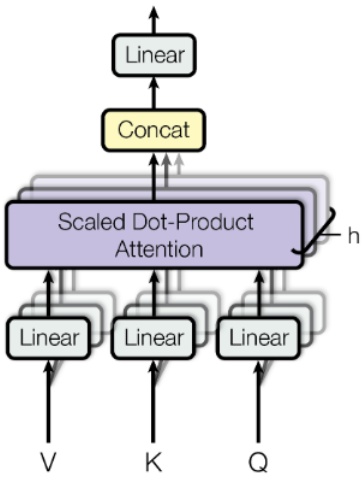


图 4.2　多头注意力机制由多个并行的注意力层组成，一般 $h = 8$。

集上进行训练。

这里分成两类情况，分别是有 mask 图监督信号和无 mask 图监督信号。

针对第一种情况，其实退化成前文所述的有监督学习。对于得到的定位分割图，只需要做逐像素点的二分交叉熵或者 L_{DICE} 即可。

针对第二种情况，有些图像只知道它的分类标签，但却并不知道它的 mask 图，此时依然可以应用本文提出的框架。只需要在计算损失函数的时候添加一个判断，分为两类：如果是真实图像的时候，不要激活分割头的任何一个像素点；而如果是虚伪或者编辑图的时候，则是其中存在至少一个像素点它的置信度（概率）大于 0.75（0.75 为超参数）。用公式表达上述两种情况即为：

$$\mathcal{L}_{\text{seg}} = \begin{cases} \left|\text{Sigmoid}\left(\text{Mask}_{seg}\right) - 0\right|, & \text{if real} \\ \left|\max\left(\text{Sigmoid}\left(\text{Mask}_{seg}\right)\right) - 0.75\right|. & \text{if fake or edit} \end{cases} \tag{4.1}$$

§4.2　实验与结果分析

尽管在第三章中的实验效果已经很出色了，但本文仍希望能够再进一步提升精度的同时，使得模型更加的轻量化。为此抛弃了 XceptionNet 的完整结构，只采用其中的 XceptionBlock，同时大量减少了使用该 Block 的数量。因此，虽然增加了一条原始图像提取特征的支流，但每条支流只使用了两个 Block，所以使得参数量减少了大约十倍。然而实验的效果却表明，见表4.1，通过使用多模态图像融合注意力机制能够使得模型的精度相比第三章使用的庞大网络结构更好。在不使用多任务学习模式的情况下，F1-score 提升了将近 1 个点，而 ACC 也有大幅度提升。同样的，本文也实验了多任务下的情况，F1-score 达到了 99.69，ACC 准确率达到了 99.76，分割定位的 DICE 系数也达到了 99.27，三者都是所有实验中最高的。

表 4.1 不同的特征融合方式和是否使用多任务学习模式的消融实验。实验表明：使用多模态图像融合注意力机制不仅参数量大大减少，模型的精度也有显著提升。

骨干网络	频域特征提取方式	特征融合方式	多任务学习模式	ACC↑	F1-score↑	DICE 系数↑	参数量 (M)
XceptionNet	FAD+LFS	MixBlock	✗	96.76	97.86	/	33.17
	FAD+LFS	Multi-Model Graft Attention	✗	**99.42**	**98.92**	/	3.141
	FAD+LFS	MixBlock	✓	98.92	98.25	98.47	35.17
	FAD+LFS	Multi-Model Graft Attention	✓	**99.76**	**99.69**	**99.27**	3.149

其中的原因，绝大多数是来自于融合注意力，它通过模拟人类观测物体的直觉本能，从全局出发到局部信息统计，最后回归于经过频域变换和滤波后的时域空间信息，高效地融合了三种模态的信息。在达到如此好的效果的同时，也解决了在第三章中提出的单纯的 FAD 收敛速度较慢的问题。如图4.3可见，运用融合注意力的网络结构收敛速度有明显的提升，在 1000 步的时候就已经下降到了较低的水准。与此同时也生成了训练时候的 ACC 和 F1-score 的时间图。可以明显的发现：所提出的融合注意力方法收敛速度快，表现效果好。

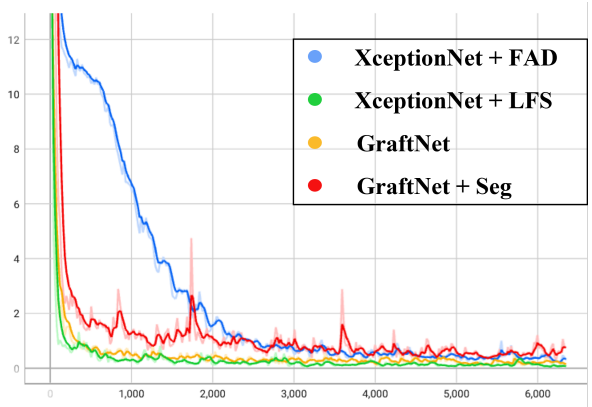
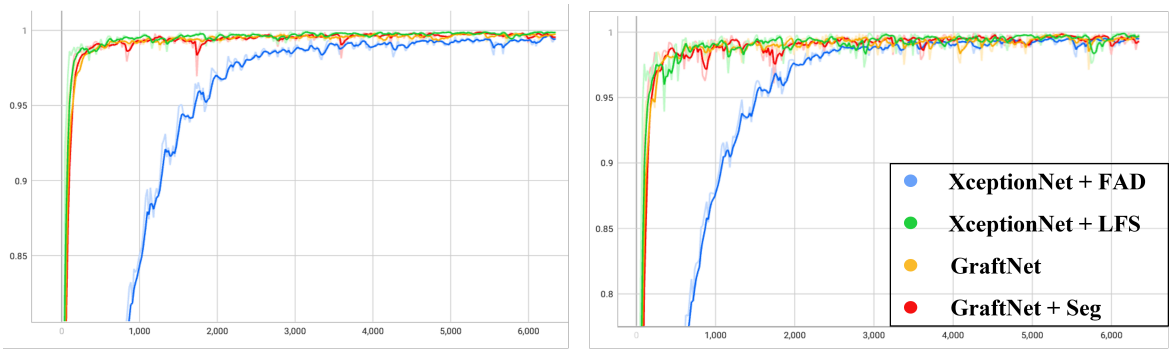


图 4.3 四种不同网络结构在训练时损失（Loss）的变化曲线。



(a) ACC **(b) F1-score**

图 4.4 训练过程：（a）ACC 精度随步数增加时的曲线；（b）F1-score 随步数增加时的曲线。

最后，本文生成了表现最好的模型的混淆矩阵，如图4.5。可以发现之前的模型在编辑伪造图像上存在约 3% 的出错率，但通过目前的改进，分类检测的准确率已经达到了 100%，同时其他两个类别的检测准确率也都在 99% 以上，这是一个很优秀的结果。

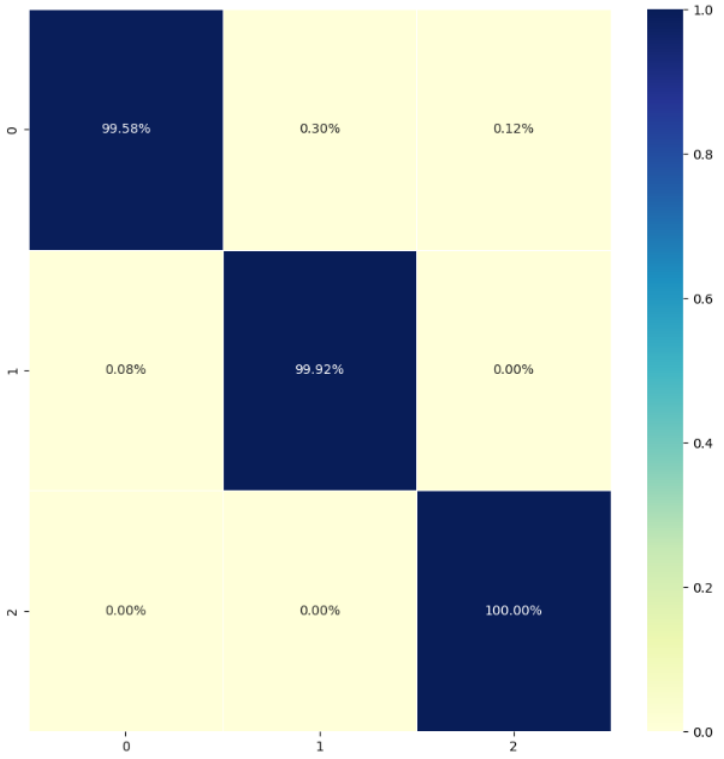


图 4.5　使用多模态图像融合注意力机制的模型的混淆矩阵

综上所述，本实验论证了采用多模态图像融合注意力机制的优越性，在收集的数据集上展现出了良好的性能表现；同时也较好地解决了模型参数量大、收敛速度慢和模块缺乏可解性这三个在第三章的遗留问题。

§4.3　本章小结

本章主要介绍基于多模态图像融合注意力机制的 Deepfake 检测算法，为了解决第三章最后提出的问题，本文重新设计了网络中的部分模块，运用了多流网络、多模态图像融合注意力机制和半监督学习，使得网络大大减少了参数量，同时也更为符合人类观测事物的模式。最后与前文的最优解进行了消融实验，论证了采用多模态图像融合注意力机制的实用性，在收集的数据集上达到了非常优秀的实验效果。

第 5 章　总结与展望

本工作在前期进行了大量的调研，深入分析了目前主流的 Deepfake 检测技术，特别的，本文将其分为四种不同的研究方向，分别是基于特定伪影的视觉深度伪造检测、基于数据驱动的视觉深度伪造检测、基于信息不一致的视觉深度伪造检测以及其他类型视觉深度伪造检测。本文探讨了其中存在的难点，并从现有的方法以及其他领域的先进方法中汲取经验，将其融合使用，提出了两个 Deepfake 检测算法。

第一种检测算法通过使用频域分析的方式，将往常在时域空间进行检测的算法推广到了频域空间上。同时，改善了以往单纯的频域变换方式，使用基于可学习的频率滤波器进行频域特征的提取，提出了频域图像分解和局部频域信息统计两种特征提取方式。之后运用双流网络、特征融合和多任务学习模式进一步提升模型的泛化性和可解释性。而第二种检测算法在前者的基础上进一步改进，运用了多模态图像融合注意力机制以及半监督学习的方式，大大压缩了模型的参数量，同时也保证了在大规模弱监督数据集上进行预训练的可行性。具体而言，本文提出的多模态图像融合注意力源于嫁接思想，同时也符合人类观测事物的直觉本能，通过增加一条原始图像的支流，并使用多流网络提取的特征进行融合注意力操作，实现了从全局到局部细节再回归到全局，即从时域出发，查询局部频域统计信息，最后再回到经过频域滤波的时域信息的过程。通过采用该方法，模型的参数量大约减少了十倍，收敛速度有了显著的增加，同时模型的精度也有了很大的提升。

对于未来的工作，本文认为可以从两点出发：(1). 将此方法运用到更大规模的数据集中。此处指的是可以不只局限于现有的数据集，由于本文提出了半监督训练的方式，因此可以采用网络爬虫大范围的搜索图像。这样的好处主要在于目前的数据集多是由具体算法得到的，一个数据集可能最多就包含 3-5 个算法，容易被针对性的攻击，而直接采用网络爬虫搜索的话可以很好的避免这一问题。(2). 探索实时视频流的检测算法。目前的 Deepfake 检测还多局限于图像或者离线的视频，很难达到实时检测。然而生活中很需要快速地鉴别哪些是伪造的、哪些是真实的。如果能融合进日常的应用中，肯定会对日常的生活有很大的帮助。因此，探索小型的、高效的检测网络模型对于实时视频的检测很有必要。

致　谢

年少的我曾经多少次幻想过如何在致谢中满腹经纶、妙笔生花，如何将其作为一篇大学四年的完美句号，但真的慢慢写到这一章时，却恍然间发现一切都是那么的平静，就好似中考时填完最后一份答题卡，高考铃响交卷那时的心境一样。我始终相信：如果你没被过程的起伏不定所击倒，那结果就一定是可期的。回首大学四年，弹指一挥间，发生了太多太多的事，但好似所有的事又都可以被“挑战”和“战胜”所概括。

大一入学时的紧张，第一次接触微积分时的不知所措，最后不也顺利的通过考试了吗？暑假时的游学交换，一个人的 20 多天旅行，面对全英文的环境，最后不也拿到了通过的证书和奖学金吗？大二时被计算机专业繁重的专业课和无数的 bug 所数次崩溃，但到头来，不也很好的打下了编程的基础了吗？大三参加了多项比赛，都数不清多少次在深夜觉得自己的脑袋是木鱼，担忧写不好代码，调不好模型，但最后不也取得了一些小成就吗？进入大四开始为未来的出路担忧，总是抱怨自己为什么没有从小生活在英语环境，考了无数次托福和 GRE，甚至崩溃到上知乎搜索什么是万念俱灰的感觉，但最后的最后，不也顺利且幸运地达到了希望的分数吗？临近春招，无数次徘徊在找工作和留学的边缘，沮丧地以为不会有院校要我，但惊喜却总是不期而遇，如果它中途告诉了你，又怎能让你在最后高兴得乐开了花呢？进入五月，醒来的每一天都在为毕设而发愁，大改了数次文章结构，但试想，若不是因为前期的不断接受挑战，克服挑战，又怎会有我现在平静地在电脑前码下这一段似乎不那么简单的经历呢？那些不能把你打败的，只会让你变得更加强壮。多年以后，当我再回想起这些的时候，我相信所有的人都会用一种略带轻蔑的语气，半开玩笑半认真的和你讲述那段多么有意思的时光，但我同时也希望自己始终铭记，我在这一路上，并不孤单。

我想感谢我的父母，在我最艰难最焦虑的时候，是我坚强的后盾，始终相信着我，替我排忧解难，没有你们的陪伴，很难有故事的下半章。我想感谢我的导师武星，在大二大三做项目的时候，以及写毕设的时候，时时刻刻地给我们提供帮助，经常忙到边吃饭边和我们开会，没有导师的指导，很难有这篇毕业论文。我想感谢我大学遇到的所有老师，是你们手把手领着一个科研小白，一步步地教授学识，慷慨地提供各种资源，我就好似一株树苗，在肥沃的土壤中肆意生长。我想感谢我的辅导员李清怡老师，在我遇到任何困难的时候，李老师都不遗余力的给我提供任何形式的帮助，无论生活还是学习，她是我们班的“保护伞”。我想感谢我的室友、朋友和我们组的其他几个优秀的同学，我们经常彻夜谈论自己的项目内容，毫无保留的

给出自己的意见，在临近截稿日期的时候，我们互相鼓励，互帮互助，营造了一种互相竞争的积极学术环境。一个人的力量终究是有限的，正如面对面霸时，我们需要复仇者联盟一样，很感谢大学四年能遇到你们，因为有你们，我才拥有了跌宕起伏、绚丽多彩的大学生活！

最后，愿我和各位前程似锦，未来可期，归来仍是少年。

参考文献

[1] Mirsky Y, Lee W. The creation and detection of deepfakes: A survey [J]. ACM Computing Surveys (CSUR), 2021, 54 (1): 1–41.

[2] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets [J]. Advances in neural information processing systems, 2014, 27.

[3] 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.

[4] 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.

[5] Li Y, Chang M-C, Lyu S. In ictu oculi: Exposing ai created fake videos by detecting eye blinking [C]. In 2018 IEEE International Workshop on Information Forensics and Security (WIFS), 2018: 1–7.

[6] Wang R, Juefei-Xu F, Ma L, et al. Fakespotter: A simple yet robust baseline for spotting ai-synthesized fake faces [J]. arXiv preprint arXiv:1909.06122, 2019.

[7] Agarwal S, Farid H, Gu Y, et al. Protecting World Leaders Against Deep Fakes. [J], 2019, 1.

[8] Tariq S, Lee S, Kim H, et al. Detecting Both Machine and Human Created Fake Face Images In the Wild [C]. In Proceedings of the 2nd International Workshop on Multimedia Privacy and Security, New York, NY, USA, 2018: 81–87.

[9] Do Nhu T, Na I, Kim S. Forensics Face Detection From GANs Using Convolutional Neural Network [M]. 2018.

[10] Afchar D, Nozick V, Yamagishi J, et al. Mesonet: a compact facial video forgery detection network [C]. In 2018 IEEE international workshop on information forensics and security (WIFS), 2018: 1–7.

[11] 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): 6.

[12] Papernot N, McDaniel P, Goodfellow I. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples [J]. arXiv preprint arXiv:1605.07277, 2016.

[13] Hussain S, Neekhara P, Jere M, et al. Adversarial deepfakes: Evaluating vulnerability of deepfake detectors to adversarial examples [C]. In Proceedings of the IEEE/CVF winter conference on applications of computer vision, 2021: 3348–3357.

[14] Carlini N, Farid H. Evading deepfake-image detectors with white-and black-box attacks [C]. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops, 2020: 658–659.

[15] Suwajanakorn S, Seitz S M, Kemelmacher-Shlizerman I. Synthesizing obama: learning lip sync from audio [J]. ACM Transactions on Graphics (ToG), 2017, 36 (4): 1–13.

[16] Wu W, Zhang Y, Li C, et al. Reenactgan: Learning to reenact faces via boundary transfer [C]. In Proceedings of the European conference on computer vision (ECCV), 2018: 603–619.

[17] Frühstück A, Singh K K, Shechtman E, et al. InsetGAN for Full-Body Image Generation [J]. arXiv preprint arXiv:2203.07293, 2022.

[18] Siarohin A, Lathuilière S, Tulyakov S, et al. First order motion model for image animation [J]. Advances in Neural Information Processing Systems, 2019, 32.

[19] Nirkin Y, Keller Y, Hassner T. Fsgan: Subject agnostic face swapping and reenactment [C]. In Proceedings of the IEEE/CVF international conference on computer vision, 2019: 7184–7193.

[20] Chen H, Hu G, Lei Z, et al. Attention-based two-stream convolutional networks for face spoofing detection [J]. IEEE Transactions on Information Forensics and Security, 2019, 15: 578–593.

[21] Kumar P, Vatsa M, Singh R. Detecting face2face facial reenactment in videos [C]. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2020: 2589–2597.

[22] Thies J, Zollhofer M, Stamminger M, et al. Face2face: Real-time face capture and reenactment of rgb videos [C]. In Proceedings of the IEEE conference on computer vision and pattern recognition, 2016: 2387–2395.

[23] Korshunova I, Shi W, Dambre J, et al. Fast face-swap using convolutional neural networks [C]. In Proceedings of the IEEE international conference on computer vision, 2017: 3677–3685.

[24] Choi Y, Choi M, Kim M, et al. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation [C]. In Proceedings of the IEEE conference on computer vision and pattern recognition, 2018: 8789–8797.

[25] Fernando T, Fookes C, Denman S, et al. Exploiting human social cognition for the detection of fake and fraudulent faces via memory networks [J]. arXiv preprint arXiv:1911.07844, 2019.

[26] Zhang D, Li C, Lin F, et al. Detecting deepfake videos with temporal dropout 3DCNN [C]. 2021.

[27] Frank J, Eisenhofer T, Schönherr L, et al. Leveraging frequency analysis for deep fake image recognition [C]. In International Conference on Machine Learning, 2020: 3247–3258.

[28] Zhou P, Han X, Morariu V I, et al. Two-stream neural networks for tampered face detection [C]. In 2017 IEEE conference on computer vision and pattern recognition workshops (CVPRW), 2017: 1831–1839.

[29] Conotter V, Bodnari E, Boato G, et al. Physiologically-based detection of computer generated faces in video [C]. In 2014 IEEE International Conference on Image Processing (ICIP), 2014: 248–252.

[30] Ciftci U A, Demir I, Yin L. How do the hearts of deep fakes beat? deep fake source detection via interpreting residuals with biological signals [C]. In 2020 IEEE international joint conference on biometrics (IJCB), 2020: 1–10.

[31] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition [C]. In Proceedings of the IEEE conference on computer vision and pattern recognition, 2016: 770–778.

[32] Chollet F. Xception: Deep learning with depthwise separable convolutions [C]. In Proceedings of the IEEE conference on computer vision and pattern recognition, 2017: 1251–1258.

[33] Qian Y, Yin G, Sheng L, et al. Thinking in frequency: Face forgery detection by mining frequency-aware clues [C]. In European Conference on Computer Vision, 2020: 86–103.

[34] Songsri-in K, Zafeiriou S. Complement face forensic detection and localization with facial landmarks [J]. arXiv preprint arXiv:1910.05455, 2019.

[35] Dang H, Liu F, Stehouwer J, et al. On the detection of digital face manipulation [C]. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern recognition, 2020: 5781–5790.

[36] Chai L, Bau D, Lim S-N, et al. What makes fake images detectable? understanding properties that generalize [C]. In European Conference on Computer Vision, 2020: 103–120.

[37] Durall R, Keuper M, Keuper J. Watch your up-convolution: Cnn based generative deep neural networks are failing to reproduce spectral distributions [C]. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020: 7890–7899.

[38] Sarlashkar A, Bodruzzaman M, Malkani M. Feature extraction using wavelet transform for neural network based image classification [C]. In Proceedings of Thirtieth Southeastern Symposium on System Theory, 1998: 412–416.

[39] Stuchi J A, Angeloni M A, Pereira R F, et al. Improving image classification with frequency domain layers for feature extraction [C]. In 2017 IEEE 27th International Workshop on Machine Learning for Signal Processing (MLSP), 2017: 1–6.

[40] Fujieda S, Takayama K, Hachisuka T. Wavelet convolutional neural networks for texture classification [J]. arXiv preprint arXiv:1707.07394, 2017.

[41] Li J, You S, Robles-Kelly A. A frequency domain neural network for fast image super-resolution [C]. In 2018 International Joint Conference on Neural Networks (IJCNN), 2018: 1–8.

[42] Karras T, Laine S, Aittala M, et al. Analyzing and improving the image quality of stylegan [C]. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020: 8110–8119.

[43] Wang S-Y, Wang O, Zhang R, et al. Cnn-generated images are surprisingly easy to spot... for now [C]. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020: 8695–8704.

[44] D'Avino D, Cozzolino D, Poggi G, et al. Autoencoder with recurrent neural networks for video forgery detection [J]. Electronic Imaging, 2017, 2017 (7): 92–99.

[45] Chen M, Sedighi V, Boroumand M, et al. JPEG-phase-aware convolutional neural network for steganalysis of JPEG images [C]. In Proceedings of the 5th ACM Workshop on Information Hiding and Multimedia Security, 2017: 75–84.

[46] Simonyan K, Zisserman A. Two-stream convolutional networks for action recognition in videos [J]. Advances in neural information processing systems, 2014, 27.

[47] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need [J]. Advances in neural information processing systems, 2017, 30.

[48] Araslanov N, Roth S. Self-supervised augmentation consistency for adapting semantic segmentation [C]. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021: 15384–15394.

[49] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale [J]. arXiv preprint arXiv:2010.11929, 2020.

[50] Yi F, Wen H, Jiang T. Asformer: Transformer for action segmentation [J]. arXiv preprint arXiv:2110.08568, 2021.

[51] Xie C, Xia C, Ma M, et al. Pyramid Grafting Network for One-Stage High Resolution Saliency Detection [C]. In CVPR, 2022.

[52] Liu Z, Qi X, Torr P H. Global texture enhancement for fake face detection in the wild [C]. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 8060–8069.

[53] Wu D, Liao M, Zhang W, et al. Yolop: You only look once for panoptic driving perception [J]. arXiv preprint arXiv:2108.11250, 2021.

[54] Karras T, Laine S, Aila T. A style-based generator architecture for generative adversarial networks [C]. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019: 4401–4410.

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

[56] Gulrajani I, Ahmed F, Arjovsky M, et al. Improved training of wasserstein gans [J]. Advances in neural information processing systems, 2017, 30.

[57] Xu R, Wang X, Chen K, et al. Positional encoding as spatial inductive bias in gans [C]. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021: 13569–13578.

[58] Rossler A, Cozzolino D, Verdoliva L, et al. Faceforensics++: Learning to detect manipulated facial images [C]. In Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019: 1–11.

[59] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation [C]. In International Conference on Medical image computing and computer-assisted intervention, 2015: 234–241.