中图分类号：TP391
学科分类号：081200
论文编号：1028716 23-S125

硕士学位论文

基于多头注意力机制的深度伪造视频鉴别

研究生姓名　　李灿东
学科、专业　　计算机科学与技术
研 究 方 向　　计算机视觉
指 导 教 师　　谭晓阳教授

南京航空航天大学
研究生院 计算机科学与技术学院
二〇二三年三月

Nanjing University of Aeronautics and Astronautics
The Graduate School
College of Computer Science and Technology

Deepfake Videos Detection Based on Multi-Head Attention Mechanism

A Thesis in

Computer Science and Technology

by

Candong Li

Advised by

Prof. Xiaoyang Tan

Submitted in Partial Fulfillment

of the Requirements

for the Degree of

Master of Engineering

March, 2023

承诺书

本人声明所呈交的博/硕士学位论文是本人在导师指导下进行的研究工作及取得的研究成果。除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得南京航空航天大学或其他教育机构的学位或证书而使用过的材料。

本人授权南京航空航天大学可以将学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存、汇编学位论文。

（保密的学位论文在解密后适用本承诺书）

作者签名：

日　　期：

摘 要

深度伪造的兴起，导致了大量针对特定人物的伪造音视频媒体的出现，近年来深度伪造的质量和数量都在不断提升，但这同时也带来了许多社会问题甚至政治军事危机，因此深度伪造检测的工作近几年也成为了研究热点。然而随着伪造技术的发展，在肉眼可见的范围内，伪造视频图像的质量越来越高，伪造痕迹开始更难被挖掘。另一方面，由于图像数据的信息具有高冗余度，大量的内容信息对伪造检测时无益的。因此，如何设计一种高效且具有良好泛化性以应对多种伪造手段的真伪检测模型成为当下的研究重点。本文基于此提出了两种不同的深度伪造检测模型，即基于多域特征网络的伪造检测模型和基于帧间滑动窗口注意力机制的伪造检测模型，从不同角度去进行伪造检测。

本文提出了一个多域特征网络，结合了频域信息和 RGB 域信息，并引入了注意力机制对特征图进行了自适应的权重分配，通过类别得分融合将不同域的预测信息融合成最终的结果。此外，本文还提出了一种基于人脸关键点，通过旋转和缩放人脸标定框来进行帧间人脸对齐的方法，并以此提高 3D 伪造检测模型的性能。通过在 Celeb-DF 和 DFDC 数据集上进行了大量的实验后，实验结果表明进行多模态信息融合后的模型性能有了显著提高，并且该模型在深度伪造视频的鉴别性能上接近当前最先进的水平，并具有很好的泛化性。

本文提出了一种滑动窗口注意力机制，分别用于通道间和帧间的自注意力计算，据此分别提出了 2D 伪造检测模型和 3D 伪造检测模型。该机制有效地降低了高维度向量带来的计算注意力复杂度过大的问题，并通过选择更关键的帧信息来提高检测的性能。此外，本文还引入了一种注意力掩码的方式去解决人脸对齐时带来的背景区域过大、冗余信息过高的问题。通过开源伪造人脸数据集上的实验可以表明该模型具有很好的真伪分类性能和不错的鲁棒性。

关键词：深度学习，深度伪造检测，自注意力，人脸替换 ，滑动窗口

ABSTRACT

The development of deepfake has led to the emergence of a large number of forged videos targeting specific characters. In recent years, the quality and quantity of deepfake have been increasing, but it has also brought many social problems, even political and military crises. Therefore, the work of deepfake detection has become a research hotspot in recent years. However, with the development of deepfake technology, the quality of forgery videos and images is getting higher and higher within the visible range, and forgery traces are becoming more difficult to be mined. On the other hand, due to the high redundancy of image data information, a large amount of content information is not conducive to deepfake detection. Therefore, how to design an efficient and well generalized deepfake detection model to deal with a variety of forgeries has become the focus of current research. Based on this, we propose two different deeepfake detection models, namely deepfake detection model based on multi-domain feature network and deepfake detection model based on frame shifted window attention mechanism, to detect forgery from different angles.

In this paper, a multi-domain feature network is proposed, which combines the frequency domain information and RGB domain information, and introduces an attention mechanism to adaptively assign the weight of the feature map, and fuses the prediction information in different domains into the final result through the class score fusion. In addition, this paper also proposes a method of face alignment between frames by rotating and scaling the face calibration frame based on the face key points to improve the performance of 3D deepfake detection model. After a lot of experiments on Celeb DF and DFDC datasets, the experimental results show that the performance of the model after multimodal information fusion has been significantly improved, and the identification performance of the model in depth forgery video is close to the current most advanced level, and has a good generalization.

In this paper, a shifted window attention mechanism is proposed, which is used to calculate the self-attention between channels and frames respectively. Based on this mechanism, a 2D forgery detection model and a 3D forgery detection model are proposed respectively. This mechanism effectively reduces the computational complexity caused by high-dimensional vectors, and improves the detection performance by selecting more critical frame information. In addition, this paper also introduces an attention mask to solve the problem that the background area is too large and the redundant information is too high when the face is aligned. Experiments on deepfake video datasets show that the model has good classification performance and robustness.

Keywords: Deep learning, Deepfake Detection, self-attention, face-swap, shifted window

目 录

图表清单

注释表

R_u	非面部区域	R_f	面部非器官区域
R_o	面部器官区域	dim	通道线性映射后的维度
win	窗口数量	win_size	窗口大小
H	图像高度	W	图像宽度
C	通道数	X	原始图像
f_r	RGB 特征	d_i	频域分量图像
$mask$	掩膜	D	频域图像
L_{rgb}	RGB 域损失	L_{freq}	频域损失
$Loss$	模型总损失	P	批处理下的输出向量
Y	批处理下的标签向量	λ_1，λ_2，λ_3	损失权重系数

缩略词

缩略词	英文全称
GAN	Generative Adversarial Networks
CNN	Convolutional Neural Networks
RNN	Recurrent Neural Networks
rPPG	Remote Photoplethysmography
LSTM	Long Short-Term Memory
LRCN	Long-term recurrent Convolutional Networks
CAN	Convolutional Attention Network
ViT	Vision Transformer
CViT	Convolutional Vision Transformer
DFT	Discrete Fourier Transform
DCT	Discrete Cosine Transform
FAD	Frequency-Aware Decomposition
LFS	Local Frequency Statistics
MLP	Multilayer Perceptron
RGB	Red Green and Blue
FSW	Frame Shifted Window
CSW	Channel Shifted Window
MSA	Multi-Head Self-Attention

第一章　绪论

1.1 课题研究背景

深度伪造（Deepfake）是指基于深度学习，生成含有虚假的人像的图片或音视频的一种合成技术[1]。其既可以进行人物面部属性的编辑，也可以通过生成模型用目标人物的面部替换掉原视频中的人的面部，并保留原始面部表情和头部姿势不变[2,3]。得益于深度学习技术特别是生成模型[4]的飞速进步，加之大规模人脸数据库的可用性，深度伪造技术也取得了巨大的进步。

深度伪造的兴起是在2017年12月，一个Reddit用户发布了一个使用深度神经网络生成的女明星的换脸视频，而在此之后，各种深度伪造方法如雨后春笋般地出现。这其中，最主流的伪造方向是人脸替换[5]（Face Swap）。目前，人脸替换主要可由两种方式得到。一种是基于深度学习，通过生成模型，如变分自编码器[6]或生成对抗网络[7]训练得到。而另一种方法是基于3D人脸模型的重建方法[8]，该方法需要细致的人为后处理手段来弥补人脸重建时产生的伪造痕迹。

除了人脸替换，通过生成模型产生一张从未出现过的全合成的图像也属于深度伪造的范畴。此前通过生成对抗网络即可以学习特定约束下的生成图像，但其由于攻击性较弱，检测需求较低，往往不是深度伪造检测的首选目标。然而，扩散模型的兴起改变了这个状况。扩散模型相比其他模型生成的图像具备更加多样化，且无需后处理等等优势，但图像生成的质量稍差。但随着2020年DDPM方法[9]的提出，扩散模型在生成图像的真实度上也超过了生成对抗网络。DALLE2[10]这样的基于文本的图像生成模型也因为扩散模型的优化得到发展，其可以根据文本描述生成匹配文本内容的高质量全合成的图像，这为深度伪造检测带来了新的挑战。

而对于伪造者来说，人脸替换也存在着诸多技术难点。目标人物的图像和原始图像不仅光照、对比度和图像压缩编码方式不同，人物的头部姿态也会有所不同。如果不对上述的图像或人脸属性进行背景适应，伪造的图像往往具有强烈的嵌入感。甚至目标人物和原始图像中的人物的面部属性差异过大也会使得伪造的图像质量欠佳。而如果是视频伪造，伪造者还需要进行帧间平滑、生物生理特征（如眨眼、牙齿纹理、唇部运动）修复等工作。由于上述原因，同一种伪造手段得到的图像或视频往往质量参差不齐，需要进行人工筛选或者后处理以提高图像质量。而目标的开源深度伪造数据集中，往往都存在着这个问题。

1.2 课题研究意义

目前，深度伪造技术可以生成全合成的伪造人脸，也可以在非常高的分辨率下为被摄体的人脸设置不同表情和动作，使其表达所需的情感或者修改其面部表情。该技术可以为

影视作品提供更多可能性，让无法参演的演员可以出现在镜头内。然而，比起深度伪造技术的应用，其带来的更多的是社会问题甚至是政治危机。

由于当下有诸多深度伪造方法都进行了开源或提供了接口，几乎任何人都可以没有门槛地使用这项技术。这意味着，只要有用某个人的视频或照片，就可以在其不知情的情况下让他出现在某段视频中，毫无疑问这将很容易触碰到人们的隐私权和肖像权。2019 年，一款名为“DeepNude”的软件使用深度伪造技术将数万名女性的面部作为目标人物图像生成了色情视频以牟取盈利[11]。更可怕的是，该软件一经上线两个月其用户量就突破了万人。

现如今互联网已经取代了纸质媒介成为了最普及传播最快的信息交互渠道，而近年来发展迅猛的直播和短视频行业更是使得信息容易以病毒式地在某个圈子传播开来。一个伪造的视频可以让一个人说出他不曾说过的话，做出不曾做过的事情。2020 年，麻省理工公布了一条采用深度伪造技术合成的尼克松总统宣布美国阿波罗 11 号人物失败的影响。伪造的视频的将清晰度还原到了上世纪 70 年代的影像水平，从中根本难以用肉眼发现伪造的痕迹。可想而知，如果伪造政客或其他公众人物的音视频，让他们发表不当言论或扭曲事实抹黑他人，而当事人即便在事后辟谣，其公众形象也难以修复。更严重的，这些伪造视频甚至可能影响选举结果、引发谣言传播和社会不安等后果。

掌握能够快速准确地判断一个视频或图像是否真实的技术，对网络安全和社会安全都有这重要意义。而近年来，越来越多的国家和机构认识到了这个问题。互联网巨头脸书、微软等公司提供了 DFDC 数据集[12]，并在此基础上联合举办了多次深度伪造检测挑战赛以鼓励深度伪造检测技术的发展。我国也在 2019 年发布的《网络音视频信息管理服务规定》表明网络音视频服务提供者应部署检测音视频真伪的鉴别技术。

深度伪造和伪造检测两种技术往往呈交替上升的趋势发展。早期的伪造视频还无法处理好眨眼和唇形等细节，伪造检测往往就可以依据这些生物生理特征去进行检测。而如今的深度伪造在质量和数量上都在不断上升，不论是人工检测还是早期的传统检测方法，都无法应付当前最先进的伪造技术。而如果伪造视频的背景是在复杂的室外场景下，检测的难度也会增大，并且会随光照方向、亮度变化等环境条件的变化表现出高敏感性。因此，当前主流的伪造检测方法大都是基于深度学习的模型。而当前的主流算法中并不具有很好的鲁棒性。例如，基于检测伪造边界的算法受背景影响程度较大，且不适用于检测全合成的伪造图像。而基于数据驱动的方法面对不同的伪造手段时，又难以提取出具有泛化性的特征或伪造痕迹。

此外，对于伪造视频的检测存在着如何由图像的检测结果得到视频的检测结果的问题。常见的做法有提取连续帧序列进行 3D 检测或通过视频关键帧的检测结果融合得到视频的检测结果。前者的缺点是需要将不同帧的人脸进行精细的对齐，且当视频中的人物运动幅度过大时会影响性能，而后者则需要采用均值法或者运动分析法来进行关键帧的筛选。因此提供一种简洁有效的方法来解决从图像级预测得到视频级预测的问题也十分重要。

1.3 国内外研究现状

1.3.1 深度伪造方法及应用

人或人像的深度伪造主要可以分为两种：传统方法和基于深度学习的伪造。传统方法需要人工筛选出肤色、年龄、姿态相近的目标人物图像和原人物图像。再通过仿射变换，将目标人物和原始人物的面部关键点对齐。再裁剪、拼接的方式将目标人物复制到原图像上。而为了使得伪造质量更高，需要进行一些后处理操作。如需要对拼接边缘进行高斯模糊以消除边界，对目标图像做渲染保持色彩和对比度一致等。然而传统方法需要耗费大量的手动操作，每张人脸的仿射变换、后期的渲染带来的工作量都十分大，不仅费时费力，而且不可以进行批量处理。

因此当前主流的深度伪造数据集都是通过深度学习的方式生成的。而目前的三种生成模型：自编码器[13]、生成对抗网络和扩散模型都可以很好地完成这个任务。其中，早期的开源软件 FakeAPP、DeepFaceLab[14]就采用了自编码器的结构。该方法需要两个“编码器-解码器”结构，分别用于训练目标人物图像与原始人物图像，且两个模型的参数会进行共享。通过“编码器-解码器”的结构让模型学习如何在给定引入噪声的条件下恢复出人脸。

而基于生成对抗网络的方法则更广泛。常见的如 Face Swapping GAN[15] (FSGAN)以及 Region-Separative GAN[16] (RSGAN)都可以被用来进行人脸替换和表情迁移。以 Face Swapping GAN 为例，该方法在训练过程中，通过两个共享编码器的自编码器来重建源图像的面部和目标图像的面部。学习后的源图像的编码器和解码器会将源图像的特征转换为目标图像的特征。

而当前火热的扩散模型则多被用于文本图像生成模型中。2021 年，OpenAI 发布了他们的文本图像生成模型 DALLE[17]，在当时就引发了热议。该模型可以基于文本描述自适应地生成对应文本内容的图像。不过该模型生成图像的质量还有待提高，且无法应对复杂的描述文本。但紧接着在 2022 年 5 月，OpenAI 又发布了 DALLE2，DALLE2 生成的图像的质量和超越人们想象的创造力使得其迅速爆火。这其中 DALLE2 的图像生成模型 GLIDE[18]抛弃了传统的自回归模型，转而使用了扩散模型并取得了更好的效果。不仅可以更快地生成了分辨率更高的图像，还支持对生成的图像进行编辑。

1.3.2 深度伪造检测方法及应用

人脸视频鉴伪一般会被建模成一个二分类问题，对视频帧中的的面部图像区域进行真伪加测，将对应帧或帧序列的真实性进行分类。主流的鉴伪的方法一般有两种，一种是基于挖掘潜在的伪造痕迹的方法，另一种是基于数据驱动的方法。

早期的伪造视频质量较低，基于挖掘潜在的伪造痕迹的方法会采用检测面部生理特征的伪影（Artifacts）的手段来辨别真伪，如通过是否眨眼或眨眼频率的真实性，以及牙齿纹理特征等来进行鉴伪[19]。而随着深度伪造技术和伪造检测技术的交替发展，很快在伪造时

这些生物生理特征便被很好地再现了，因此这些基于一些简单的生物生理特征去检测视频真伪的方法存在着时效性，且不具备很好的泛化性。

而一些更复杂的生理特征是很难伪造的，例如目前还不能在伪造时使得唇部运动的连贯性达到真实视频的标准，因此即便是在伪造技术高度发展的 2021 年，Haliassos 等人[20]依然通过学习高层的唇部时序语义特征来判断视频的真实性，并在跨数据集测试时展示了模型的高泛化性。此外，Yang 等人[21]根据头部 3D 姿态的一致性进行判别。因为在进行人脸替换时，头部的 3D 姿态会出现伪影或不一致性，作者通过人脸关键点估计得到的头部姿态进行分类进而判断视频的真假。

随着伪造视频的精细化，现在的伪造视频一般不再具有这些容易被识别的生理特征上的伪影，基于挖掘潜在的伪造痕迹的方法把研究重心转移到了更深层的语义特征上，而这往往需要同基于数据驱动的方法结合起来。

Zi 等人[22]认为伪造痕迹容易出现在面部器官这种高频部分，遂通过构建人脸面部器官的注意力掩膜，将掩膜与原始图像做点乘，引导网络关注面部器官部分。随后，Li 等人[23]提出了 Face X-ray 算法，通过自生成的数据训练并定位伪造视频的换脸边界，该方法仅针对换脸的伪造视频特别有效。

近年来，人们试图将频域特征引入到视频鉴伪的工作中来。Durall 等人[24]通过方位角平均算法提取频域特征，并在向量空间上用 SVM 等常见的分类器进行分类，也取得了不错的分类效果。2020 年商汤团队提出了 F3Net 网络，通过了引入两种频域特征提取方法 FAD (Frequency-Aware Decomposition) 和 LFS (Local Frequency Statistics)设计出了一种基于频域特征的双流网络用于深度伪造的检测。文献[25]和文献[26]采用双流网络融合了频域特征和 RGB 特征，并分别提出了新的损失函数以达到减少类内差距加大类间差距的目的。Wang 等人[27]将特征多尺度的思想引入到 RGB 图像的处理中，通过注意力机制结合了不同尺度的特征学习，提高了模型的泛化能力。

纯粹的基于数据驱动的方法则往往是一些有高泛化性和可扩展性的模型。Afcha 等人[28]根据常见的几种伪造方法提出了一种轻量化的伪造检测网络 MesoNet，并参考 Inception 网络的结构优化了 MesoNet 模型。Chollet 通过解耦通道相关性和空间相关性推导出深度可分离卷积，并设计出了 Xception 网络[29]进行深度伪造检测，该网络被作为性能优秀的 baseline 沿用至今。Rossler 等人[30]提供了大型深度伪造数据集 FaceForensic++，并比较了几种简洁的卷积神经网络在深度伪造鉴伪上的性能。Dang 等人[31]利用注意力机制在定位伪造区域后，引导网络关注关键区域并对伪造区域进行了定位。

1.4 本文的研究工作

本文主要研究的是通过提取多维度的图像特征来对伪造的视频和图像进行真实性的判断。事实上，本文主要通过强化伪造痕迹的挖掘、从多维度抽取图像特征和优化特征提取的方式，设计出有更高性能、更好泛化性的伪造检测模型。通过本文的实验以及相关项目

的落地，可以表明本文的工作具有良好的应用价值。

现有的伪造检测算法大都从单个域中取挖掘伪造痕迹或全局一致性特征，然而某些伪造算法生成的图片可能会在某个域内的质量非常高，也即无法在这个域内有效挖掘伪造痕迹，且不同位置方法也会在不同的域内产生不同程度的伪造痕迹。因此单单只考虑时序特征，或只考虑频域特征这样的方法在检测特定伪造方法产生的图片时可能会出现精度较差的情况，且泛化性较差。本文通过了离散傅里叶变换提取频域特征并结合空域特征，设计了一个双流网络，并通过类别得分融合两个分支的信息，且两个分支的权重可以进行调节。

在视觉任务中，滑动窗口的注意力机制相比传统的注意力机制，学习了卷积神经网络的层次性，不再一次性地进行全局建模，而是层次化的建模。此外，滑动窗口的注意力机制通过划分并移动多个二级窗口，有效减少了计算复杂度。事实上基于滑动窗口注意力机制的 SWIN Transformer[32]在视觉领域的各种任务上都取得了优异的表现。而在深度伪造检测任务中，需要更加关注可表达全局一致性的语义特征并且需求更加轻量化的检测模型。因此本文通过 CNN 进行特征提取，并基于特征图设计了通道间和帧间的滑动窗口注意力，针对时序信息、空域信息等多个维度让模型自适应地学习注意力权重。

在进行视频级别的真伪检测时，往往需要先进行单帧的预测，再根据帧级别的预测得到视频级别的预测。通过帧级别的预测得到视频级别的预测一般需要先对视频进行关键帧的选择，再优先对被选择的关键帧进行预测，这不仅带来了额外的开销，还需要根据不同的任务制定不同的关键帧选择策略。另一种方法则是对视频中的所有图像帧进行随机采样，如对于 3D 伪造检测模型，则可以采样数段连续的帧序列。但这样的做法可能会丢失关键的帧语义信息，使得视频真伪预测的精度下降。本文提出了一种基于图像结构信息强化和时序窗口注意力机制的模型，该模型提供了更加简洁的关键帧选择功能，并对图像的结构信息进行了增强，使得模型更关注容易产生全局一致性语义的结构信息。

1.5 本文的内容安排

本文可分为五章，每章的结构和内容如下所示：

第一章：绪论。主要介绍选题的背景和意义，阐述深度伪造技术和伪造检测技术的发展历程和国内外的研究现状。

第二章：深度伪造检测技术理论基础。介绍深度伪造和深度伪造检测的基础理论，包括了一些经典的伪造和检测的算法，针对深度伪造的过程介绍了一些可以被用于检测的先验知识。根据伪造检测的主要难点给出一些常见的解决方法，并分别给出了 CNN 和 Transformer 的构造和改进方案。

第三章：基于多域特征网络的深度伪造视频检测。提出了一个多域特征网络，结合了频域信息和 RGB 图像，并引入了注意力机制对特征图进行了自适应的权重分配，减少了冗余特征；通过类别得分融合将不同域的预测信息融合成最终的结果。在 Celeb-DF 和 DFDC 数据集上的实验结果表明，进行多模态信息融合后的模型性能有了显著提高，并且本文的

模型在深度伪造视频的鉴别性能优秀，并具有很好的泛化性。

第四章：基于滑动窗口注意力的深度伪造视频检测。提出一种基于帧间滑动窗口的自注意力计算方式以减少计算复杂度，并将其用于 3D 伪造检测模型，通过 Transformer 表征时序特征，代替了关键帧提取的工作，在视频级预测上有着更好的表现。同时，该模型还采用了通道间的滑动窗口注意力机制用于 2D 伪造检测。另外，该模型还引入了一种层级式注意力掩码来解决背景区域冗余度过高的问题。

第五章：总结与展望。对本文的工作进行概括性的总结和评价，并分析相关工作中未工科的难点，提出后续的研究方向。

第二章　深度伪造检测技术理论基础

2.1 人脸伪造

2.1.1 人脸伪造基础

深度伪造技术可以生成全合成的伪造人脸，也可以在非常高的分辨率下为被摄体的人脸设置不同表情和动作，使其表达所需的情感或者修改其面部表情。一般而言，人脸伪造方法可以分为四类：表情交换、身份交换、属性篡改和全脸合成[33]。

表情交换是指将原始人物的实时表情递到目标人物的面部，使其显露出于原始人物一致的表情。其中，以Face2Face[34]为代表的三维人脸重建和动画方法被广泛用于表情交换。

身份交换是指将原始人物的脸部替换为目标人物的脸。这是运用最多也是伪造数量最多的伪造方式。常见的方法如 FaceSwap[11]，该方法可用于影视拍摄，让无发进行拍摄的演员出现在影视作品中。

属性篡改是指编辑人物面部中的单个或多个属性，例如年龄、肤色、毛发和眼镜等。其一般通过 GAN 来完成，如 StarGAN[35]就是一个经典的用来进行面部属性修改的模型。FaceApp 将面部属性操作作为一个消费者级应用程序进行了推广，它通过提供 28 个过滤器来修改特定的面部属性，包括添加眼睛修改发型等。

全脸合成是指在没有目标身份的情况下，通过随机噪声合成人脸。以大量人脸数据为基础，生成对抗网络或是扩散模型能够生成一个完全合成的人脸图像，其真实性使得人类难以辨别。全脸合成往往被用于 AI 创作、电子游戏和 3D 建模领域中，其伪造手段目的性较弱，带来的社会危害也不如其他几种伪造方式。

2.1.2 基于传统方法的人脸伪造

人脸变换[36](Face Morphing)是一种经典的人脸伪造方法。它可以将原始人物的脸以一种渐变的形式换成目标人脸。该过程中的任一张图都是由原始人脸 S 和目标人脸 T 通过参数控制来混合叠加得到的，通过改变参数，可以得到过程中的某一张脸 M，使其看起来既具备原始人脸 A 的特点 又具备目标人脸 T 的特点。

人脸变换通过需要先在视频或图像中检测出人脸，进而对人脸中的关键点进行标定，如通过 OpenCV 或 dlib 工具可以标定出人物面部的 68 个关键点。再对人脸区域按照人脸关键点划分成多个不相交的多边形区域。由于原始人脸和目标人脸的特征点通常位于不同的位置，在进行人脸之间的变换时，必须对图像进行仿射变换，使两者的特征点匹配。否则，变形后的图像将出现面部器官的重叠。最终得到的面部图像是介于目标图像和原始图像之间的一个中间图像。原始人脸 A 和目标人脸 T 的权重分别表示为为 α 和 1-α。对于图像 A

中的一个特征点 a，以及图像 T 中对应的特征点 b，可以采用线性插值的方法生成新的特征点 f 的位置。如公式 2.1 所示。

$$f = \alpha a + (1-\alpha)b \tag{2.1}$$

事实上人脸变换并没有创造出一个新的人脸，而只是通过调节原始图像和目标图像的权重得到了一个介于两者之间的人脸。当然，也可以将 α 设置为 0，从而形成人脸替换。

除此之外，为了提高图片质量，一些后处理也经常被用于伪造图像的生成中。通过高斯模糊可以对区域的边缘做虚化的操作，从而减少突兀的伪造痕迹的产生。即便如此，该方法想要生成质量较高的伪造图像，对原始人物和目标人物的面部的相似性要求也比较高，包括了年龄、肤色、光照和头部姿态等，显得十分不灵活。因此，该方法不论是应用范围还是潜在的危害程度都比较有限。

2.1.3 基于深度学习的人脸伪造

2.1.3.1 基于 GAN 的人脸伪造

对抗生成网络（Generative Adversarial Networks，GAN）是一种应用广泛发展成熟的生成模型，其最早由 Goodfellow 在 2014 年提出。GAN 网络主要由生成器 G 和分类器 D 组成。生成器对数据分布建模，以高斯分布采样并生成图片，并使得生成的图片尽可能地接近真实样本以使得分类器无法做出正确判断；而分类器则对样本来自训练数据的概率进行估计，需要学习区分图片是否由生成器生成的虚假的图片还是真实图片。在训练时需要轮流更新定生成器和分类器的参数，同时保持另一部分的参数固定。GAN 网络的优化目标如公式 2.2 所示：

$$min_G max_D V(G,D) = E_{x\sim Pdata}[\log D(X)] + E_{z\sim P_z}[\log(1-D(G(z)))] \tag{2.2}$$

在完成对生成器的训练后，损失函数将近似为真实分布与当前分布的 JS 散度，再对生成器训练以最小化两个分布间的距离。当两个网络达到平衡，即达到纳什均衡时训练结束。理论上，最终生成网络将再现真实数据分布，判别网络将退化为随机预测。

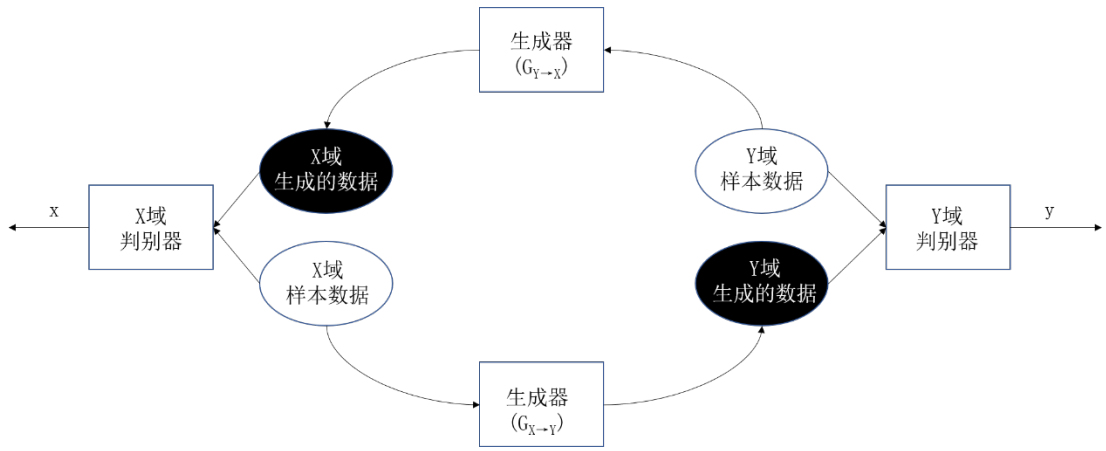


图 2.1 Cycle-GAN 的计算流程

Cycle-GAN[37]是在 2017 年被提出的一种风格迁移和域自适应算法，是一种基于 pix2pix 改进的方法。pix2pix 需要成对的数据，而现实中很难确保两个域中存在成对的图像，Cycle-GAN 的提出实现了无需配对样本即可实现不同域之间的转换，而这个特性使得其很可以很好地被用于人脸替换中。Cycle-GAN 由两个 LS-GAN[38]组成，因此包括了两个生成器和两个判别器，每对生成器-判别器各自处理域 X 到域 Y 的转换和域 Y 到域 X 的转换。首先将图像从域 X 转换到域 Y，然后再经过另一个 LS-GAN 转换回来。通过学习让两次转换时减小误差，最终经过两次转换后的图像会与原始输入图像保持近似。通过这样的一个循环，可以将转换后图像和原始图像的配对。其结构图如图 2.1 所示。Cycle-GAN 的损失函数表示为：

$$L_1(G,F) = E_{x\sim Pdata(x)}[\| F(G(x)) - x \|_1] + E_{y\sim Pdata(y)}[\| G(F(x)) - y \|_1] \tag{2.3}$$

$$L_2(G,F,D_X,D_Y,X,Y) = L_{GAN}(G,D_Y,X,Y) + L_{GAN}(F,D_X,X,Y) \tag{2.4}$$

其中L_1和L_2分别表示对抗损失和循环一致损失。循环一致损失描述了重建图像和原始图像之间像素级别的差异，是一种比较严格的损失。

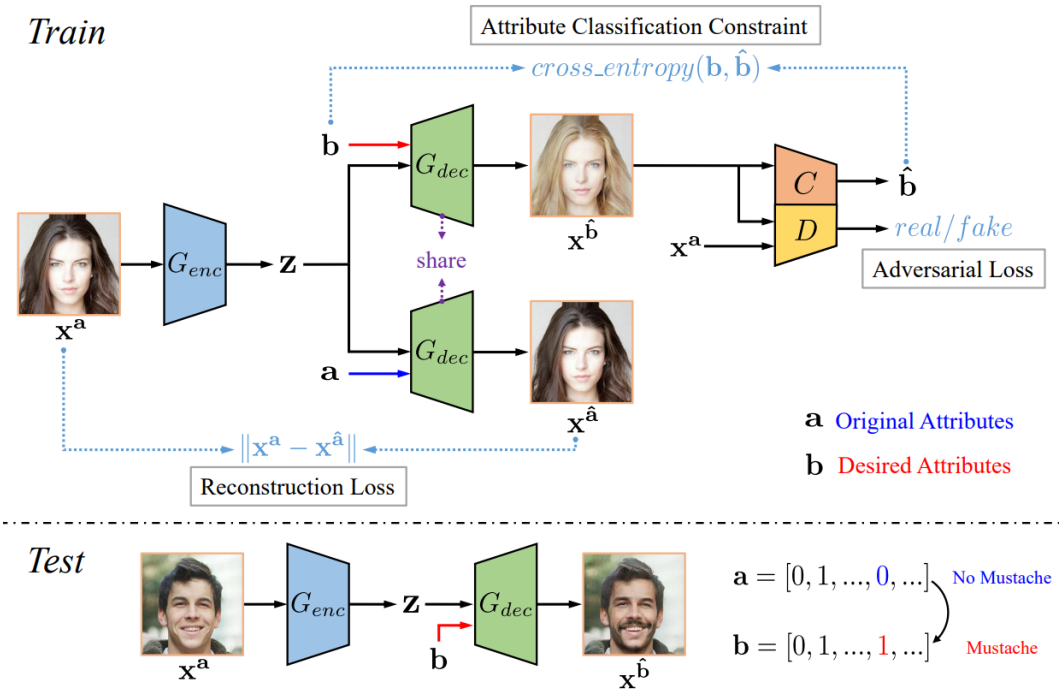


图 2.2 AttGAN 的结构图[37]

事实上，在使用 GAN 进行伪造人脸的生成时，由于面部属性间具有较强的关联性，之前的 GAN 系列方法并没有考虑到这一点，生成的图像会存在许多瑕疵。例如，因为训练集中大部分拥有“光头”这一属性的对象是男性，即“光头”和“男性”这两个属性具有高度关联性，因此模型在进行“编辑成光头”的操作时具有将女性转换成男性的倾向。AttGAN[39]为此提出了属性分类约束（Attribute Classification Constraint）的思想，即通过改变属性约束

来实现人脸属性的编辑，并通过设计一个分类器来对编辑后的伪造图片做出属性上的分类，以保证在进行人脸编辑的时候，人脸的其他属性不被错误的转变。其损失函数包含了对抗损失和重构损失。前者保证了生成的图像的真实性，后者则保证了人脸属性不被错误地进行修改。其本质上是由属性分类约束、对抗学习和重构学习构成的一个多任务学习模型。AttGAN 的结构图如图 2.2 所示，其由两个子网络组成，包括了编码器 G_{enc} 和解码器 G_{dec}，以及属性分类器和判别器。

上述方法大都是用于人脸编辑或表情迁移等工作中的，而 FSGAN 还可以进行身份交换。其结构图如图 2.3 所示，模型分成了重演与分割、图像修复和图像融合三个部分。其中 I_s 表示原始人脸， I_t 表示目标人脸。G_r 为重演模块，会将原始人脸重制成目标人脸的姿态，并得到重演后的图 I_r。在这个过程中，为了使得重演的人脸更加贴近目标人脸的姿态，该方法会参考整个目标人脸对应的整个视频的序列信息； G_s 为分割模块，对目标人脸进行背景分割与实例分割，将背景与前景分割后，再将前景分割为面部、头发等区域，得到分割图 S_t； G_c 为图像修复模块，将 重演后的图I_r根据 分割图 S_t 进行更加细致的修复； G_b为图像融合模块，将修复后的图像 I_c 与目标图像融合，最终得到换脸后的图像 I_b 。

FSGAN 提供了一种真实度高、泛化性好的人脸替换模型，同时通过重演人脸很好地解决了侧脸的问题。但它并不是一个端到端的模型，整个算法需要设计四个生成器，也就意味着需要单独训练多个模型。并且，当面对复杂的人脸姿态时，生成的图像质量会显著下降。并且，FSGAN 在捕捉人脸细节特征上做的并不如一些三维模型。

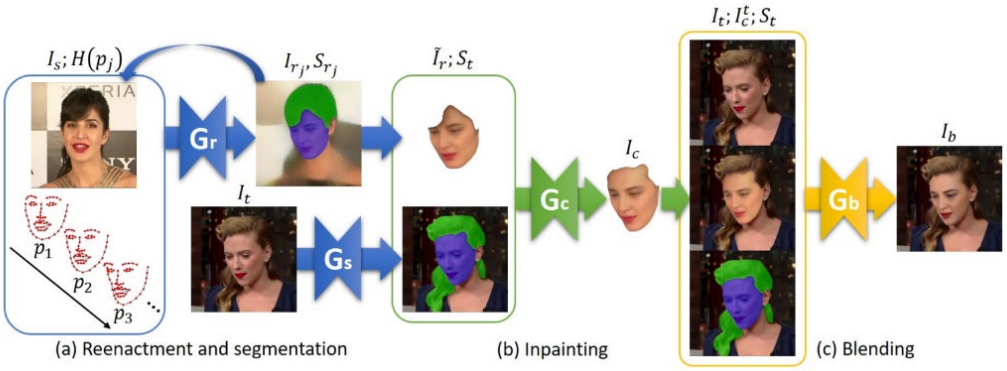


图 2.3 FSGAN 的流程图[15]

2.1.3.2 基于 3D 人脸重建的伪造

Face2Face 算法是一个经典的基于 3D 人脸重建的伪造算法，是一种基于单目标 RGB 图像的面部密集标记表情捕获方法。然而，Face2Face 不是将面部表情转移到目标人物上，而是实时的面部重现。具体来说，先分别将 2D 的原始对象和目标对象的人脸关键点隐射到 3D 人脸上，得到原始对象的 3D 点集 A 和目标对象的 3D 点击 B。再根据 A 和 B 之间的映

射关系转换回2D坐标中，并以此进行仿射变换或透视变换，从而完成表情的编辑。其方法的结构图如图 2.4 所示。该算法提出了全局非刚性模型约束（Global Non-rigid Modelbased Bundling Approach）来重建目标角色的外形特征。缓解了重构面部时出现的几何模糊问题。而对于图像合成，该方法通过转移的表情系数重新渲染目标的脸部，并将其与目标视频的背景进行合成，并且这个合成过程同时考虑了光照的情况。

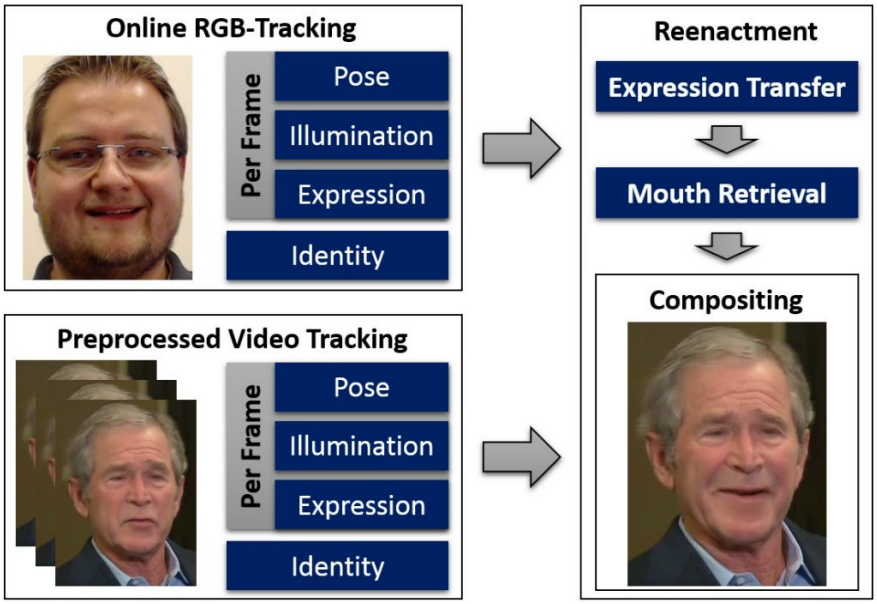


图 2.4 Face2Face 的结构图[34]

此外，该方法还提出了一种新嘴部合成方法，通过从离线样本中检索出最佳匹配口型并进行变换来产生逼真的嘴部图像，并且该方法不会改变唇部的形状。Face2Face 算法可以生成逼真的人脸编辑后的图像和视频，并且在面部器官中很少存在伪影。然而如果要保证生成视频的质量和真实度，则需要有足够充分且足够多样的被编辑对象的数据。

2.2 基于生物信号或生理特征的伪造检测方法

2.2.1 基于眨眼检测的伪造检测方法

眨眼是指眼睑快速闭合和张开的动作。它主要分为三种类型，自发眨眼、反射性眨眼和自愿眨眼。自发眨眼是指在没有外部刺激和内部努力的情况下眨眼，由运动前脑干控制，在没有意识努力的情况下发生。自发眨眼是一种重要的生物功能，它可以通过眼泪来保湿，去除角膜和结膜表面的刺激物。对于一个健康的成年人来说，每次眨眼之间的时间间隔一般为 2 至 10 秒，实际的眨眼频率因个人而异。

Li 等人[40]提出了一种基于眨眼真实性检测的伪造检测方法。使用人脸检测器在视频的

每一帧中定位人脸区域后，从每个检测到的人脸区域中提取人脸标志，这些标志是人脸上承载重要结构信息的位置，如眼睛、鼻子和嘴巴的尖端以及脸颊的轮廓等。视频帧中头部的运动和脸部方向的变化会影响面部分析。因此，作者使用了基于关键点的人脸对齐算法将人脸区域对齐到统一的坐标空间。变换后的人脸应该位于图像中心，眼睛处于水平线上，并且被缩放到相似的大小。从对齐的人脸区域中，可以提取出一个基于人眼轮廓的标志点形成的矩形区域，构成一个新的帧序列，如图 2.5 模块（b）所示。矩形区域是通过提取每只眼睛的标志点的边界框，并在水平和垂直方向上将边界框放大 1.25×1.75 倍得到的，这样以保证眼睛区域被包含在裁剪区域中。

人类眨眼表现出很强的时间依赖性，因此作者采用长期循环卷积网络[41]（Long-term Recurrent Convolutional Networks，LRCN）模型来捕捉这种时间依赖性。LRCN 模型由三部分组成，即特征提取、序列学习和状态预测。

特征提取的结果被输入到序列学习中，序列学习由带有 LSTM 单元[42]的 RNN 实现。LSTM-RNN 的使用可以增加 RNN 模型的存储容量，避免产生梯度消失。LSTM 控制何时和如何忘记此前的隐藏状态以及何时和如何更新隐藏状态的内存单元。在最后的状态预测阶段，每个 RNN 神经元的输出被进一步发送到由一个完全连通层组成的神经网络，该层接收 LSTM 的输出并产生眼睛打开和关闭状态的概率。

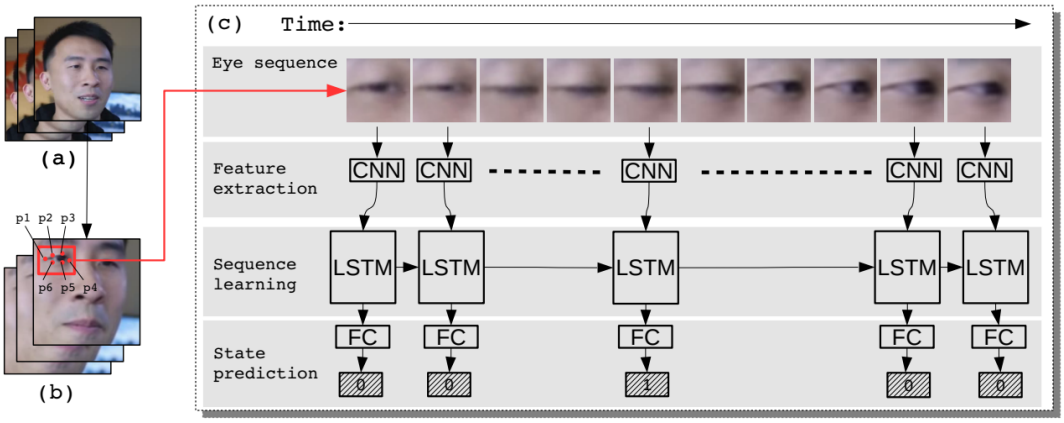


图 2.5 基于眨眼真实性检测的伪造检测方法的结构图[40]

该方法只使用人脸伪装视频中是否出现眨眼作为检测的线索，并不具有很强的泛化性。而随着伪造技术的发展，当前的伪造技术产生的视频完全可以真实地再现人的自然眨眼的过程，因此该方法只在早期对于一些特定的伪造方法具有很好的识别效果。

2.2.2 基于生物心率信号的伪造检测方法

事实上，人脸伪造的直接目的是模仿真实人脸的视觉外观，但人类的一些肉眼不易察

觉的生理学特征却是难以复现，例如心率、血氧或呼吸频率等等。在真实的视频中，人物的面部存在的正常心跳节律在伪造视频中会遭到破坏，这使得心率成为一个潜在的强大的伪造检测指标[43]。氧气浓度引起的颜色和光照的变化非常微妙，人眼无法看见，而基于远程视觉光容积描记术（Remote Photoplethysmography，rPPG）可以通过监测脸部血液泵送引起的微小周期性肤色变化估计心率。

而近年来由于 rPPG 技术，远程心率测量技术取得了巨大进步，也出现了许多基于 rPPG 进行人脸伪造检测的工作。Hernandezo 等人[44]提出了一种基于人类心率信号的深度伪造检测模型 DeepFakesON-Phys。DeepFakesON-Phys 的初始架构基于 DeepPhys 模型，使用 DeepPhys 层的权重作为初始化，并使用微调使其适应人脸伪造检测。与 DeepPhys 模型从头开始训练相比，它不需要大量的样本。

DeepPhys 则是一个卷积注意网络[45]（Convolutional Attention Network，CAN），由两个并行的 CNN 网络组成。该模型从视频帧中提取并共享时间和空间信息估计人类心率。它通过血液中氧浓度变化引起的人脸颜色变化提取特征，信号处理方法用于将血液引起的颜色变化与其他可能由外部照明、噪声等因素引起的变化隔离开来。DeepPhys 在处理非均匀光照或低分辨率等具有挑战性的条件时显示出很高的精确度。

其结构如图 2.6 所示。在第一预处理阶段之后，卷积注意网络由两个并行的 CNN 分支组成：

1）运动模块：设计用于检测连续帧之间的变化，即对视频进行短时分析以检测伪造。为了完成这个任务，时刻 t 的输入由单个帧组成，该帧被计算为当前帧 $I(t)$和上一帧 $I(t\text{-}1)$的归一化后的差值。

2）外观模块：重点分析每个视频帧的静态信息。该分支的功能是向运动模型提供关于当前帧的哪些点可能包含用于检测深度伪造的最相关信息，即在 CNN 的不同层上共享的一批注意力掩码。该分支在时间 t 的输入是视频 $I(t)$的归一化后的原始帧。

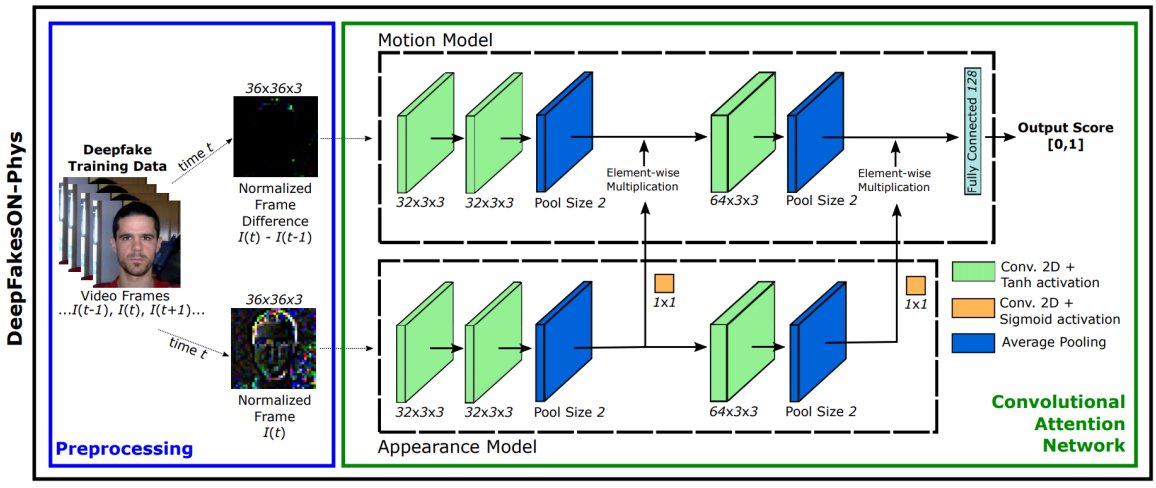


图 2.6 DeepFakesON-Phys 的结构图[44]

2.2.3 基于唇部运动规律的伪造检测方法

即便伪造技术发展的势头很快，如今可以生成肉眼不可区分的伪造视频图像的技术层数不穷。但即便如此 Haliassos 等人[46]还是基于唇部运动规律设计提出了一种伪造检测模型，LipForensics。LipForensics 的目标是学习到唇部运动中的高级语义规则。作者先对唇读任务上预训练一个逐帧的 ResNet-18[47]特征抽取网络和一个时序网络，通过固定逐帧的特征抽取网络，而只训练时序网络。从而抽取能够决定真假的唇部运动的语义特征。实验表明该方法具有很好的泛化性和抗攻击性。

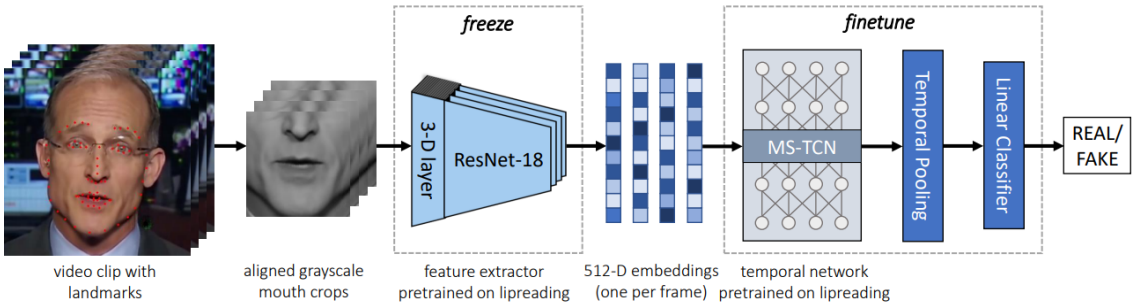


图 2.7 LipForensics 的结构图[46]

2. 3 基于伪造痕迹挖掘的伪造检测方法

受限于面部伪造算法，伪造的人物视频中，往往会出现可以被挖掘的伪造痕迹。伪造痕迹一般可以分为纹理痕迹和语义痕迹。纹理痕迹包扩了伪影和几何失真等。例如对于 Face2Face 算法，由于存在着面部几何体的尺寸误差，目标人物在覆盖到原图像上时会出现沿遮盖边界的伪影。这些伪影表现为强边缘或高对比度的斑点或阴影，常常出现在面部器官或面部遮挡区域的边界。而语义痕迹则需要通过提取深层的语义特征，对全局一致性作出判断，常见的包括色彩、光照、人物属性、GAN 指纹的一致性分析。

2.3.1 基于 GAN 指纹的伪造检测方法

在 GAN 网络的训练过程中，因为目标函数非凸以及 GAN 中生成网络和判别网络之间对抗式平衡的不稳定性，模型权重值对随机初始化非常敏感，在每次训练都有可能收敛到不同的值。因此，两个经过良好训练的 GAN 模型即使性能相当，也会生成不同模式的高质量图像。GAN 生成的图像是大量固定的滤波和非线性过程的结果，这些过程在相同的 GAN 实例中会存在共同的和稳定的模式，而在不同的 GAN 实例中其所存在的模式却截然不同，这一模式可以被称为 GAN 指纹，利用 GAN 指纹可以同时解决图像检测和溯源的问题。

Yu 等人[48]通过真伪分类神经网络的最后一层来表示 GAN 指纹，并通过解码器对每张图像中的 GAN 指纹的进行还原。该算法不仅通过 GAN 指纹来对图像进行了真伪分类，对于伪造图像的 GAN 指纹进行溯源，判断其是哪张 GAN 模型生成的图像。

Yang 等人[49]对于伪造图像的溯源任务建模成一个多分类问题，分别进行了结构分类和参数分类并对模型进行了训练。对于结构分类，训练数据来自 4 种不同的 GAN 模型，包括 ProGAN[50]、MMDGAN[51]、SNGAN[52]和 InfoMaxGAN[53]，四个 GAN 模型均在 CelebA[54] 数据集上训练得到。而对于参数分类，训练数据来自四个不同参数的 GAN 模型，具体来说是 ProGAN 在 4 个不同初始化种子下对 CelebA 数据集训练得到的 4 个模型。该工作通过实验论证了 GAN 的参数指纹在图像不同位置上是不一致的，以及结构指纹可以通过全局一致性的伪造痕迹来对图像进行真伪分类和溯源的。

2.3.2 WildDeepfake

2020 年，Zi 等人[22]提出了一种基于面部注意力掩膜的伪造检测方法 WildDeepfake。作者认为，伪造的图像在面部往往会出现伪造痕迹，而这样的伪造痕迹在眼睛、鼻子、嘴巴等面部器官中出现的概率更高。该算法分别对面部图像和面部图像对应的注意力掩膜设计了特征提取网络，并在对面部图像提取特征的过程中，不断地将同一层的面部图像对应的特征图和注意力掩膜对应的特征图做按元素乘法。最后提取到的带掩膜的特征将通过一个判别网络对图像真实性进行分类。除此之外，作者还分别设计了 2D 和 3D 的网络分别用于处理对图像和视频的真伪分类。

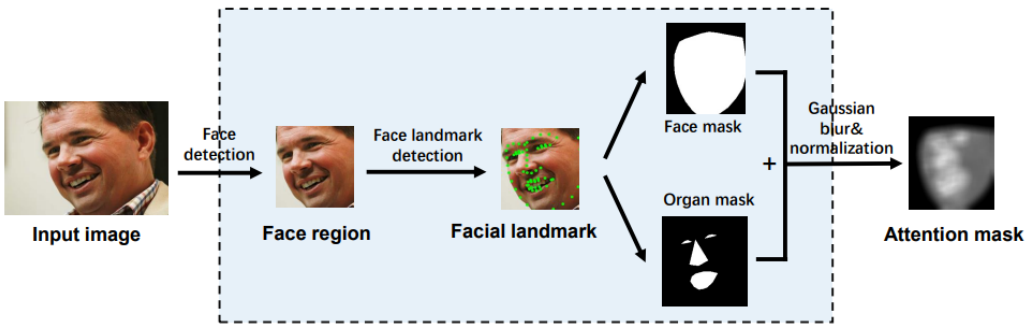


图 2.8 注意力掩膜生成示意图[22]

通过面部图像生成对应的注意力掩膜的过程如图 2.8 所示。作者先对人脸图像进行 68 个人脸面部关键点的标注，再通过面部关键点分别划定面部掩膜和面部器官的掩膜。又因为伪造痕迹在面部不同区域出现的是呈不同概率分布的，简单的 0-1 掩膜不能很好地表达伪造痕迹的分布情况，因此作者通过高斯模糊使得注意力掩膜在边界区域的变化更加平滑。

具体来说，注意力掩膜的像素值按照面部器官区域、面部区域和非面部区域的顺序依次减小。

2.3.3 Face X-ray

2020 年，微软亚洲研究院提出了 Face X-ray 算法[23]，该算法可以判断输入图像是否可以分解为两张不同源图像的混合的灰度图，理论上可以在伪造图中展示融合边界，在真实图中呈现空白。因此，它不仅可以确定图像是否伪造，而且能够进行伪造区域的定位。Face X-ray展示出了很好的泛化性，因为它只依赖于伪造时原始图像和目标图像融合步骤的存在性，而不基于任何特定面部伪造技术的伪影知识，对训练集中不存在的伪造技术生成的图像也有很高的分类精度。此外，Face X-ray可以利用大量由真实图像合成的混合图像进行训练，而不需要任何实际上由人脸伪造方法生成的伪装图像作为训练数据。

典型的人脸伪造算法通常包括检测面部区域、合成目标面部、将目标面部融入原图这三个步骤。当前的人脸伪造检测方法一般都集中在第二阶段，基于包含人脸伪造视频和真实视频的数据集训练有监督的二分类器。虽然它们在测试数据集上获得了近乎完美的检测精度，但当将训练模型应用于新的伪造算法生成的图像时，性能就会显著下降。Face X-ray不再执着于第二步中特定方法生成的伪影，而是尝试定位第三步中更加通用的融合边界。

对于给定输入的面部图像 I，Face X-ray 会检测图像是否可以通过两个来源的图像 I_F 和 I_B 而组合成伪造图像 I_M。则图像合成过程可定义为公式 2.5 所示：

$$I_M = M \odot I_F + (1 - M) \odot I_B \tag{2.5}$$

其中，⊙表示按元素乘法，I_F 表示具有所需面部属性的前景伪造人脸图像，即目标人物对应的图像。I_B 表示背景图像，即原始图像。M 表示掩膜，用以限制伪造的边界，掩膜的像素的取值范围在 0 和 1 之间。

将 Face X-ray 的输出定义为图像 B。如果输入一个伪造的图像，那么 B 将通过像素值显示伪造区域；如果输入是一个真实的图像，那么 B 则会是一个像素值为全 0 的图像。对于输入图像 I，其对应的 Face X-ray 输出图像可定义为：

$$B_{i,j} = 4M_{i,j}\left(1 - M_{i,j}\right) \tag{2.6}$$

其中，下标(i, j)表示像素位置的索引，M 是由输入图像 I 确定的掩模。如果输入图像是真实的，则掩模 M 是像素值全为 0 或全为 1 的图像。否则，掩模 M 将是可表示前景图像区域的非平凡图像。$M_{i,j}\left(1 - M_{i,j}\right)$的值不大于 0.25，并且仅在$M_{i,j} = 0.5$时达到最大值 0.25。因此，Face X-ray 的输出图像的像素$B_{i,j}$的值总是介于 0 和 1 之间。掩模 M 和输出图像 B 之间的关系如图 2.9 所示。

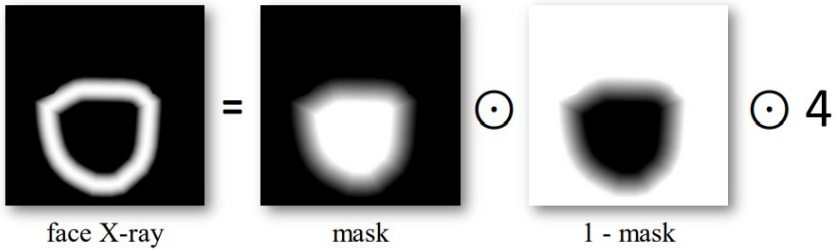


图 2.9 掩模和输出图像之间的关系示意图

2.4 基于语义特征的伪造检测方法

由于原始图像与目标图像的颜色、光照、噪声等等图像属性的不同，生成模型得到的伪造图像或视频中会出现伪造痕迹，可能是一块噪声或阴影，也可能在某处存在明显的突变。基于伪造痕迹挖掘的方法可以通过捕捉这些“瑕疵”来对图像的真实性进行判断。而这些伪造痕迹本质上属于纹理特征，是图像存在像素的区域性的突变或图像像素存在有特征的变化规律。它们可以是肉眼可见的，也可以是肉眼不可见的，但是神经网络可以通过比较少的神经网络层学习到。

然而。随着深度伪造技术变得越来越成熟，伪造出来的图像和视频的真实度和还原度也变得越来越高。高质量的伪造图像在人眼可见的 RGB 域上很难再出现易于被挖掘的纹理特征，因此需要更加复杂的模型通过语义特征来进行判断。

2.4.1 RNN

Güera 等人 [55]提出了一个基于实习特征的两阶段人脸伪造检测模型，其流程示意图如图 2.10 所示。作者先将一串连续的图像帧作为输入，通过一个 Inception v3 网络以提取帧级特征，该网络是一个在 ImageNet 上预训练好的网络。然后通过 LSTM 模块以捕捉由人脸交换过程引入的帧间不平滑所导致的时序不一致性。这种时域上的不一致包括了帧间抖动和帧间光源不一致导致的面部区域闪烁现象。之后，作者使用了 600 个视频的集合来测试训练好的模型，其中的伪造视频数量大约是一半，且伪造视频是源自于不同伪造方法生成的的视频。

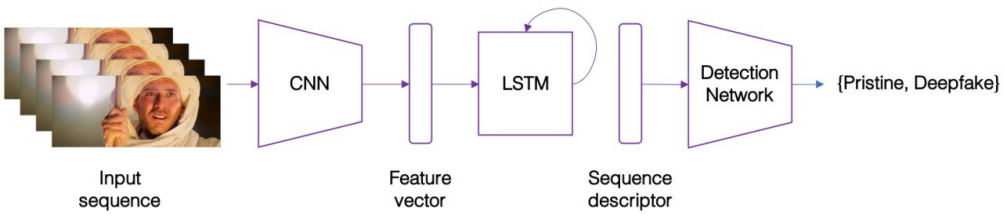


图 2.10 基于时序特征的深度伪造检测算法结构图

2.4.2 XceptionNet

2017 年，Chollet 等人根据 Inception v3 的基础上设计出了 Xception 网络[29]，其网络结构如图 2.11 所示。Inception 的一系列网络的主要思想是解耦卷积过程中的空间信息交互与通道间的信息交互，而一般的卷积过程中，这两者是高度耦合的。Xception 网络通过深度可分离卷积模块替换掉了原有的 Inception 模块，进一步优化了空间信息交互与通道间的信息交互的解耦过程。此外，Xception 网络还引入了残差连接增加了网络的鲁棒性。事实上，Xception 网络做的这些优化使得网络剔除了部分冗余参数，达到了更高效的实现。在参数量相近的情况下，Xception 网络在各类任务中往往都具有更高的性能。

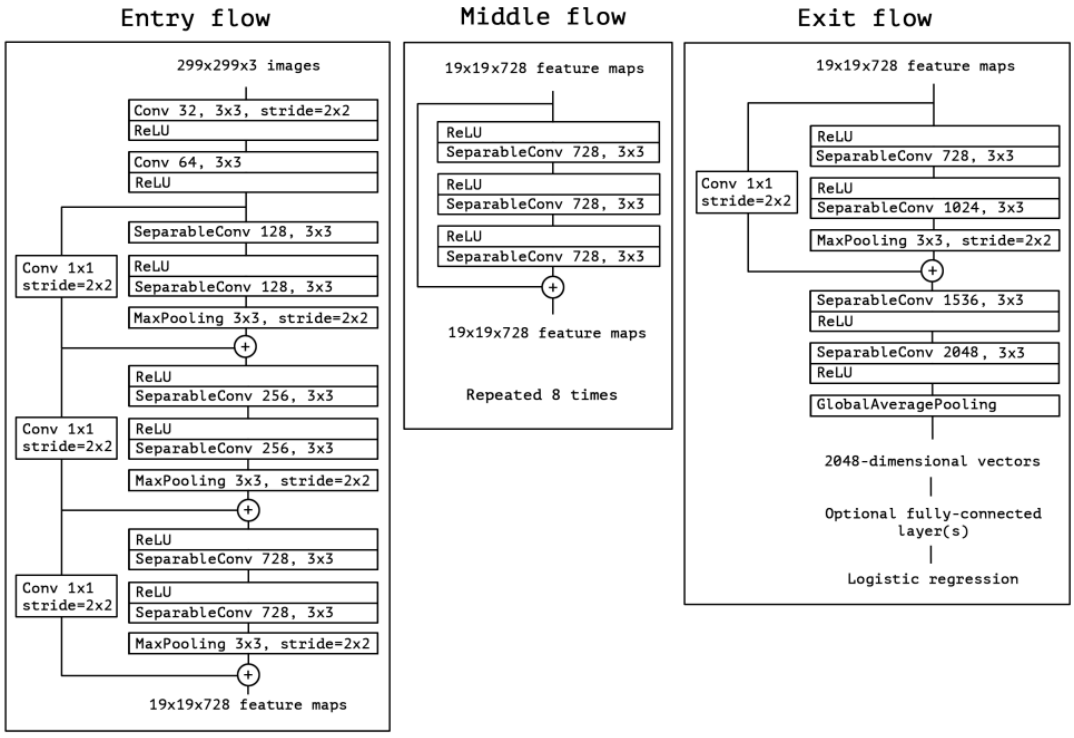


图 2.11 Xception 的结构图[29]

2.4.3 MesoNet

Afchar 等人提出了一种基于数据驱动的专门用于检测伪造视频图像的网络 MesoNet[28]。该方法专门用于检测由当时最常见的两种伪造方法，即 Deepfakes 和 Face2Face 这两种方法生成的伪造图像。该方法对 Deepfakes 和 Face2Face 生成的伪造图像识别的准确率分别达到了 98.4%和 95.3%。MesoNet 提出了两种网络结构，MesoNet-4 和 MesoInception-4。

MesoNet-4 的结构如图 2.12 所示。其结构简单，性能高效，并没有像 Xception 选择了深度可分离卷积，虽然相比 Xception 具有更多的参数量和计算量，但是深度可分离卷积需

要做大量的单通道的卷积，造成大量的内存访问，受限于内存带宽和数据 IO，深度可分离卷积在 GPU 上的速度往往并不会更快。而 MesoNet-4 因其简洁的设计很快成为了高效的数据驱动类方法的开山之作。

MesoInception-4 的结构如图 2.13 所示，是作者基于 Inception 网络对 MesoNet 的改良后的模型。其先并行计算恒等映射和、普通卷积和不同步长的膨胀卷积，最终将不同类型卷积得到的结果拼接在一起。这样的做法使得网络在每一层都拥有不同尺寸的感受野的信息，同时类似于分组卷积的原理，也减少了一定的计算量。

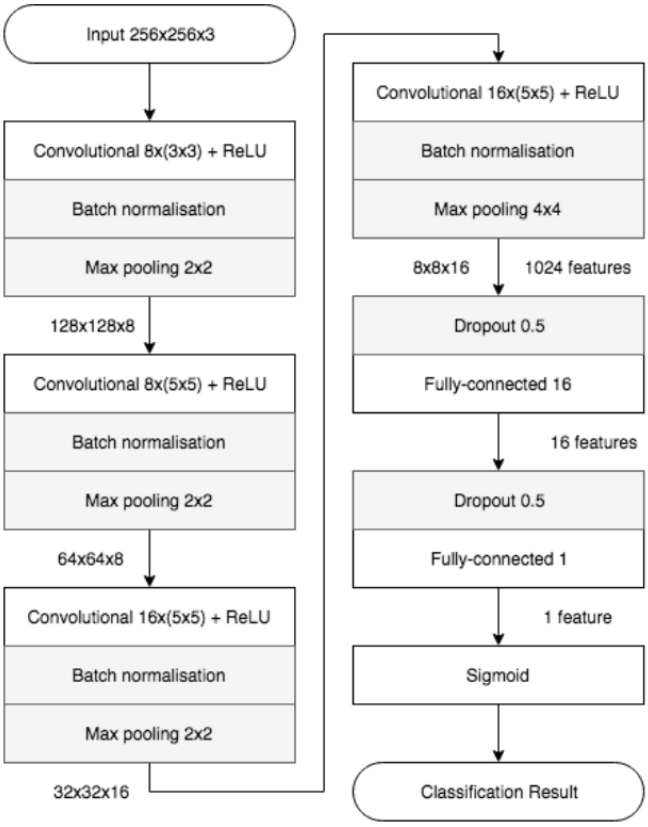


图 2.12 MesoNet-4 的结构图

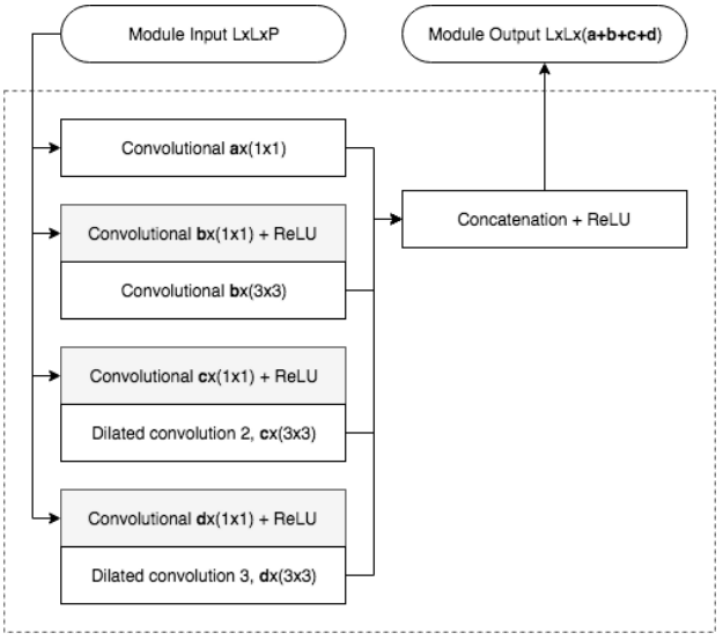


图 2.13 MesoNet-4 的结构图

2.4.4 F3Net

2020 年商汤团队提出了 F3Net 模型[56]，通过 FAD 和 LFS 两种不同的频域特征提取方法设计出了一种基于频域特征的双流网络用于深度伪造的检测，其结构如图 2.14 所示。

FAD 通过离散余弦变换[57]将图像转换到复数域上的频域信息，并通过设计掩码将频域信息进行分离，再通过逆离散余弦变换转换为实数域。尽管 FAD 提取到了频域特征，但最终依然通过逆离散余弦变换转换回了 RGB 域上，这些信息并不是直接的频域信息，因此作者还使用了 LFS，对图像采用了滑动窗口离散余弦变换（Silde Window DCT)，从而提取局部的频率响应，通过计算一系列可学习频带的频率响应，将频率统计信息重新组合为与输入图像共享相同布局的多通道空间映射。LFS 能满足原始 RGB 图像的平移不变性以及局部一致性的特点。

通过 Xception 网络提取出 FAD 和 LFS 的特征后，将 FAD 和 LFS 对应的特征通过卷积层分别求出对应的 attention map，将特征与 attention map 做按元素乘法后与 FAD 和 LFS 的特征相加，作为带注意力的特征作为额外信息，最终将特征通过特征融合工作将双流的频域特征融合进行图像的真伪分类。

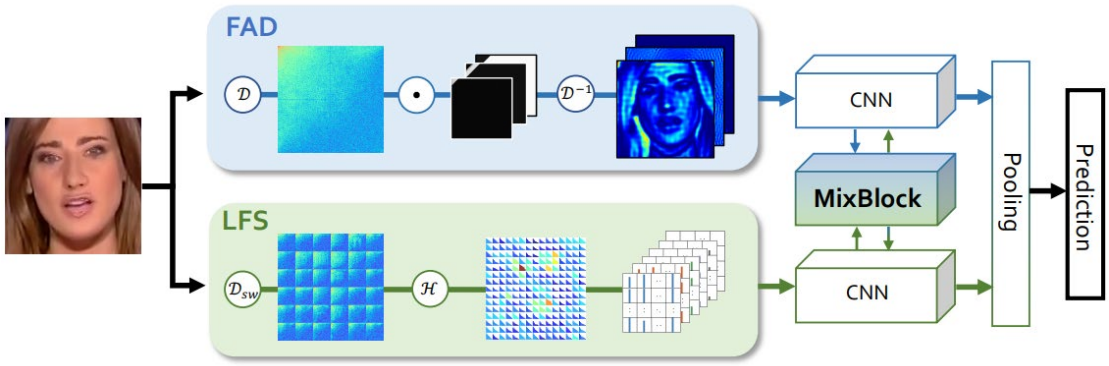


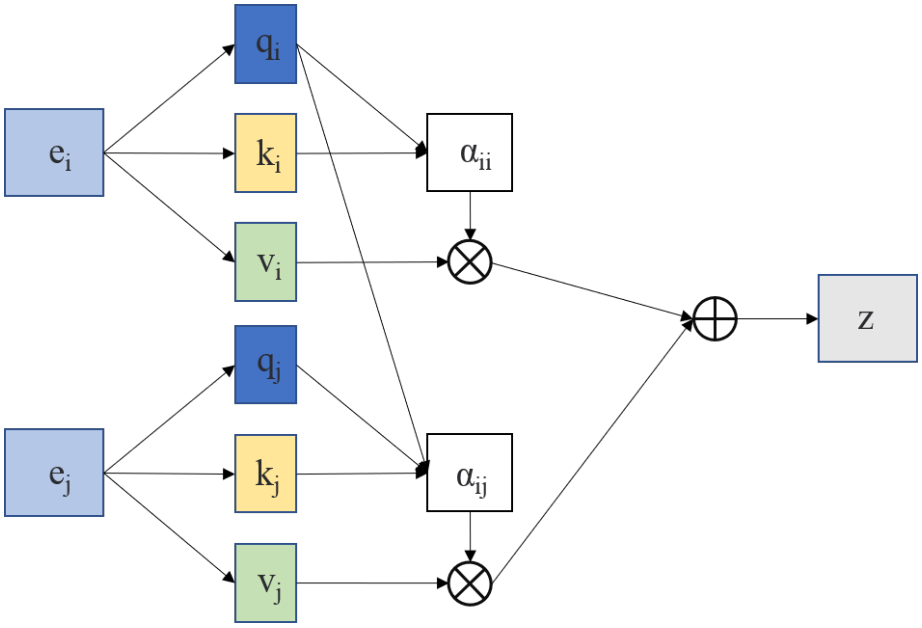
图 2.14 F3Net 的结构图[56]

2.4.5 Vision Transformer

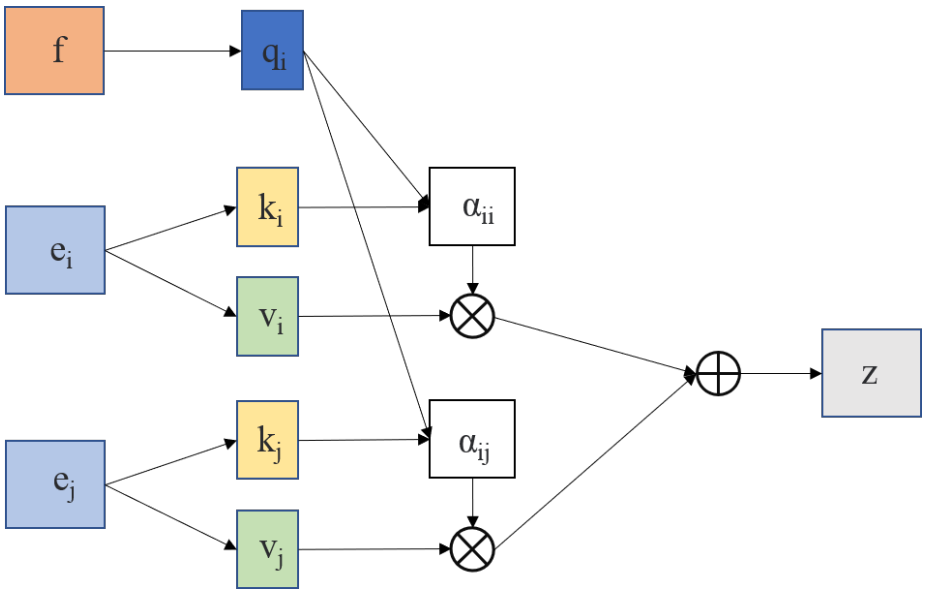
图像是一个像素构成的矩阵，而视频又由连续的图像帧构成，对于绝大多数视觉任务来说，图像本身的信息是有一个极高的冗余度的，对本文的工作来说也是如此。在深度伪造检测这一任务中，按照区域进行划分，前景的信息比背景更加重要，而前景之中，面部器官又比头发更加重要；而按照语义来划分，纹理中的伪造痕迹和深层特征图中的一些重要的语义特征则更加重要。

注意力机制则是在自然语言处理领域中被提出的一种让模型去学习关注更重要的信息的方法。事实上，人类在获取信息时，也会优先获取或更加重视某些关键的信息。例如人类在观察一张图像时，往往会优先观察图像的前景部分，观察前景的时候又容易先去观察前景中的重要的部分。而注意力机制正是参照生物认知机制，让模型在整体信息中选择关键信息，并不同程度地削弱甚至忽略不重要的信息。

对于一些序列化数据，以编码器-解码器结构为基础的 LSTM 和 RNN 等时序模型存在一个问题：不论输入长短都将输入编码成一个固定长度的向量表示。这使模型对于长输入序列的学习效果很差。而注意力机制则克服了上述问题，在模型输出时会选择性地专注考虑输入中的对应相关的信息。具体而言，对于编码器编码的向量，上一步的解码器的输出将会和隐藏状态做内积，再通过 Softmax 函数计算每个隐藏状态的注意力权重，最终求得所有隐藏状态的带权和后才会送入解码器。



(a) self-attention 的计算示意图



(b) cross-attention 的计算示意图

图 2.15 self-attention 和 cross-attention 的计算对比图

注意力机制可以分为self-attention和cross-attention，两者的原理示意图如图2.15所示。事实上，注意力机制可以被抽象成公式 2.7 所示：

$$output = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V \tag{2.7}$$

式中 Q 表示查询矩阵，K 表示键矩阵，V 表示值矩阵，d 表示中间向量的维度。其中，self-attention 中的查询矩阵和键矩阵由同一个中间向量通过线性映射获得，而 cross-attention 中的查询矩阵和键矩阵由两个不同的向量通过线性映射获得。相比之下，前者关注了序列内

部的内在关联性，而后者关注了两个向量的关联性。

注意力机制使相比较传统的 RNN、LSTM 等模型来说，其可以进行并行计算，解决了传统时序模型的序列化计算带来的效率低下的问题。同时，向量间的内积计算使得其回避了长期记忆的问题，不再需要像传统序列模型一样在每个时刻考虑应该更新或保留哪些信息。使用注意力机制的方法被广泛应用在各种序列预测任务上，包括文本翻译、语音识别等。

2017 年，在注意力机制的基础上，一种基于编码器-解码器结构和多头注意力机制的新的深度学习模型 Transformer 被提出[58]。最早其被用于自然语言处理领域中，而在 2019 年，Dosovitskiy 等人将 Transformer 引入了计算机视觉领域中并提出了 Vision Transformer[59]（ViT）。ViT 只能以一维的序列作为输入，为了处理 2D 图像，作者将图像 $\mathbf{X}\in\mathbf{R}^{H\times W\times C}$ 切分并重塑为一个二维的小块序列 $\mathbf{X}_p\in\mathbf{R}^{N\times(P\times P\times C)}$，其中(H, W)是原始图像的分辨率，C 是通道数，(P, P)是每个图像补丁的分辨率，$N = HW/P^2$ 是切分后的得到的小块数，也是 ViT 的有效输入序列长度。ViT 在其所有层中都使用固定维度 D 的潜在向量，也即用一个可训练的线性映射将每个小块展平成一维向量后再映射到 D 维。

在 token 序列前会额外添加一个可学习的 token，用以调节所有 token 的权重。此外，同自然语言处理中的 Transformer 一样，ViT 也采用了位置编码。位置编码会作为补充信息添加到每个小块展平的向量中以保留小块所对应原始图像的位置信息。Transformer 使用了余弦位置编码，如公式 2.8 和公式 2.9 所示：

$$PE(pos，\ 2i) = \sin\frac{pos}{1000^{2i/dim}} \tag{2.8}$$

$$PE(pos，\ 2i+1) = \cos\frac{pos}{1000^{2i+1/dim}} \tag{2.9}$$

其中，pos 表示每个元素对应的位置，dim 表示整个序列的总长度，i 表示当前元素的序号。而 ViT 的作者则使用了标准的可学习一维位置编码，添加位置编码后得到的嵌入向量序列作为编码器的输入。事实上，后续有工作已经指出，对于 Transformer 来说位置编码是必要的，没有位置编码相比添加位置编码模型性能会有显著降低，但采用不同的编码方式对性能几乎没有影响。

ViT 由多头自注意(Multi-head Self Attention，MSA)和多层感知机(Multilayer Perceptron，MLP)组成。在每个模块之前都会使用 LayerNorm 进行标准化，在每个块之后使用了残差连接以提高模型的鲁棒性。MLP 则包含了两个使用 GELU 为激活函数的线性层。

ViT 为计算机视觉领域提供了一种新的网络结构，其不再使用卷积操作去提取特征，拥有更优秀的全局建模能力。之后也出现了的一系列对 ViT 进行优化的工作。DeiT[60]通过知识蒸馏有效提高了 ViT 的训练速度。DAT[61]引入了可变性卷积，将注意力计算集中在重要的区域上，使得模型获得了可变的感受野。SWIN ViT 通过二级式的窗口注意力机制，降低了模型的计算复杂度，使得 ViT 可以接受更大分辨率的图像，并且学习了 CNN 中的层级

式建模的思想，使得 ViT 的性能在多项视觉任务中取得了优异的性能。

注意力机制可以让网络关注到更为重要的信息，这对于冗余信息较高的图像来说是十分有效的。2021 年，Wodajo 等人提出了 CViT 算法[62]，将 ViT 用于了深度伪造检测领域。CViT 是一个利用 CNN 提取特征，再使用 ViT 对 512 维的 7×7 特征图进行注意力计算的轻量化的伪造检测模型，具有很好的可拓展性，其结构如图 2.16 所示。具体来说，由于在进行人脸伪造检测的时候，需要从图像帧中标注并裁剪出人脸，这让截取出来的人脸的分辨率一般都比较小，因此通过 CNN 提取的特征图也会较小。对特征图计算注意力时不会再将其划分成小块，这使得在进行注意力矩阵计算时，矩阵的维度会较小，模型具有更快的推理速度。

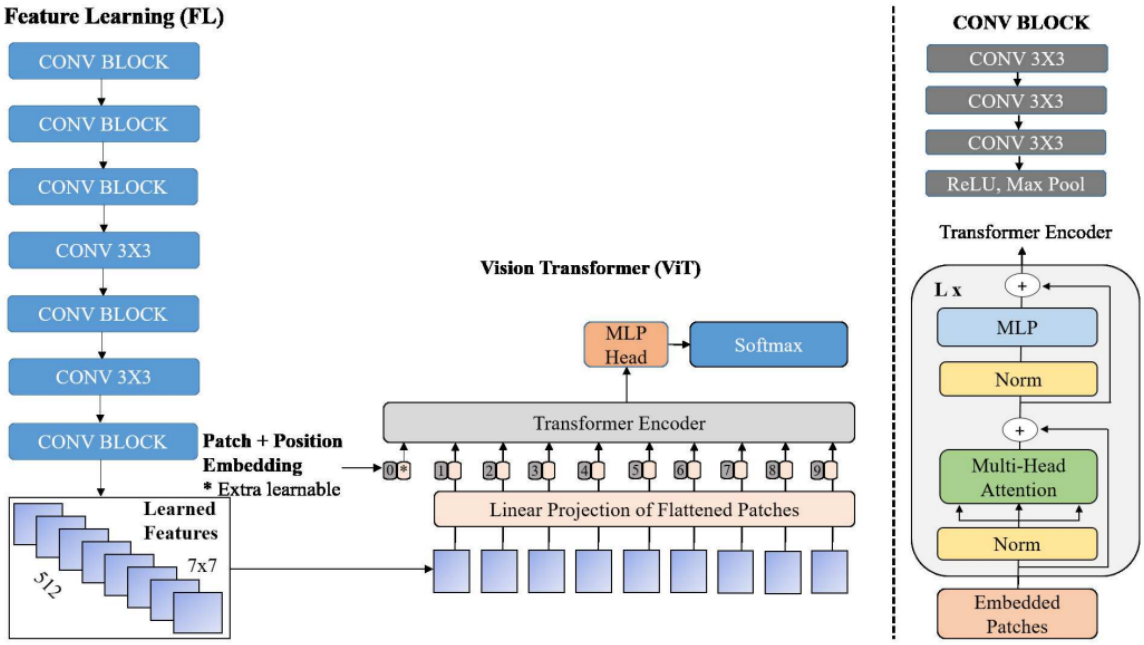


图 2.16 CViT 的结构图[62]

2021 年，Wang 等人提出了 M2TR 算法[27]，作者将多尺度的思想引入到 Transformer 结构中并结合频域信息进行伪造检测。具体来说，再多头注意力计算时，每个头会将图像划分成不同小块时，作者尝试按照 10、20、40、80 四个不同的尺寸去进行划分。每一个尺度都会在各自尺度内分别计算自注意力，得到 4 种尺度的输出，最后将不同尺寸的块拼接起来并通过一层卷积层消除混叠效应。这样的做法使得每个 token 表征的信息量不同，相当于具有大小不一的感受野。多尺度的思想有利于 Transformer 模型在面对数据集中存在尺寸相差较大的伪影或其他伪造痕迹时，这有更鲁棒的检测性能

2.5 本章小结

本章介绍了深度伪造的种类和基础理论，并分别对传统方法和基于深度学习的伪造技术进行了介绍和分析。之后本章按照基于生物信号和生理特征的伪造方法、基于伪造痕迹挖掘的伪造检测方法和基于语义特征的伪造检测方法对现有的伪造检测算法进行分类，并按照算法提出的时间顺序进行介绍和分析。

第三章　基于多域特征网络的深度伪造视频检测

3.1 引言

深度伪造技术的发展使得各种针对特定人物的攻击层出不穷，这种对特定人物的伪造对人员识别系统和安保系统造成了极大的挑战。人工检测深度伪造已经不能满足当前社会需求，也无法应对日趋精细的伪造方法。因此，为了避免深度伪造技术被滥用，各种检测深度伪造产物的方法也应运而生。

由于伪造时分辨率、色彩和光照等因素的不同，伪造的图像上容易出现极具辨识度的伪造痕迹，也即伪影。常见的卷积网络可以比较好的发掘这种伪影。而对于没有产生明显伪影的图像，则需要更加复杂更加高效的网络去学习语义上的特征，进而学习伪造过程中能表达出目标图像与原始图像的不一致性的特征。伪影这种在伪造面部时产生的分辨率不一致的异常特征。但是一些伪造视频在被压缩时，伪影特征会被破坏，这种情况在低质量视频中尤为明显。但在频域特征中，伪影依然容易被检测到。本文希望通过增加频域信息作为神经网络的输入，提高模型的分类性能和鲁棒性。除此之外，每个人的面部神态、头部运动姿态等都具备特有的模式，这种独特的模式被称为软生物特征模型，这也是目前伪造出来的视频不具备的特征，因为伪造得到的生物运动方式和神态表情都是事先设计好的。

另一方面，在人脸替换中，被替换的只有视频中的人脸部分，而神经网络所需要关注的有两类特征：一类是纹理特征，例如人脸替换时经常出现的伪影和一些生物生理特征。另一类是语义特征，是神经网络深层所学习到的伪造视频图像所共有的特征，它可以是语义不一致性或是一些软生物特征。而随着深度伪造技术发展得越来越成熟，易被神经网络学习的纹理特征越来越少出现在伪造视频中了。而如何让神经网络高效准确地学习这些特征，本质上是让网络关注到伪造视频所具有而真实视频所不具有的信息，并尽可能地减少冗余参数的过程。然而事实上视频或图像的内容信息并不能对分辨其真假提供帮助，我们在鉴伪时只关心其部分结构信息，这也意味着图像中绝大部分是冗余信息。如何引导网络高效学习有用特征也是一个关键。

为此，本文设计了一个结合 RGB 特征和其频域特征的多域特征网络模型。具体而言，本文的工作如下：

1) 提出了一个多模态深度伪造检测算法，对于图像结合空域特征和频域特征进行伪造检测，而对于视频则还会结合其时域信息，在一定程度上解决了上文中部分伪造痕迹难以在 RGB 图像中被捕捉的问题。
2) 提出了一种以 CNN 提取特征，Transformer 强化特征的模型用于伪造检测，设计了通道间的自注意力计算有效减低了计算复杂度。
3) 采用了类别得分融合[63](class score fusion)对不同域的预测结果进行了融合，并通过

增加得分差损失函数进行了优化，较直接拼接法、基于注意力机制的融合等方法取得了更好的效果，并对此进行了实验验证。

4) 在开源数据集 Celeb-DF[4]和 DFDC[64]上与现有的算法进行比较，可以表明该模型取得了接近目前最先进算法的性能。

3.2 多域特征网络

本节主要介绍基于多域特征的伪造视频检测模型。其中检测模型的框架如图 3.1 所示。模型主要由处理 RGB 图像的上分支和处理频域特征的下分支构成。

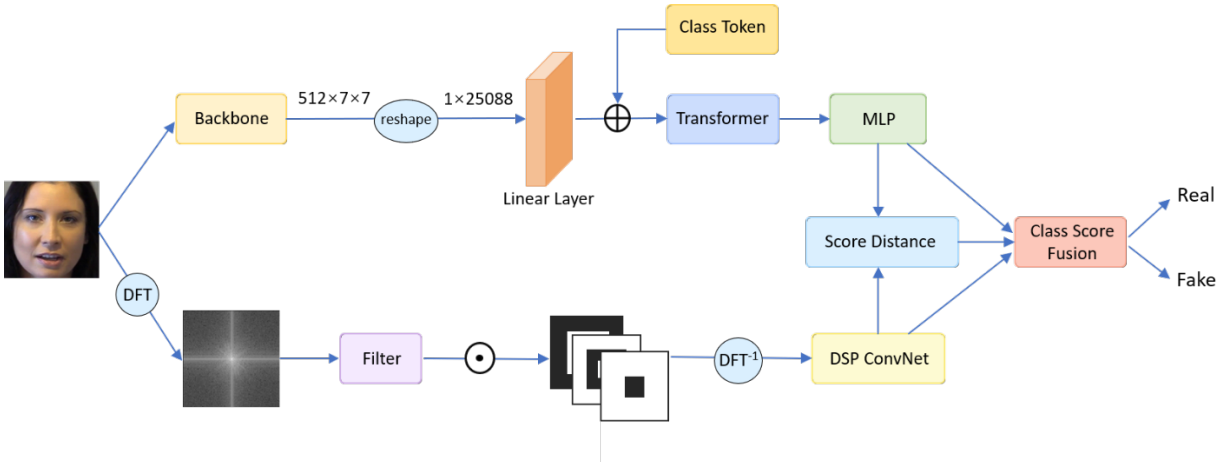
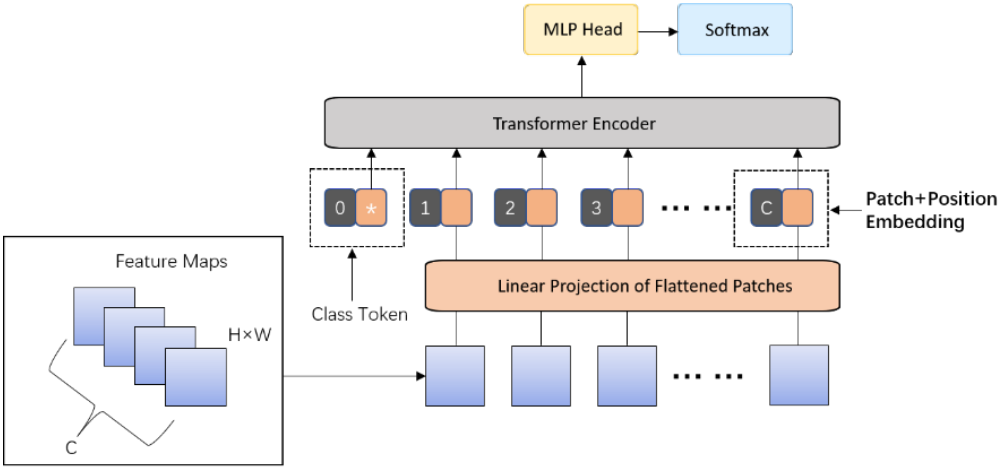


图 3.1 多域特征网络的结构图

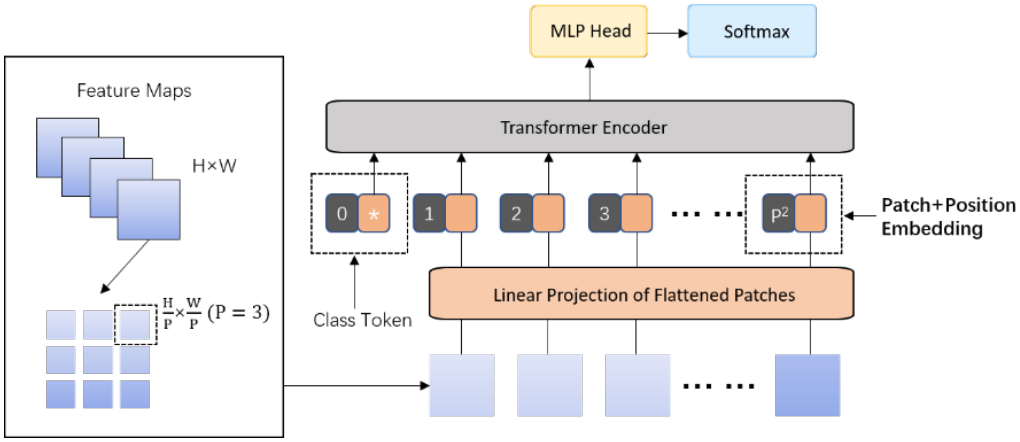
3.2.1 双流特征

在神经网络中，网络中的浅层特征具有更强的纹理信息，而深层特征具有更强的语义信息。深层的特征图一般更小且具有更多的通道数。本章通过 RESNET-38 网络对 RGB 特征进行提取。然而，最终提取的特征中包含的信息并非全部是有效的，视频中不同帧所提取的特征的效果也不尽相同。Montserrat [64]提出了一种自动加权机制，在视频级预测中尽可能关注更重要的帧。而本章则通过在模型中使用通道间的自注意力机制来对不同特征通道进行了注意力加权。

输入网络的原图像表示为 $X \in R^{3\times224\times224}$,经过 ResNet-38 提取得到的特征图表示为 $f_r \in R^{512\times7\times7}$。传统的 ViT 算法将原始图像进行切分后，将每一小块表示成向量形式送入 Transformer 进行处理。在切分后，每个小块在原图像上的位置信息会丢失，虽然可以通过 position embedding 来对这种相对位置信息的丢失进行补偿，但这种对原图像的硬性划分会割裂开面部重要区域，影响特征的提取。本章对特征图通道间的自注意力机制，以较小的计算开销为每一块特征图计算了学习时的注意力权重。



(a) Application of Transformer between feature maps



(b) Application of Transformer in feature maps

图 3.2 两种对特征图应用 Transformer 的结构比较

常见的对图像或者特征图应用 Transformer 的方式如图 3.2 所示。即对较大的图像或者特征图进行切分，在切分后的块间使用注意力机制，适合浅层特征或图像。但对较大的图像或特征图直接进行全局建模容易使得问题过于复杂，往往需要更多的数据和更久的训练时间，在某些视觉任务中，特别是密集预测的视觉任务里，Transformer 的这种全局建模的特征学习方式反而不如卷积神经网络。同时，为了避免对特征图进行硬性切分而对特征图的信息产生破坏，本章采用了另一种方式。卷积神经网络提取的特征图 f_r 以高为 H，宽为 W，通道数为 C 的大小进过一层线性映射后，每一个特征图将再加上表示位置信息编码的位置编码部分后再被送入 Transformer 网络。而 class token 则作为“空类”辅助计算出单个特征图在网络中的权重。最终特征图被送入多层感知机模块进行二分类。

对于频域特征，将原图 X 像做离散傅里叶变换(Discrete Fourier Transform，DFT）后得到频域图像 D。由于作离散傅里叶变换后的图像具有将低频分量保留在了图像中心，而将

高频分量保留在了图像边缘的特性。如图 3.1 的频域特征分支网络所示，借助这 DFT 图像的这一特点，可以通过设计环形的掩膜，将频域图像 D 与环形掩膜做按元素乘法，可对图像的频域信息划分成高频、中频和低频三个部分:

$$d_i = D \odot mask_i, i = \{1,2,3\} \tag{3.1}$$

其中，$mask_i$ 为对应频域分量的掩膜，⊙表示按元素乘法。DFT 图像和频域量分离后的效果示意图如图 3.3 所示。相较于把图像信息以 RGB 的形式划分成三个通道，这种划分方式使得图像信息以频域差异被划分，由于伪造图像一般在肉眼可视的 RGB 域内会将伪造痕迹修复得很好，这样有利于模型从不同角度提取特征。并且，将图像按照频域差异划分成三分可以保证其输入与 RGB 网络分支的输入大小相同，同时避免划分部分过多导致的信息稀疏。同时我们发现，相较于 RGB 特征，深度可分离卷积可以更好地提取频域特征。因此本章采用了一个深度可分离卷积网络用作频域特征的提取，该网络相当于一个轻量化的 Xception 网络。最终得到的频域特征通过一个全连接层进行分类。

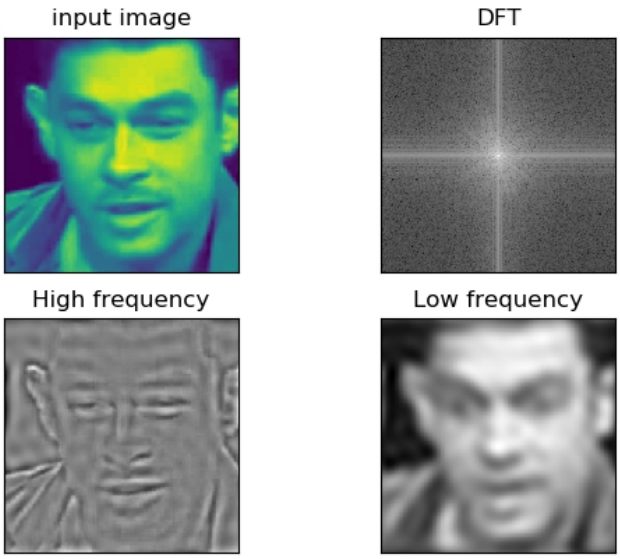


图 3.3 DFT 图像和频域量分离后的效果示意图

3.2.2 多模态融合

多模态融合的思路一般有两种：一种是在特征层面进行融合，如直接连接、attention 融合等；另一种是在决策层面进行融合。在本章中，对 RGB 特征和频域特征采用的特征提取方式差异较大，且不同域的特征由于表征形式不同，在拼接时容易产生混叠效应。虽然在融合后继续进行卷积操作可以一定程度上减轻混叠效应的影响，但又容易产生过拟合。因此本章采取了一种决策层面的融合方式，即类别得分融合。

具体来说，最终的预测图像为伪造的概率是两个网络分支分别预测图像为伪造的概率

的平均数。而两个分支网络将作为整体进行训练，即两个分支网络的损失函数将被共同作用到整个网络的反向传播的过程中。最终的损失函数 Loss 定义表示为：

$$Loss = \lambda_1 L_{rgb} + \lambda_2 L_{freq} + \lambda_3 L_{dis} \tag{3.2}$$

$$L_{rgb} = -y \log y_{rgb} - (1-y) \log(1 - y_{rgb}) \tag{3.3}$$

$$L_{freq} = -y \log y_{freq} - (1-y) \log(1 - y_{freq}) \tag{3.4}$$

$$L_{dis} = \left|| \boldsymbol{P} - \boldsymbol{Y} \right|\Big|_2 \tag{3.5}$$

其中 L_{rgb} 和 L_{freq} 分别表示两个分支网络中的交叉熵损失。L_{dis} 表示两个分支网络的分类得分差损失，该损失有助于加速模型收敛。P 表示模型的一个批处理中输出的向量，Y 则表示对应的标签向量。λ1、λ2、λ3 为权重参数，实验中 λ1=λ2=1，λ3 在实验过程中由 0.5 逐渐减小到 0.01。

3.3 实验分析

3.3.1 评价指标

在本实验中，我们采用准确率和 AUC 两项指标来衡量模型的性能。本节主要介绍这些指标。

准确率 (Accuracy) 是最常用的评价模型分类性能的指标之一。我们定义 TP 为真正例，即真实值和预测值都为正例的样本；定义 TN 为真负例，即真实值和预测值都为负例的样本；定义 FN 为假负例，即真实值为正例但预测值为负例的样本。定义 FP 为假正例，即真实值为负例但预测值为正例的样本。则准确率的定义如公式 3.6 所示。它度量了模型分类正确样本所占总样本的比例。

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{3.6}$$

然而，如果数据集的正负样本不平衡，准确率就不能很好地评价模型分类的性能。如正样本远多于负样本的情况下，模型会倾向于将更多的样本预测为正例，极端的情况下所有的样本都会被预测为正例，但此时的模型准确率依然很高，这显然是不合理的。因此，本文同时采用 AUC[66](Area Under Curve)作为度量指标。AUC 被定义为受试者工作特征曲线 (Receiver Operating Characteristic Curve) 与坐标轴围成的面积。AUC 可以根据正负类的样本量进行标准化地衡量模型的性能。其物理含义为任取一对正负样本(S+,S-)，正类样本 S+的分类得分高于负类样 S-的分类得分的概率。 AUC 的计算公式如公式 3.7 所示。

$$AUC = \frac{\sum_{i=1}^{M} S_i - \frac{M \times (M-1)}{2}}{M \times N} \tag{3.7}$$

式中 Si 表示依分类概率得分从小到大排列的第 i 个样本，M 表示正样本的数量，N 表示负样本的数量。

3.3.2 数据集

本章采用了两个常用深度伪造数据集 Celeb-DF 和 DFDC，并在原始数据集中随机抽取了约 20000 帧图像进行训练。

Celeb-DF 包含了 590 个从 YouTube 上下载的真实视频和 5639 个高质量的深度伪造视频。其覆盖了不同性别、不同种族和不同年龄。视频场景包括了室内外的采访、主持等场景。Celeb-DF 中的伪造视频质量较高，面部伪影较少，同时进行了帧间平滑，多数视频无法用肉眼区分真假。

DFDC 数据集是由 Facebook 联合微软、亚马逊等研究机构发布的开源深度伪造视频数据集，并被用于 DFDC 竞赛。为了保证数据的多样性，该数据集由约 430 个人拍摄的 19197 个真实视频和 100000 个伪造视频构成。DFDC 数据集的伪造视频由多种伪造算法伪造而成，包括 Face2Face、FaceSwap 和 NeuralTextures[67]等。本实验选用了其中 7%的数据用于训练和测试。

需要注意的是，在进行帧级别的预测时，我们在 Celeb-DF 数据集中对每个视频采样了 5 帧左右的数据，对于 DFDC 数据集，由于其数据充足，只对每个视频采样了 1~2 帧进行训练和测试。而在进行视频级预测时，本实验对每个视频随机采样连续的间隔为 2 的 16 帧图像。数据集的统计如表 3.1 所示。

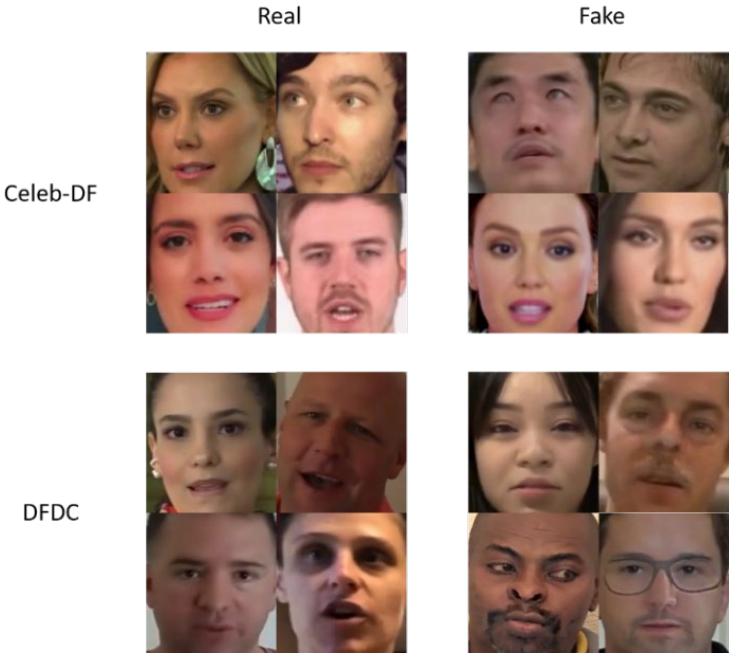


图 3.4 从数据集中提取的人脸

表 3.1 实验划分的数据集大小

Datasets	训练集	验证集	测试集	伪造视频占比
Celeb-DF	17076	2073	2136	79.2%
DFDC	16860	2115	2040	75.3%

3.3.3 数据处理与实验过程

事实上，该模型可以进行 3D 网络训练，此时网络的输入是一串连续的人脸图像帧。而在进行帧序列人脸标定时，需要额外进行人脸对齐的操作，使得相邻帧的人脸尽可能对齐。但由于存在视频中人物面部移动幅度过大的问题，这一般表现为距离摄像头远近方向的大幅移动和水平方向上的大幅移动。每一帧都进行独立人脸标注时会使得帧序列难以对齐，也即一些面部关键点的伪造偏差较大，这会比较严重地影响模型的性能。另一方面，人脸的转动也是一个会影响帧间对齐的问题，当相邻两帧的眼角关键点的连线与水平线的夹角相差过大时，会严重影响模型性能。为此，本文采用了一种基于面部关键点的带约束的人脸标注方法。

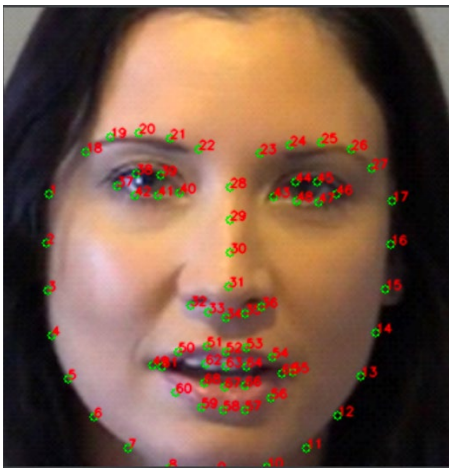


图 3.5 人脸 68 个关键点标定示意图

如图 3.5 所示，通过 dlib 库对人脸进行 68 个人脸关键点标定，由于唇部关键点的伪造常常变化较大，而眼部和鼻部关键点在人脸上的位置一般比较稳定，因此先选择眼角关键点，即第 37 和第 46 号关键点，将图像依据这两个关键点将人脸旋转到使得第 37 号关键点和第 46 号关键点水平的位置后，再进行裁剪。特别的对于侧脸情况，此时只能捕捉到一个眼角关键点，将不再对图像进行旋转。旋转示意图如图 3.6 所示。红色标定框为初始人脸标定框，其眼部关键点与水平线的夹角为 θ，旋转后的标定框将绕标定框中心点旋转 θ，得到的标定框如绿色框所示。

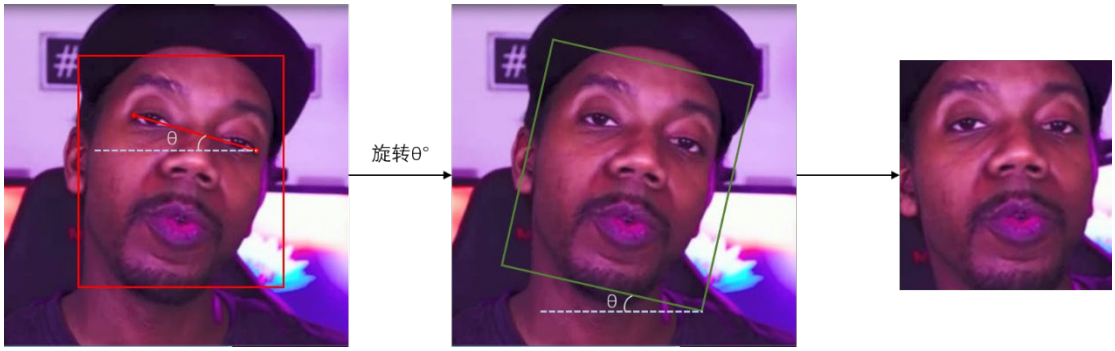


图 3.6 人脸对齐操作中的基于眼部关键点的人脸标定框旋转

此外，在得到图3.6中的绿色人脸标定框后，会基于上一帧的人脸标定框对当前帧的人脸标定框进行微调，以缓解帧序列的抖动问题。具体来说，水平方向的边框，将根据上一帧的 37 号点和 46 号点所在的图像中水平方向的位置进行缩放，竖直方向上的边框将根据上一帧的 37 号点和 31 号点所在图像中水平方向的位置进行缩放。并且每个方向上缩放的幅度不会超过边框长度的 20%。

对视频中将每个采样帧进行面部检测，并对面部区域进行裁剪，再通过上述仿射变换以完成帧间对齐。将面部帧或者帧序列送入检测模型，检测器将对输入数据进行真伪分类。面部伪造检测流程如图 3.7 所示。

图 3.7 人脸视频伪造检测流程示意图

3.3.4 实验环境及参数设置

为了验证本算法对深度伪造视频检测的有效性，我们对数据集中的视频随机挑选可检测出人脸的帧进行测试，并将其和现有的方法进行分析。本文先进行帧级别的预测，再将多帧预测结果融合成视频级别的预测。具体而言，所有受测帧的平均类别得分将作为整个视频的预测结果。实验中超参数的设置如表 2 所示。其中学习率在训练过程中从 0.0001 逐渐降低为 0.00001。

表 3.2 模型训练使用的超参数

Hyperparameters	Value
学习率	0.0001~0.00001
权重衰减	0.00001
批大小	4
训练轮次	50
Dropout	0.3

3.3.5 性能实验

本节将通过与几个经典伪造检测算法对比，进行伪造检测模型的性能实验，以比较模型的分类精度。

其中对比算法包括 MesoNet、Xception 等经典深度伪造检测模型以及 CViT 这一较新提出且具有很好延展性和泛化性的方法，以此进行对比实验。所有算法都将在 DFDC 数据集和 Celeb-DF 数据集上训练并进行测试。在单个数据集内训练并测试的结果如表 3.3 所示。由实验结果看，本章的模型较几个经典的深度伪造取识别算法以及近两年新提出的模型都有着更好的性能。

表 3.3 不同模型在各个数据集上的分类性能

Methods	Accuracy (Celeb-DF)	AUC (Celeb-DF)	Accuracy (DFDC)	AUC (DFDC)
MesoNet[28]	82.6%	0.714	83.3%	0.719
CViT[62]	95.6%	0.984	88.8%	0.940
Xception[29]	93.3%	0.977	92.5%	0.956
Ours	97.1%	0.996	93.7%	0.968

为了检测模型的泛化性，实验中使用了在 DFDC 数据集上训练的模型对 Celeb-DF 数据集上的数据进行测试。实验结果如表 3.4 所示。考虑到 DFDC 数据集和 Celeb-DF 数据集在伪造方法、视频清晰度和拍摄环境等方面都有较大差异，实验结果中的各类方法的泛化性能都不是特别高，AUC 均在 0.7 以下，但本章提出的模型的泛化性较现有的高性能算法仍然有一定的提高。

表 3.4 不同模型在 DFDC 数据集上训练并在 Celeb-DF 上训练的分类性能

Methods	Accuracy	AUC
MesoNet[28]	76.4%	0.610
CViT[61]	78.7%	0.653
Xception[29]	79.1%	0.645
Ours	79.5%	0.663

3.3.6 消融实验

此外，为了测试模型设计的合理性，本章还对模型结构和损失函数进行了消融实验。表 3.5 展示了不同网络在 Celeb-DF 数据集上的性能。其中，第一行表示仅使用本章模型中的 RGB 流网络；第二行表示仅使用该模型中的 RGB（3D）流网络；第三行表示仅使用该模型中的频域特征流网络。

值得一提的是，由于 Celeb-DF 数据集里的数据帧间平滑做的十分精细，加之实验中面部区域帧间不对齐的问题并没有得到完全解决，通过 3D 卷积网络提取时域特征并没有让分类性能得到提升。

表 3.5 对于模型结构的消融实验

Models	Accuracy	AUC
RGB	94.6%	0.978
RGB(3D)	93.9%	0.972
Frequency	95.4%	0.984
Two Stream	97.1%	0.996

表 3.6 展示了在 Celeb-DF 数据集上采用不同的融合方式得到的分类性能。实验结果表明双流网络的设计和多模态融合起到了很好的效果。

表 3.6 在 Celeb-DF 数据集上的不同融合方式的性能比较

Methods	Accuracy	AUC
Directly Concat	96.5%	0.992
Attention	96.6%	0.992
Class Score Fusion	97.1%	0.996

3.4 本章小结

本章提出了一种基于多域特征的双流网络来进行深度伪造检测，针对部分伪造痕迹难以在 RGB 图像中被识别的问题，有效处理了频域信息，提取了频域特征辅助检测，并通过了类别得分融合的方式实现了双流网络“1+1>2”的目的。在 Celeb-DF 和 DFDC 数据集上

的实验结果可以表明，该模型在识别精度和泛化性上都接近于目前最先进的水平。

第四章　基于滑动窗口注意力的深度伪造视频检测

4.1　引言

在深度伪造视频中，伪造者不仅需要考虑单帧画面的质量和真实度，还需要保证帧间画面的流畅性。因此，从时序特征的角度去检测视频的真伪是可行的。事实上，Guera 和 Ekraam 等人就尝试过使用 LSTM 和 RNN 等结构去检测一段视频帧的真伪性。然而，当问题变成了检测一段视频帧的真伪，就需要考虑时序模型了。相比于 RNN 等一系列时序模型，Transformer 可以更好的进行并行计算，但是过长的帧序列会导致 Transformer 中的向量维度过大，从而带来不可接受的计算量。因此本章基于“分组”的思想，提出了一种通道间的滑动窗口注意力机制，既可以对单张图片进行训练，也可以使用连续的帧序列进行训练。在这个基础上，层次化的分组注意力计算可以有效降低计算量，避免计算量的爆炸。

另一方面，在进行视频级的预测时，许多工作的做法都是先对单帧图像进行训练，再根据视频中的多张图像帧的预测结果得到视频的真伪预测。这种朴素做法是存在一定道理的，当视频中的图像帧采样足够丰富时，可以获得一个比较准确的结果。然而，对于一些特定的视频，由于人脸姿态的不同，鉴别真伪的难度和精度也会产生较大差异。如图 4.1 所示，图中同一人物分别处于侧脸和正脸部位时，前者需要进行的伪造区域更小，更不易产生伪造痕迹。同时当人脸在镜头内大幅移动，或者做出夸张表情时，伪造起来相对更加困难，相对其他帧的质量可能更低，因此理论上来说，这些帧是更有利于被用于提供真伪检测任务所需的信息的。然而，根据这些情况，事先进行关键帧选择相当费时费力。

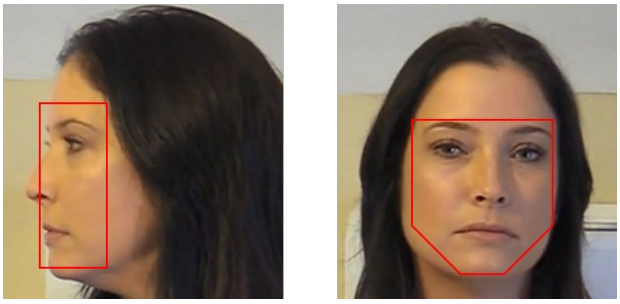


图 4.1 不同视角下的面部伪造区域对比

事实上，对于以连续的序列帧为输入的视频伪造检测模型来说，输入的信息量是较大的。对于 Transformer 来说，每个 token 的维度不可以过大，因为计算注意力时的计算复杂度与 token 维度的平方呈正相关。因此本章提出了一种通道间的滑动窗口自注意力算法，不仅可以对单张帧计算通道间的注意力，还可以通过“分组”降低计算注意力时的复杂度，使得模型在一定算力内对连续的帧序列特征计算帧间特征的注意力。

4.2 模型结构

模型结构如图4.2所示。模型的处理流程为将输入图像与其生成的注意力掩码分别再进行卷积操作后，进行两次按元素乘法，通过 ResNet 进行特征提取，再计算通道间窗口注意力（2D 模型）或帧间滑动窗口注意力（3D 模型）来进行特征强化，最终通过一个 MLP 进行真伪二分类。

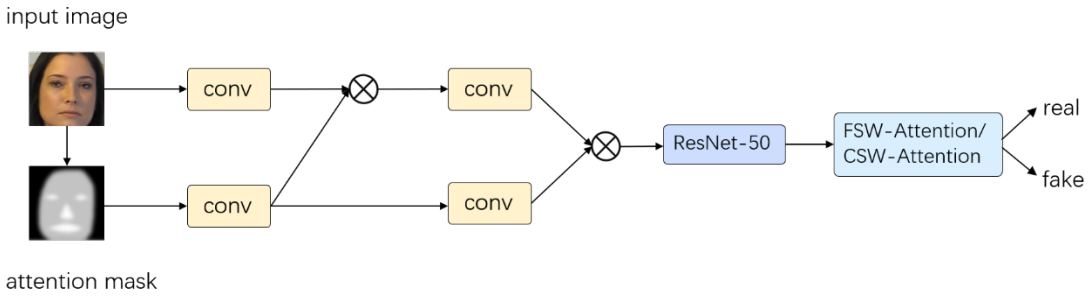


图 4.2 基于滑动窗口注意力的深度伪造检测模型的结构图

4.2.1 层级式注意力掩码

本章在第三章中提出了一种人脸序列对齐方法，其中包括了通过对人脸标定框的缩放，使得当前帧的标定框中的人脸关键点与上一帧对齐。然而，一定程度的缩放可能存在着人脸标定框不准确的问题，其可能将部分非人脸区域截取在内。然而一般意义上来说，背景信息是对伪造检测没有直接帮助的，因为伪造者几乎不会对背景做出篡改。只有在考虑前景背景的语义一致性的时候可能会用到背景信息。

此外，在人脸不同的区域中，伪造痕迹出现的概率也不相同，一般来说伪造痕迹也更易于出现在面部器官上。因此，为了缓解人脸对齐阶段的人脸标定框的误差，同样也基于伪造痕迹在不同区域出现的概率，本章设计了一种层级式的面部注意力掩码，旨在能够屏蔽图像中大量的冗余信息。

一张输入网络的图像可以被划分为三个区域，即非面部区域 R_u、面部非器官区域 R_f和面部器官区域 R_o，这三个区域的重要性按由高到底排序，重要性月高的渔区注意力掩码的值就越大。其中非面部区域 R_u 包括了背景、人物头发等区域；面部器官区域 R_o 包括了眼睛、嘴巴和鼻子的区域；面部非器官区域 R_f 则是面部剩余的其他区域。本章设计了三种不同的区域掩码值如表 4.1 所示，并在实验部分给出了不同掩码值的效果。

表 4.1 层级式注意力掩码的不同区域赋值

组别	R_u	R_f	R_o
a	0	0.5	1
b	0	0.7	1
c	0.3	0.7	1

4.2.2 通道间滑动窗口注意力机制

Transformer 网络在自然语言处理领域里大放异彩，并很快以 Vision Transformer 的形式被引入到计算机视觉领域中。受益于其优秀的全局建模能力。ViT 在检测和分割等多个视觉任务中取得了非常好的效果。在以 ViT 为基础的 DETR 目标检测模型中，因为其直接对整个图像或特征图建模，它的大目标检测精度非常高。然而，传统的 ViT 同时也存在着局部建模能力较差、计算复杂度高、需求数据量较大等缺点。

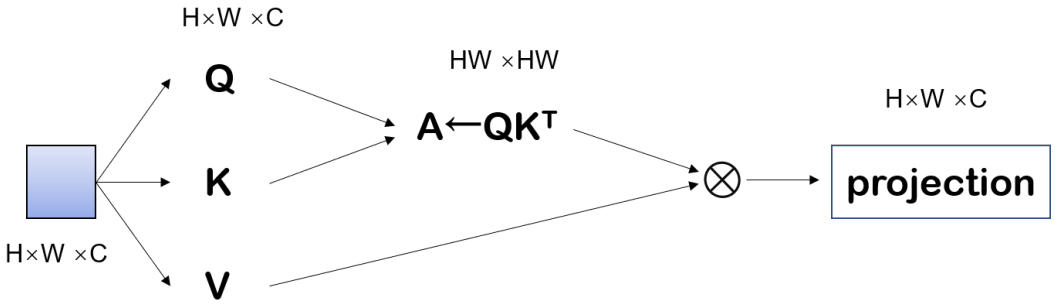


图 4.3 多头自注意力的计算流程

首先，根据 Transformer 的计算流程可得知，其计算复杂度与图像的边长的平方成正比。其中注意力计算的过程如图 4.3 所示。将图像或特征图的高宽分别表示为 H 和 W，记通道数为 C。则在由特征图计算 Q、K、V 时的计算复杂度都为HWC^2；由 Q 和 K 计算注意力权重时是由 Q 乘上 K 的转置，复杂度为$(HW)^2C$；自注意力权重和 K 的按元素乘法复杂度同样为$(HW)^2C$；最后经过一个复杂度为HWC^2投射层就完成了自注意力的计算。因此多头自注意力的计算复杂度可表示为公式 4.1。

$$\Omega(MSA) = 4HWC^2 + 2(HW)^2C \tag{4.1}$$

因此受限于当前的算力，ViT 往往不能处理分辨率较大的图像。事实上早期人们经常通过分组卷积的方式来缓解算力有限的问题，然而分组卷积一定程度上会减少不同通道间的信息融合，降低了模型特征提取的能力。Swin Transformer 采用了分组卷积的“分组”的思想，将图像划分成多个窗口，再将每个窗口划分成多个小块，现在每个窗口内计算自注意力，再在图像上对每个窗口计算自注意力，形成了一个层级式的模型。同时，再下一次计算时会将窗口平移，使得原始的窗口划分下，不同窗口直接也能有信息交互，很好的弥补了

“分组”带来的缺点。

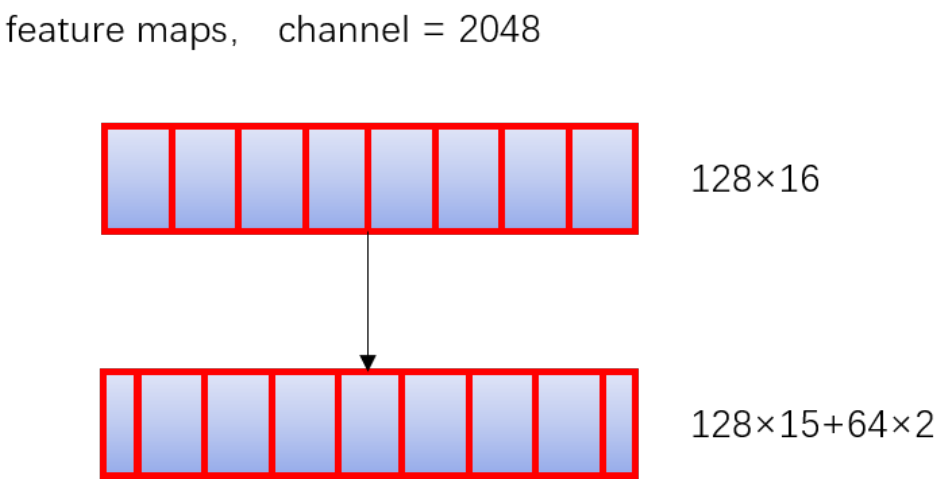


图 4.4 通道间滑动窗口注意力计算示意图

本章使用通道间的自注意力算法，由于通过 ResNet-18 提取的特征图足够的小，没有必要对其进行划分，而是将每个特征通道作为一个 token 展平成向量去计算 Q、K、V，这使得计算自注意力的成本大大减少了。然而如图 4.4 所示，所有的 token 在进行线性投射时，投射过后的维度决定了可以表征的信息量，然而维度过大时依然会使得在计算注意力时的时间代价很大。因此我们采用了通道间的滑动窗口注意力机制以降低计算复杂度。

本文中，通道间的滑动窗口注意力一共会计算两次注意力，如图 4.4 所示。第一次会将通道分成 8 个窗口，每个窗口会在窗口内单独计算通道间的注意力。经固定的线性映射后的向量维度为 2048，将窗口数量设置为 $win = 16$，则窗口大小 win_size =128，即每个窗口内包含 128 个特征通道，图 4.4 中每个红色边框的矩形内即表示一个窗口。第二次计算时会将窗口像右移动 $2/win$ 的距离，则会产生 15 个包含 128 个通道的窗口，并在两侧各留下一个包含 64 通道的窗口。

通道间的注意力的计算类似于图 4.3 中的计算流程，区别在于通道的维度已经被固定成了 dim，而不再是图像的宽高$H \times W$。因此通道间的注意力的计算复杂度可以表示为公式 4.2:

$$\Omega(A-MSA) = 4dimC^2 + 2dim^2C \tag{4.2}$$

而通道间的滑动窗口注意力计算则相当于进行了“分组”操作，可以有效降低计算复杂度，每个窗口内的计算复杂度为$\frac{4dimC^2}{win} + \frac{2dim^2C}{win^2}$，且一共有 win 个不重叠的窗口，因此其复杂度可表示为公式 4.3：

$$\Omega(AW-MSA) = 4dimC^2 + \frac{2dim^2C}{win} \tag{4.3}$$

事实上，在计算注意力时，主要的计算存在于 Q 矩阵和 K 矩阵做乘法求注意力矩阵，

以及注意力矩阵和 V 矩阵的乘积，也即公式 4.3 的第二项，由此可知，窗口数 *win* 越多，计算复杂度越小。然而，事实上是不可能无限的增大窗口数量的，因为窗口的大小越小，Transformer 所关注的越局限，其全局建模能力越差。当窗口尺寸很小时，单个窗口内的信息就越有限，而窗口数量增多时，需要更多的去弥补窗口之间的信息建模能力。因此，窗口的大小与模型建模能力是一个动态取舍的关系。

在计算第二次注意力时，会由于窗口的移动产生尺寸不一的窗口。为了保证窗口内的特征通道数量一直，同时减少计算开销，将两个较小的窗口合并，并通过注意力掩码去消除掉合并处的注意力。如图 4.5 所示，将一侧的大小为 *win_size*/2 的窗口 A 移动到另一侧与窗口 B 组成一个新的窗口，即可同一窗口的数量和大小。然而事实上，窗口 AB 本身并不相邻，其对应的语义信息也不在空间上近似，因此理论上窗口 A 和窗口 B 并不适合进行自注意力计算。本来采用了一种掩码，使得 A 和 B 组成的窗口即使存在来源不同的区域，也能通过一次前项过程计算注意力，并且 A 和 B 之间的特征通道不会互相干扰。

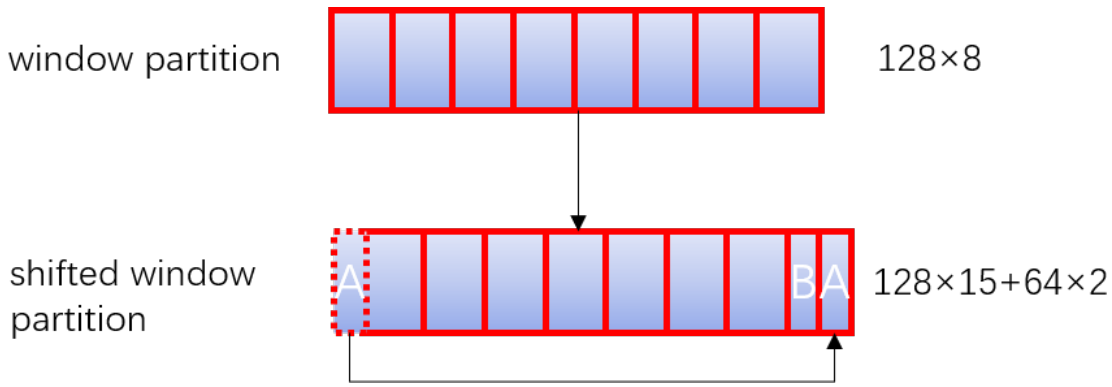


图 4.5 通道间滑动窗口注意力计算示意图

如图 4.6 所示，窗口 A 和窗口 B 中的维度都是 64，将两个窗口合并得到一个新的大小为 128 的窗口。计算新的窗口自注意力时，由于新窗口存在旧窗口 A 和旧窗口 B 两个部分，注意力掩码可以得到 AA、AB、BA、BB 四个区域，其中 AA 区域表示原窗口 A 内部的自注意力计算，BB 区域表示窗口原窗口 B 内部的自注意力计算。而 AB 区域和 BA 区域由于 A 和 B 原本并不处于同一窗口内，不应该做自注意力计算，需要将其屏蔽掉。最终的注意力掩码如图 4.7 所示。

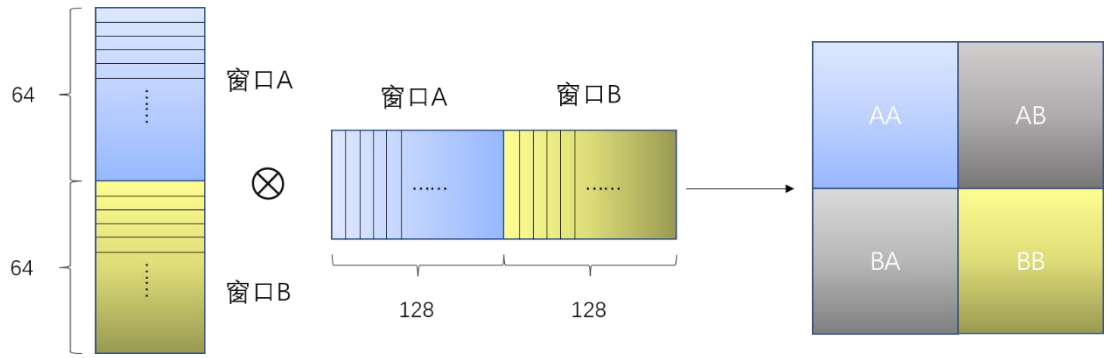


图 4.6 窗口 A 与窗口 B 合并后的注意力计算示意图

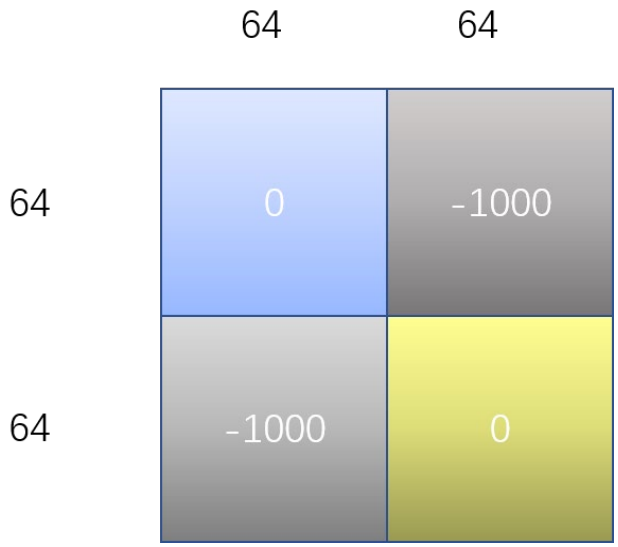


图 4.7 窗口 A 与窗口 B 合并后的注意力掩码

最终的注意力掩码是一个 64×64 的矩阵，其中需要保留的区域会被置为 0，需要被消除的区域会被置为-1000。注意力掩码会被加到注意力矩阵中，由于被覆盖的区域的注意力值会被加上一个极小的常数，这使得在通过 Softmax 函数时，需要被覆盖的区域的注意力值将是一个接近于 0 的极小整数。可以近似看作被该区域的自注意力值被覆盖掉。

4.2.3 帧间滑动窗口注意力机制

由于伪造检测的输入图像一般是裁剪出来的人脸，输入人脸图像的尺寸会比较小，一般采用 224×224 尺寸，这个尺寸不使用滑动窗口注意力机制，也是也可以完成注意力计算的。但是本章讨论将连续 16 帧图像作为网络输入，引入额外的时序特征时，就意味着每个 token 的信息来自于单帧图像的特征。此时 token 的维度过大，则会产生计算复杂度爆炸，而维度过小，又会限制其表征能力。因此本节将讨论，如何使用帧间的滑动窗口注意力机制，在低计算开销下，让网络关注更加重要的帧对应的特征信息。

通过 CNN 提取的特征序列的维度为 16×512×7×7，其中 16 为时间序列长度，512 为特征通道数，7×7 为特征图的宽高。此时将每个 token 对应的张量展平成一维向量后，其维度为 25088，将其通过线性映射层映射为 8192 维向量，将 8192 维向量划分为 64 个窗口后，每个窗口内单独计算注意力，之后再通过移动窗口的方式去计算新窗口下的注意力。其中，窗口移动方式和注意力掩码的设计同 4.2.2 节。

由于单个 token 既不是 ViT 模型中的切分后的小块，也不是通道间滑动窗口注意力计算的单个通道，而是单帧图像提取的特征信息。因此帧间滑动窗口注意力计算可以为不同帧的特征信息计算权重，提高关键帧的学习比重。

4.3 实验分析

4.3.1 数据处理与实验环境

本章同样采用 Celeb-DF 和 DFDC 数据集进行训练与测试。其中在对图像进行训练测试的时候，采用通道间的滑动窗口注意力机制；在对图像帧序列进行训练测试的时候，采用帧间滑动窗口注意力机制。

实验中使用的参数如表 4.2 所示，其中学习率采用线性衰减。

表 4.2 模型训练使用的超参数

Hyperparameters	Value
学习率	0.0001~0.00001
权重衰减	0.00001
批大小	16
训练轮次	100

4.3.2 图像级伪造检测实验

本章同样对各种 2D 检测模型在两个数据集上独立进行训练与测试，以比较模型的伪造检测的性能。

不同模型在两个数据集上的真伪检测精度和 AUC 如表 4.2 所示。可以看到，本章提出的基于通道间的滑动窗口注意力机制的伪造检测模型有着优秀的分类性能，在 DFDC 这样的多伪造方法数据集上也展示出了不错的鲁棒性。

表 4.3 不同 2D 伪造检测模型在各个数据集上的分类性能

Methods	Accuracy (Celeb-DF)	AUC (Celeb-DF)	Accuracy (DFDC)	AUC (DFDC)
MesoNet[28]	82.6%	0.714	83.3%	0.719
MesoInception[28]	81.8%	0.707	83.9%	0.727
CViT[62]	95.6%	0.984	88.8%	0.940
Xception[29]	93.3%	0.977	92.5%	0.956
Ours	96.7%	0.990	93.3%	0.962

4.3.3视频级伪造检测实验

本节对各种 3D 检测模型在两个数据集上独立进行训练与测试，以比较 3D 伪造检测模型的检测性能。

不同 3D 伪造检测模型在两个数据集上的分类性能对比结果如表 4.3 所示。事实上，当前由于人脸序列对齐精度的影响，以及实习特征提取能力受限。3D 检测的模型一直没有 2D 检测模型的精度高。例如，CViT 算法中 3D 模型比 2D 模型在 Celeb-DF 数据集上的精度下降了 2.7%；WildDeepfake 算法中 3D 模型比 2D 模型在 DFD 数据集上的精度下降了 2.6%。而本章基于带约束的人脸关键点对齐方法，以及帧间滑动窗口注意力机制，在 3D 伪造检测模型中仍然取得了优秀的分类性能。

表 4.4 不同 3D 检测模型在各个数据集上的分类性能

Methods	Accuracy (Celeb-DF)	AUC (Celeb-DF)	Accuracy (DFDC)	AUC (DFDC)
CViT-3D[62]	94.2%	0.974	86.9%	0.909
Conv-LSTM	92.3%	0.961	87.5%	0.913
Xception-3D[29]	90.6%	0.953	90.8%	0.933
Ours	95.1%	0.979	91.7%	0.942

4.3.4 其他对比实验

本章对于注意力掩码设计了三种不同的区域赋值，在同一个模型下 Celeb-DF 数据集上的性能如表 4.5 所示。

表 4.5 在 Celeb-DF 数据集上的不同注意力掩码赋值的性能比较

组别	R_u	R_f	R_o	AUC
a	0	0.5	1	0.982
b	0	0.7	1	0.984
c	0.3	0.7	1	0.990

4.4 本章小结

本章提出了一种基于滑动窗口注意力机制的深度伪造模型，通过“分组”的思想，有效降低了高维度向量计算注意力时的复杂度，并分别基于通道间的窗口注意力机制和帧间注意力机制设计了 2D 伪造检测和 3D 伪造检测模型。在 Celeb-DF 和 DFDC 数据集上进行了性能测试，不论是 2D 检测模型还是 3D 检测模型都展现出了很好的分类精度和良好的泛化性。

第五章　总结与展望

5.1 工作总结

本文主要就深度伪造检测工作中出现的问题和难点进行了原理性分析，并就提出的问题和难点或提出了相应的解决方法，或设计了可以应对的模型结构。

第一，由于伪造者的目的是让伪造的视频图像看起来足够逼真，因此伪造痕迹一般很难在肉眼可见的 RGB 域中被挖掘。本文从频域和空域分别提取特征，设计出双流网络模型进行伪造检测。并且尝试了几种不同的特征融合和逻辑融合的方式，最终使得模型的预测性能高于两个分支网络的任一分支的性能。

第二，图像数据中包含了大量的冗余信息，而对于伪造检测任务而言，重要的只是图像中的部分结构信息，而剩余的大部分内容信息是冗余的，如何高效地提取特征并筛除冗余信息也是一大难点。对此，本文设计了一种通过 CNN 提取特征，通过 Transformer 进行强化特征选择的模型。再次基础上，在通道间的自注意力机制上提出通道间的滑动窗口自注意力机制，通过分组的思想，简化了计算自注意力时的复杂度。

第三，在 3D 伪造检测模型中，受限于人脸标注的精度和视频中人物面部运动幅度过大，视频帧很难做到人脸很好的对齐。本文根据人脸的几个较为稳定的关键点，包括眼部的 37 号关键点和 46 号关键点以及鼻子的 31 号关键点来进行人脸标定框的优化。首先根据眼部关键点将图像中的人脸旋转至两个眼角的连线在水平方向，再根据上一帧的人脸关键点对当前标定框进行缩放以尽可能地进行对齐。

第四，由于为了尽可能对齐上述人脸标定框序列，可能会在截取时比一般人脸标定算法截取的面积过大，产生更大的背景区域，而背景信息在进行伪造检测时一般是没有用的。本文使用了一种层级式的面部注意力掩码，将面部划分为不同区域以使得模型可以关注到更加重要的区域，也即前景区域特别是前景区域中的面部齐全区域。同时，本文还测试了几种不同区域赋值的掩码在同一模型中的效果。

第五，在视频级伪造检测任务中，不同帧中人物的表情和姿态的不同会导致检测难度的不同，如何选择重要的帧上的信息也是 3D 伪造检测模型的一个重点。本文通过 Transformer 进行时序特征的提取，然而以图片或特征图为单位的序列化数据会使得计算注意力时的计算开销十分大，而简单地通过线性映射降维又会降低其表征能力。本文提出了帧间滑动窗口的注意力机制，在一定的计算复杂度内完成了关键帧信息的强化。

本文基于上述的工作，分别提出了两种深度伪造检测模型。并在 Celeb-DF 数据集和 DFDC 数据集上进行了测试，验证了模型的高分类的性能。通过不同数据集上的交叉测试，体现了模型在未知伪造方法下也有着优秀的泛化性和鲁棒性。此外，本文还进行了多项消融实验以验证模型设计的合理性。

5.2 未来展望

本文提出了基于多域特征网络的深度伪造检测模型，在此基础上，仍然有以下待改进或有待探究的工作点，未来的研究重点如下：

1） 小样本问题

对于伪造检测任务来说，目前仍然存在样本不足的问题，特别是高质量伪造视频的缺失。而在一个常见的开源数据集中，往往存在一些视频图像具有肉眼可见的伪造瑕疵，也存在一些高性能模型也无法区分的优质伪造视频。因此，提高伪造图像生成模型的稳定性，消除伪造瑕疵并生成大量的开源数据集是一个研究的热点。另一方面来说，数据集的多样化也是有利于伪造技术和伪造检测技术的发展的。

2） 困难样本问题

本文在实验中发现，多个模型分类错误的样本就有很高的重合度。这意味着对于这些数据集来说，其中存在着一些容易被判别成为负例的正样本和容易被判别成正例的负样本。研究困难样本被误分类的原因有利于对模型进行优化，从而拜托目前达到的分类精度的瓶颈。

3） 更好的帧间防抖动算法

基于 dlib 或是更准确的 RetinaFace 算法[68]标注的人脸标定框和面部关键点依然存在着帧间抖动的问题，这会直接影响到模型的性能。本文虽然采用了一些旋转、缩放的方式去进行帧间对齐，但是抖动依然存在。如何以较低的代价进行帧间的人脸对齐有利于 3D 检测模型的性能提高。

本文提出了基于滑动窗口注意力的深度伪造检测摸，基于本文已有的工作，未来的其他研究方向和改进工作可能如下：

1） 区域注意力掩码的泛化性

本文采用的面部注意力掩码是一种人为预设的掩码，相当于引入了特定的先验到模型中。尽管本文尝试了几种不同的区域掩码值进行对比，但是由于存在不同的伪造方法，每种伪造方法理论上都应该有不同的重点区域。例如，对于只在眼部进行伪造的视频图像而言，眼部区域和嘴部鼻子等区域的重要性就并不相同。对所有的伪造方法生成的视频图像都采用同一种预设好的注意力掩码，会在一定程度上影响模型的泛化性。因此，能够应对不同伪造方法的高鲁棒性伪造检测模型是未来研究的重要方向。

2） 基于特定人物的伪造检测

事实上，不同人物不仅在面部神态、行为动作上存在着比较特殊的表现模式，这种模式是伪造者难以高质量伪造的。而在伪造攻击变得更频发的如今，伪造一些大国政要和公众人物的视频是最亟待解决的一类，对特定人物的伪造检测也将成为一个研究热点。解决这样一类的问题，需要对这些特定的人物进行行为表情的模式分析并与待检测的视频进行对比。

3） 更全面的评价指标

当前的伪造检测任务一般会被建模成一个二分类问题，也即去对图像或视频进行真伪的二分类。然而一个视频中被伪造的可能是音频而不是画面，可能是物体而不是人物。而单一的伪造检测方法一般只会对其中一种进行伪造检测。因此，理论上来说一个视频需要有包含其多个属性的真伪标签，而想要完备地检测一个视频的真实性，需要分级检测多项视频属性的真实性，进行多标签的检测。

参考文献

[1] Chesney R, Citron D. Deepfakes and the New Disinformation War[J]. Foreign affairs, 2019, 98(1): 147-155.

[2] Shu Z, Sahasrabudhe M, Guler R A, et al. Deforming autoencoders: Unsupervised disentangling of shape and appearance[C]. Proceedings of the European conference on computer vision (ECCV). 2018: 650-665.

[3] 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.

[4] Li Y, Yang X, Sun P, et al. Celeb-DF: A Large-scale Challenging Dataset for DeepFake Forensics[C]. Proc of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2020: 3204-3213.

[5] Korshunova I, Shi W, Dambre J, et al. Fast Face-Swap Using Convolutional Neural Networks[C]. Proceedings of IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017. IEEE Computer Society, 2017. 3697–3705.

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

[7] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets[C]. Proc of the 27th Int Conf on Neural Information Processing Systems. La Jolla, CA :NIPS, 2014: 2672-2680.

[8] 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.

[9] Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models[J]. Advances in Neural Information Processing Systems, 2020, 33: 6840-6851.

[10] Ramesh A, Dhariwal P, Nichol A, et al. Hierarchical text-conditional image generation with clip latents[J]. arXiv preprint arXiv:2204.06125, 2022.

[11] Torzdf, Faceswap: Deepfakes software for all[EB/OL]. http://www.github.com/faceswap.

[12] Dolhansky B, Bitton J, Pflaum B, et al. The DeepFake Detection Challenge (DFDC) Preview Dataset[J]. arXiv preprint arXiv: 1910.08854, 2019.

[13] Vilaplana J M, Batlle J F, Coronado J L. Connectionist models of cortico-basal ganglia adaptive neural networks during learning of motor sequential procedures[C]. International Work-Conference on Artificial Neural Networks. Springer, Berlin, Heidelberg, 2001: 394-401.

[14] Perov I, Gao D, Chervoniy N, et al. DeepFaceLab: Integrated, flexible and extensible face-swapping framework[J]. arXiv preprint arXiv:2005.05535, 2020.

[15] Nirkin Y, Keller Y, Hassner T. FSGAN: Subject Agnostic Face Swapping and Reenactment[J].

International Conference on Computer Vision, 2019, 7183-7192

[16] Ryota N, Tatsuya Y, Shigeo M. Rsgan: face swapping and editing using face and hair representation in latent spaces[J]. arXiv preprint arXiv: 1804.03447, 2018.

[17] Ramesh A, Pavlov M, Goh G, et al. Zero-shot text-to-image generation[C]. International Conference on Machine Learning. PMLR, 2021: 8821-8831.

[18] Nichol A, Dhariwal P, Ramesh A, et al. Glide: Towards photorealistic image generation and editing with text-guided diffusion models[J]. arXiv preprint arXiv:2112.10741, 2021.

[19] Li Y, Chang M C, Lyu S. In Ictu Oculi: Exposing AI Generated Fake Face Videos by Detecting Eye Blinking[C]. Proc of the 2018 IEEE International Workshop on Information Forensics and Security (WIFS). Piscataway, NJ: IEEE, 2018: 1-7.

[20] Haliassos A, Vougioukas K, Petridis S, et al. Lips don't lie: A generalisable and robust approach to face forgery detection[C]. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 5039-5049.

[21] 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.

[22] Zi B, Chang M, Chen J, et al. Wilddeepfake: A challenging real-world dataset for deepfake detection[C]. Proceedings of the 28th ACM international conference on multimedia. 2020: 2382-2390.

[23] Li L, Bao J, Zhang T, et al. Face x-ray for more general face forgery detection[C]. Proc of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2020: 5001-5010.

[24] Durall R, Keuper M, Pfreundt F J, et al. Unmasking deepfakes with simple features[J]. arXiv preprint arXiv: 1911.00686, 2019.

[25] Masi I, Killekar A, Mascarenhas R M, et al. Two-branch recurrent network for isolating deepfakes in videos[C]. Proc of the 2020 European Conference on Computer Vision. Berlin, German: Springer, 2020: 667-684.

[26] Li J, Xie H, Li J, et al. Frequency-aware Discriminative Feature Learning Supervised by Single-Center Loss for Face Forgery Detection[C]. Proc of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2021: 6458-6467.

[27] Wang J, Wu Z, Chen J, et al. M2TR: Multi-modal Multi-scale Transformers for Deepfake Detection[J]. arXiv preprint arXiv: 2104.09770, 2021.

[28] 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.

[29] Chollet F. Xception: Deep Learning with Depthwise Separable Convolutions[C]. Proc of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ: IEEE, 2017: 1251-1258.

[30] Rossler A, Cozzolino D, Verdoliva L, et al. FaceForensics++: Learning to Detect Manipulated Facial Images[C]. Proc of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway, NJ: IEEE, 2020: 1-11

[31] Dang H, Liu F, Stehouwer J, et al. On the detection of digital face manipulation[C]. Proc of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2020: 5781-5790.

[32] Liu Z, Lin Y, Cao Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 10012-10022.

[33] Tolosana R, Vera-Rodriguez R, Fierrez J, et al. Deepfakes and beyond: A survey of face manipulation and fake detection[J]. Information Fusion, 2020, 64: 131-148.

[34] Thies J, Zollhöfer M, Nießner M. Deferred neural rendering: Image synthesis using neural textures[J]. ACM Transactions on Graphics, 2019, 38(4): 1-12.

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

[36] Raghavendra R, Raja K B, Venkatesh S, et al. Face morphing versus face averaging: Vulnerability and detection[C]. 2017 IEEE International Joint Conference on Biometrics (IJCB). IEEE, 2017: 555-563.

[37] Kaneko T, Kameoka H, Tanaka K, et al. Cyclegan-vc3: Examining and improving cyclegan-vcs for mel-spectrogram conversion[J]. arXiv preprint arXiv:2010.11672, 2020.

[38] Mao X, Li Q, Xie H, et al. Least squares generative adversarial networks[C]. Proceedings of the IEEE international conference on computer vision. 2017: 2794-2802.

[39] He Z, Zuo W, Kan M, et al. Attgan: Facial attribute editing by only changing what you want[J]. IEEE transactions on image processing, 2019, 28(11): 5464-5478.

[40] Li Y, Chang M, Lyu S. Exposing AI Generated Fake Face Videos by Detecting Eye Blinking[J]. arXiv preprint arXiv:1806.02877, 2018.

[41] 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.

[42] Greff K, Srivastava R K, Koutník J, et al. LSTM: A search space odyssey[J]. IEEE transactions on neural networks and learning systems, 2016, 28(10): 2222-2232.

[43] Qi H, Guo Q, Juefei-Xu F, et al. Deeprhythm: Exposing deepfakes with attentional visual heartbeat rhythms[C]. Proceedings of the 28th ACM international conference on multimedia. 2020: 4318-4327.

[44] Hernandez-Ortega J, Tolosana R, Fierrez J, et al. Deepfakeson-phys: Deepfakes detection based on heart rate estimation[J]. arXiv preprint arXiv:2010.00400, 2020.

[45] Allamanis M, Peng H, Sutton C. A convolutional attention network for extreme summarization of source code[C]. International conference on machine learning. PMLR, 2016: 2091-2100.

[46] Haliassos A, Vougioukas K, Petridis S, et al. Lips don't lie: A generalisable and robust approach to face forgery detection[C]. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 5039-5049.

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

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

[49] Yang T, Huang Z, Cao J, et al. Deepfake Network Architecture Attribution[J]. arXiv preprint arXiv:2202.13843, 2022.

[50] 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.

[51] Li C L, Chang W C, Cheng Y, et al. Mmd gan: Towards deeper understanding of moment matching network[J]. Advances in neural information processing systems, 2017, 30.

[52] Miyato T, Kataoka T, Koyama M, et al. Spectral normalization for generative adversarial networks[J]. arXiv preprint arXiv:1802.05957, 2018.

[53] Lee K S, Tran N T, Cheung N M. Infomax-gan: Improved adversarial image generation via information maximization and contrastive learning[C]. Proceedings of the IEEE/CVF winter conference on applications of computer vision. 2021: 3942-3952.

[54] Liu Z, Luo P, Wang X, et al. Large-scale celebfaces attributes (celeba) dataset[J]. Retrieved August, 2018, 15(2018): 11.

[55] 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.

[56] Qian Y, Yin G, Sheng L, et al. Thinking in frequency: Face forgery detection by mining frequency-aware clues[C]. European conference on computer vision. Springer, Cham, 2020: 86-103.

[57] Lam E Y, Goodman J W. A mathematical analysis of the DCT coefficient distributions for images[J]. IEEE transactions on image processing, 2000, 9(10): 1661-1666.

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

[59] 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.

[60] Touvron H, Cord M, Douze M, et al. Training data-efficient image transformers & distillation through attention[C]. International Conference on Machine Learning. PMLR, 2021: 10347-10357.

[61] Xia Z, Pan X, Song S, et al. Vision transformer with deformable attention[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 4794-4803.

[62] Wodajo D, Atnafu S. Deepfake video detection using convolutional vision transformer[J]. arXiv preprint arXiv:2102.11126, 2021.

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

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

[65] Montserrat D M, Hao H, Yarlagadda S K, et al. Deepfakes detection with automatic face weighting[C]. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops. 2020: 668-669.

[66] Lydick E, Epstein R S, Himmelberger D, et al. Area under the curve: a metric for patient subjective responses in episodic diseases[J]. Quality of life research, 1995, 4(1): 41-45.

[67] 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.

[68] Deng J, Guo J, Ververas E, et al. Retinaface: Single-shot multi-level face localisation in the wild[C]. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 5203-5212.

致 谢

首先，我要衷心感谢我的导师谭晓阳教授，感谢在研究生期间他对我的耐心培养和悉心指导，在研究生期间的组会和讨论中我收获颇丰。此外，我要感谢我的实验室师兄师姐和同门，是他们在生活和学业上给了我很多建议，特别感谢蒋珂师兄和张敏师姐在我刚刚入门时解决了很多我的疑问。

其次我要感谢我的室友和朋友，我们一起生活和学习，既可以一起讨论学习和工作，又会在一起运动休闲，感谢他们陪我度过精彩的研究生生活。最后我要感谢我的父母，是他们从小到大在背后一直默默关心和支持我。

在学期间的研究成果及发表的学术论文

攻读硕士学位期间发表（录用）论文情况

李灿东，谭晓阳. 基于多域特征网络的深度伪造视频检测[C]. 第九届中国数据挖掘会议.2022.

攻读硕士学位期间参加科研项目情况

国家自然科学基金(61976115,61732006)；

南航人工智能+项目(NZ2020012)；

装备共性研究项目(50912040302).