武漢理工大學

（申请工程硕士学位论文）

基于光流运动补偿的视频插帧技术研究

培 养 单 位 ：计算机与人工智能学院

学 科 专 业 ：软件工程

研 究 生 ：李智

指 导 教 师 ：胡燕 教授

副指导教师 ：李伟 高级工程师

2022 年 5 月

基于光流运动补偿的视频插帧技术研究 李智 武汉理工大学

分类号_____ 密　　级_____

UDC_____ 学校代码 10497

武漢理工大學

学位论文

题　目 基于光流运动补偿的视频插帧技术研究

英　文 Research on Video Frame Interpolation Based on

题　目 Optical Flow Motion Compensation

研究生姓名 李智

指导教师 姓　　名 胡燕 职称 教授 学位 博士

单位名称 计算机与人工智能学院 邮编 430070

副指导教师 姓　　名 李伟 职称 高级工程师 学位 硕士

单位名称 伟乐视讯科技股份有限公司 邮编 516000

申请学位级别 硕士 学科专业名称 软件工程

论文提交日期 2022 年 4 月 论文答辩日期 2022 年 5 月

学位授予单位 武汉理工大学 学位授予日期 2022 年 6 月

答辩委员会主席 饶文碧 评阅人 教育部学位中心盲审

教育部学位中心盲审

2022 年 5 月

独创性声明

本人声明，所呈交的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。尽我所知，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，也不包含为获得武汉理工大学或其他教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

签名：李智 日期：2022.05.21

学位论文使用授权书

本人完全了解武汉理工大学有关保留、使用学位论文的规定，即学校有权保留并向国家有关部门或机构送交论文的复印件和电子版，允许论文被查阅和借阅。本人承诺所提交的学位论文（含电子学位论文）为答辩后经修改的最终定稿学位论文，并授权武汉理工大学可以将本学位论文的全部内容编入有关数据库进行检索，可以采用影印、缩印或其他复制手段保存或汇编本学位论文。同时授权经武汉理工大学认可的国家有关机构或论文数据库使用或收录本学位论文，并向社会公众提供信息服务。

（保密的论文在解密后应遵守此规定）

研究生（签名）：李智 导师（签名）：[illegible] 日期 2022.05.21

摘 要

随着视频直播技术的日益成熟，越来越多的直播内容与形式出现在各大视频直播平台，在满足内容需求的前提下，用户对于视频质量的需求也越来越高。受视频采集设备和网络带宽等条件的限制，各大视频直播平台上的低帧率视频与现有的高刷新率显示设备无法匹配，这就迫切地需要将低帧率视频转化为高帧率视频，给用户提供更好的视觉体验。本文的研究旨在设计与实现一种可以有效提高低帧率视频插帧质量的高效视频插帧方法，具体研究工作如下：

（1）针对现有基于光流运动补偿的视频插帧方法中，插帧结果过于依赖光流的准确性，光流提取中的特征点匹配误差会造成原始图像中纹理细节的损失，同时采用 U-Net 结构的帧合成网络只能传递单一尺度下的特征信息，精细的语义信息会随着卷积操作而逐渐丢失等问题，提出了一种基于全尺度纹理感知的视频插帧方法。首先，引入纹理提取网络提取多尺度纹理特征以及纹理匹配损失以弥补损失的纹理细节信息。其次，在帧合成网络中引入全尺度跳跃连接模块，将各层同尺度间的连接拓展为编码器层与解码器子网络的连接，充分利用编码器子网络的特征信息。最后在 Vimeo90K 、UCF101、Middlebury 以及 HD 四个公开数据集上进行实验与验证，结果表明，该算法能有效恢复中间帧合成过程中损失的纹理细节，并且整体性能优于其他视频插帧方法。

（2）针对多个特征提取网络中的重复卷积操作，为网络模型增加了大量的矩阵乘法运算，同时静态卷积网络中对样本的每一个像素均采用相同大小卷积核而导致模型容量较低等问题，提出了一种基于可分离动态卷积的高效视频插帧方法。首先，引入动态卷积方法替换掉光流特征、上下文特征以及纹理特征提取网络中的静态卷积，使用动态卷积将图像特征划分为若干个区域并为每个区域生成不同大小的卷积滤波器，在降低模型参数量的同时，提升网络特征表达能力。其次，通过引入深度可分离卷积，将动态卷积生成的二维卷积滤波器进行分解，大大减少模型参数量。最后在 Vimeo90K 与 UCF101 数据集上进行实验与验证，结果表明，基于可分离动态卷积的高效视频插帧方法计算效率更高，可以应用到实时场景中。

（3）本文将提出的视频插帧方法应用到音视频编解码服务器（CMP308）的视频插帧子模块中。根据视频直播平台的应用场景需求，设计了一套合理的视频插帧任务工作流程，实现了实时视频插帧算法，并对其进行了实验分析以

及功能展示。

关键词：视频插帧；纹理特征；全尺度跳跃连接；深度可分离卷积；动态卷积；

Abstract

With the growing maturity of live video technology, more and more live broadcast content and forms appear on major video live broadcast platforms. Under the premise of satisfying content demands, users have higher and higher demands for video quality. Restricted by conditions such as video capture equipment and network bandwidth, the low frame rate video on the existing major video live broadcast platforms cannot match the high refresh rate display equipment, which urgently needs to convert low frame rate video into high frame rate video, to provide users with a better visual experience. The purpose of this thesis is to design and implement an efficient video frame interpolation method that can effectively improve the quality of low frame rate video frame interpolation. The specific research works are as follows:

(1) In view of the existing video frame interpolation methods based on optical flow motion compensation, the results of frame interpolation rely too much on the accuracy of optical flow, and the feature point matching error in optical flow estimation will cause the loss of texture details in the original image. At the same time, the frame synthesis network using the U-Net structure can only transmit the feature information at a single scale, and the fine semantic information will be gradually lost with the convolution operation. A video frame interpolation method based on full-scale texture-aware is proposed. Firstly, a texture extraction network is introduced to extract multi-scale texture features and texture matching loss to compensate for the lost texture details. Secondly, a full-scale skip connection module is introduced into the frame synthesis network, and the connection between the same scale of each layer is extended to the connection between the encoder layer and the decoder sub-network, and the feature information of the encoder sub-network is fully utilized. Finally, experiments and validation were conducted on four publicly available datasets, Vimeo90K, UCF101, Middlebury, and HD. The results show that the algorithm can effectively restore the texture details lost in the intermediate frame synthesis process, and the overall performance is better than other video frame interpolation methods.

(2) Aiming at the repeated convolution operation in multiple feature extraction networks, a large number of matrix multiplication operations are added to the network model, and the low model capacity is caused by using the same size convolution kernel for each pixel of the sample in the static convolution network. An efficient video frame interpolation method based on separable dynamic convolution is proposed. Firstly, the dynamic convolution method is introduced to replace the static convolution in the optical flow feature, context feature and texture feature extraction network, and dynamic convolution is used to divide the image features into several regions and generate convolution filters of different sizes for each region. It can improve the network feature expression ability while reducing the amount of model parameters. Secondly, by introducing depth separable convolution, the two-dimensional convolution filter generated by dynamic convolution is decomposed, which greatly reduces the amount of model parameters. Finally, experiments and validation were conducted on the Vimeo90K and UCF101 datasets. The results show that the efficient video frame interpolation method based on separable dynamic convolution is more computationally efficient and can be applied to real-time scenes.

(3) In this thesis, the proposed video frame interpolation method is applied to the video framing sub function module of audio and video codec server (CMP308). According to the application scenario requirements of the live video platform, a set of reasonable video frame interpolation task process is designed, the real-time video frame interpolation algorithm is implemented, and the experimental analysis and function display are carried out.

Keyword: Video frame interpolation; Texture feature; Full-scale skip connection; Depth separable convolution; Dynamic convolution;

目 录

摘 要 I

Abstract III

目 录 V

第 1 章 绪论 1

1.1 研究背景与意义 1

1.2 国内外研究现状 2

1.2.1 基于卷积核的视频插帧方法 3

1.2.2 基于光流运动补偿的视频插帧方法 4

1.3 主要研究内容 7

1.4 论文组织结构 8

第 2 章 基于全尺度纹理感知的视频插帧方法 10

2.1 FSTA-RIFE 视频插帧模型的构建 10

2.1.1 RIFE 实时中间流估计网络结构 10

2.1.2 FSTA-RIFE 视频插帧网络的构建 12

2.2 FSTA-RIFE 方法关键技术及算法 14

2.2.1 基于 VGG-16 网络提取多尺度纹理特征 14

2.2.2 基于全尺度跳跃连接进行高效特征传递 15

2.2.3 损失函数 17

2.2.4 FSTA-RIFE 训练算法 18

2.3 实验结果及分析 20

2.3.1 实验环境与实验数据 20

2.3.2 评价指标 20

2.3.3 实验设置 22

2.3.4 有效性验证 23

2.3.5 网络总体性能对比 24

2.4 本章小结 27

第 3 章 基于动态卷积的高效视频插帧方法 28

3.1 问题的提出 28

3.2 FSTA-DCRIFE 方法关键技术及算法 29

3.2.1 基于深度可分离卷积的视频插帧方法.....29
3.2.2 基于动态卷积的视频插帧方法.....30
3.2.3 基于深度可分离的动态卷积结构.....33
3.2.4 FSTA-DCRIFE 训练算法.....34
3.3 实验结果与分析.....35
3.3.1 实验环境与实验数据.....35
3.3.2 评价指标.....35
3.3.3 实验设置.....36
3.3.4 有效性验证.....36
3.3.5 网络总体性能对比.....37
3.4 本章小结.....39
第 4 章 高效视频插帧算法的应用.....40
4.1 视频直播系统架构.....40
4.2 视频插帧模块的设计与实现.....42
4.2.1 视频插帧模块工作流程.....42
4.2.2 视频插帧模块的实现.....43
4.3 实验结果分析.....44
4.4 应用功能展示.....46
4.5 本章小结.....48
第 5 章 总结与展望.....49
5.1 论文工作总结.....49
5.2 未来工作展望.....50
致 谢.....52
参考文献.....53
攻读硕士学位期间相关的工作.....58

第 1 章 绪论

1.1 研究背景与意义

随着多媒体技术的高速发展以及各种硬件显示设备的更新换代，人们对多媒体内容质量的要求也愈发提高。视频作为当前最主流的多媒体呈现形式，具有直观、信息量大、声色兼备等特点。视频由若干张连续的图像帧组成，单位时间内连续显示的图像帧数称为帧率（FPS，Frame Per Second），帧率是评判视频质量优劣的指标之一，帧数越高，视频的视觉效果越流畅，同时所需的存储空间及传输带宽也越大。近年来，随着各种依托视频为载体的技术得以应用，短视频、视频直播、在线教育、在线会议等视频服务已经渗透到人们的日常工作与生活中，研究人员对视频帧率的优化也在不断地探索。

视频插帧技术（Video Frame Interpolation）作为计算机视觉领域中的一项重要且基础的工作，是视频标注、动作识别、目标检测等高级任务的重要基础。视频插帧技术是指根据视频中连续的两帧或者多帧图像生成中间帧的方法。视频插帧的应用场景广泛，可用于帧率上转换[1]、视频编码中的帧恢复和帧间预测[2]，慢动作生成[3]以及视图合成[4]。

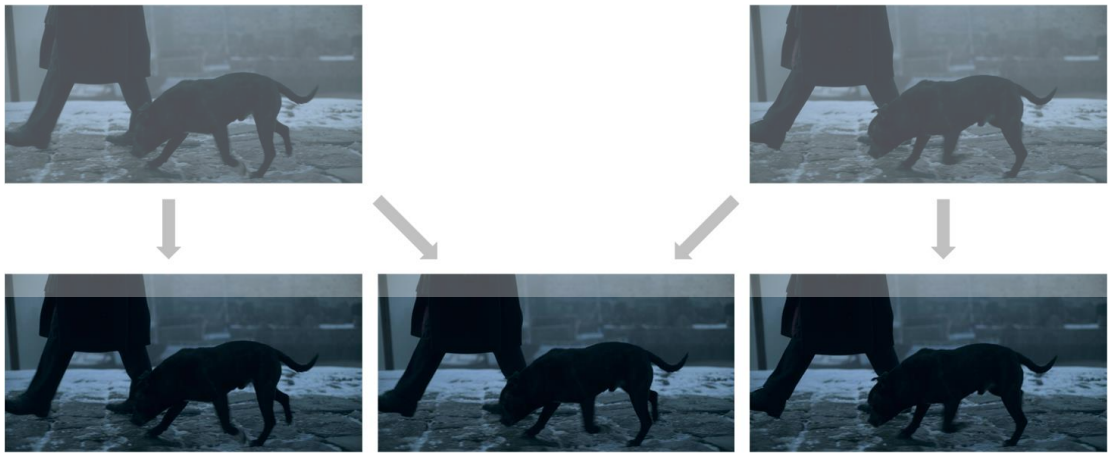

图 1-1 视频插帧技术示意图

受制于视频采集设备、视频存储成本以及通信传输带宽等因素，目前各大视频直播平台的视频流帧率最高只支持 60FPS。近年来，随着各种硬件显示设备的更新换代，用户开始倾向于选择高帧率视频以获得更好的观看体验。屏幕刷新率（单位，Hz）代表硬件显示设备在单位时间内刷新图像帧的次数。目前，移动显示设备刷新率最高可以支持 120Hz，桌面显示设备刷新率最高可以支持 240Hz。现有的视频流资源远没有发挥出硬件显示设备的全部性能，这就迫切地需要使用视频插帧算法实现将低帧率视频转换为高帧率视频。

近年来，得益于卷积神经网络在各种图像处理任务中取得的进展，大量研究人员开始将其应用于光流估计与视频插帧等任务中，并取得了不俗的成果，然而将基于光流运动补偿的视频插帧方法应用到具有丰富纹理细节的视觉场景中，插帧结果往往会出现局部模糊块，同时光流等额外特征的提取需要消耗大量的计算资源，限制了视频插帧算法在实时场景中的应用。因此，如何改善基于光流运动补偿的视频插帧方法的插帧质量，减少插帧结果中纹理细节的丢失，同时提高视频插帧算法的推理效率，是一个极具价值与挑战的研究。

1.2 国内外研究现状

传统的视频插帧技术往往采用简单的帧拷贝、帧平均等策略[5]直接从原始帧序列中计算得到中间帧。帧拷贝的策略是将原始帧序列中每一帧拷贝一次作为补帧方法，这种方法无需额外的计算消耗，并且在相对静止的场景下表现良好，但在转场以及大量像素运动的场景下，会有明显的画面顿挫感。帧平均的策略是将前后两帧作为输入，计算前后两帧的像素加权平均来获取插帧结果，这种方法的计算复杂度较低且利用了更丰富的原始帧信息，但是仅仅采用简单的加权平均算法，在大位移运动场景下，中间帧的运动像素同时保留了前后两帧的部分残像，导致中间帧画面模糊甚至出现伪影等问题。

鉴于深度学习在图像重建、图像合成等领域的成功，越来越多的研究人员开始尝试将深度学习方法应用到视频插帧领域中。目前基于深度学习的视频插帧方法主要可分为两类：（1）基于卷积核的视频插帧方法；（2）基于光流运动补偿的视频插帧方法。

1.2.1 基于卷积核的视频插帧方法

基于卷积核的视频插帧方法将光流提取视为一种中间过程，通过卷积核与输入帧进行卷积操作直接获取插帧结果。Long 等人[6]首次将深度卷积神经网络引入到视频插帧领域中，通过对两张输入帧进行卷积操作，直接预测出插帧结果，但是这种方法得到的插帧结果趋于模糊。为了得到更清晰的插帧结果，Niklaus 等人[7]提出了 AdaConv 方法，该方法通过神经网络估计每个输出像素的一对空间自适应卷积核，通过对输入帧进行局部卷积来生成插值像素。将像素插值转化为对像素块的卷积，并且能够从相对较大的邻域中合成像素。为了减少大量的内存空间消耗，Niklaus 等人[8]提出了 SepConv 方法，将每个二维卷积核分解为两个维度上的一维卷积核，在一定程度上提高了性能，但它无法处理大位移的运动场景。为了解决这个问题，Hyeongmin 等人[9]提出了 AdaCof 方法，该方法参考任意位置，以及任意数量像素点来合成中间帧，作者计算了指向参考位置的多个抵消向量，并对他们进行采样，最终通过线性合成采样值获取目标像素。Cheng 等人[10]提出 DSepConv 方法，采用可变形的分离卷积替换传统的标准卷积，自适应计算卷积核大小及偏移量，使得网络能够以更少但相关性更大的像素获取特征信息。这种方法能够对大位移运动场景建模，但是仍然没有解决运动遮挡造成的模糊问题。

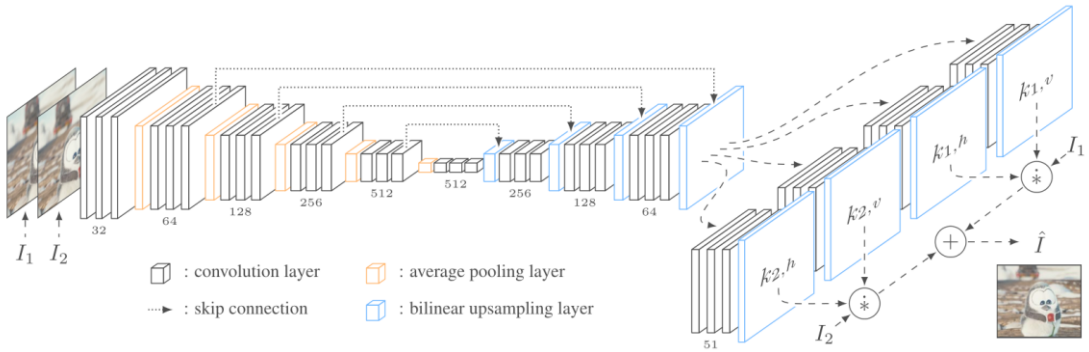

图 1-2 SepConv[8]网络总体结构

为了解决运动遮挡问题，Choi 等人[11]提出了一种特征重构方法，利用 PixelShuffle 操作将运动相关的信息逐步分布到多个通道中，与通道注意力结合，捕获两帧之间的运动信息以合成中间帧，该方法能有效处理运动遮挡问题。Kalluri 等人[12]提出了一种端到端的 3D 时空卷积网络对非线性运动以及时序信

息进行建模，但是由于 3D 卷积核需要大量的输入帧来建模时序信息，因此该模型的空间复杂度较大。Peleg 等人[13]提出了 IM-Net 方法，该模型使用卷积滤波器计算前后帧的运动矢量以及运动遮挡图，然后将 IMVF 遮挡图与输入帧一起传递给中间帧合成网络生成插值结果，该模型有效解决了高分辨率视频下的运动遮挡问题。最近，Ding 等人[14]提出了一种压缩驱动的帧插值网络，通过稀疏诱导优化利用模型减枝来显著减小模型尺寸，同时引入一个多尺度变形模块，提高了多级图像细节的一致性。Shi 等人[15]提出了 GDConv 方法，该方法采用了一种广义变形卷积机制，允许采样点在连续的时空域中自由移动，所以在处理大位移运动场景时具有很大优势。这些基于卷积核的方法[6-19]网络结构设计相对简单，且均使用卷积核对帧间运动信息进行隐式建模，无法建模现实世界中的复杂运动。同时受卷积核大小的限制，位于感受野范围之外的像素没有作为插值像素的参考，如果试图在高分辨率图像中增加对大位移运动的覆盖，则需要很高的内存占用以及计算复杂度。

1.2.2 基于光流运动补偿的视频插帧方法

基于光流运动补偿的视频插帧方法通常包含光流提取以及运动补偿两个步骤，首先提取两帧之间的双向光流特征，然后将输入帧与光流进行变形合成中间帧。Liu 等人[20]提出了一种名为 DVF 的全卷积编解码器结构网络，用于估计跨越空间和时间的 3D 流，然后通过三线性采样变形得到中间帧。Jiang 等人[21]提出的 SuperSlomo 方法使用了两个 U-Net[22]网络联合建模运动估计和遮挡分析过程，首先使用 U-Net 网络提取前后两帧的双向光流，使用另一个 U-Net 网络预测软可见性图以去除运动边界产生的伪像，该方法支持多倍插帧并有效解决了伪像问题。Liu 等人[23]提出的 CyclicGen 网络额外增加了图像边缘信息，该模型将输入帧映射到插帧结果，然后再将其映射到输入帧，引入循环一致性损失建模输入帧和映射帧之间的相似性，解决了插帧结果过于平滑的问题。Park 等人[24]提出了 BMBC 网络，通过计算双向匹配代价卷来估计前后两帧的双向运动信息，并引入了动态滤波器生成模块参与中间帧的合成，这种方法能够有效提高插帧质量，然而其算法时间复杂度过高。马等人[25]提出了一种基于多尺度光流的视频插帧方法，该方法从多个尺度上提取光流并结合注意力机制增强特征表达能力，通过特征融合网络将多尺度光流聚合以生成插帧结果，这种方法避免了额外特征的提取，具有较小的时间复杂度。

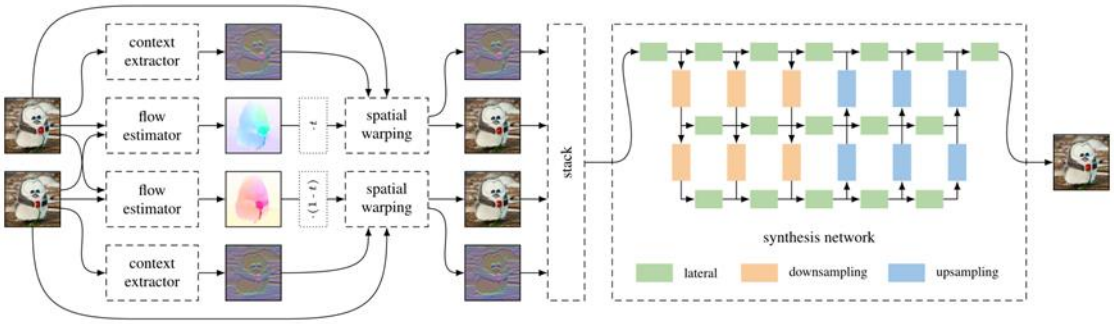

图 1-3 CtxSyn[33]网络总体结构

为了捕获更精确的像素运动信息，一些方法[26-34]在其模型中采用更有效的光流提取网络作为子模块。Xue 等人[29]提出的 ToFlow 网络利用 SPyNet[30]获取光流信息，该模型包含三个子网络，光流提取网络、遮挡推理网络以及中间帧合成网络，这项工作证明了不同的视频处理任务需要不同的光流。MEMC-Net[31]方法选择 FlowNet[32]进行运动估计，将光流与输入帧的变形操作与基于学习的运动补偿滤波器集成到自适应变形层，该模块是完全可微的，适用于多个视频处理任务。Niklaus 等人[33]提出了 CtxSyn 网络从输入帧中额外提取像素级上下文信息，取代了直接根据光流生成中间帧的做法，将上下文映射与双向光流引导的输入帧作为中间帧合成网络的输入，采用一种可学习的局部自适应合成方法来取代标准的加权混合方案，这种方法有效改善了遮挡错误。DAIN[34]方法使用 PWC-Net[35]网络提取双向光流，使用深度估计网络[36]提取图像深度信息，合成中间帧时给予深度值较大的像素以更小的权重，有效解决了运动遮挡问题。

由于提取光流特征需要消耗大量的计算资源，中间帧合成远达不到实时推理的效果，一些研究[37-40]开始提出简化视频插帧模型的方法。Choi 等人[38]提出了一种动作感知的动态架构网络，引入了一种尺度深度探测器 SD-finder，通过预测输入规模和以移动量作为复杂度准则的模型深度来有效地分配合适的计算量。该框架在推理 2K 分辨率图像帧时，在性能几乎没有损失的前提下，可以节省接近 50%的计算量。Huang 等人[39]提出了一种实时中间流估计模型，通过多尺度金字塔结构由粗到细的细化光流并去除现有光流估计网络中计算消耗极大的匹配代价卷以及金字塔特征变形操作，大大减小了光流估计的时间消耗。Malwina 等人[40]在 RIFE 的基础上进行优化并提出了 FastRIFE 方法，采用 Gunnar-Farnebäck 算法[41]以及 Lucas-Kanade 算法[42]替换掉原文中 IFNet 光流估

计网络，通过多组对比实验给出了两种算法的最佳实践，对比 IFNet 方法，该方法的光流估计时间缩短了 4 倍左右。Lee 等人[43]提出的基于 DNN 架构的 ECM-Net 方法采用递归金字塔结构，在每个金字塔层之间追踪具有最大相关系数的位置来递归地细化光流。与其他方法相比，该方法在 4K 视频数据集上，以最少的模型参数量，实现了最优的效果。

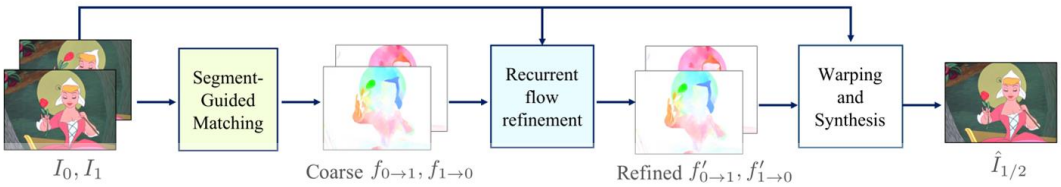

图 1-4 AnimeInterp[49]网络总体结构

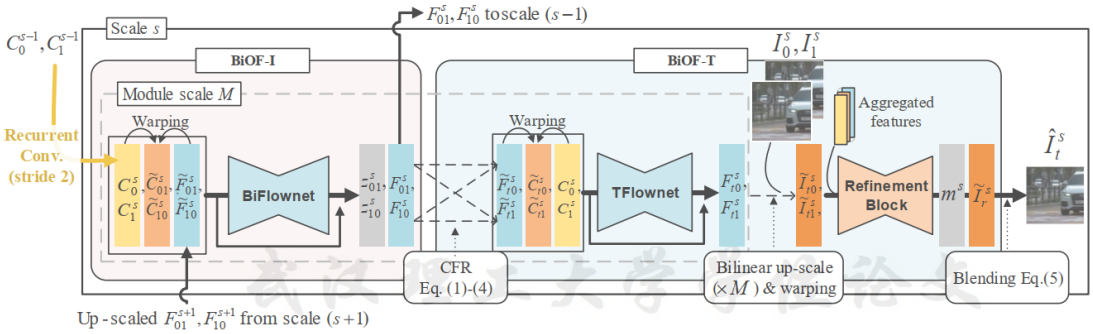

图 1-5 XVFI[50]网络总体结构

另外一些方法则将研究重点放在具体的应用场景中[44-51]。南等人[48]提出了针对立体视频的帧率提升方法，结合深度信息提升运动估计的准确性，并利用显著性检测以及视频分割对运动矢量进行优化。Li 等人[49]提出了 AnimeInterp 方法用于对动画视频进行插帧，针对动画视频插帧中存在的纹理缺失以及非线性超大位移运动等问题给出解决方案，构建了 ATD-12K 动画视频数据集。Sim 等人[50]提供了超高质量视频数据集 X4K1000FPS，该数据集由 1000FPS 的 4K 大位移运动视频组成，并基于递归多尺度结构提出了 XVFI 方法，该方法能够建模大位移像素运动和复杂纹理细节。Shen 等人[51]提出了 BIN 方法用于对模糊视频进行插帧，利用金字塔间的递归模块连接时间序列模型以利用时序关系，通过可调的空间感受野和时序边界控制计算复杂度。尽管基于光流运动补偿的方法已经被证明是有效的，但是这类方法过于依赖光流的准确性，若光流估计不够精确，特征点匹配误差较大，最终插帧结果的质量将会大打折扣。

1.3 主要研究内容

本文的研究内容来自某科技公司的音视频编解码服务器（CMP308）中视频插帧子模块的应用需求，目前主流直播平台上的分辨率能够达到 4K，但是视频帧率最高只能支持到 60FPS，而现有的桌面显示设备刷新率最高能够支持 240Hz，现有的视频资源远没有发挥出硬件显示设备的全部性能。该公司在音视频处理设备（CMP308）中的视频插帧子模块采用的是传统的帧拷贝策略，经过帧率上转换的视频显示效果往往不够理想。受某视频直播平台委托，为了减少插帧过程中纹理细节的损失，提高插帧后视频的图像质量，本文针对现有视频插帧方法存在的问题，提出了一种基于全尺度纹理感知的高效视频插帧方法并将其应用到实时视频插帧模块中。本研究的技术路线如下图所示。

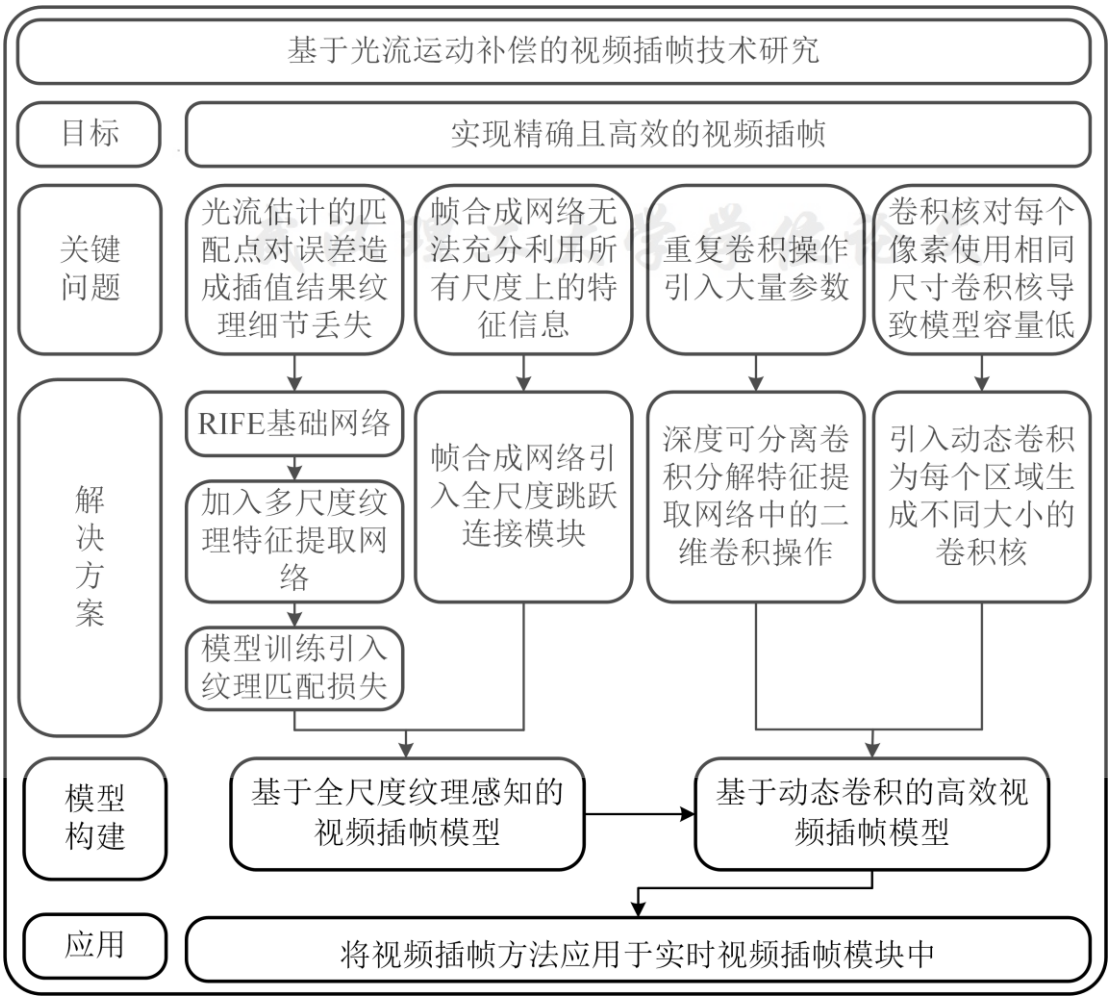

图 1-6 视频插帧技术路线图

本文主要研究工作如下：

（1）针对现有基于光流运动补偿的视频插帧方法中，插帧结果过于依赖光流提取的准确性，光流提取中的特征点匹配误差会造成原始图像中纹理细节的损失，以及采用 U-Net 结构的帧合成网络无法充分利用所有尺度上的特征信息，精细的语义信息会随着卷积操作而逐渐丢失等问题。提出了一种基于全尺度纹理感知的视频插帧方法。首先，引入多尺度纹理提取网络提取多尺度纹理特征以弥补损失的细节信息。其次，在中间帧合成网络中引入全尺度跳跃连接模块，将各层同尺度之间的跳跃连接拓展为编码器与解码器子网络的连接，充分利用编码器子网络的特征信息。最后，在 UCF101、Vimeo90K、Middlebury 以及 HD 这四个视频插帧任务常用的公开数据集上进行实验与验证，并与其他方法进行对比分析。

（2）针对多个特征提取网络中的重复卷积操作，为网络模型增加了大量的矩阵乘法运算以及静态卷积网络中对样本的每一个像素均采用相同大小卷积核而导致的模型容量较低等问题，提出了一种基于可分离动态卷积的高效视频插帧方法。首先，引入动态卷积方法替换光流特征、上下文特征、纹理特征提取网络中的静态卷积，使用动态卷积将图像特征划分为若干个区域并为每个区域生成不同大小的卷积滤波器，降低网络模型参数量，同时提升网络特征表达能力。其次，通过引入深度可分离卷积，将动态卷积生成的二维卷积滤波器进行分解，大大减少模型参数量。最后，在 UCF101 与 Vimeo90K 数据集上进行实验与验证，并与其他方法进行对比分析。

（3）将提出的视频插帧方法应用于音视频编解码服务器（CMP308）的视频插帧子功能模块中，根据实时的应用场景需求，设计了一套合理的视频插帧任务工作流程，实现了实时视频插帧算法，并对其进行实验分析及功能展示。

1.4 论文组织结构

本文组织结构如下：

第 1 章 绪论。首先，介绍了视频插帧技术的应用前景并分析了其研究意义。然后，介绍了视频插帧技术的研究现状，对比分析了不同方法的特点。最后，阐述了本文的主要研究内容以及论文组织结构。

第 2 章 基于全尺度纹理感知的视频插帧方法。本章提出了基于全尺度纹理

感知的视频插帧方法。首先，简单回顾了实时中间流提取方法的整体框架，并针对实时中间流提取方法中存在的问题进行优化，构建了本文提出的全尺度纹理感知的视频插帧方法网络框架。其次，对全尺度纹理感知的视频插帧方法中的关键技术进行详细分析。然后，对视频插帧方法使用的实验环境、实验数据、评价标准以及模型训练设置进行介绍。最后，在四个公开数据集上进行对比实验，并对实验结果进行展示与分析。

第 3 章 基于动态卷积的高效视频插帧方法。本章提出了基于动态卷积的高效视频插帧方法。首先，分析了现有视频插帧方法中存在的问题，并针对现有的问题引入了深度可分离卷积和动态卷积的方法。其次，将两种方法的特点结合构建了基于深度可分离的动态卷积结构并引入到模型中。然后，对视频插帧使用的评价标准以及模型训练设置进行介绍。最后，在两个公开数据集上进行对比实验，并对实验结果进行展示与分析。

第 4 章 高效视频插帧算法的应用。本章将视频插帧算法应用到实时的工作场景中。首先，介绍了视频直播系统框架。其次，针对音视频编解码服务器设计了实时视频插帧子模块的工作流程，并实现了实时视频插帧算法。最后，通过实验验证了视频插帧方法在视频直播场景下应用的可行性，并将该方法部署到音视频编解码服务器的视频插帧子模块中，同时对视频插帧模块的工作流程进行展示。

第 5 章 总结与展望。对本文提出的基于光流运动补偿的视频插帧算法做出总结，并展望了未来的工作方向。

第 2 章 基于全尺度纹理感知的视频插帧方法

在现有的基于光流运动补偿的视频插帧方法中，光流提取的特征点匹配误差会造成原始图像中纹理细节的丢失，采用 U-Net 结构的帧合成网络缺乏从所有尺度上捕获特征信息的能力，精细的纹理信息会随着卷积操作而逐渐丢失。针对这些问题，首先在已有的 RIFE 网络[39]（Real-Time Intermediate Flow Estimation，RIFE）中引入基于 VGG-16 编码器结构[52]的纹理特征提取网络，提取多尺度纹理特征，在帧合成网络中引入全尺度跳跃连接模块，充分利用编码器子网络的特征信息，以此提出了一种基于全尺度纹理感知的视频插帧方法 FSTA-RIFE（Full Scale Texture-Aware，FSTA）。最后在 UCF101、Vimeo90K、Middlebury 以及 HD 四个视频插帧常用的公开数据集上进行了实验，结果表明该方法的整体性能优于其他视频插帧方法。

2.1 FSTA-RIFE 视频插帧模型的构建

2.1.1 RIFE 实时中间流估计网络结构

现有的视频插帧方法往往首先提取帧间的双向光流，然后采用线性组合方式近似中间流，然而这种处理方式会在运动边界区域产生伪像问题。RIFE 方法提出的 IFNet 光流提取网络采用由粗到细的方式，通过连续两帧的输入直接预测中间流信息；然后利用估计的中间流与输入图像进行后向变形，最后采用 U-Net 结构的帧合成网络生成最终的插帧结果。RIFE 网络结构主要由 3 个模块组成：（1）IFNet 光流提取网络；（2）上下文特征提取层；（3）中间帧合成网络。

（1）IFNet 光流提取网络

给定视频的前后两帧 I_0, I_1，不同于传统光流估计方法提取前后两帧的双向光流 $F_{0\to1}, F_{1\to0}$，IFNet 网络[39]直接估计中间时刻 t 到前后两帧的双向光流 $F_{t\to0}, F_{t\to1}$。IFNet 去除了现有光流估计网络中计算消耗极大的匹配代价卷以及金字塔特征变形操作，大大减小了光流估计的所需时间，能够实现实时的光流估计。IFNet 采用残差思想，由三个 IFBlock 进行残差连接组成，IFBlock 结构

如下图 2-1 中虚线框部分。为了建模大位移运动，从低分辨率到高分辨，每个 IFBlock 依次接受不同分辨率的输入，从低分辨率输入中提取大范围运动信息，从高分辨率输入中迭代细化光流。其计算公式如下：

$$F^{i}=F^{i-1}+g^{i}(F^{i-1},\hat{I}^{i-1}) \tag{2-1}$$

其中 $\hat{I}^{i-1},F^{i-1}$ 是当前 IFBlock 层的输入，$\hat{I}^{i-1}$ 表示预变形帧，由上一层输出光流与输入帧后向变形得到，F^{i-1} 为上一层 IFBlock 的输出光流。g^{i} 表示当前 IFBlock 层，当前 IFBlock 层的输出光流与上一层输出进行叠加得到该层的输出光流 F^{i}。IFNet 网络结构如下图 2-1 所示，其中右侧虚线框部分为 IFBlock 的详细结构。

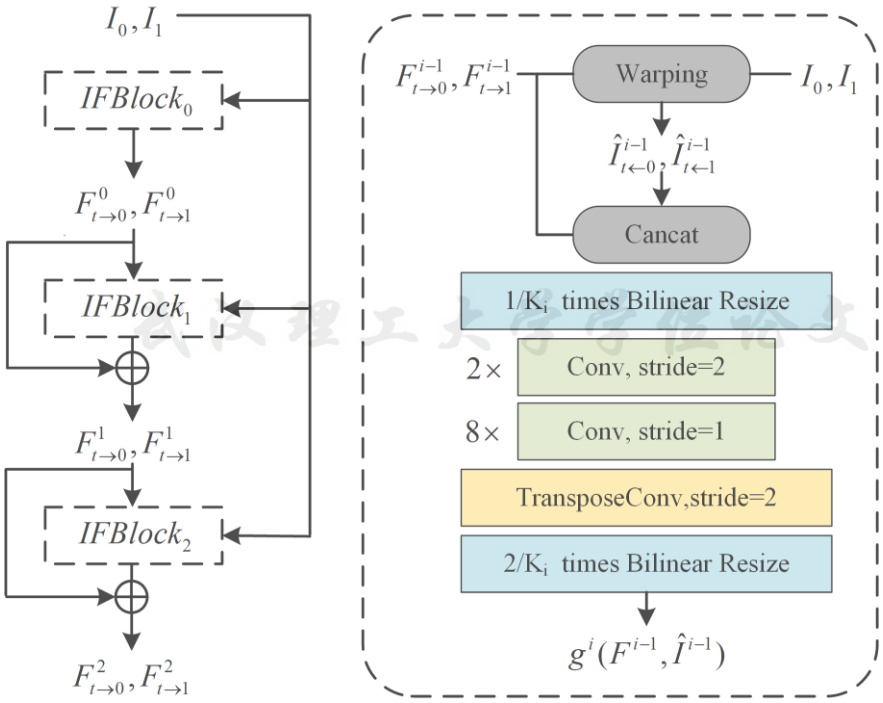

图 2-1 IFNet 网络结构[39]

（2）上下文特征提取层

上下文特征提取层结构与 U-Net 的编码器子网络结构相似，由四个卷积块组成，其中每个卷积块均由两个大小为3×3的卷积核构成。上下文提取层首先从输入两帧中提取多尺度上下文金字塔特征，表示为 $C_0{:}\{C_0^1,C_0^2,C_0^3,C_0^4\}$ 和 $C_1{:}\{C_1^1,C_1^2,C_1^3,C_1^4\}$。然后利用估计的光流对这些特征进行后向变形，产生对齐的金字塔上下文特征 $\hat{C}_{0\to t}$ 和 $\hat{C}_{1\to t}$，该特征连同其他特征输入到中间帧合成网络

生成插帧结果。

（3）中间帧合成网络

中间帧合成网络为 U-Net 编码器-解码器结构，其中编码器子网络由四个卷积块构成，卷积块中均包含两个大小为3×3的卷积核，解码器子网络则由四个反卷积块组成。由于预变形的帧已经被光流对齐，因此帧合成网络侧重于增强细节，该网络负责在四个尺度下合成原始帧、光流、预变形原始帧以及预变形上下文特征并输出特征融合图和重建残差项以减少图像噪声的引入。最后使用特征融合图和重建残差项对预变形帧进行微调，其计算公式如下：

$$\hat{I}_t = M \odot \hat{I}_{0\to t} + (1-M) \odot \hat{I}_{1\to t} + \Delta \tag{2-2}$$

其中 M 表示特征融合图，用于融合两个预变形帧，Δ 表示重建残差项，用于细化图像细节，$\odot$ 表示矩阵乘法，$\hat{I}_{0\to t}, \hat{I}_{1\to t}$ 表示使用中间流对齐后的预变形帧，$\hat{I}_t$ 为最终的插帧结果。

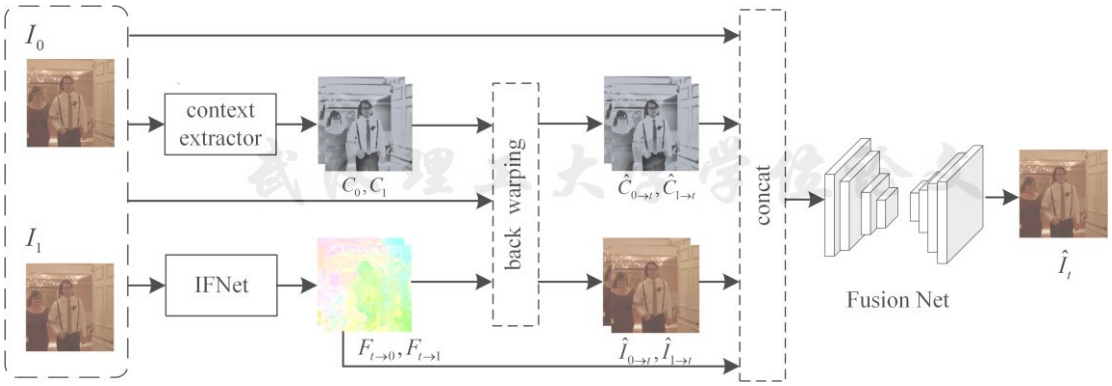

图 2-2 RIFE 网络模型总体结构

2.1.2 FSTA-RIFE 视频插帧网络的构建

针对基于光流运动补偿的视频插帧网络中，光流估计过程中的特征点匹配误差会造成原始图像中纹理细节的丢失，编码-解码器结构的帧合成网络无法有效利用全尺度下的特征信息等问题。本文提出了一种将纹理提取网络与全尺度跳跃连接模块引入到 RIFE 模型中的视频插帧方法（FSTA-RIFE）。FSTA-RIFE 网络总体架构如图 2-2 所示，该方法由 5 个模块组成：（1）多尺度纹理特征提取层；（2）全尺度跳跃连接模块；（3）IFNet 光流特征提取层；（4）上下文特征提取层；（5）中间帧合成网络。其中，额外引入的模块为：（1）多尺度纹理

特征提取层；（2）全尺度跳跃连接模块。

（1）多尺度纹理特征提取层

本文额外加入纹理特征提取网络以弥补中间帧合成时丢失的纹理细节，如图 2-3 所示，纹理特征提取网络采用 VGG-16 网络[52]中的前 13 个卷积层作为基础结构，分别在第 4、7、10、13 层输出四个尺度的纹理特征。与上下文特征类似，使用光流估计网络提取多个尺度下的光流，并与多尺度纹理特征进行后向变形，得到预变形纹理特征，该特征作为中间帧合成网络的输入。

（2）全尺度跳跃连接模块

原有的中间帧合成网络中，通过跳跃连接将编码器的上下文特征作为相同尺度下解码器层的额外输入，这种方式为解码器提供了丰富的粗糙语义特征，然而这些图像语义信息会随着卷积与反卷积操作而逐步丢失。因此，本文将额外引入的多尺度纹理特征也通过跳跃连接传递到解码器子网络中。为更有效的利用编码器子网络的特征信息，本文在编码器与解码器间加入全尺度跳跃连接模块[53]（Full-Scale Skip Connections，FSC），对编码器子网络中多个不同尺度下的特征进行重组并传递给解码器层，以增强图像纹理细节表达。在下一节内容中将着重对这些改进的技术细节做进一步的分析与阐述。

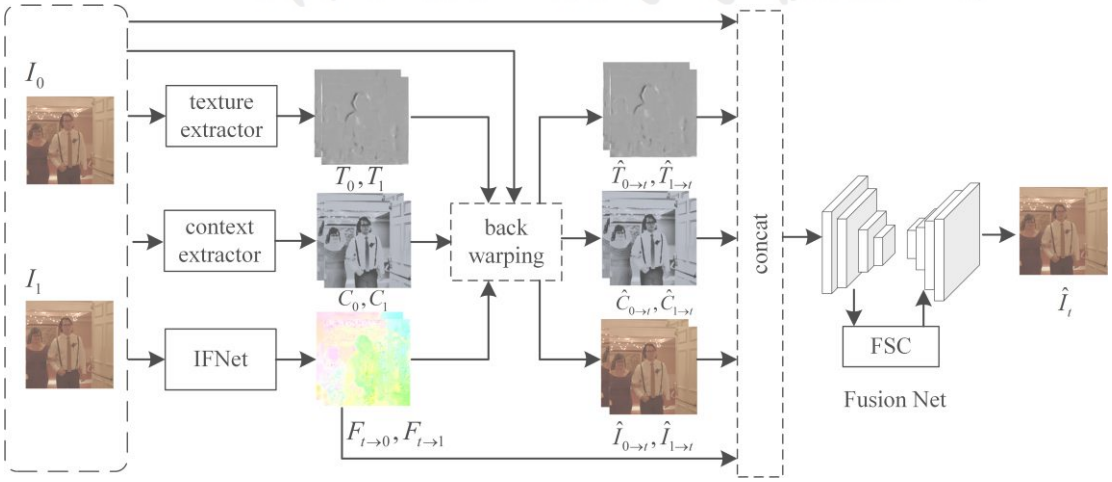

图 2-3 FSTA-RIFE 网络模型总体框架

2.2 FSTA-RIFE 方法关键技术及算法

2.2.1 基于 VGG-16 网络提取多尺度纹理特征

相较于基于卷积核的视频插帧方法，基于光流的视频插帧方法能够通过提取光流信息来捕获视频中的像素运动，本文采用的 IFNet 光流提取方法通过去除现有光流估计网络中计算消耗较高的匹配代价卷以及金字塔特征变形操作实现了实时光流估计的效果。然而，由于失去匹配代价卷层来计算前后两帧图像逐像素的匹配程度，得到的特征点匹配误差较大，在具有丰富动态纹理的场景下，会导致最终插帧结果的局部模糊。基于此，本文通过引入纹理特征提取层来弥补中间帧合成时丢失的精细纹理细节。VGG-16 网络采用大小为3×3卷积核串联的方式代替传统特征提取网络中的5×5卷积核或7×7卷积核，这种小卷积核串联的方式不仅拥有与大卷积核相同的感受野，而且具有更小的网络参数量，有效避免了模型过拟合问题。同时由于模型引入了更多卷积层，增强了对图像特征的学习能力。因此，本文使用 VGG-16 网络作为纹理特征提取网络，具体来说，本文使用 VGG-16 网络的前 13 个卷积层作为特征提取主干网络，由于最后三层的全连接层参数量较大，占原模型参数量的四分之三，为保证模型的计算效率，本文将其舍弃。网络结构如下图所示。

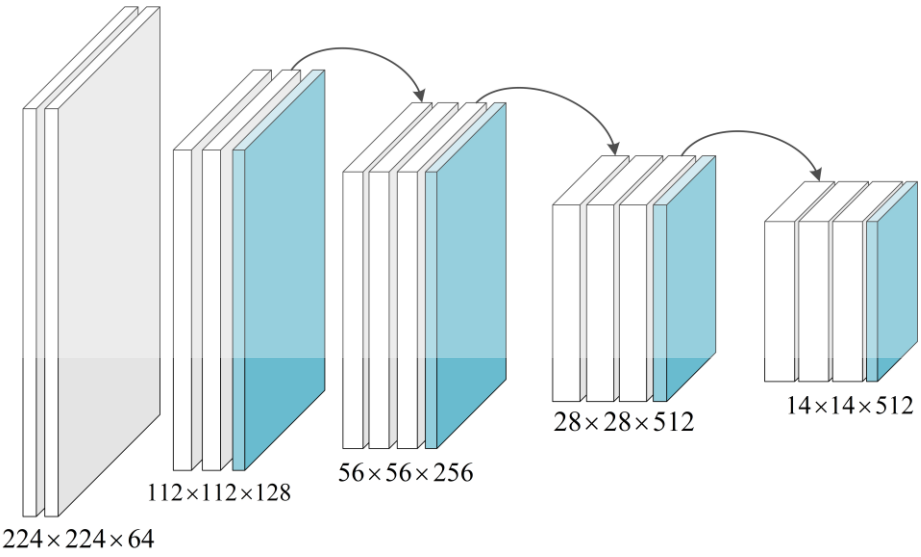

图 2-4 基于 VGG-16 的纹理特征提取网络

如图 2-4 所示，输入前后两帧原始图像 I_0^0, I_1^0，分别从网络第 4、7、10、13 层输出四个尺度下的特征图，同时在每个输出层后增设一个 1×1 卷积层以降低通道方向上的维度，输出前后两帧的多尺度纹理特征，记为 $T_0:\{T_0^1,T_0^2,T_0^3,T_0^4\}$ 和 $T_1:\{T_1^1,T_1^2,T_1^3,T_1^4\}$，然后利用 IFNet 网络实时估计相应尺度下的双向光流，记为 $F_{t\to0}:\{F_{t\to0}^1,F_{t\to0}^2,F_{t\to0}^3,F_{t\to0}^4\}$ 和 $F_{t\to1}:\{F_{t\to1}^1,F_{t\to1}^2,F_{t\to1}^3,F_{t\to1}^4\}$，最后将纹理特征与中间流分别进行后向变形，其计算公式如下：

$$\hat{T}_{0\to t}^i=\omega(F_{t\to0}^i, f^i(I_0^{i-1}))\ ,\ i=1,2,3,4 \tag{2-3}$$

$$\hat{T}_{1\to t}^i=\omega(F_{t\to1}^i, f^i(I_1^{i-1}))\ ,\ i=1,2,3,4 \tag{2-4}$$

以第一帧为例，I_0^{i-1} 表示纹理提取网络中上一层输出的纹理特征，该特征为经过 1×1 卷积核降维后的特征，f^i 表示纹理提取网络中的当前层，输出第 i 个尺度下的纹理特征。$F_{t\to0}^i$ 表示当前尺度下的光流特征。ω 表示后向变形操作，最终输出第 i 个尺度下的预扭曲纹理特征 $\hat{T}_{0\to t}^i$，该特征连同预变形上下文特征、原始帧、预变形帧以及光流作为中间帧合成网络的输入。

2.2.2 基于全尺度跳跃连接进行高效特征传递

低层次精细特征描述了图像的空间信息，突出了像素的局部边界；而高层次语义特征则包含了图像中像素的整体分布信息。采用 U-Net 结构的中间帧合成网络虽然可以通过跳跃连接模块来组合相同尺度下的高级语义特征和编码器相应的低级详细特征映射，然而，这种方式缺乏从所有尺度上捕获特征信息的能力，当使用卷积操作进行向下采样或者向上采样时，这些精细的细节信息会逐渐丢失。本文在中间帧合成网络中引入全尺度跳跃连接模块[53]，重新设计中间帧合成网络中编码器和解码器之间的跳跃连接以捕获全尺度下的精细纹理细节和粗糙语义特征。具体而言，本文将 U-Net 中各层同尺度之间的跳跃连接拓展为编码器层与解码器子网络的连接，即中间帧合成网络中的每个解码器层都包含来自编码器的相同和较小尺度的特征，以及来自解码器的较大尺度的特征。解码器 X_{de}^4 接收来自编码器 $X_{en}^1,X_{en}^2,X_{en}^3,X_{en}^4$ 的输出，X_{de}^3 接收来自编码器 $X_{en}^1,X_{en}^2,X_{en}^3$ 和上一级解码器 X_{de}^4 的输出，以此类推，X_{de}^2 接收来自 $X_{en}^1,X_{en}^2,X_{de}^3,X_{de}^4$ 的输出，X_{de}^1 接收来自 $X_{en}^1,X_{de}^2,X_{de}^3,X_{de}^4$ 的输出。每个解码器均利用了来自编码器和上级解码器的多尺度特征。

解码器端的特征传递方式如图 2-5 所示，以解码器 X_{de}^2 为例，与 U-Net 类似，