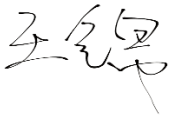
中图分类号：TP391　　论文编号：1028716　23-S138
学科分类号：083900

硕士学位论文

基于深度学习的帧插值算法研究

研究生姓名　　王锦
学科、专业　　网络空间安全
研 究 方 向　　视频处理
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

二〇二三年五月

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

Research on Frame Interpolation Algorithms Based on Deep Learning

A Thesis in
Cyberspace Security

by
Jin Wang

Advised by

Prof. Xiaoyang Tan

Submitted in Partial Fulfillment

of the Requirements

for the Degree of

Master of Computer Engineering

May, 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

视频帧插值技术是视频处理领域的核心技术之一，它被广泛的应用于诸如动画电影的增强，体育赛事的慢动作视频以及视频的编解码当中。但随着多媒体硬件的高速发展，人们对高分辨率与高帧率的需求越来越大，这也意味着我们迫切地需要更加优越的帧插值算法来应对这些需求。随着深度学习在视觉领域取得了巨大成功，大量基于深度学习的帧插值研究也涌现了出来，并且取得了惊人的效果。受基于深度学习的帧插值方法的启发，本文针对帧插值任务中卷积神经网络上采样过程会出现的模糊结果问题以及卷积神经网络无法很好的提取视频间时序特征的问题提出了两个新的算法模型。具体而言，本文的主要贡献包括以下两个方面：

1. 目前基于光流的帧插值方法已经取得了非常不错的效果，但是它们也成为了限制帧插值进一步发展的桎梏，光流估计的误差会传播至下游任务，这导致了现有方法在一些困难场景下精度低的问题，于是我基于 3D 卷积神经网络开发了一种不需要光流的端到端的模型，利用 3D 卷积神经网络提取视频序列的时空信息，并结合 Pixel Shuffle 和通道注意力进一步增强网络对运动信息的提取，最后引入了用于纹理增强的后处理模块，增强插值帧的视觉效果。

2. 卷积神经网络在视觉领域一直长盛不衰，因为它具有局部性和平移不变性的归纳偏置，可以帮助对图像特征的提取，促进网络的收敛，但是这也导致了其面临着性能上的瓶颈，而 Transformer 的出现让我看到了打破这种性能瓶颈的希望。传统的卷积神经网络无法充分地提取视频序列中的时序上的运动信息，所以往往要引入通道注意力机制去加强特征提取能力，但依然不能完全解决这个问题，我提出利用 Transformer 直接对视频序列信息进行特征提取，结合 3D 转置卷积进行上采样来完成帧插值任务，这样既能利用到 Transformer 对特征更强的提取能力，又利用了 3D 卷积实现对特征的上采样，保证了插值帧图像的平滑度，并且沿用了基于纹理增强的后处理模块，进一步提升插帧质量。

关键词：视频帧插值，3D 卷积神经网络，金字塔级联网络，变换神经网络，像素重组

ABSTRACT

Video frame interpolation technology is one of the key technologies in the field of video processing, which is widely used in areas such as enhancement of animated movies, slow motion videos of sports events, and video encoding and decoding. But with the rapid development of multimedia hardware, the demand for high resolution and high frame rate is increasing, which also means that we urgently need more superior frame interpolation algorithms to meet these needs. With the great success of deep learning in the field of vision, a large number of frame interpolation studies based on deep learning have emerged and achieved astonishing results. Inspired by deep learning based frame interpolation methods, we have developed two new network frameworks for frame interpolation problems. Specifically, the main contributions of this article include the following two aspects:

1. At present, frame interpolation methods based on optical flow have achieved very good results. However, they have also become a constraint on the further development of frame interpolation. The error of optical flow estimation can propagate to downstream tasks, leading to low accuracy in some difficult scenarios. To address this issue, we developed an end-to-end model based on 3D convolutional neural networks that does not require optical flow. Our model utilizes 3D convolutional neural networks to extract spatiotemporal information from video sequences and combines Pixel Shuffle and channel attention to further enhance the network's extraction of motion information. Additionally, we introduced a post-processing module for texture enhancement to improve the visual effect of interpolated frames.

2. Convolutional neural networks have been thriving in the field of vision because they have the inductive bias of locality and translation invariance, which can help extract image features and promote network convergence. However, this has also led to a performance bottleneck. The emergence of Transformers has given us hope to break this bottleneck. Traditional convolutional neural networks cannot fully extract motion information from video sequences over time, so channel attention mechanisms are often introduced to enhance feature extraction capabilities. However, this still cannot completely solve the problem. We propose using Transformers to directly extract features from video sequence information and combine 3D transposed convolution for upsampling to complete the frame interpolation task. This not only utilizes the stronger feature extraction capabilities of Transformers but also uses 3D convolution to achieve feature upsampling, ensuring the smoothness of the interpolated frame image. Additionally, we use a post-processing module based on texture enhancement to further improve the quality of interpolated frames.

Keywords: Video Frame Interpolation, 3D Convolutional Neural Network, Pyramid Cascade Network, Transformer, Pixel Shuffle

目 录

图表清单

注释表

I,I_t,I_{t-1},I_{t+1}	像素矩阵	x,y	像素坐标
u,v	光流偏移量	t	时刻
α_0	遮挡系数	$F_{0\to1}$，$F_{1\to0}$	光流偏移矩阵
$g(\cdot,\cdot)$	绘制函数	$V_{t\leftarrow0}$，$V_{t\leftarrow1}$	遮挡映射矩阵
K	卷积核	⌊x⌋	向下取整
H	图像高度	W	图像宽度
C	通道数	r	Shuffle 因子
$\hat{I}_t$	t 时刻插帧结果	I_{gt}	对应于插帧的真实帧

缩略词

缩略词	英文全称
VAE	Variational Auto-Encoder
GAN	Generative Adversarial Network
CNN	Convolutional Neural Network
ViT	Vision Transformer
LN	Layer Normalization
MLP	MultiLayer Perceptron
PCN	Pyramid Cascade Network
PSNR	Peak Signal-to-Noise Ratio
MSE	Mean Square Error
SSIM	Structural SIMilarity
NLP	Natural Language Processing
RNN	Recurrent Neural Network
LSTM	Long Short Term Memory

第一章 绪论

1.1 课题研究背景

近年来随着多媒体领域进一步的发展，我们的生活已经离不开视频技术了，短视频，影视剧，监控录像等应用已经充斥在我们的身边。但是显示设备的迭代速度太快，也导致市场上帧率以及帧分辨率的显示都出现了高低版本并存的现象，这严重的影响了用户体验，为了解决这个问题，给用户带来更好的体验，帧插值技术应运而生，并且被广大的科研人员以及企业工程师们所关注。

同时，随着显卡性能的提升，与视频相关的计算机视觉领域在过去的几年取得了长足的进步，其中视频帧插值，超分辨率，多模态任务等领域都有大量成果在不断涌现，而随着深度学习的逐渐成熟和它在计算机视觉领域所表现出来的巨大潜力，使得与深度学习相结合的各种视频帧插值方法层出不穷，也因此衍生出了大量应用。

因此，在视频处理方面，帧插值任务是一个非常重要、非常值得研究的一个问题。它的主要任务就是为视频帧序列进行扩展，即通过对连续帧之间的帧预测实现数据的扩充，从而改善视觉效果。帧插值技术已经被广泛地应用于诸如动画电影插帧，体育赛事的慢镜头视频生成等生活场景当中，目的是用于生成在两个相邻的帧间的过渡。帧插值的最终目标是生成更流畅的视频，以满足观众的视觉需求。在技术上，帧插值可以通过利用计算机视觉算法以及复杂的运动模型来实现。

帧插值技术的发展受到计算机视觉、图像处理、运动估计和深度学习等领域的影响，这些领域的技术都可以用来帮助算法更好地插入连续的视频帧。同时，帧插值也为视频处理的其他方面，如视频编辑和视频压缩，提供了有用的支持。

在实际应用中，主流方法是基于光流法的帧插值方法，这类方法有比较强的先验假设，即要求运动物体的灰度在较短时间内是维持不变的以及给定邻域内光流的变化是缓慢的。这是两个非常强的假设，因为它要保证像素点的光流是可导的。但是生活中充斥着大量的场景无法满足这些假设，例如比较大幅度的运动变化，光照的不稳定导致像素点灰度的剧变以及有遮挡的情况使得前后帧之间某些像素点的灰度值发生了变化。这些场景会导致基于光流法的帧插值技术的精度大幅下降。近年来，随着深度学习的火热，将深度学习与光流法相结合的算法也如雨后春笋般不断涌现，深度学习大幅的提升了基于光流法的帧插值技术的效率，同时研究者们也针对极端情况进行了改进。随着自编码器在机器学习领域的大火，另一种基于特征提取的自编码器结构也开始被广泛关注。基于特征提取的帧插值技术可以有效地改善在上述情况下的帧插值质量。但是此类方法由于没有建立比较强的假设，相对而言，解空间会更大一些，这也就导致该类方法在速度上不如基于光流法的帧插值方法，且依赖于模型的特征提取能力，在复杂环境下的稠密运动对该类方法而言也是一项严峻的挑战。

针对复杂环境下的稠密运动场景下的帧插值问题，可以通过设计网络提高特征提取能力以及对某些特定特征进行针对性的提取进行改进，从而提升模型的泛化性能。

我们对帧插值任务进行一个更加清楚的定义，假设原始视频的帧率为 30 帧每秒，帧插值任务需要对该帧序列中的每两个连续帧进行中间帧预测，可以使单帧预测也可以是多帧预测，这里假设为单帧预测，则我们可以得到一个数量与原视频序列相等的插值帧序列，按照时序关系将其插入原视频中，我们即可得到插帧后的视频，其帧率为 60 帧每秒。

1.2 研究意义

随着显示器材，摄影器材以及互联网的高速发展，观众们已经不再满足于低分辨率以及低帧率的视频源，但是在发展初期人类已经制造了大量珍贵的影像，它们虽然在帧率和分辨率的质量上不够先进，但其中所蕴含的内容和带来的深远意义是不可磨灭的，为了更好的保存这些珍贵的视频，并且传播其中丰富的内容信息，因此对图像进行帧率的提升是一个非常重要且具有深远意义的研究方向。

同时我们也可以应用帧插值技术去获得慢动作视频，受制于人眼和大脑的机理结构，我们往往无法准确地捕获一些微动作和微表情，因此会产生许多谬误，例如在羽毛球场上，许多边裁在羽毛球的落点上的判断不够准确，在经济原因的影响下，我们有时候会受制于摄像器材的帧率，而视频帧插值技术可以轻易的将视频插帧至千帧以上，使得我们可以准确判断落点，从而进一步保证体育赛事的公平公正。微表情的重要性不言自明，许多心理学研究人员都在研究微表情与人类行为之间的关系，并得出了很多有效的技巧去进行行为的判别和预测乃至于测谎，但是这需要大量的专家知识，在人工智能技术发展蓬勃的今天，搭建一个成熟的人工智能去进行行为的判别或者预测将会是一个非常好的应用方向，但微表情需要的数据集必然需要更高的帧率，这也意味着更大的成本，而帧插值技术正可以解决该类问题，为更多的视频类任务提供帧率更高的数据，从而促进相关科研乃至工程任务的发展。

另外，视频帧插值技术也被广泛的应用于其他的视频处理任务，例如视频超分辨率任务以及视频编解码任务等，由于视频超分辨率任务的特殊性，为该任务提供更加丰富的时域信息也会提高此类任务的算法效果，而针对视频编解码任务，引入帧插值技术后，可以让压缩包更小，且获得更高的帧率，这对于网络间的视频信息传输而言无疑是一个十分有意义的工作。

当前主流的帧插值算法一部分是基于光流法的，光流法具备更高的推理速度，而且基于深度神经网络的光流估计往往可以带来了泛化性更好，插值帧质量更高的性能，但是由于光流法本身所做出的强假设限制，理想场景中，光流法当然是行之有效的，但是在复杂的生活场景中其效果就很难得到保证，例如光照的突然变化和大运动以及遮挡等问题都是生活场景中经常出现但是光流法又难以处理的问题。还有一部分算法是基于特征提取-融合-重构架构的，这类算法相比于光流法在复杂场景下会有更好的鲁棒性，但是受制于神经网络提取特征的能力，对于稠密运动它往往无法取得好的性能，且提取特征过程中往往会损失一些细节信

息，这也会导致最终插值帧会有一些模糊伪影等问题。于是如何对复杂场景下的视频帧插值具有一定的研究意义。

1.3 国内外研究现状

最初的帧插值算法思想十分简单，一般采用帧复制或者帧平均的的方法去获得插值帧，这样做的优势在于其计算速度非常快，是可以满足实时性的需求的，但是缺点也很明显，这种帧插值的算法完全没有考虑场景和目标的运动，这就会导致插值帧生成过程中会出现许多模糊或者伪影的情况。

在此之后，许多基于运动估计(ME,Motion Estimation)和运动补偿（MCI,Motion Compens--ation Interpolation）的算法被提了出来，而这种流程也逐渐成为了传统帧插值算法的标准流程。那什么是运动补偿呢，我们假设视频中特定的局部块内的每个像素都有相同的位移。我们根据第 t-1 帧的局部块信息，在第 t 帧中与每个块进行比较，找到最相似的局部块之后，将它们之间所发生的偏移作为两个局部块之间的运动向量。当我们得到了运动向量之后，我们就可以对利用一些线性运动的假设去近似出插值帧的运动向量，然后，结合运动向量的信息和局部块的语义信息对中间帧进行插值。而这种插值的方式非常多样，对运动的表示和估计方法也各有不同，所以形成了现在各种各有千秋的帧插值方法。基于运动补偿的帧插值方法往往可以实现在图像整体的平滑性，并且保持了运动信息，可以获得更加流畅的自然地插值帧，但提高了精度之后，不可避免的是计算复杂度会增高。所以如何改进帧插值算法的计算复杂度，并且提高精度是帧插值问题所亟待研究的问题。接下来，我们将主要介绍运动估计与运动补偿这类标准方法的发展现状。

既然我们想要估计出运动信息，那么首先我们要考虑的问题就是如何高效的对运动信息进行建模和表示。总的来说，我们有两种方法去构建运动的表示。第一种被称为拉格朗日表示法，它聚焦于像素点本身在运动过程中在位置上所产生的偏移，代表性的方法是基于光流场的方法。另一种被称为欧拉表示法，它聚焦于挖掘时序信息上信息的流动，即它不关心位置，而只考虑随着时间的变化，相同位置处的信息变化过程，代表性的方法是通过对图像进行分解，关注其在亮度或者相位上的变化过程。从以上的描述中我们也可以看出，拉格朗日表示法更加贴近人类直观认识的方法，所以更多的研究人员基于拉格朗日表示法对运动信息进行表示，此类方法一般都会对运动进行显式地匹配，根据匹配方式的不同，我们又可以将它细分为以下三种：前向运动匹配、后向运动匹配和双向运动匹配。

刚刚介绍过的两种运动表示方法：拉格朗日方法和欧拉方法，它们以完全不同的着眼点去解释运动的变化。拉格朗日方法着眼于物体在运动后所产生的空间位移，而欧拉方法着眼于同一位置上运动所带来的在时序上的信息变化。根据这两种角度，运动估计算法可以分为基于运动匹配的显式运动估计和专注时序上的局部信息变化的隐式运动估计两类。所谓的运动表示其实是将运动映射到人类或者机器可以理解的表示空间当中，它们的本质依然是运动，所以不管用哪种运动表示方法，我们最终都可以根据这种映射关系进行相互的转换。这

也是为什么我们采用隐式运动估计的方法也可以通过一系列映射操作获得例如光流等基于拉格朗日显示运动表示的原因。很容易我们就可以证明这一点，像是许多最近提出的基于卷积神经网络的方法并没有用到显式地运动匹配，而只是通过特征的变换实现隐式的运动估计，却可以得到光流这样的拉格朗日运动表示。

详细来说，我们可以参照 2011 年，Baker 等人[1]发表的对于传统光流法的解释进一步地阐述基于显示运动估计的方法。首先，光流法需要对运动进行建模，为了得到更好地理论性质便于推导和求解，光流法必须要遵循亮度值不变的假设，即在运动过程中，物体的亮度值不会随时间发生变化，如公式（1.1）所示：

$$I(x,y,t) = I(x+u, y+v, t+1) \tag{1.1}$$

其中$I(x,y,t)$的含义是：取坐标为(x,y)的像素点，获取该像素点在时刻 t 的亮度值，(u, v)的含义是像素点在运动后所产生的位置偏移信息，即光流信息。对公式(1.1)进行泰勒展开可以得到公式(1.2)如下：

$$I(x+u, y+v, t+1) = I(x,y,t) + u\frac{\partial I}{\partial x} + v\frac{\partial I}{\partial y} + \frac{\partial I}{\partial t} \tag{1.2}$$

又根据公式（1.1），我们可以轻松得到以下的光流约束公式：

$$u\frac{\partial I}{\partial x} + v\frac{\partial I}{\partial y} + \frac{\partial I}{\partial t} = 0 \tag{1.3}$$

根据这些公式，基于拉格朗日运动表示的方法，我们可以去寻找合适的光流满足上述公式（1.3）并结合能量方程全局最小化的算法即可得到每个像素点的光流值，而后结合光流进行插帧。但是光流法在实际中存在一些问题，由于一开始做出了非常强的假设，需要运动保持连续，亮度恒定即光照不能发生突变，而其中最重要的问题就是“孔洞问题”以及“数据歧义”问题，前者是因为计算光流过程中可能会有像素点间的映射存在“一对多”或者“多对一”的问题，此类映射问题在应用中就表现为会有一些像素点我们无法确定其光流，可能有多个光流我们需要进行加权求和，也可能没有光流经过，我们需要根据近邻点的光流对其进行近似；后者是因为实际获得的视频帧物体像素点所发生的亮度变化以及相似的像素点亮度引起的。为了解决这些问题，一般都会引入正则化项，从而得到更加精确平滑的光流。

由于计算每个像素点的光流时间复杂度较高，许多研究人员提出了块匹配算法[2,3,4,5]。这是一种基于稀疏光流的方法，相比于稠密光流，它可以大大减少计算量，而这种提升是以损失精度为代价的，而且由于不重叠的块，块与块之间的光流信息可能存在偏差，这会导致出现块效应，也就是在块的边界处会出现明显的不连续。近年来，许多研究[6,7]致力于在保持基于块方法效率的同时提高运动矢量的精度。随着研究的深入，基于显式运动估计的方法已经取得了长足的进步，但是由于光流只是运动偏移的信息，而不包含图像的原始信息，所以在构建插值帧时，我们往往还需要对每个像素点进行翘曲计算，这会影响帧插值算法的整体计算复杂度。此外，对于有遮挡、亮度剧变、大运动以及复杂运动等特殊场景，基于光流的方法存在先天的缺陷，而大量的研究者也在不断地提出新方法去解决这些问题。

对于隐式运动估计的方法， Wu 等人[8]提出了一种新方法，他们着眼于运动过程中相位

幅值所发生的变化，通过图像 Laplacian 金字塔上对相位幅值信息进行提取来实现对运动的隐式估计。近年来，许多研究者们也从频域的角度出发去解决帧插值问题，他们提出基于相位的隐式运动估计方法用相位信息去编码运动信息[9,10,11,12,13]。相位信息不同于并不像频谱信息一样关注像素值的变化趋势，而更多的聚焦于图像中物体边缘的位置信息，所以它对亮度变化具有鲁棒性，也因为这样，通过相位信息对运动进行隐式估计的方法往往可以应对光照亮度的剧烈变化，但由于视频信息在时序上的离散结构，相位信息并不足以表示出大运动。Meyer 等人[9]利用从粗到细(coarse-to-fine)的思想构建的帧插值算法在基于相位的基础上增加了能够表示的运动范围。

近来，深度学习在视觉领域大火，大量基于深度学习的算法证实了深度学习在处理图像与视频信息上的强大能力，于是许多研究者们聚焦于利用卷积神经网络去实现隐式的运动估计，因为卷积神经网络并不具备直接进行运动匹配的能力，它只能通过对特征的不断变换和融合提取混合的特征流，而后再经过特征流的分解和聚合输出类似光流或是其他形式的流信息[14,15]。还有一部分基于卷积神经网络的运动估计算法使用了相关性操作[16,17]，某种意义上来说这种相关性的计算也是在进行一种匹配搜索的操作。但是，由于卷积神经网络带有局部性与平移不变性的归纳偏置，所以这种相关性计算得到的特征往往会和图像的语义特征交织在一起，后续还需要从混合特征中将运动特征隐式地提取出来，所以一般将基于卷积神经网络的运动估计归为隐式运动估计一类。

Ranjan 等人[18]考虑到了直接对光流进行估计的难度，受到一些人体姿态估计方法的启发，提出了利用级联金字塔网络去提取光流的方法，这种方法的核心思想是将困难任务分解，不执着于直接生成足够好的光流，而是通过级联金字塔对光流进行一步步细化，这种思想取得了巨大的成功，相比 FlowNet[16]该方法的参数量减少了 96%，但是光流精度只是略有下降。Sun 等人[19]在前述研究的基础上又引入了代价量设计了一个全新的光流估计网络的框架。Hui 等人[20]沿用了之前的由粗到细地光流提取框架，并进一步对在大运动情况下估计不准的问题进行了优化，设计了一个全新的称为 LiteFLowNet 的光流估计网络，与先前的工作 F1owNet2 [17]相比不仅在精度上取得了提升，在速度上也是不遑多让。但是所有这些通过卷积神经网络对运动进行隐式估计并提取光流的方法在训练的过程中都需要通过真实光流进行监督，才能取得良好的效果，但是真实光流并不是容易获得的数据，这也是这类方法目前所面临的问题之一。

Zhou 等人[21]提出了一种新的方法，他们以基于卷积深经网络的自编码器为骨干网络对运动进行隐式的估计，而后从中提取出了一种叫做表观流的流信息，然后设计了一种基于表观流的图像生成方法，基于输入帧信息和表观流实现对插值帧的生成。Liu 等人[22]同样基于隐式运动估计的思想提出利用卷积神经网络估计一种体素流(voxel flow);它是一种与表观流概念相似的流，我们可以将它们都当作是一种面任务的流[23](task-oriented flow)。这类方法的优势在于它们不需要用到真实的光流作为训练时的监督信号，而是以真实帧作为自监督信

号，这样就可以将容易获取的大量原始视频数据直接应用于模型训练当中。

还有一些做法在原始图片之外，还利用相位信息作为参考去进行运动估计，例如，Pinte等人[11]先估计出连续帧之间的局部相位信息，联合这种局部的相位信息再利用卷积神经网络进行运动估计。

接下来主要介绍有关运动补偿的部分。一般来说，由于运动估计中会出现一些固有的问题，例如“孔洞问题”，多个光流流经同一个像素点或者没有光流流经的像素点，都会导致插值帧中的一些像素点的值无法确定，同时如果出现一些大运动、亮度突变等特殊情况，运动估计的结果就更不准了，此时我们需要通过运动信息进一步的对运动估计进行细化，使其能覆盖至每一个像素点，并且保证每个像素点上一定的精度，总之运动补偿其实就是结合输入帧的图像结构信息以及运动估计所得到的流信息对运动表示进行进一步的优化，这种技术已经被广泛应用于在视频中的高效编解码[24,25,26]。从另一个角度看来，运动补偿亦相当于是一种生成方法[27,28,29]，只不过它利用了光流信息进行像素点的生成。

一般来说，运动补偿技术在使用时都是基于光流这样的拉格朗日表示[30,31,32,33]。可以想象，假如我们使用不同的方式去表征运动信息，在最后进行像素级别的生成时所需要采用的方法也会大不相同。即使是对光流信息，现有的方法也分为两类，一类是只获取单向光流，利用输入帧作为参考，仅使用单向光流对输入帧进行翘曲（即像素点位置的偏移）得到插值帧。但是这种仅使用单向光流的方法就必然会出现我们之前所提过的多个光流流经同一个像素点或是没有光流流经像素点的“孔洞”问题，而之所以会出现“孔洞”问题的原因是运动过程中的遮挡使得解空间增大[34,35]。为了解决在生成插值帧过程中遇到的多个光流流经同一个像素点或是没有光流流经像素点的“孔洞”问题，重叠块加权的方法[31]，空洞填充的方法[33]以及运动场平滑[6]的方法被相继提出并得到了广泛的应用。还有一类方法是基于双向光流的它们可以通过对双向光流进行对齐去解决在插帧过程中可能出现的“孔洞”问题[36,37,38,39]。具体来说，在得到双向光流之后，通过对连续的两个输入帧分别进行翘曲，再将两个翘曲帧送入卷积神经网络进行对齐，从而改善光流法中的“孔洞”问题。基于双向光流的方法是以待插值帧为基准，但是实际上我们并没有该帧，所以其实此类方法是基于线性运动的假设实现的，它们假设插值帧指向两个输入帧的运动矢量相反，且大小与插值帧和输入帧之间在时序上的距离成正比。通常为了处理遮挡问题，会为了两张输入帧的像素点分别计算遮挡因子，通过遮挡因子作为权重对前后帧进行加权[33,40]。对于帧插值的任务，一般来说都采用基于双向光流的方法[14,15,37,41]，它相比于单向光流的方法精度更高，鲁棒性更好。

传统的运动补偿方法是利用光流信息以及输入帧的信息对插值帧进行像素级的合成。但是还有一些研究人在探索利用一些其他的信息比如相位信息来实现帧插值，而想要利用这些结构并不相同的信息，所要采取的像素级合成方法也不相同，它们往往都会根据情况设计特定的像素级合成方式。比如说，Mahajan 等人[42]提出了一种基于相位梯度信息的方法，同样是基于卷积神经网络对运动进行隐式地估计然后提取出相位梯度信息，而后根据泊松方程，

可以有效地根据相位梯度信息对图像进行像素级的生成。如果我们要进行隐式运动估计，并用欧拉运动表示法对运动进行表示的话，如何选择像素级合成的方法就与具体的图像分解息息相关。具体来说，之前提到的 Wu 等人[8]通过获得图像的相位信息的方式对运动进行隐式的估计，基于图像的 Laplacian 金字塔去提取图像相位信息，并利用逆过程对插值帧进行像素级的生成。而后，Wadhwa 等人[43]又提出了利用复数可操纵金字塔[44]来分解图像信息，最后同样采用可操纵金字塔的塌缩过程去实现对插值帧的像素级合成。再之后，Meyer 等人[10]又进一步的考虑图像中蕴含的相位信息，提出了一种用卷积神经网络提取图像的相位和幅值的方法，而后根据相位与幅值构建可操纵金字塔，实现对插值帧的像素级合成。这类基于欧拉表示法的帧插值方法虽然不会出现光流法因为遮挡而出现的“孔洞”问题，但是它们在进行运动估计时可能因为与运动信息之间存在一些分布偏移，从而导致了一些估计错误，而这些错误同样会在特定的像素级合成方法中向后传递，从而造成运动模糊的问题。

随着卷积神经网络的大火，生成模型方法也取得了巨大的进步，而这种图像生成的模式或许也可以为给帧插值任务中的插值帧生成带来启发。其中变分自编码器[44](Variational Auto-Encoder, VAE)和生成对抗网络[45] (Generative Adversarial Network, GAN)是被研究者们广泛研究和应用的生成模型标准范式，最近，扩散模型[46]的提出也使得生成模型的研究愈发火爆，成为了生成模型的新范式。生成模型是一种用来生成现实中不存在的图像的模型，它通过对图像在自然界中的分布进行建模，建立起一个低维概率分布到高维图像数据的映射，而后通过在低维分布上的采样获得高质量的合成图片。而这种由低维当高维图像的映射，是我们在帧插值网络最后生成插值帧的过程中可以用到的，像很多现有的帧插值方法，都采用了类似自编码器的架构，而生成模型正可以用于对特征的解码。有一些研究人员已经利用生成模型的思想取得了不错的帧插值效果[47,48]。但是直接使用卷积神经网络由于在特征编解码过程中的细节丢失，因为卷积神经网络往往更关注于图像的语义信息，所以会导致出现模糊的结果[49,50]。考虑到这些问题，许多研究者们借鉴了传统的基于光流的方法[22,41,51]，在通过卷积神经网络获得了光流信息之后，基于光流信息和输入帧信息进行翘曲计算得到插值帧。Niklaus 等人[14]提出了一个新的看法，他将运动估计和像素合成融合起来交给一个自适应卷积核来完成，这样我们只需要通过卷积神经网络为每个像素点计算出一个自适应卷积核即可，卷积核直接对输入帧的局部块进行操作，这样可以有效避免细节的丢失，从而防止出现模糊的结果。除了基于自适应卷积核的方法以外，还有一些方法为了得到更加清晰地插值帧，引入了感知损失[52]作为损失函数，感知损失是一种在不同层级的特征图之间计算差异的损失函数，这样可以促使插值帧对人类视觉友好。Niklaus 等人[53]在感知损失的启发下，先对图像提取不同分辨率的特征图，然后利用光流信息对特征图进行翘曲，再送入卷积神经网络当中进行图像生成。

基于卷积神经网络的学习方法在这些年不断地拔高着帧插值的精度，取得了传统方法许多年都未曾达到的成就，但这并不意味传统方法就没有研究意义了，我们可以看到，当下最

有效的深度学习方法往往都是基于传统方法提出的运动估计流程，对传统方法的研究可以继续启发我们通过不同的特征映射去表示运动并高效的合成插值帧。

同时，目前最主流的方法依然是基于光流方法，由于光流信息是运动信息一种非常有效的表示，所以提取光流拟合运动信息是非常有效的。但是基于光流的方法一般都需要先进行光流估计，然后再将光流信息送入不同结构的像素生成模块生成插值帧，所以光流估计当中估计误差就会继续延展到后续的像素生成模块当中，且光流估计对于复杂运动以及大遮挡的情况面临许多技术难点，所以近年来也有很多学者展开了大量的端到端的直接进行像素合成的算法，严格来说，这种方法所提取的是一种面向任务的特征流。

1.4 本文主要研究工作

本文主要的研究内容是基于纹理增强的帧插值方法研究以及基于 TransFormer 的帧插值方法研究，主要是为了解决一些例如光照突变或是遮挡物或是大运动等特殊场景下的帧插值问题。

考虑到基于光流的方法对光流过度依赖，从而导致其网络输出依赖于光流特征的提取，这使得帧插值任务沦为光流估计的下游任务，理想的情况是帧插值算法的特征可以反过来去作为光流估计的监督信息以促进光流网络或是其他下游任务的发展。现有的基于直接特征表示的方法在结果上往往会出现模糊，我认为这是由于在下采样的过程中不可避免的出现了一些纹理细节信息的丢失，基于这种想法，我引入了基于纹理增强的后处理模块，增强插值帧的视觉效果，它先使用金字塔级联网络对输入帧进行对齐，再通过一个融合模块提取出纹理信息，并作为残差与生成的插值帧相加。同时我也注意到了利用 Pixel Shuffle 的方法对输入进行特征重塑，在不引入额外参数的情况下，促进了网络学习运动信息，于是引入了 Pixel Shuffle 进行浅层的下采样和最后阶段的上采样。并通过充分的对比实验验证了该方法的有效性。

由于卷积神经网络固有的局部性和平移不变性的归纳偏置，在对视频信息进行提取时往往并不尽如人意，于是我想到了利用 Transformer 进行特征提取，先利用 Transformer 进行特征提取获得特征图，再通过 3D 转置卷积进行上采样，最后加入了简化版的纹理增强模块，进一步提升插值帧视觉质量。并经过一系列的对比实验以及消融实验证明了该算法在帧插值任务上的有效性。

1.5 本文的内容安排

全文共分为五章，其中每章的内容安排如下:

1. 第一章是绪论，简单地介绍了一下本文的课题研究背景、研究意义，并阐述了一下本文主要的研究工作和本文的内容安排。
2. 第二章是相关工作，在简单介绍了帧插值问题后，详细介绍了领域内的一些重要工作，以及一些对本文工作具有强启发性的方法。

3. 第三章提出了一个基于纹理增强的帧插值方法，本章考虑到基于特征表示的方法现在面临的细节模糊的问题，提出了一个基于金字塔级联网络进行帧对齐再提取纹理残差增强插值帧视觉效果的方法。
4. 第四章提出了一个基于 Transformer 的帧插值方法，考虑到卷积神经网络的归纳偏置，提出用 Transformer 作为骨干网络进行视频帧特征的抽取，并辅以后处理模块，其结果可以打败基于卷积神经网络的方法。
5. 第五章是总结与展望，简要概括了本文所做的研究工作，并且给出了当前研究方向依然面临的问题以及可以继续突破的难点。

第二章 相关工作

帧插值领域中近些年来取得了长足的进步，传统方法以运动估计与运动补偿为主，而随着近年来机器学习、深度学习的飞速的发展，基于学习的方法也越来越为大家所关注，也涌现出了许多新的思路，当下基于学习的方法主要可以分为四类，基于光流的方法；基于相位的方法；直接方法以及基于核估计的方法。得益于深度卷积神经网络在计算机视觉上的卓越效果，这些基于深度学习的帧插值算法都取得了不错的效果。

本章首先简单介绍了一下帧插值，而后分别介绍基于光流的方法，基于核估计的方法，光流与核估计相结合的方法，基于卷积神经网络的直接方法，基于 Transformer 的方法。

2.1 帧插值

近年来随着多媒体领域进一步的发展，我们的生活已经离不开视频技术了，短视频，影视剧，监控录像等应用已经充斥在我们的身边。但是显示设备的迭代速度太快，也导致市场上帧率以及帧分辨率的显示都出现了高低版本并存的现象，这严重的影响了用户体验，为了解决这个问题，给用户带来更好的体验，帧插值技术应运而生，并且被广大的科研人员以及企业工程师们所关注。

帧插值技术是一种极具应用价值的技术，它能够在两个给定的帧之间生成一个新的帧，使视频更加流畅、细腻。这种技术在各个领域都有广泛应用，比如视频摘要、自动事件检测等。它能够提取关键帧，增强视频的表现力和可读性。在电影制作中，它能创造慢动作效果；在体育比赛直播中，它能生成超级慢动作回放；在游戏领域，它能提高画面流畅度。此外，在视频压缩和传输中，它还能减少数据量，提高传输效率。总之，帧插值技术为我们提供了更好的视觉体验。

而对研究工作而言，帧插值问题可以被建模为输入 t-1 时刻的帧I_{t-1}和 t+1 时刻的帧I_{t+1}，输出 t 时刻的帧I_t，如图 2.1 所示。

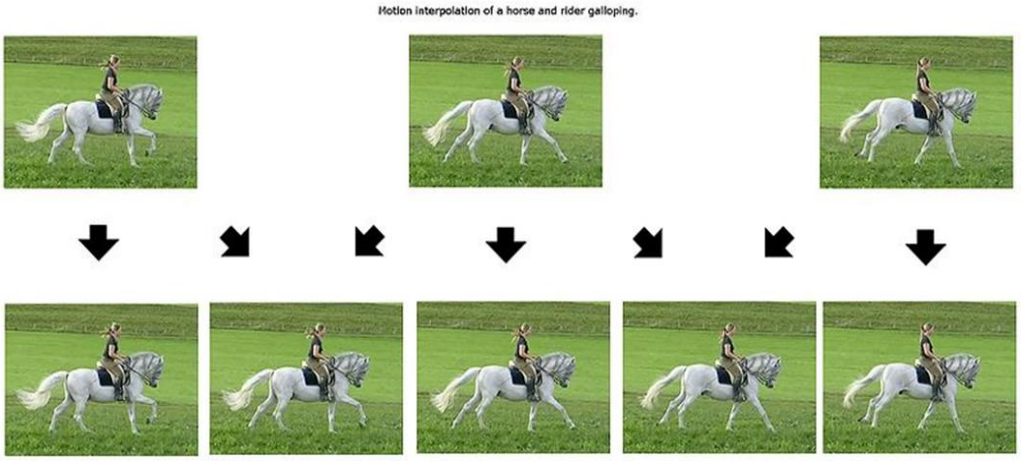


图 2.1 帧插值过程

2.2 基于光流的帧插值方法

2.2.1 Super Slomo

Super Slomo[51]是一种基于光流的高效帧插值方法，该算法基于传统的光流法假设，根据双向光流将两个输入帧翘曲到某一特定时间点，根据时序上的距离对双向光流分配权重进行融合。基于这样的思想，它允许模型对输入帧中的任意时刻的运动信息进行建模，并实现对两帧间任意时刻进行插帧操作。Super Slomo 的具体做法可以分为以下几个步骤：

1. 计算光流

首先，对于当前帧的前一帧和后一帧，需要计算它们之间的光流。光流是一种描述相邻帧之间像素位移的技术，可以帮助模型理解相邻帧之间的运动信息。在 Super Slomo 中，采用 FlowNet 2.0 来计算光流。

2. 提取特征

使用卷积神经网络来学习相邻帧之间的空间和时间上的关系，并提取特征。在 Super SloMo 中，使用 ResNet-50 来提取特征，其中输入包括当前帧的前一帧和后一帧，以及它们之间的光流信息，经过多个残差块后，输出中间帧的特征表示。

3. 生成中间帧

根据前一帧、后一帧和提取的特征，使用卷积神经网络来生成中间帧。在 Super Slomo 中，使用了一种基于学习的方法来生成中间帧，即将视频帧插值任务转化为像素级别的回归问题。具体来说，通过学习相邻帧之间的空间和时间上的关系，可以预测中间帧的像素值，从而实现视频帧之间的平滑过渡。

4. 运动补偿

为了更好地利用运动信息，Super SloMo 还使用了一种运动补偿技术，即利用光流算法将前一帧和后一帧对齐，从而消除运动的影响。具体来说，在生成中间帧时，将前一帧和后一帧按照光流对齐，然后再进行插值操作。

如图 2.2 所示，即为 Super Slomo 的网络模型结构，该方法提出主要考虑了两个问题。

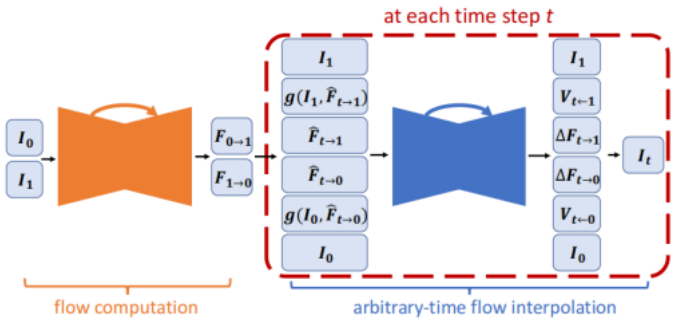


图 2.2 Super Slomo 模型结构图

首先，之前的方法都是只能实现 t=1/2 处的中间帧估计，如果想要在两帧中间插入更多帧就需要多次预测，比如要估计出 t=1/4 处的中间帧，就需要先估计出 t=1/2 处的，再用 t=0

和 t=1/2 的估计，因此会比较耗时，而且这种方法也没有办法估计 t=1/3 的中间帧，因此作者希望能够加入时间变量 t 来控制中间帧的生成过程。自然而然地，作者想到了将时间变量引入到中间帧的生成过程当中，于是提出了一种线性估计的方法去估计 t 时刻的双向光流。计算方法如公式(2.1)和(2.2)所示，先用光流预测网络估计出$F_{0\to1}$和$F_{1\to0}$，再利用时间 t 对$\hat{F}_{t\to1}$和$\hat{F}_{t\to0}$进行估计

$$\hat{F}_{t\to0} = -(1-t)tF_{0\to1} + t^2F_{1\to0} \tag{2.1}$$

$$\hat{F}_{t\to1} = (1-t)^2F_{0\to1} - t(1-t)F_{1\to0} \tag{2.2}$$

具体而言，利用双向光流$F_{0\to1}$和$F_{1\to0}$预测时刻 t 的双向光流过程如图 2.3 所示

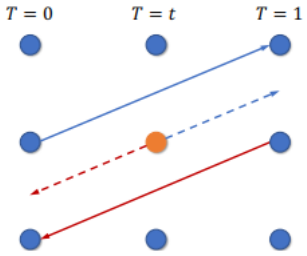


图 2.3 Super Slomo 的中间光流近似过程

时刻 t 的光流可以被近似为：

$$\hat{F}_{t\to1} = (1-t)F_{0\to1} \tag{2.3}$$

或：

$$\hat{F}_{t\to1} = -(1-t)F_{1\to0} \tag{2.4}$$

而后再根据时刻 t 同时利用上述两种近似，对它们进行加权求和即可获得本文的中间光流计算公式(2.1)和(2.2)，当我们获得了 t 时刻的双向光流之后，我们就可以通过两个输入帧和双向光流分别绘制出 t 时刻的两个近似帧，并加权求和得到中间帧图像，计算过程如下：

$$\hat{I}_t = \alpha_0 \odot \mathrm{g}(\mathrm{I}_0, F_{t\to0}) + (1-\alpha_0) \odot \mathrm{g}(\mathrm{I}_1, F_{t\to1}) \tag{2.5}$$

这里的α_0代表着近似帧的权重，它由时序依赖关系和遮挡推理这两个元素控制，g(·,·)代表着一个后向绘制函数，它一般是用双线性插值来实现的，相比于前向绘制函数，它更好计算而且被证明是可微的，⊙代表着对应元素相乘。时序依赖关系在前述的计算中间光流的公式中也有所体现，其中蕴含的意义是当时刻 t 越趋近于 0 时，则I_0对$\hat{I}_t$贡献越趋于 1，同理，当时刻 t 越趋近于 1 时，则I_1对$\hat{I}_t$贡献越趋于 1。同时，对于此类基于光流的方法，遮挡是必须要处理的一个问题，因此，该方法引入了一个可见性映射$V_{t\leftarrow0}$和$V_{t\leftarrow1}$来表示从 T = 0 到 T = 1 像素移动过程中是否可见。$V_{t\leftarrow0}(p) \in [0,1]$表示从 T = 0 到 T = t 像素 p 是否可见，越趋近于 1 则代表着可见程度越高，趋于 0 则意味着完全遮挡，作者利用卷积神经网络预测出了$V_{t\leftarrow0}$和$V_{t\leftarrow1}$，并使得它们之间满足约束：

$$V_{t\leftarrow0} + V_{t\leftarrow1} = 1 \tag{2.6}$$

这个约束也可以看作是对遮挡参数的建模，可以看做是物体不遮挡的部分与遮挡的部分相加就是物体的全部了。由此，也可以进一步细化公式(2.5)，得到如下公式：

$$\hat{I}_t = \frac{1}{Z} \odot \left((1-t)V_{t-0} \odot g(I_0, F_{t\to 0}) + tV_{t+1} \odot g(I_1, F_{t\to 1})\right) \tag{2.7}$$

基于以上的改进，Super Slomo 取得了更好的插帧效果，而且基于学习的方法也使得它的泛化性更好，但显而易见的是，这样的做法在获得的双向光流足够精确地情况下可以取得良好的效果，同样也导致了它依赖于光流提取的精确度，而光流法本身在有遮挡、光照变化以及复杂运动的情况下都会出现估计不准的情况，而且光流提取与插值帧生成的模块之间是以级联的方式构建的，这也会影响到计算速度。

2.3 基于核估计的帧插值方法

2.3.1 Adaconv

视频插帧通常涉及两个步骤：运动估计和像素合成。这种两步法的效果很大程度上取决于运动估计的质量。Adaconv[15]提出了一种鲁棒的视频插帧方法，该方法将这两个步骤合并为一个过程。具体来说，该方法将插值帧的像素合成视为两个输入帧上的局部卷积。卷积核捕获输入帧之间的局部运动以及用于像素合成的系数。该方法采用全卷积的深度神经网络来估计每个像素的空间自适应卷积核。可以直接使用大量的可用视频数据直接端到端地训练此深度神经网络，而无需像光流这样的难以获得的标准数据。实验表明，将视频插值表述为单个卷积过程可以使该方法很好地处理诸如遮挡，模糊和亮度突然变化之类的问题，并实现高质量的视频插帧。

给定输入帧I_1和I_2，传统的基于运动估计与像素合成的方法通过先为插值帧中的像素点坐标(x,y)找到对应于输入帧I_1和I_2中的像素点坐标(x_1,y_1)和(x_2,y_2)，并根据(x_1,y_1)和(x_2,y_2)所对应的像素值进行加权求和得到插值帧像素点(x,y)对应的像素值，同时由于运动估计得到的结果不一定为整数，而像素点坐标需要为整数，所以一般还需要对像素点坐标(x_1,y_1)和(x_2,y_2)在输入帧中进行采样才能获得最终的插帧像素值，其流程如图 2.4 所示：

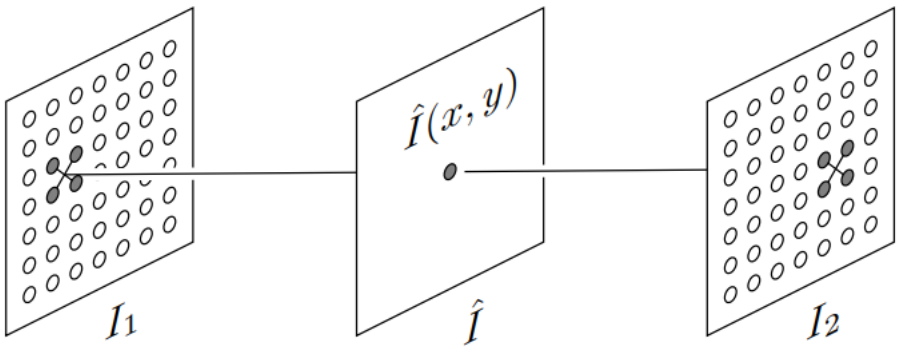


图 2.4 运动估计像素合成两步流程图

但是这样做光流法所固有的在有遮挡、光照剧烈变化、复杂场景下估计不准的问题无可避免，但是如果将两步合为一步，将帧插值的问题视作对前后两张图像对应区域的卷积，则可以很好的避免这些问题，卷积中隐含的包括了运动估计、遮挡推理等一系列任务，得出了

一个针对帧插值任务完全数据驱动的深度学习方案，而不需要额外的光流信息等无法直接获得的标签信息，一步实现插帧的建模可视化流程如图 2.5 所示：

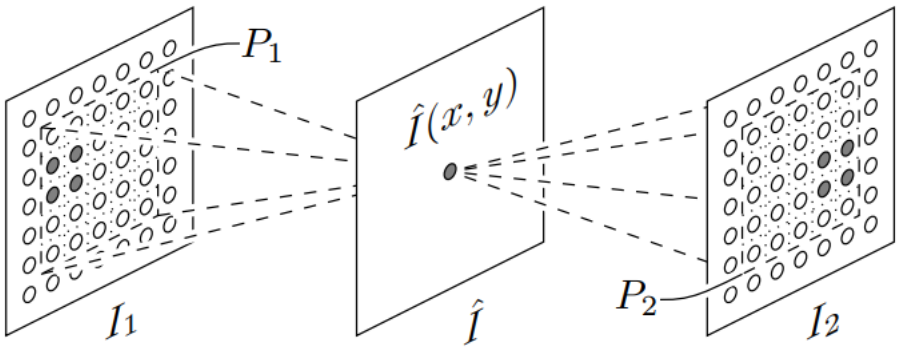


图 2.5 基于核的插帧方法一步流程图

具体来说，对于插值帧中的像素 p(x, y)，深度神经网络将以该像素为中心的两个感受野块(receptive field patches) R_1和R_2作为输入并估计卷积核 K，该卷积核用于与输入块P_1和P_2卷积来合成输出像素，处理过程如图 2.6 示。

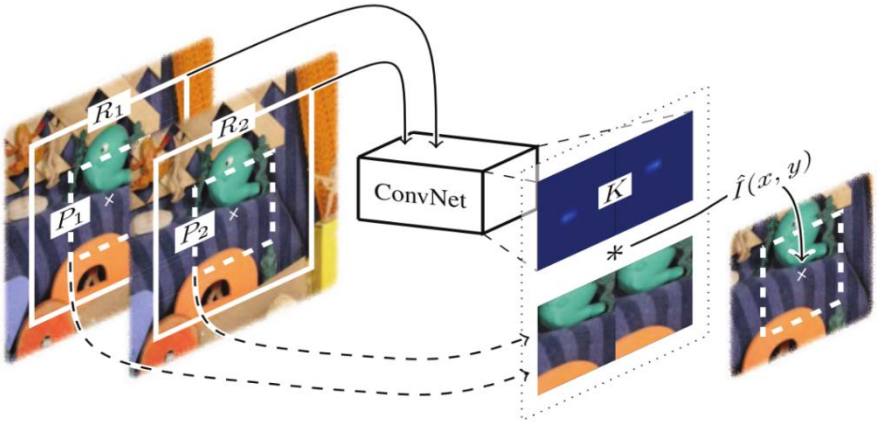


图 2.6 Adaconv 模型可视化流程图

这里的输入块P_1和P_2大小为41×41，感受野块 R_1和R_2尺寸为79×79，输入块P_1和P_2经过拼接后与卷积核 K 进行卷积操作生成插值帧像素点，故卷积核 K 尺寸为41×82，感受野块设置大于输入块主要是为了解决运动估计中的孔径问题。

通过卷积表示像素内插具有一些优点。首先，将运动估计和像素合成结合到一个步骤中提供了比两步过程更为强大的解决方案。其次，卷积内核提供了解决诸如遮挡之类的困难情况的灵活性。例如，在遮挡区域中的光流估计是一个本质上很困难的问题，这使得典型的两步法难以进行，必须采取其他基于启发式的步骤。利用这种数据驱动的方法来直接估计卷积核，可以为被遮挡区域产生视觉上合理的内插结果。第三，如果进行了正确的估计，则该卷积公式可以无缝地与高级的重新采样技术(如边缘感知滤波)相结合以提供清晰的插值结果。

该方法同样也有其局限性，特别是卷积核中隐含着运动估计的内容，但是受限于卷积核的局部性，对于大动作而言，卷积核将无法捕获超过卷积核尺寸的运动。

2.3.2 Sepconv

原作者继提出方法 Adaconv 之后，充分考虑了 Adaconv 的不足之处，因为要处理幅度比较大的运动，用空间自适应卷积核估计中间帧需要较大的核，而且要为每个像素生成一个卷积核，因此计算量是十分的大的，于是又针对以上问题进行了进一步的改进。在 Sepconv[14] 中提出了自适应可分离的卷积，大大减小了运算量，用一对 1D 卷积核替代原先的 2D 卷积核，大大减小了参数量，由n^2减小到了$2n$，并且加入了感知损失以促进神经网络对语义信息的学习。

具体而言，不同于 Adaconv 中直接为每个像素点生成一个 2D 卷积核，而是用一对 1D 卷积核代替它，从而实现了参数量量级的锐减，从而可以直接生成高分辨率图像，而不会受限于内存大小。Sepconv 网络结构图如图 2.7 所示：

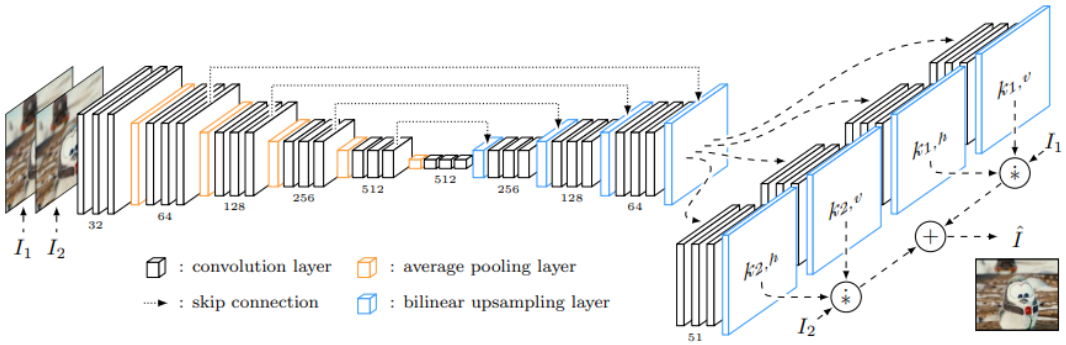


图 2.7 Sepconv 网络结构图

简单来说，Sepconv 的网络结构前面是 encoder 的结构，用于提取特征，而后经过 decoder 生成预测，通过 skip connection 连接，使得中间特征包含更为丰富的语义信息，再之后通过四个子网络得到两对插帧要用的卷积核。基于卷积的插值方法原先的单像素生成公式为：

$$\hat{\mathrm{I}}(\mathrm{x},\mathrm{y}) = \mathrm{K}_1(\mathrm{x},\mathrm{y}) * \mathrm{P}_1(\mathrm{x},\mathrm{y}) + \mathrm{K}_2(\mathrm{x},\mathrm{y}) * \mathrm{P}_2(\mathrm{x},\mathrm{y}) \tag{2.8}$$

其中$\mathrm{K}_1(\mathrm{x},\mathrm{y})$和$\mathrm{K}_2(\mathrm{x},\mathrm{y})$代表在$(\mathrm{x},\mathrm{y})$处自适应估计出来的两个卷积核，$\mathrm{P}_1(\mathrm{x},\mathrm{y})$和$\mathrm{P}_2(\mathrm{x},\mathrm{y})$代表输入两帧中以$(\mathrm{x},\mathrm{y})$为中心，大小与卷积核大小相同的像素块。Sepconv 则将K_1拆解成了了$\mathrm{k}_{1,h}$和$\mathrm{k}_{1,v}$，然后按垂直方向和水平方向依次对对应输入帧的像素块进行卷积，进一步加快了运算速度。在损失函数方面，将L_1损失和感知损失相结合进行训练，损失函数的公式如下：

$$L_1 = \| \hat{\mathrm{I}} - \mathrm{I}_{\mathrm{gt}} \|_1 \tag{2.9}$$

$$L_{\mathrm{F}} = \| \phi(\hat{\mathrm{I}}) - \phi(\mathrm{I}_{\mathrm{gt}}) \|_2^2 \tag{2.10}$$

其中L_{F}代表感知损失，$\phi(\cdot)$代表从图像中提取的特征。

Sepconv 在空间复杂度和时间复杂度上相较之前的方法都取得了提升，但是在对视频进行处理的时候，运动的物体旁边还是会有伪影出现，而且可分离卷积并不是万能的，它并不能完全替代卷积，有些卷积核是无法被可分离卷积所表示的，同时可分离卷积在图像生成时

可能存在棋盘伪影的问题，也对模型的设计提出了更高的要求，需要选择合适的上采样方法，所以 Sepconv 采用了双线性插值的方法构建上采样模块。

2.4 结合光流估计网络与核估计的帧插值方法

2.4.1 MEMC-net

2018 年，Wenbo Bao 等人[37]在论文中提出一种由运动估计与运动补偿联合驱动的神经网络模型 MEMC-net 来解决帧插值问题。它用一个新颖的自适应变形层结合光流和插帧核，来合成目标帧的像素。这个层是完全可微分的，使得流和核的估计网络可以被同时优化。

传统的基于运动估计与运动补偿的方法首先会估计出前后帧之间的动作向量，沿着动作轨迹，参考帧的像素被用于插值出中间帧，传统的运动估计方法使用基于块的算法，例如 3D 循环搜索，这是硬件友好且运算高效的。基于块的方法通常把图片切分成小的像素块，依据例如最小化绝对块间差之和的搜索标准，利用特定的搜索算法，例如在空间上或时间上的搜索，层次搜索等，计算这些小块的动作向量。而后是运动补偿，利用重叠块来应对像素块的错误动作向量，也有一些方法利用光流来保证流场的真实性，通过图片融合或重叠块构建的补偿滤波器被用于解决遮挡或块状效应。除此之外，还有一步是后处理，用以最小化伪影并且提升视觉质量，由于相对运动和不同深度物体间的遮挡，估计到的流向量可能产生带有空洞区域的错误插值结果，对于这些空洞区域，Kim 等人利用空洞插值方法来恢复缺失的像素，Wang 等人提出了一个三边滤波方法来填充空洞和平滑时空域上的补偿误差。

基于流的方法则需要预测双向光流，或者在线性动作模型的基础上使用双线性插值来对齐输入帧。为了合成输出图片，一个通用的技巧是估计一个遮挡掩码(occlusion mask)去自适应混合变形的帧。由于双线性插值混合了邻近像素，当输入帧没有被很好地对齐时，流方法不可避免地产生了伪影或者模糊。

基于核的方法不依赖于光流，插帧任务可以被转化为局部卷积操作，前述的 AdaConv 和 Sepconv 都需要对每个输出像素估计空间自适应的卷积核，在这类方法中，一个大的核被用于处理大动作，这就需要很大的内存来处理高分辨率图片，尽管 Sepconv 用可分离卷积做出了改进，显著减少了内存需求，但是它依然不能处理超过预定义的核大小的动作。

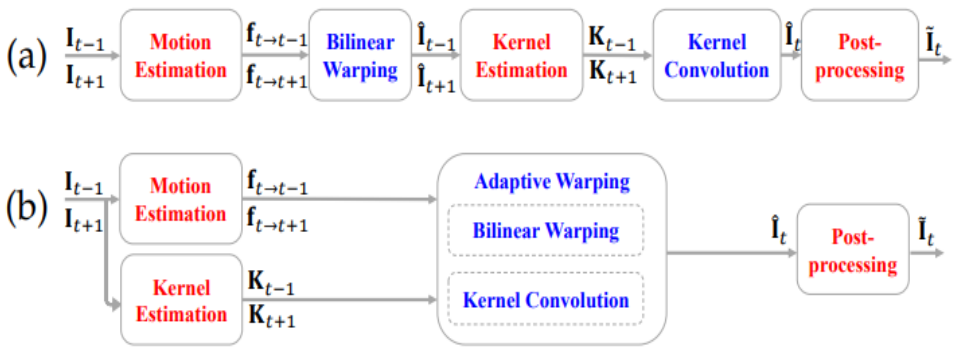


图 2.8 顺序结合(a)与 MEMC 结合(b)对比图

根据上述方法各自的局限性，MEMC-net 考虑将基于核的深度学习方法与基于流的深度学习方法相结合，由于直接顺序结合的效果并不好，光流估计的低质量限制了后续的核估计、后处理，以致于无法得到观感良好的结果，因此，该方法设计了自适应变形层，使得核估计网络与光流估计网络呈并联状态，将光流估计网络和核估计网络紧密的结合在了一起，取长补短，起到了互补的作用，两种结合方式流程如图 2.8 所示。

该方法的自适应变形层的在一个单步中融合了双线性插值和核卷积，其前向传播过程如公式(2.11)所示。

$$\hat{I}(x) = \sum_{r\in[-R+1,R]^2} k_r(x) I(x + \lfloor f(x) \rfloor + r) \tag{2.11}$$

其中x代表在 RGB 图像中像素点的位置，$x \in [1,H] \times [1,W]$，$\lfloor f(x) \rfloor$代表经过流投影层得到的光流并对其进行向下取整的操作，$k_r(x)$是k_r^l和k_r^d的乘积，k_r^l代表补偿核参数，k_r^d代表双线性系数，它们的可视化如图 2.9 所示。

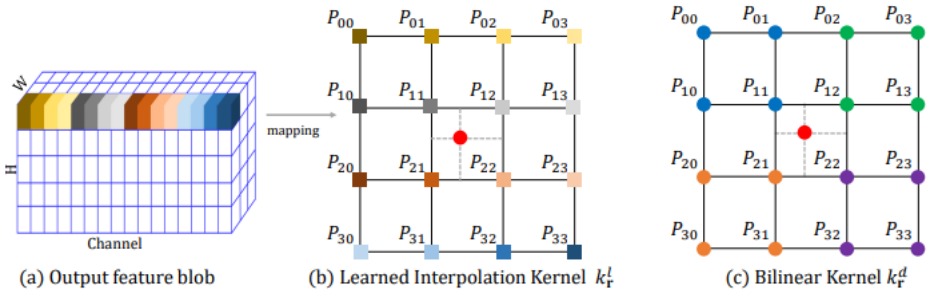


图 2.9 MEMC 自适应卷积层参数可视化

对于插值帧的每一个像素点，核估计网络都会输出一个1×16的特征，可以与以当前像素点为中心的周围 16 个像素点形成一一对应的关系，即得到了补偿核参数k_r^l，而后根据光流对齐预测的像素点位置（即图中红点位置），根据与周围四个点的距离计算出双线性权重k_r^d，为使其能和不同大小的补偿核相匹配，位于相同方位的像素点采取相同的权重，如图 2.9 (c)所示，相同颜色的像素点具有相同的双线性权重。

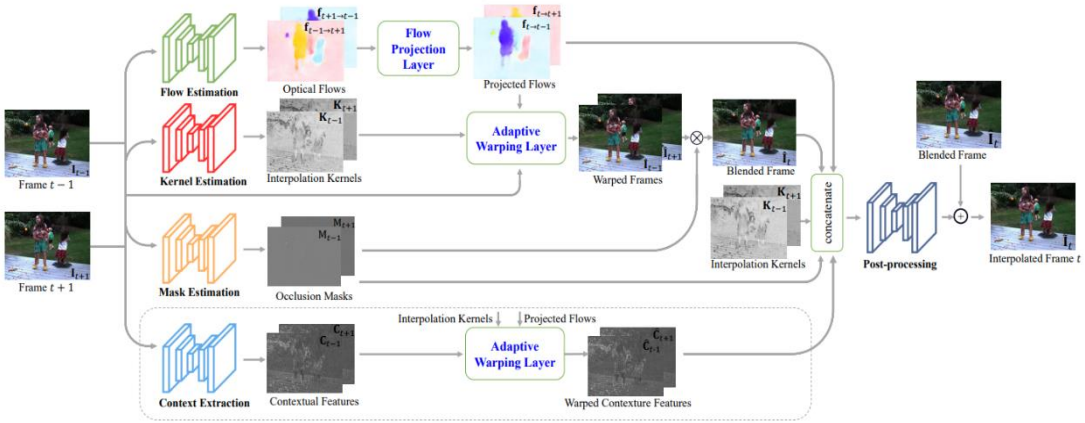


图 2.10 MEMC-net 算法模型示意图

该方法具体的算法模型结构如图 2.10 所示，该模型利用了传统运动估计与运动补偿框架的优点来解决大动作的问题，也利用数据驱动的基于学习的方法来提取有效特征。提出了名为自适应变形层和流投影层的两个网络层，它们被用于紧密结合所有的子网络，使得模型可以端到端训练。

2.4.2 DAIN

2019 年，Wenbo Bao 等人[54]基于 MEMC-net 的工作进行了进一步的改进，在之前对帧插值任务的研究中虽然已经取得了较大的提升，但当出现大物体的运动或遮挡时，插值质量往往不尽人意，出于对这个问题的考虑，该方法的研究人员们想到了利用视频帧的深度信息来检测遮挡。这基于一种直觉，前景对象的运动会优先于背景对象，即浅层对象的光流应该得到优先考虑。为此，他们提出了一种引入深度信息的算法模型。不同于之前的 MEMC-net，他们利用深度信息和光流信息一起经过流投影层来生成中间流，如图 2.11 所示。

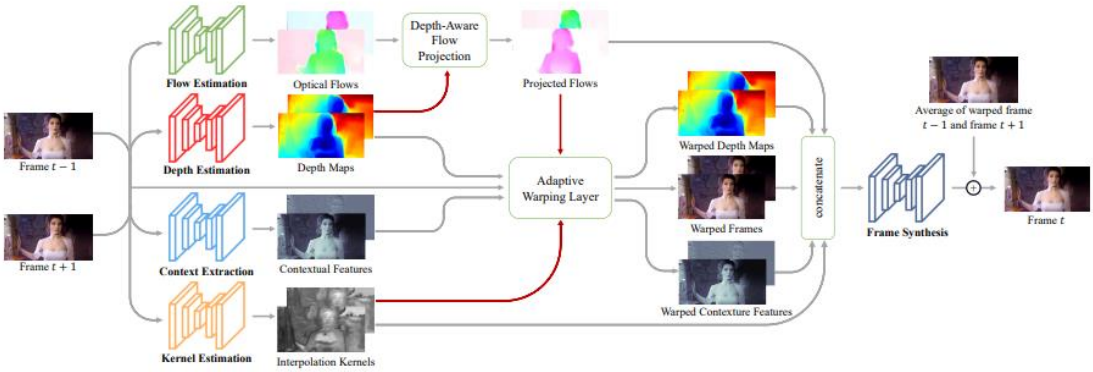


图 2.11 DAIN 算法模型示意图

相对于简单平均的光流，该算法提出的深度感知流投影层由于深度的影响而生成了更清晰的运动边界，其可视化效果如图 2.12 所示，相比于 MEMC-net 基于平均混合的流投影，

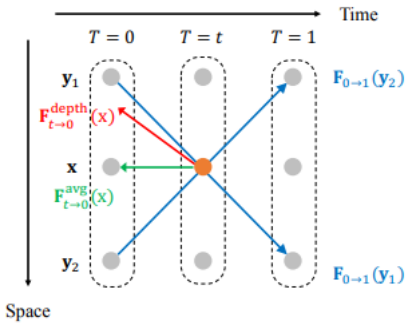


图 2.12 深度感知流与平均混合流对比图

结合了深度信息的感知流会更加精确，这里深度信息估计网络用的是 Megadepth 中的模型，并用已训练好的权重对参数进行初始化。论文中结合深度信息对光流进行扭曲的计算方法如公式(2.12)所示。

$$F_{t\to0}(x) = -t \cdot \frac{\sum_{y\in S(x)} w_0(y) \cdot F_{0\to1}(y)}{\sum_{y\in S(x)} w_0(y)} \tag{2.12}$$

其中$w_0(\cdot)$代表对应像素点深度的倒数，$D_0(\cdot)$代表对应像素点的深度，如公式(2.13)所示。

$$w_0(y) = \frac{1}{D_0(y)} \tag{2.13}$$

而$S(x)$代表在时刻 t 经过像素点x的像素集合。

可以看出当有多个光流流经同一像素位置时，由于前景信息的深度较浅，所以会拥有更大的权重，使得光流的预测更加准确。但作者也提到了使用深度图信息的弊端在于有些特殊场景下深度信息也会预测不准，而这会导致最后的插帧结果中出现模糊的边界。

2.5 基于深度神经网络的直接帧插值方法

2.5.1 CAIN

由于目前流行的视频帧插值技术通常依赖于光流估计，这会增加模型的复杂度和计算成本，并且在存在大运动和严重遮挡的复杂场景中容易出现误差的传播。为了解决这些限制，作者[55]提出了一种简单而有效的深度神经网络视频帧插值算法，它是端到端可训练的，并且不需要运动估计网络组件。该算法采用了一种特殊的特征重塑操作，称为 "PixelShuffle"，并结合通道注意力，用于替代光流计算模块。设计背后的主要思想是将特征映射中的信息分布到多个通道中，并通过关注通道来提取运动信息，以实现像素级帧合成。该算法在复杂的运动和遮挡情况下表现出色。

首先介绍一下 PixelShuffle 的机制，如图 2.13 所示，我们可以看到该机制其实是通过一

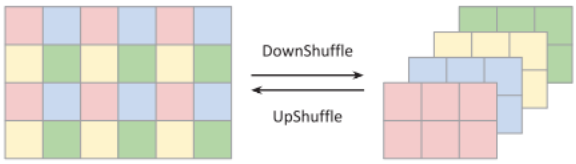


图 2.13 PixelShuffle 示意图

Shuffle 因子 s，以这个 Shuffle 因子作为步长对原图进行分割或是重组，DownShuffle 是将原图分割为多个通道，UpShuffle 则是将多个通道合并为一个通道，这样做既可以保持特征图的信息又可以减少冗余信息，同时它也不需要什么参数，便于计算。这个操作使得获得多通道信息所需的参数量和显存都大大降低，同时引入了通道注意力机制可以捕获运动信息，经通道注意力对运动信息强化之后，配合卷积操作即可实现对像素点的重构，该模型网络结构如图 2.14 所示。

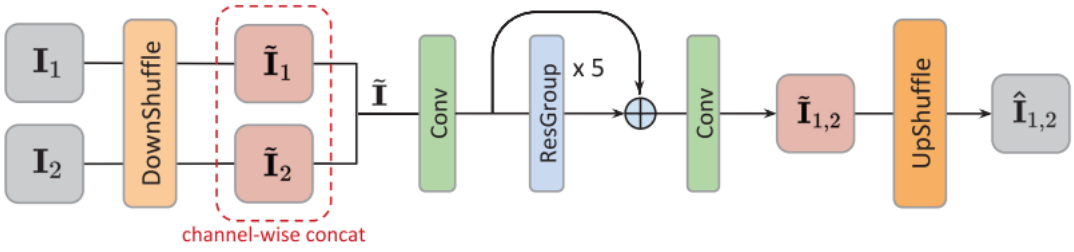


图 2.14 CAIN 模型结构图

该模型经过实验后选择的 Shuffle 因子 s=8，即一个尺寸为$H \times W$的 RGB 三通道图像经

过 DownShuffle 操作后可以得到$3\times8\times8=192$个通道的特征图每个特征图的尺寸为$H/8\times W/8$，ResGroup 的结构如图 2.15 所示。

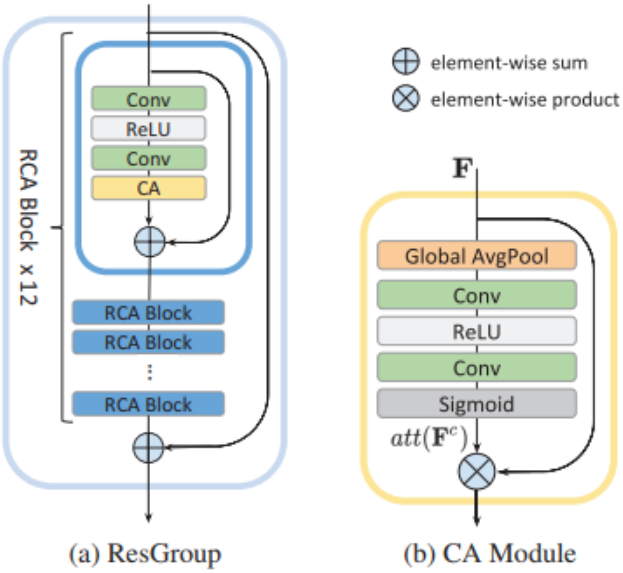


图 2.15 ResGroup 结构图

通道注意力机制的实现是先对每个通道进行空间上的全局平均池化，再通过两个卷积层进行通道注意力的计算，最终在分辨率为1280×720的视频上进行测试，CAIN 模型所需的显存仅为 492MB，而 DAIN 模型所需显存则高达 6944MB。

2.5.2 FLAVR

FLAVR [56]提出了一种新型视频帧插值方法。它不依赖光流，而是使用 3D 时空卷积来实现端到端的学习和推理。这种方法能够有效地学习非线性运动、复杂遮挡和时间抽象，从而在视频插值方面取得了更好的性能，而且不需要额外的输入，如光流或深度图。FLAVR 在多帧插值方面比当前最准确的方法快 3 倍，同时保持了插值精度。而且这种直接方法结构简单易扩展，甚至可以将中间特征用于下游任务当中，比如动作识别、光流估计等任务。

随着深度卷积神经网络在图像领域展现出了强大的学习性能，研究者们也陆续使用基于学习的方法对帧插值问题进行改进，这些方法大致可以分为四类，基于相位的方法，基于核的方法，直接方法，基于流的方法。Long 等人训练深度 CNN 直接预测中间帧。其输出通常是模糊的，并且缺乏细节。原因就是深度模型不能捕捉自然图片和视频的多模态分布。基于相位的方法在一个多尺度金字塔中使用像素相位信息来插帧。然而这种方法不能很好地处理复杂场景中的大动作。基于核的方法，不依赖像素级别的光流，插帧可以被转化为局部块上的卷积操作。在这些方法中，一个大的核被用于处理大动作，这就需要很大的内存来处理高分辨率图片。而直接使用卷积网络的方法，其输出通常是模糊的，并且缺乏细节，原因在于深度模型不能捕捉自然图片和视频的多模态分布。基于流的方法，凭借深度 CNN 在光流估计上的优势，一些方法通过预测双向光流，或者在线性动作模型的基础上使用双线性插值来

对齐输入帧。由于双线性插值基于子像素位移混合了邻近像素，当输入帧没有被很好地对齐时，流方法不可避免地产生了伪影或者模糊。而且由于光流在有遮挡情况下估计不准，为了合成高质量的中间帧，往往还需要设计遮挡模块去优化插值帧的视觉效果。

本文作者为了提高深度模型对视频特征的提取能力，使用 3D 时空卷积来提取特征，使得提取特征的能力增强，除了来自图像空域的信息以外，也能获取来自于帧间的时序依赖关系，从而获得了一个对视频图像更加全面的特征表示，并且选择了 U-net 的网络结构，通过短跳连接，可以使得特征在编解码的过程中可以保留更多的细节信息。FLAVR 网络结构如图 2.16 所示。

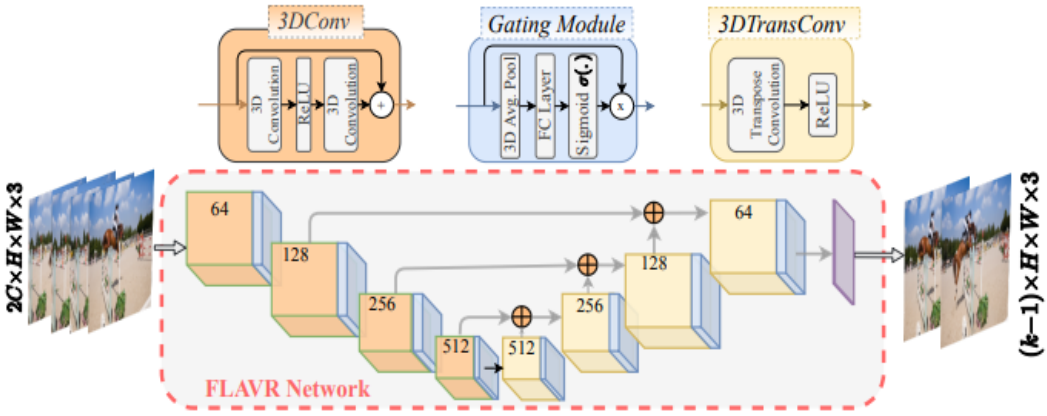


图 2.16 FLAVR 网络结构图

文中编码器结构使用的是 Resnet3D 的结构，除了使用 3D 时空卷积之外，为了增强对运动信息的提取能力，还引入了 3D 通道注意力模块，以更多的关注运动区域，最后在解码器部分，使用了 3D 转置卷积模块，由于转置卷积操作时内核会出现不均匀重叠的情况，当转置卷积的步长大于 1 时，有些输出像素只受单个输入像素的影响，而其他像素则受多个输入像素的影响。这种不均匀重叠会在输出图像中产生高频伪影，呈现为棋盘图案会产生棋盘状的伪影，所以解码器中作者也加入了 3D 卷积模块去减少棋盘状伪影。

作者同时考虑到之前的算法往往只能进行单帧插值，为了实现多帧的插值，作者提出了一种新的从原始视频中进行采样的方法，即首先确定插帧因子 k，将原始视频（假设为 f 帧每秒）进行处理，每隔 k 帧取一帧生成一个低帧率的帧序列V，此时我们需要对V_i和V_{i+1}进行插帧，则选取以V_i和V_{i+1}为中心的 2C 个帧作为输入，这里 C 代表我们选择的输入通道倍数，可以自定义进行设置，文中实验表明，C=2 即可实现较好的插帧效果，以此为例，其采样可视化图如图 2.17 所示。

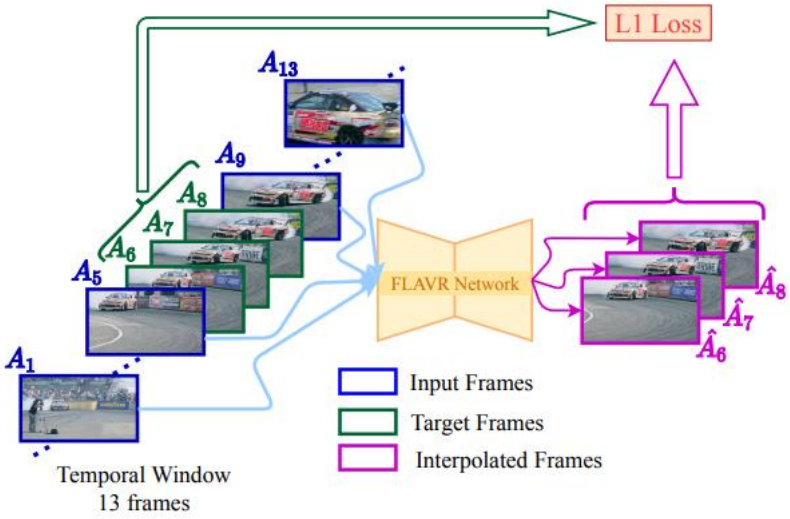


图 2.17 FLAVR 采样示意图

2.6 Transformer 方法

2.6.1 ViT

Transformer[58]架构最初是为自然语言处理任务而设计的，但近年来它在计算机视觉领域也取得了巨大的成功。Transformer 能够捕捉数据中的长距离依赖关系，并且能够处理大量数据，模型容量很大，这使得它在图像分类、目标检测和语义分割等任务中表现出色。在视觉领域，究竟是卷积更好还是 Transformer 更好也一度引发争议。ViT[57]可以说是 Transformer 在视觉领域进行应用的里程碑式著作，因为其模型简单，易扩展而且效果好，所以引爆了后续的研究。而 ViT 原论文中最核心的结论就是：当拥有足够多的数据进行预训练的时候，ViT 的表现就会超过 CNN，突破 Transformer 缺少归纳偏置的限制，可以在下游任务中获得较好的迁移效果，但是当训练数据集不够大的时候，ViT 的表现通常比同等大小的 ResNets 要差一些，因为 Transformer 和 CNN 相比缺少归纳偏置，即一种先验知识，提前做好的假设。CNN 具有两种归纳偏置，一种是局部性，即图片上相邻的区域具有相似的特征；一种是平移不变形，由于 CNN 具有以上两种归纳偏置，提供了很多先验信息，所以只需较少的数据就可以学习一个比较好的模型，但是这也意味着在面临大数据时，CNN 模型容量是不足以学习到足够的知识的，容易陷入过拟合。

ViT 将输入图片分为多个局部块，再将每个局部块投影为固定长度的向量送入 Transformer。出于对图片分类问题的考虑，ViT 还在输入序列中加入了一个特殊的 token，该 token 对应的输出即为最后的类别预测，其工作流程如图 2.18 所示。

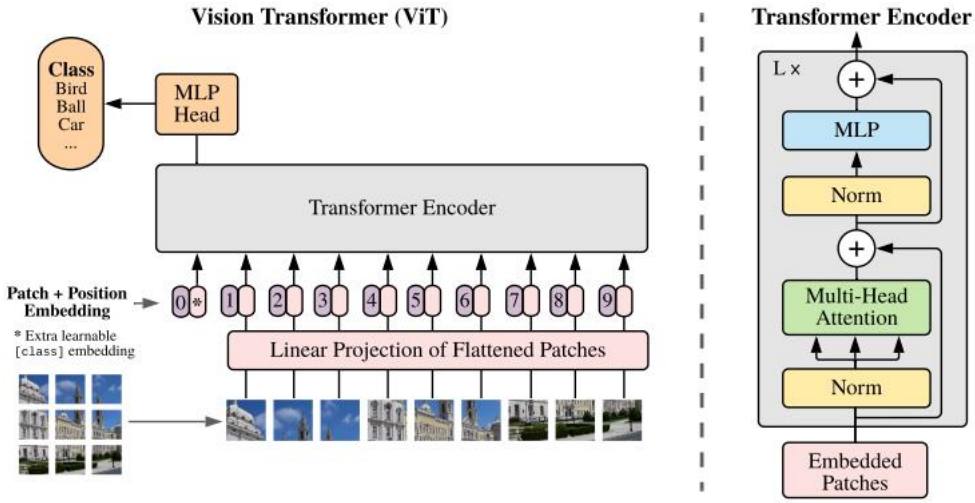


图 2.18 ViT 流程图

根据上述流程图，一个 ViT block 可以分为以下几个步骤：

Patch Embedding：例如，对于一个大小为 224x224 的输入图像，将其分割为固定大小的 patch，每个 patch 的大小为 16x16。这样，每张图像会生成 224x224/16x16=196 个 patch，即输入序列的长度为 196。每个 patch 的维度为 16x16x3=768，线性投射层的维度为 768xN（N=768）。因此，输入通过线性投射层后的维度仍然为 196x768，即共有 196 个 token，每个 token 的维度为 768。此外，还需要加入一个特殊字符 cls，因此最终的维度为 197x768。到目前为止，我们已经通过 patch embedding 将一个视觉问题转化为了一个 seq2seq 问题。

Positional Encoding（标准可学习的一维位置嵌入）：ViT 同样需要加入位置编码。位置编码可以理解为一张表，表共有 N 行，N 的大小与输入序列长度相同。每一行代表一个向量，向量的维度与输入序列嵌入的维度相同（768）。而且位置编码的操作是求和而不是拼接，所以加入位置编码信息后，维度仍然是 197x768。

LN/multi-head attention/LN：LN (Layer Normalization)输出维度仍然是 197x768。在多头自注意力时，先将输入映射到 q、k、v。如果只有一个头，则 qkv 的维度都是 197x768；如果有 12 个头（768/12=64），则 q、k、v 的维度是 197x64，共有 12 组 q、k、v。最后再将 12 组 qkv 的输出拼接起来，输出维度是 197x768，然后再过一层 LN，维度仍然是 197x768。

MLP：将维度放大再缩小回去。197x768 放大为 197x3072，再缩小变为 197x768。一个 block 后维度仍然与输入相同，都是 197x768，因此可以堆叠多个 block。最后会将特殊字符 cls 对应的输出作为 encoder 的最终输出，即为最终的图像特征（另一种做法是不加 cls 字符，对所有 token 的输出求平均）。

文中实验表明相比于 CNN，在经过大量数据进行预训练后，再迁移至其它小数据集上时，ViT 所需的计算资源更少，并且获得了比最先进的 CNN 更好的结果。

2.6.2 Swin Transformer

不同于自然语言处理，视觉任务中图像往往具有局部性，直接应用 Transformer 在高分

辨率图像上会是比较困难的，这会带来平方级的计算复杂度增长，出于这些考虑，Swin Transformer[59]提出了一种新颖的 Transformer 结构，它通过采用滑窗操作和分层设计来解决在图像领域应用中遇到的挑战，它包括不重叠的局部窗口和重叠的交叉窗口，将注意力计算的重心聚焦于这些窗口，既能保留注意力机制，同时也可以实现类似 CNN 卷积的分层结构，节省计算量。

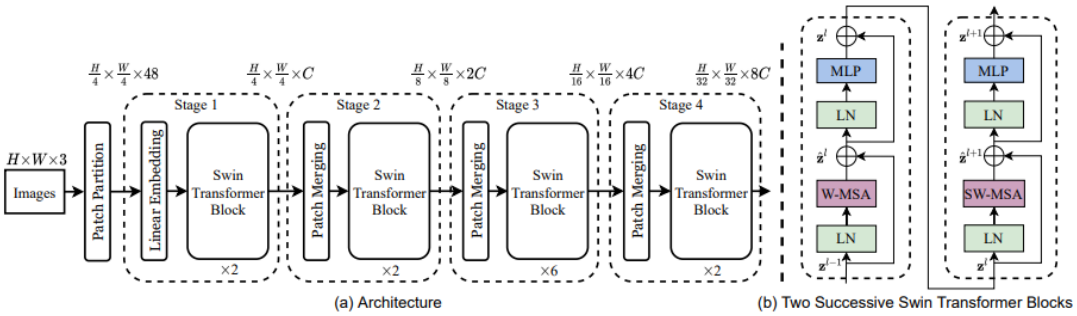


图 2.19 Swin Transformer 结构图

如图 2.19 所示整个模型采用分层设计，共有 4 个阶段。每个阶段都会减小输入特征图的分辨率，逐层扩大感受野，类似于 CNN。在输入开始时，模型执行 Patch Embedding，将图像切成多个局部块并嵌入到 Embedding 中，该操作与前文 ViT 中的做法类似，但是该方法在获得嵌入向量时没有引入分类标记，而是通过最后进行求和平均实现分类功能（类似于卷积的全局平均池化），也没有引入一开始的位置编码，而是在后续引入了相对位置编码用于窗口注意力机制。每个阶段由 Patch Merging 和多个 Block 组成。其中，Patch Merging 模块主要用于在每个阶段开始时降低图像分辨率。

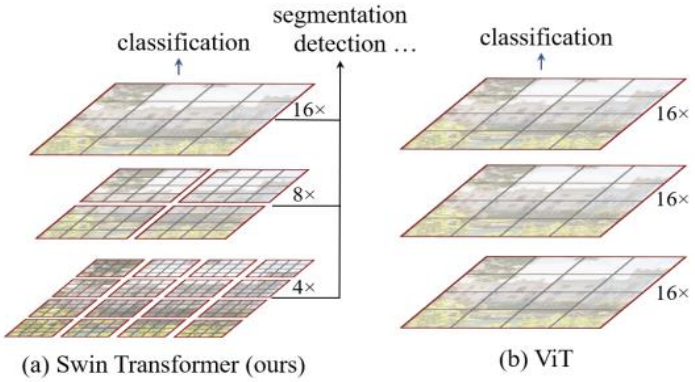


图 2.20 Swin Transformer 分层结构示意图

如图 2.20 所示，Patch Merging 采用了类似 Pixel Shuffle 的操作，将特征图的尺寸缩小四倍，即由$H \times W$变为$H/2 \times W/2$，同时通道数增加至四倍$4C$，再经过一个全连接层将通道数缩小至$2C$，这样，一个类似于 CNN 的 Transformer 骨干网络就已经初具雏形了。

之前的例如 ViT 的方法，其注意力计算的复杂度都与图像大小呈二次关系，Swin Transformer 可以将计算的复杂度降至与图像大小呈线性相关，而这正是得益于窗口注意力机制的引入，这同时也是在进一步对图像结构所具有的局部性与全局注意力之间做出权衡，而且窗口的大小是固定的，也使得前述的分层的 Transformer 骨干网络更好训练。但是只有窗口的话，注意力机制无法捕获到跨窗口的信息，自然地，作者提出了移动窗口的分割方法，

并在两个连续的 swin transformer 基本块中交替使用。具体来说，窗口注意力机制中引入了相对位置编码，即对窗口中每个像素位置互相之间求得相对位置，并在计算 q、k、v 时加上相对位置编码，相当于是将一开始的位置嵌入向量转移到了计算窗口注意力的位置，公式如(2.14)所示。

$$Attention(Q,K,V) = Softmax\left(\frac{QK^T}{\sqrt{d}} + B\right)V \tag{2.14}$$

其中 B 代表相对位置编码。

移动窗口注意力的实现则依赖于作者提出的分割以及计算的方法，如图 2.21 所示。

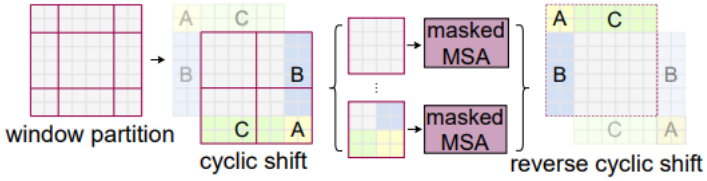


图 2.21 移动窗口注意力分割机制

通过这种移位方式，可以使得跨窗口的像素点之间也能进行交互，相当于增大了感受野，但是同时它也使得窗口的数目变化了，于是作者又对其进行了适当的移位重新进行了分区，并计算了此时不同窗口所得注意力的有效区域，并据此设计了掩码，从而使得移动窗口注意力的计算能够行之有效。

2.6.3 VFIformer

随着 ViT 以及 Swin Transformer 等基于 Transformer 的工作展现出了该结构在视觉任务上的巨大潜力，Tansformer 在视觉领域上也开始爆火，但是由于图像数据本身包含大量冗余信息，所以直接将图像切块作为嵌入向量的方法在许多视觉的下游任务中都无法取得最先进的性能，而在自然语言处理中，词嵌入向量本身拥有很好的性质，它的冗余信息更少，是比较成熟的词特征，我们也就自然而然的会想到先对图像进行特征提取，得到冗余信息更少的的特征作为图像的嵌入向量是否会产生更好的效果，于是大量将卷积神经网络与 Transformer 相结合的工作也如雨后春笋般冒了出来，本文也是基于这一思想针对帧插值任务设计出了新的模型[60]。

本文沿用了基于光流的方法的框架，先通过一个光流估计网络估计出光流，再通过光流以及输入帧即可绘制出对应的中间帧$\widetilde{I_0}$和$\widetilde{I_1}$，而后经过遮挡推理，对中间帧$\widetilde{I_0}$和$\widetilde{I_1}$进行对齐，再经过一步后处理，通过计算残差进一步提高插值帧的生成质量和视觉效果。但该方法利用 Transformer 结构合并了遮挡推理和后处理模块，直接获得了遮挡蒙版H和残差ΔI_t，结合之前获得的中间帧$\widetilde{I_0}$和$\widetilde{I_1}$，即可计算得到最终的插帧结果，计算公式如(2.15)所示。

$$\boldsymbol{I}_t = \boldsymbol{H} \otimes \tilde{\boldsymbol{I}}_0 + (1 - \boldsymbol{H}) \otimes \tilde{\boldsymbol{I}}_1 + \Delta \boldsymbol{I}_t \tag{2.15}$$

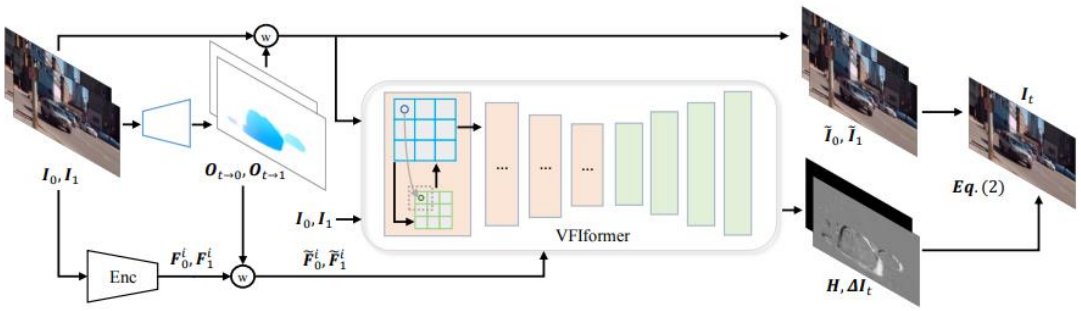


图 2.22 VFIformer 模型结构图

VFIformer 模型结构如图 2.22 所示，文中也用到了多尺度信息，通过一个编码器，获得了四个不同尺度的特征，并沿用 U-net 结构以提取更丰富的语义信息，并减少细节的丢失，将多尺度特征作为编码器不同阶段的输入，从而增强对大动作的捕获能力。该模型设计的 VFIformer 网络中基本块的结构如图 2.23 所示。

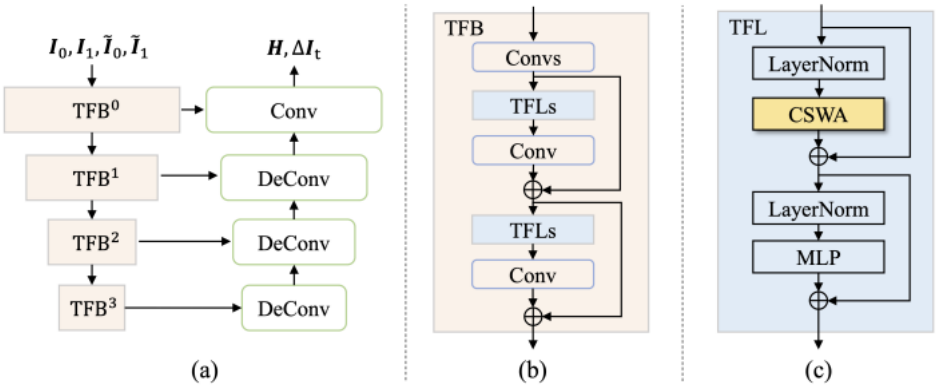


图 2.23 VFIformer 基本块结构

其中 Transformer 部分主要的变化在多尺度窗口注意力模块上，作者设计这个模块正是为了解决窗口无法获得跨窗口信息的问题，不同于 Swin Transformer 提出的滑动窗口注意力，由于引入了多尺度的信息，本文充分利用了多尺度的信息，先对低分辨率的特征进行填充，填充至于高分辨率同等大小，然后再使用相同大小的窗口对不同尺度的特征计算注意力图。有些殊途同归，不同点在于 Swin Transformer 增加了深度，而 VFIformer 增加了宽度。

如图 2.24，多尺度窗口注意力机制由两个支路组成。左边的支路是在同一尺度内进行 Transformer 操作，其中查询键 Q 来自当前尺度的滑窗，而 K 和 V 都来自同一窗口。右边的支路是在不同尺度之间进行 Transformer 操作，其中 Q 仍然来自当前尺度，而 K 和 V 来自下一层尺度的特征。

基于这样的结构，VFIformer 在结构上精简了以往光流框架的模块，将遮挡模块与后处理模块合并，同时也减少了模型的参数，并且提高了插值帧的精度，这也体现了 Transformer 结构在帧插值领域上的潜力。

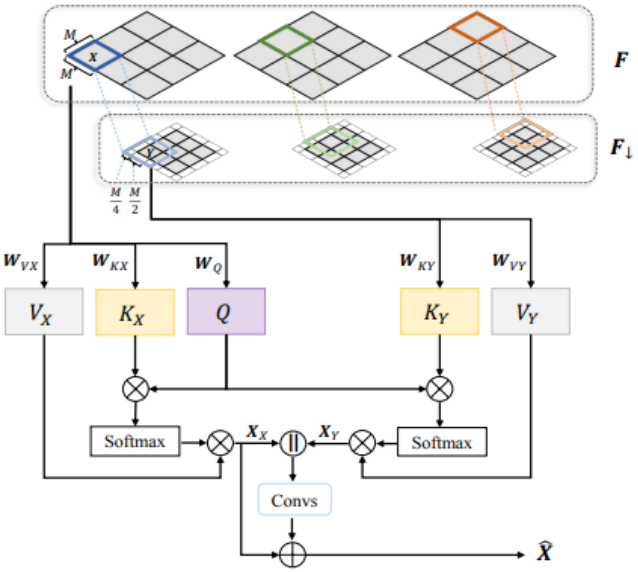


图 2.24 多尺度窗口注意力机制

2.7 本章小结

这一章详细介绍了本文所涉及的相关工作。本文的研究重点是基于深度学习的视频帧插值算法。已有研究表明，一些经典的 CNN 算法和 Transformer 算法在插值任务中表现出色。本章对这些方法进行了详细描述，并全面总结了插值算法从基于光流的对齐绘制，到基于插值内核的卷积合成，再到基于视频特征表示的直接方法，以及最后基于 Transformer 的全局自注意力计算方法的演变过程。这为本文的工作奠定了基础，并为我的方案提供了新思路和启发。

第三章 基于纹理增强的帧插值方法

3.1 引言

帧插值是一种计算机视觉技术，它利用已知的两个或多个帧之间的运动信息来生成新的帧，从而增加视频的帧率。这种技术可以用来提高视频的流畅度和清晰度，也可以用来创造不同的时间效果。帧插值的难点在于如何准确地估计运动信息，并且如何处理运动模糊、遮挡和不连续等问题。帧插值是一种重要的计算机视觉技术，它涉及到多个子领域，如光流估计、图像配准、图像合成等。研究人员正在不断开发新的帧插值方法，并且取得了显著成果。随着技术的不断发展，帧插值将能够为使用者提供更加精彩、更加逼真的视频体验。

基于光流的方法是视频帧插值任务常用的一种方法，它通过计算视频相邻帧之间的双向光流来解决视频帧插值问题，然后用合适的 warping（翘曲）算法来生成输出帧。但是，依靠光流的方法往往不能直接从视频中模拟遮挡物和复杂的非线性运动，并且由于 warping 过程的存在会引入不适合实时部署的问题。更重要的是，此类方法会过于依赖光流估计的准确性，这使得帧插值任务只能沦为光流估计的下游任务，而不能为其它视觉任务提供有效地特征表示。

基于核估计的方法也是一种比较常用的方法，如 Sepconv，是一种用于视频帧插值的自适应核估计方法。该方法通过使用一维卷积核来近似二维卷积核，大大减少了计算量，并且能够一次性生成高分辨率的视频帧。此外，该方法还采用了感知损失函数来进一步提高插值结果的视觉效果。但是为每一个像素都要估计一个卷积核是非常消耗计算资源的，而且为了捕获大的动作，此类方法往往还采用了比较大的卷积核尺寸，这会加剧对显存和计算资源的消耗。后续例如 MEMC-net 的方法提出了将基于光流的方法与基于核估计的方法相结合，通过光流估计出插值帧像素的大概位置，再通过自适应卷积核对近邻像素点进行卷积操作，即可将两种方法的优势相结合，使得每个自适应卷积核的尺寸可以降到4×4，相比于之前的方法，一定程度上对原有方法进行了改善。但同时依赖光流以及计算复杂的问题并没有得到充分地解决。

于是，为了能使得帧插值任务不依赖于光流，并且免去翘曲以及对高分辨率图像每个像素点都要计算一个卷积核等时间复杂度高的操作，一些研究人员开始着眼于能不能用类似自编码器等基于表示的方法实现帧插值任务。例如 CAIN 引入了 Pixel Shuffle 的操作进行下采样和上采样，进行初步的特征重塑，相比于直接卷积，该操作时间复杂度更低，同时在残差块中引入了通道注意力机制，促进网络对运动信息的提取，从而更加高效的实现了帧插值。而 FLAVR 则更进一步的想到 2D 卷积并不能足够好的捕获到视频序列中所蕴含的时序信息，于是引入了 3D 时空卷积，以捕获帧间时序关系，同时也引入了通道注意力的机制，促使网络能更好地学到运动信息。但上述方法依然没有完全解决直接方法所面临的问题，即由于网络缺乏对视频帧序列中所蕴含的复杂的多模态信息的建模能力，所以经过编解码所得的插帧

结果在精度上往往略逊于基于光流的方法。

近年来，基于光流的方法在帧插值领域更为流行，不同于基于直接特征表示的方法通过显示地进行光流估计再进行像素合成，这样的分步做法使得网络参数易于学习，同时由于其基于光流进行推理绘制，也使得此类方法泛化性更强，而直接基于特征表示的方法往往在训练数据集以外的数据集上泛化性不强，考虑到两种方法的特点，本章在基于直接特征表示的方法基础上引入了后处理模块，通过纹理增强的手段，促使网络能学习到更有效的运动信息，同时也增强它的视觉效果。同时考虑到 CAIN 中利用 Pixel Shuffle 进行特征重塑的成功经验，我也引入这一操作来完成最后的上采样和下采样，最后，受一些基于光流的方法的启发，考虑到直接提取特征生成插值帧的方法在结果上往往会存在一些模糊与细节的缺失，本章还引入了一个后处理模块，用来计算纹理残差，以加强插值帧的视觉效果。

现有的基于卷积神经网络的帧插值方法往往会在一些运动边缘出现模糊的情况，以 FLAVR 为例，本章以 DAVIS480p 中舞蹈类动作中的第一帧和第三帧作为输入，得到的输出结果如下：

第一帧　　第三帧　　第二帧（真实值）　　FLAVR 插帧结果

图 3.1 FLAVR 插帧结果

为了应对这样的纹理模糊问题，本章方法引入了金字塔级联网络为插帧结果生成纹理残差，从而增强插帧效果。

3.2 网络结构

本章设计了一个用于视频帧插值的基于纹理增强的模型。其整体结构如图 3.所示。模型的输入包含两帧，$t-1$ 时刻的帧I_{t-1}和$t+1$ 时刻的帧I_{t+1}，模型的输出是这两帧之间的中间插值帧，即t时刻的帧I_t。具体来说，本方法的模型包括两个子模块：生成模块、帧后处理模块。

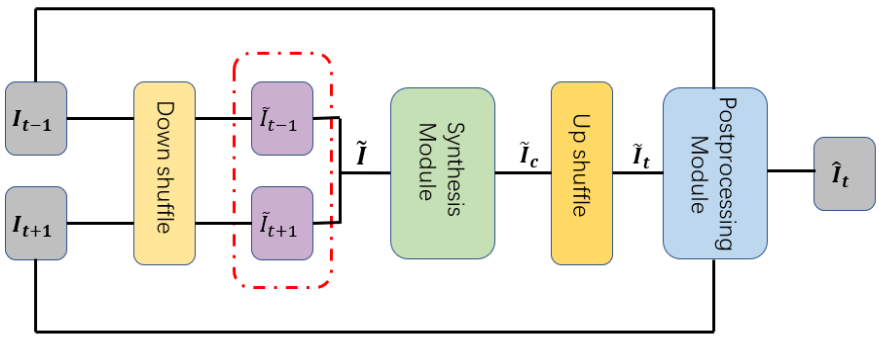


图 3.2 基于纹理补偿的帧插值网络结构

3.2.1 Pixel Shuffle

受到 CAIN 的启发，通过 Pixel Shuffle 对特征进行重塑可以促进深度网络对视频信息的特征抽取，故我也引入了这个模块。Pixel Shuffle 原本是一种上采样方法，可以对缩小后的特征图进行有效的放大。它可以代替插值或反卷积的方法实现上采样，它的主要功能是将低分辨的特征图，通过卷积和多通道间的重组得到高分辨率的特征图。这一方法最初是为了解决图像超分辨率问题而提出的。我这里主要应用了它在多通道间进行分解和重组的思想构建了 Down Shuffle 和 Up Shuffle。具体来说，Down Shuffle 将形状为(C, H, W)的张量中的元素重新排列为形状为$(C \times r^2, H/r, W/r)$的张量，其工作流程如图 2.13 所示，Up Shuffle 反之。我这里将 r 设为 4，即通过 Down Shuffle 后，对一张 RGB 图像，我可以获得一个通道数为$3 \times 4 \times 4 = 48$的特征图，再通过一个 2D 卷积操作我将通道数转换至 64，即可得到$\tilde{I}$。

3.2.2 3D CNN

3D CNN（三维卷积神经网络）是一种用于处理视频和三维数据的深度学习模型。它通过在三维数据上应用卷积运算来捕捉空间和时间上的依赖关系。与传统的二维卷积神经网络不同，3D CNN 可以同时处理数据的深度、高度和宽度三个维度，从而提取出数据中的时空特征。3D CNN 可以有效地处理视频序列、医学图像、三维物体等数据类型，具有广泛的应用前景。3D CNN 的原理和结构与传统的二维卷积神经网络类似，都包括卷积层、池化层和全连接层。不同之处在于，3D CNN 的卷积核是三维的，能够同时处理深度、高度和宽度三个维度。例如，一个大小为$c_i \times c_o \times t \times h \times w$的 3D 卷积核可以对一个大小为$c_i \times T \times H \times W$的输入数据进行滑动窗口运算，得到一个大小为$c_o \times (T - t + 1) \times (H - h + 1) \times (W - w + 1)$的输出数据，其中 t 是时间维度，(h, w)是空间维度。c_i和c_o分别是输入和输出通道的数量。一般来说，3D CNN 采用金字塔式的结构，即随着网络层数的增加，数据在空间和时间上都进行下采样，从而减少计算量和内存消耗。同时，在每一层中，3D CNN 可以利用不同尺度的时空特征来进行特征提取和融合。例如，在 FLAVR 中，作者提出了一种基于 3D U-Net 的特征提取器，它可以从多帧视频中提取时空上下文特征，再结合特征通道门控模块引入注意力机制，用于视频帧插值任务。在每一层中，3D CNN 可以使用不同的方式来进行卷积运算。

3D CNN 已经被广泛地应用于各种带有三维特征的数据场景当中，而视频帧插值由于其在时间上的相关性，正好可以充分地利用到 3D CNN 的多维度特征提取能力。在帧插值发展的过程中，许多应用了 3D CNN 的方法被提了出来，像 FLAVR 就提出了一种新颖且高效的视频帧插值方法，使用一个基于 3D U-Net 的特征提取器从四帧输入中生成任意位置的中间帧，这种在多帧上的扩展性是以前的工作所不具备的，该方法通过使用多尺度特征上采样和特征融合来从编码器捕获的深层潜在表示中构建输出帧。还有研究者[61]提出了一种基于变形卷积的视频帧插值方法（VFI-Net），它使用粗到细的 3D CNN 来增强多流预测。该方法首先使用 3D CNN 在多个尺度上提取时空特征，并以粗到细的方式使用这些特征估计多流，然

后，使用估计的多流对原始输入帧和上下文图进行扭曲，扭曲结果由像素合成网络融合以产生最终输出。该方法通过使用可变形卷积来适应不同运动模式，并且使用了一个循环一致性损失函数来增强插值结果。

3.2.2 生成模块

考虑到连续的视频帧在时间上的依赖关系，我选择使用 3D 卷积来构建我的生成模块，为了减少细节的丢失，采用 U-net 架构，其中编码器部分由四个阶段构成，每个阶段包括两个残差块和一个基于通道注意力的通道门控模块，解码器则由 3D 转置卷积和通道门控模块作为基本块，最后再经过一个 2D 卷积，将输出的通道数减小至 48 即得到了$\tilde{I}_c$，等于输入$\tilde{I}$通道数的一半。如图 3.所示，在每个 3D 卷积块和 3D 转置卷积块之后都连着一个基于通道注意力的通道门控模块，它包括一个 3D 全局平均池化层，一个全连接层组成，对注意力图进行加权求和后再与之前的特征图相乘获得注意力加权后的特征图。

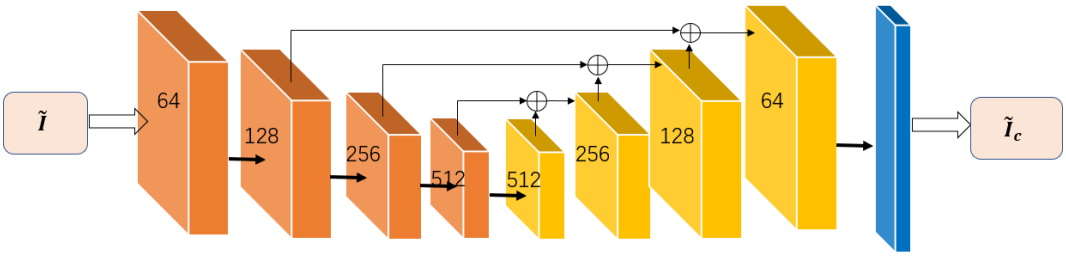


图 3.3 生成模块网络结构图

3.2.3 金字塔级联网络

金字塔级联网络是一种用于多人姿态估计的新型网络结构。它能够在复杂的场景中，准确地检测出人体的各个关键点，如头部、手臂、腿部等。金字塔级联网络由两个阶段组成：GlobalNet 和 RefineNet。GlobalNet 是一个特征金字塔网络，它能够在不同的尺度上提取人体的特征，并定位出“简单”的关键点，如眼睛和手。但是，GlobalNet 可能无法处理一些“困难”的关键点，如被遮挡或不可见的关键点。RefineNet 则是一个细化网络，它能够利用 GlobalNet 中所有层次的特征表示，并结合在线硬关键点挖掘损失，来显式地优化“困难”的关键点的位置。

金字塔级联网络在多人姿态估计方面取得了巨大的成功，并且在解决遮挡关键点、不可见关键点和复杂背景等难题方面表现出色。在 COCO 关键点基准测试中取得了最先进的结果，在 COCO test-dev 数据集上平均精度为 73.0，在 COCO test-challenge 数据集上平均精度为 72.1。

GlobalNet 的结构如图 3.所示。它使用 ResNet 作为骨干网络，将不同卷积层$conv1\sim5$中的最后一个残差块获得的特征图分别表示为 C1、C2、…、C5，这些特征层具有不同的空间分辨率和语义信息。为了保持高空间分辨率和高语义信息，GlobalNet 采用了一个 U 形结构，

将浅层特征和深层特征进行融合。具体来说，在上采样过程中，GlobalNet 在每个元素求和操作之前应用 1×1 卷积核，以减少通道数目，并在每个特征层上应用 3×3 卷积滤波器来生成关键点的热力图。GlobalNet 的输出是一个特征金字塔，其中每个级别都包含一个关键点热力图。

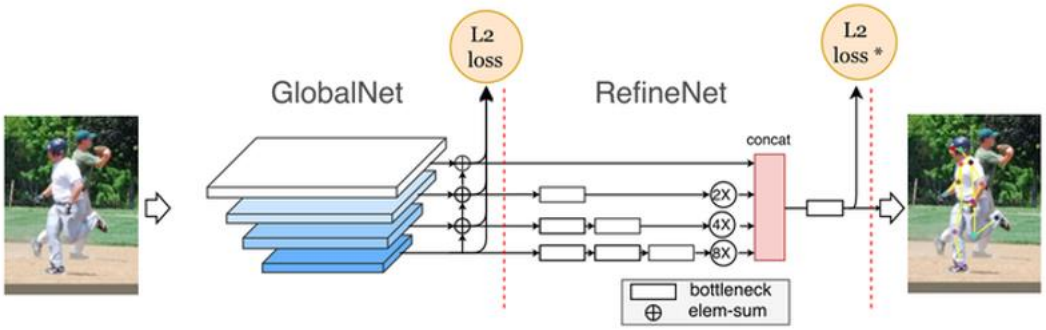


图 3.4 金字塔级联网络结构图

RefineNet 的结构可参考图 3.。它是一个细化网络，它能够利用 GlobalNet 中所有层次的特征表示，并结合困难关键点挖掘损失，来显式地优化“困难”的关键点的位置。RefineNet 的输入是 GlobalNet 的输出，即一个特征金字塔。为了提高效率并保持信息传输的完整性，RefineNet 跨不同级别传输信息，并最终通过上采样和连接来集成不同级别的信息。与堆叠沙漏网络等细化策略不同，RefineNet 连接所有金字塔特征，而不是仅使用沙漏模块末端的上采样特征。RefineNet 的输出是一个细化后的关键点热力图。

3.2.4 后处理模块

受 Super Slomo、MEMC-net、DAIN 等基于流的方法的启发，针对帧插值任务，为了增强插值帧的视觉效果，在经过生成模块以及 Up Shuffle 获得第一阶段的插帧结果$\tilde{I}_t$后,将原始输入帧I_{t-1}和I_{t+1}分别送入两个金字塔级联网络(PCN, Pyramid Cascade Network)，金字塔级联网络可以有效地获取关键点信息，分别送入两个结构相同的金字塔级联网络可以实现初步的对齐，再将两个初步对齐后的特征图与$\tilde{I}_t$一起送入一个残差生成器当中，残差生成器由三个残差块组成，最后将得到的残差与$\tilde{I}_t$相加即可获得最终的插帧结果$\hat{I}_t$。后处理模块流程如图 3.所示。

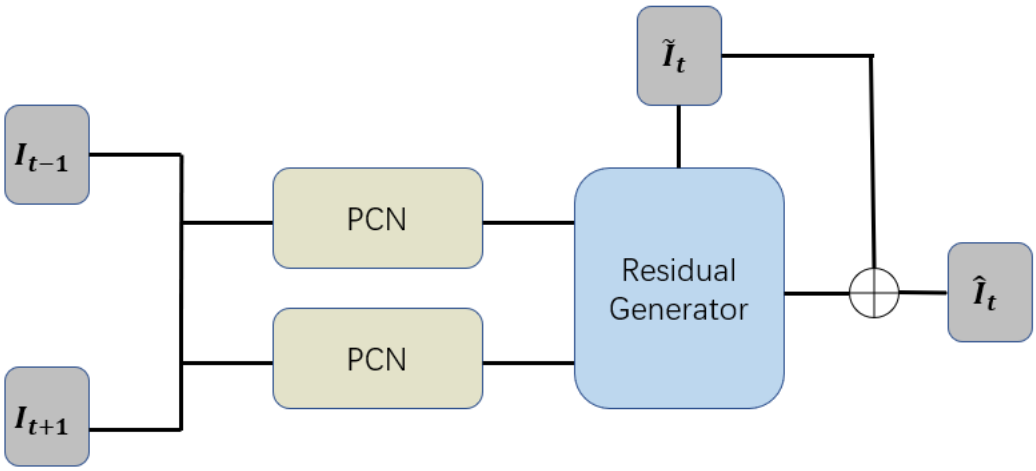


图 3.5 后处理模块网络结构图

3.3 实验细节及参数设置

3.3.1 实验细节

本章模型的损失函数包括两个部分：一部分是用来计算第一阶段生成结果$\tilde{I}_t$与真实帧I_{gt}之间的像素级损失，记为表示损失L_p，另一部分是用来计算经过纹理增强之后的插值帧$\hat{I}_t$与真实帧I_{gt}之间的像素级损失，记为纹理损失L_t。最终训练所用的损失函数可以表示为：

$$L_{total} = \rho(\tilde{I}_t - I_{gt}) + \sigma * \rho(\hat{I}_t - I_{gt}) \tag{3.1}$$

其中，σ取值为 0.2，$\rho(\cdot)$代表求一范数的函数，如公式(3.2)所示。

$$\rho(\hat{I}_t - I_{gt}) = || \hat{I}_t - I_{gt} ||_1 \tag{3.2}$$

之所以采用了一范数，是因为在帧插值任务当中，二范数往往会导致模糊的结果。本章方法在训练过程中采用的是 Adam 优化器，其中β_1和β_2分别取 0.9 和 0.999，学习率设置为 0.0002，此外，本章的训练数据集使用的是 Viemo90k 数据集，并将其分为训练集和验证集。其中，训练集包含 45000 个三元组，验证集包含 6313 个三元组，每帧的分辨率为 448 × 256。三元组的中间帧作为真实值，其余两帧作为输入数据。为了增强数据，我还将输入序列的时间顺序颠倒。在训练过程中，批量大小设置为 4，并在 GTX TITAN X 上进行实验。经过大约 200 个 epoch 的训练后，训练过程收敛。

3.3.2 数据集

此外，我还在以下数据集上对该方法进行了测试，以下是数据集的详细介绍。

1. Vimeo-90K：

Vimeo-90K 是一个用于低层次视频处理的大规模高质量视频数据集，可以用于视频超分辨率的问题。数据集由约 90000 个视频片段组成，每个片段包含 3 帧，分辨率为 448x256。数据集包含了从 vimeo.com 下载的约 90000 个视频片段，涵盖了多种场景和动作。该数据集由 Xue 等人[62]提出，可以用于解决四个视频处理任务：视频帧插值、视频去噪、视频去块和视频超分辨率。该数据集可以在 https://toflow.csail.mit.edu/网站上获取。

这个数据集是为了解决现有的视频超分辨率数据集缺乏多样性和真实性的问题而提出的。Xue 等人同时还一并提出了一个基于深度学习的视频超分辨率方法，称为深度视频超分辨率网络。该方法利用了视频中的时间一致性和子像素移动信息，取得了当时最好的效果。Vimeo-90K 数据集已经被广泛地用于评估和比较不同的视频超分辨率方法，例如 TOF2、RBPN3、EDVR 等。这些课题都在 Vimeo-90K 数据集上取得了不同程度的改进，但现有方法仍然面临一些挑战，例如处理快速运动、复杂场景和低质量输入等。

2. UCF101：

UCF101[63]是一个由 YouTube 收集的真实动作视频的动作识别数据集，是一个用于视频动作识别的数据集。数据集由 101 种动作类别组成，每种类别包含 100 多个视频片段，总共有 13320 个视频片段，时长约 27 小时。该数据集由 Soomro 等人提出，可以用于评估和比较不同的动作识别方法。这个数据集可以访问 https://www.crcv.ucf.edu/data/UCF101.php 网页进行下载，是当下最流行的公开数据集之一。

UCF101 数据集是 UCF50 数据集的扩展，其提出初衷是为了提供更加多样化和更具挑战性的动作识别任务。文章中一并提出了一个基于词袋模型的动作识别基准方法，准确率为 44.5%。UCF101 数据集已经成为动作识别领域最流行和最被引用的数据集之一，吸引了许多研究者提出了各种基于深度学习或其他技术的动作识别方法，例如 Two-stream、C3D、TSN、I3D 等。很多方法借助 UCF101 数据集中的样本，取得了显著的提升，当然，也有一些尚未攻克的难题，例如对复杂背景、相似动作和多人场景的处理等。

3. DAVIS480p：

DAVIS480p[64]是一个用于视频目标分割的数据集，由 50 个高质量视频序列组成，每个序列包含 30 到 100 帧，分辨率为 480p。这个数据集是 DAVIS2016 数据集的扩展，增加了更多的目标类别和更难的场景，并覆盖了许多常见的视频对象分割挑战，如遮挡、运动模糊、外观变化和多个对象。该数据集由 Perazzi 等人提出，可以用于评估和比较不同的视频对象分割方法。该数据集提供了 50 组分辨率为 854 × 480 的三帧连续帧和感兴趣对象的像素级注释。该数据集下载地址为：https://davischallenge.org/davis2017/code.html。

DAVIS480p 数据集在 2017 年由 Pont-Tuset 等人发表在 CVPR 上，并一起提出了一个基于区域和边界信息的视频目标分割评估指标，称为 Jaccard 稳定性。该指标综合了目标区域和轮廓之间的一致性和稳定性，并与人类主观评价具有高度相关性。自首次公开以来，DAVIS480p 数据集已经被广泛地用于评估和比较不同的视频目标分割方法，例如 OSVOS、OnAVOS、PReMVOS 等，它们全部都在 DAVIS480p 数据集上取得了显著的结果，但仍然存在一些限制，例如对快速变化过程、遮挡难题和目标消失问题的处理等。

3.3.3 评价指标

1. PSNR（Peak Signal-to-Noise Ratio，峰值信噪比）

PSNR 是一种客观的图像质量评估指标，它反映了原始图像与失真图像之间的相似度。

它基于这样一个假设，即图像中的信号与噪声是可分离的，而且噪声是由于图像或视频的压缩、传输或处理而引入的。因此，PSNR 可以用来衡量压缩或处理后的图像与原始图像之间的失真程度。PSNR 的定义是基于均方误差（MSE），即原始图像与失真图像之间每个像素值的差的平方的平均值。MSE 越小，说明两幅图像越相似，PSNR 越高，说明图像质量越好。PSNR 通常用分贝（dB）表示，以便处理图像的动态范围较大的情况。PSNR 的计算公式如下：

$$PSNR = 10log_{10}\left(\frac{MAX^2}{MSE}\right) \tag{3.3}$$

其中，MAX 是图像可能的最大像素值，例如对于 8 位灰度图像，MAX 为 255。MSE 是原始图像与失真图像之间的均方误差。

PSNR 虽然是一种简单而快速的图像质量评估方法，但它也有一些局限性。首先，它不能很好地反映人眼对图像质量的主观感受，因为人眼对不同区域和不同类型的失真有不同的敏感度。其次，它不能区分不同来源和特征的噪声，例如块效应、环形效应或模糊等。因此，PSNR 并不总是与主观评价一致，有时需要结合其他更复杂或更感知相关的图像质量评估方法。

2. SSIM（Structural SIMilarity, 结构相似度）

SSIM 是一种用于评估图像质量的方法，它通过测量两个图像之间的相似性来预测感知质量。SSIM 是一种完全参考度量，这意味着它需要一个未压缩或无失真的图像作为参考。SSIM 是一种基于感知的模型，它将图像退化视为结构信息的感知变化，同时还包括重要的感知现象，如亮度掩蔽和对比度掩蔽项。与其他技术（如 MSE 或 PSNR）不同，SSIM 不是估计绝对误差，而是考虑了像素之间的相互依赖性。这些依赖关系携带了有关视觉场景中物体结构的重要信息。亮度掩蔽和对比度掩蔽分别描述了图像失真在明亮区域和纹理区域内不太可见的现象。

SSIM 指数由三个部分组成：亮度、对比度和结构部分。这些部分分别评估图像亮度、对比度和结构变化的视觉影响。SSIM 指数在局部计算，通常使用一个移动窗口在整个图像上移动，每一步都计算一个局部 SSIM 得分。整个图像的最终得分是局部得分的算术平均值。总而言之，SSIM 通过测量两个图像之间的相似性来预测感知质量，可以用于评估图像质量。它考虑了像素之间的相互依赖性，并包括重要的感知现象，如亮度掩蔽和对比度掩蔽项。SSIM 指数由三个部分组成：亮度、对比度和结构部分，并在局部完成计算。

3.4 实验结果

表 3.1 展示了本章方法与一系列最先进的方法在 Vimeo90K、UCF101、Davis480p 数据集上的定量比较，其中包括了基于光流的方法如 SuperSlomo，基于核估计的方法如 SepConv，基于光流与核估计结合的方法如 DAIN、MEMC，基于特征表示的方法 CAIN、FLAVR，基

于纹理流的方法 FeFlow。通过 PSNR 和 SSIM 两个指标对这些方法进行了评估，我发现在大部分数据集本章所提出的方法都能达到更好的效果，总体上来说，该方法是最好的，而且在模型结构上也是相对简单的。表中标记为红色的数值代表了在对应数据集及评价指标下最佳的图像插帧结果，标记为蓝色的数值代表了在对应数据集及评价指标下次优佳的图像插帧结果，结果显示在这三个数据集上，除 UCF101 上的 PSNR 取得最优和次优以外，本章方法均取得了最高的 PSNR 和 SSIM 值。

表 3.1 在 Vimeo90K、UCF101、UCF101 数据集上的定量比较

方法	Vimeo90K		UCF101		UCF101	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
DAIN	34.04	0.958	35.26	0.963	27.31	0.870
CAIN	33.93	0.964	32.28	0.965	26.46	0.856
SuperSlomo	32.90	0.957	32.33	0.960	25.65	0.857
FLAVR	36.19	0.975	33.29	0.971	27.41	0.874
FeFlow	35.28	0.976	32.66	0.955	27.11	0.857
SepConv	33.60	0.944	31.97	0.943	26.21	0.857
MEMC	34.20	0.959	35.16	0.963	27.27	0.865
本章方法	36.32	0.979	34.21	0.976	27.60	0.877

为了进一步验证所提出方法的有效性，我进行了充分地消融实验对本文的网络结构进行分析。在整体网络架构上，分别移除了 Pixel Shuffle 模块（用 2D CNN 进行替换）和后处理模块。用完全一样的设置对重组后的模块进行训练，并在多个数据集上进行了测试。定量测试的结果如表 3.2 所示。

表 3.2 在 Vimeo90K、UCF101、UCF101 数据集上的消融研究

方法	Vimeo90K		UCF101		UCF101	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
G	36.20	0.975	33.40	0.971	27.45	0.874
G+TC	36.29	0.975	34.08	0.975	27.58	0.876
G+PS	36.23	0.977	33.60	0.972	27.47	0.876
G+TC+PS	36.32	0.979	34.21	0.976	27.60	0.877

其中 G 表示生成模块，PS 代表 Pixel Shuffle 模块，TC 代表后处理模块，观察表 3.2 第一行与第三行结果的差距，可以看出仅有生成模块与 Pixel Shuffle 操作的效果提升十分有限，但是观察表中第二行可以发现，在加入了纹理增强模块后，该模型在 Vimeo90K、UCF101、UCF101 数据集上分别提升了 0.09db、0.68db、0.13db，可以看出，纹理补偿模块对于模型性能的整体增长起到了更好的效果，而 Pixel Shuffle 对于增强图像生成质量也起到了一定作用，

PSNR 在三个数据集均获得了增长。

3.5 实验可视化分析

本节针对一些困难情境下的帧插值任务进行了结果的可视化及分析。

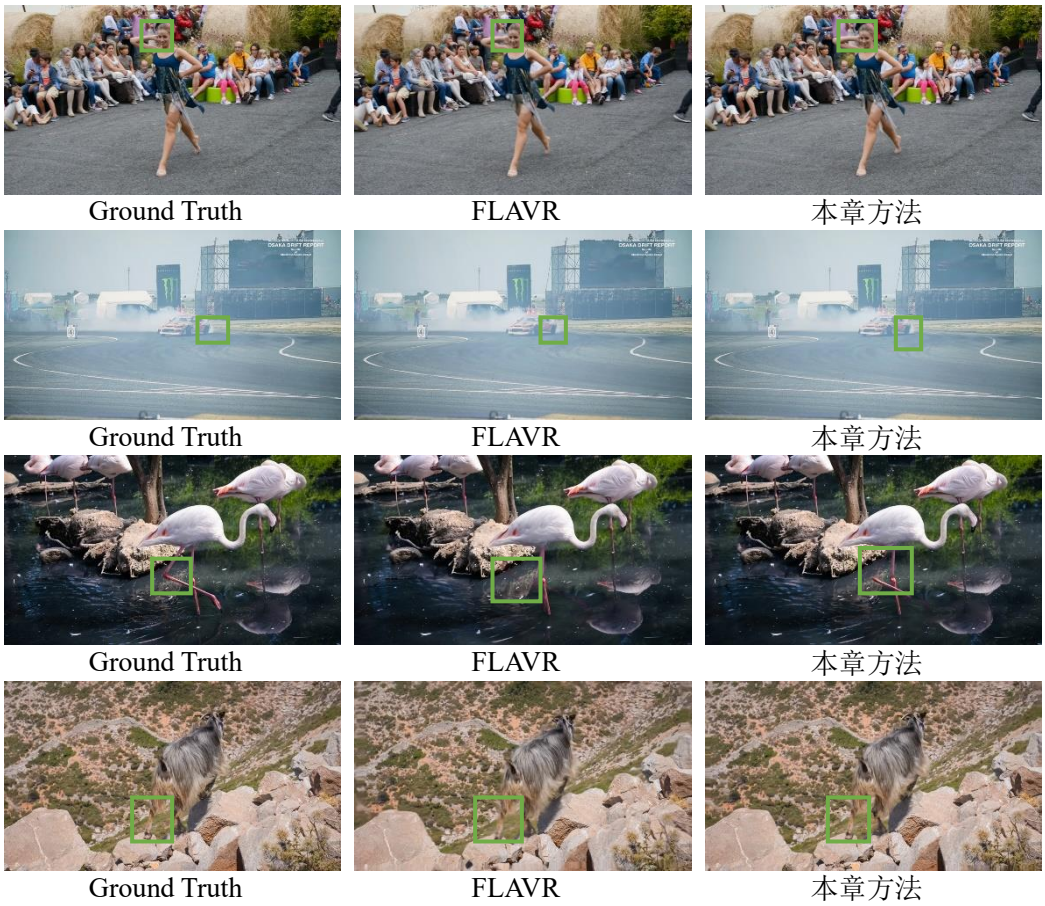


图 3.6 第三章方法可视化比较

如图 3.6 所示，展示了在一些大运动的情境下 FLAVR 方法与本章方法的比较，绿框标注了主要的运动部位，从图中可以看出本章的方法在运动边缘处的纹理更加清晰，展现了本章所提出的纹理增强方法的有效性，可以有效去除运动模糊，获得了更加清晰地插帧结果。

同时，为了验证本章方法所提出的纹理增强模块的有效性，本文还对纹理增强模块进行了消融实验，消融实验的结果如下，其中 Backbone 代表本章方法的骨干网络，T 代表纹理补偿模块。

图 3.7 第三章方法消融实验可视化

如图 3.7 所示，在加入了纹理补偿模块后，插值帧在不同场景下纹理都更加清晰，尤其是对一些运动模糊的区域，可视化结果表明本章加入的纹理增强模块是有效的。

3.6 本章小结

本章提出了一种新的帧插值方法，由于卷积神经网络受限于模型容量，容易出现过拟合的情况，在原有的 3D 卷积神经网络的基础上，考虑到在转置卷积上采样过程中容易出现伪影的问题，选择 Pixel Shuffle 进行上采样，同时为了使得输入的特征图保留更多的信息，通过 Down Shuffle 对输入进行处理，从而帮助模型通过通道注意力捕获更多运动信息，最后借鉴基于光流的方法的处理过程，在模型最后引入了后处理模块，对插帧结果进行一些细节上的补充，从而获得更好的视觉效果。本章我还在三个数据集上进行了测试，结果显示，本章所提出的方法表现出了较强的竞争力。通过消融实验，也充分验证了所提出模块的有效性。

第四章 基于 Transformer 的帧插值方法

4.1 引言

视频帧插值是计算机视觉领域的一个重要分支，它在许多领域都有广泛的应用，例如视频摘要、处理和索引、电影制作、体育直播、视频压缩和传输等，主要是用来提高视频的质量、表现力、可读性和观赏效果。它的目的是通过计算连续输入帧之间不存在的帧来提高视频序列的帧率，从而使视频更加流畅。

过去，大多数使用深度学习进行视频帧插值任务的研究都采用了卷积神经网络。例如，基于光流的帧插值方法利用参考帧相对于目标帧的运动信息进行光流估计而后进行插值来计算目标像素。基于核估计的方法则将帧插值任务视为两个参考帧中相应图像块的卷积，并使用深度卷积神经网络来估计空间自适应卷积核。基于特征表示的直接插帧方法则通过引入通道注意力机制实现对运动信息的隐式估计并融合在特征表示中，这一系列的方法虽然思路不同，但都是通过卷积神经网络作为模型的基本结构。

最近，人们发现在计算机视觉领域中，原本用于自然语言处理（NLP）领域的 Transformer 也是非常有效的。ViT(Vision Transformer)中的实验结果表明，在大规模数据集下，直接将 Transformer 应用到图像块序列上就可以很好地完成图像分类任务，甚至表现出了比卷积神经网络更具有竞争力的效果。于是，后来大量的研究者们开始关注于提高 Transformer 的计算效率，使得 Transformer 可以更好地应用于像素点稠密的图像数据。其中 Swin Transformer 提出了基于窗口的注意力模块和基于移动窗口的模块来缓解 Transformer 计算量大的问题，并且设计了分层的结构使得 Transformer 可以直接对以前的基于卷积神经网络的骨干网络进行替换，并引入了一种移位窗口注意力模块，与基于窗口的注意力模块在连续块中的两个分区之间交替使用，以增强跨窗口的注意力流动。

当下虽然也有一些基于 Transformer 的方法应用于帧插值任务当中，但是它们往往都遵循传统的运动估计和运动补偿的框架，先用光流估计网络进行运动估计，再利用 Transformer 设计运动补偿模块，这样的做法会导致插帧结果十分依赖于光流网络的精度。考虑到 Transformer 在特征提取中展现出的强大能力，可以同时对空域和时域进行建模，因此，本章结合 Swin Transformer 所提出的窗口注意力机制以及移动窗口注意力机制，提出了一个基于 Transformer 和 3D 转置卷积的网络框架，它先通过一个 Transformer 编码器获取时空特征，再通过 3D 转置卷积进行解码得到粗糙的插帧结果，最后再通过我设计的后处理模块，对插帧结果进行纹理增强。

当前主流的基于光流的帧插值算法在处理大运动、复杂运动、有遮挡情境下的效果并不好，以 DAIN 为例，在处理大运动的情况下其效果非常差，本文选取了 DAVIS480p 中的“狗”图片作为示例，选取第一帧和第七帧作为输入 DAIN 模型当中。

第一帧　　第七帧　　第四帧　　DAIN 插帧结果

图 4.1 基于 Transformer 特征抽取的帧插值方法网络结构

从图中绿框可以看出此类基于光流的算法在大运动的情况下会出现伪影和模糊，分析认为有两点原因影响了插帧结果，一是因为光流法有严格的假设前提，要满足小运动且亮度恒定，二是卷积神经网络有局部性的归纳偏置，无法构建长距离的像素点之间的依赖关系，基于这样的分析，本章提出用 Transformer 来实现对运动特征的提取，以获取像素点间的长距离依赖关系，从而改善插帧效果。

4.2 网络结构

本章方法包含的网络结构主要包括三个部分，Transformer 特征抽取模块，基于 3D 转置卷积的特征聚合模块，以及后处理模块。完整的网络结构图如图 4.所示。

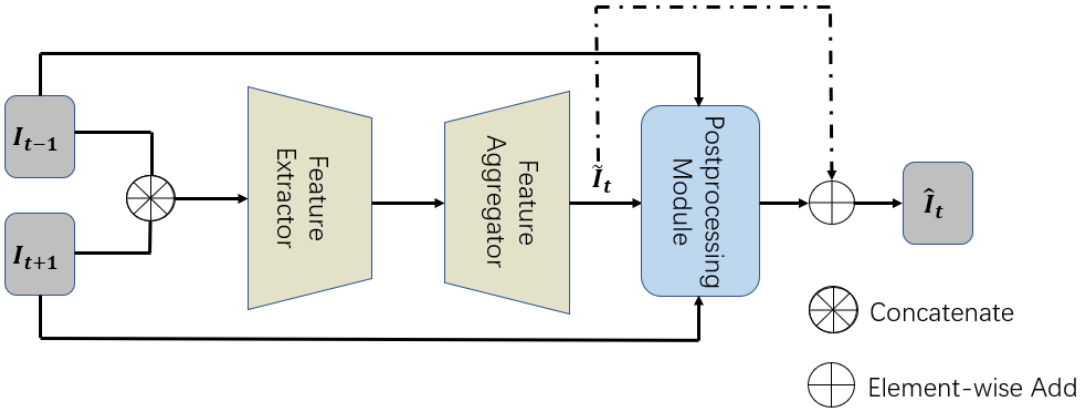


图 4.2 基于 Transformer 特征抽取的帧插值方法网络结构

接下来介绍一下这三个模块。

4.2.1 Transformer

Transformer 是一种基于深度学习的模型，它最初被设计用于处理自然语言处理中的序列到序列问题，例如机器翻译。它的特点是完全使用注意力机制来建模序列中的元素之间的关系，而不依赖于循环神经网络（RNN）或长短期记忆网络（LSTM）这样的递归结构。Transformer 模型由编码器和解码器两部分构成，每个部分都包含多个重复的层，每个层都有两个或三个子层。Transformer 模型在许多 NLP 任务上都取得了显著的效果，并且也被扩展到了图像生成、图像识别和目标检测等领域。

Transformer 模型的核心是注意力机制，它是一种计算输入序列中每个元素与其他元素之间相关性的方法。注意力机制可以让模型在生成输出时考虑到输入序列中所有位置的信息，从而更好地捕获长距离依赖关系。Transformer 模型使用了多头注意力（Multi-Head

Attention）的方式，即将输入序列投影到多个不同的子空间中，并在每个子空间中分别计算注意力权重，然后将所有子空间中的结果拼接起来。这样可以让模型同时关注不同方面的信息，并增加模型的表达能力。

模型的编码器由 N 个相同的层组成，每个层有两个子层：一个多头自注意力（Multi-Head Self-Attention）子层和一个前馈神经网络（Feed-Forward Neural Network）子层。多头自注意力子层用于计算输入序列中每个元素与其他元素之间的相关性，并生成新的表示。前馈神经网络子层则用于对每个元素进行非线性变换，并增加模型的复杂度。在每个子层之后，还有一个残差连接（Residual Connection）和一个层归一化（Layer Normalization）操作，以提高模型的稳定性和收敛速度。编码器的作用是将输入序列编码成一个固定长度的向量，这个向量包含了输入序列中所有元素的信息。

同理，Transformer 模型的解码器也由 N 个相同的层组成，每个层有三个子层：一个多头自注意力子层，一个多头编码器-解码器注意力（Multi-Head Encoder-Decoder Attention）子层和一个前馈神经网络子层。多头自注意力子层用于计算输出序列中每个元素与其他元素之间的相关性，并生成新的表示。多头编码器-解码器注意力子层则用于计算输出序列中每个元素与编码器输出向量之间的相关性，并生成新的表示。前馈神经网络子层与编码器中相同，用于对每个元素进行非线性变换，在每个子层之后，也有一个残差连接和一个层归一化操作。解码器的作用是根据编码器输出向量和已生成的输出序列来生成新的输出元素。

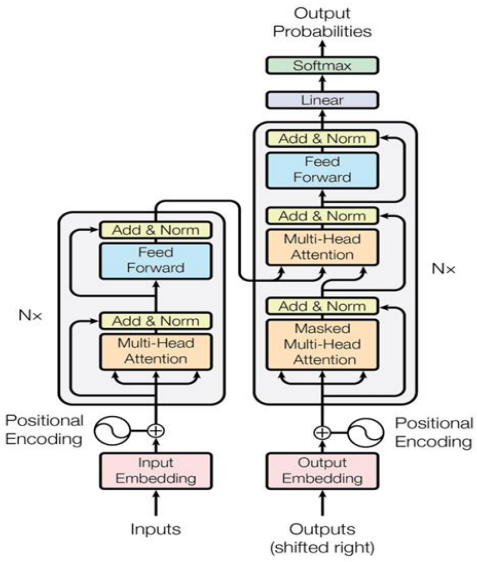


图 4.3 Transformer 结构图

4.2.2 基于 Transformer 的特征抽取模块

考虑到卷积神经网络具有局部性以及平移不变性的归纳偏置，从而无法很好的学习时序上的运动信息，引入了 Transformer 作为特征抽取模块，同时为了减少计算复杂度，沿用了 Swin Transformer 的架构，其中的窗口注意力机制以及移动窗口注意力机制可以在减少计算复杂度的同时，增大感受野，并且为了进一步捕获序列信息，我们在每个 swin transformer 块

中都加入了一个通道注意力块，从而更好地建模运动信息，网络架构如图 4.所示。

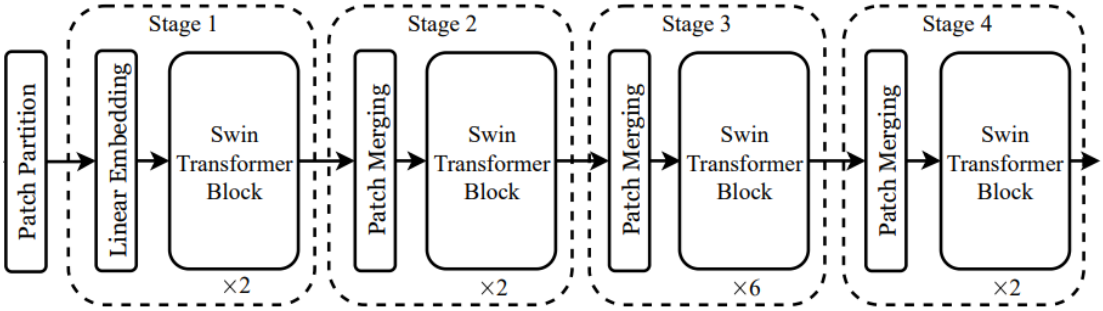


图 4.4 Feature Extractor 网络结构图

需要注意的是我将输入，$t-1$ 时刻的帧I_{t-1}和$t+1$ 时刻的帧I_{t+1}拼接在一起得到了一个尺寸为$2H \times 2W \times 3$的输入，再经过4×4的 Patch 分割后，得到了$2H/4 \times 2W/4 \times 48$的特征图，经过四个阶段的特征抽取后，得到$2H/32 \times 2W/32 \times 512$的特征图，图中的 Patch Merging 会合并相邻 Patch，从而减少特征图的尺寸，增加通道数。

4.2.3 基于 3D 转置卷积的特征聚合模块

当特征抽取模块抽取好特征后，就要选择如何对抽取好的特征进行上采样，考虑到 Transformer 所提取出来的特征在通道之间的相关性，我选择 3D 转置卷积进行上采样操作，这样可以更好地恢复出图像细节，由于转置卷积可能会存在棋盘状的伪影，所以我在该模块最后一层使用的是一个 3D 卷积，去改善出现棋盘状伪影的情况。

4.2.4 后处理模块

受到基于流的方法的启发，针对帧插值任务，为了增强插值帧的视觉效果，我加入了一个后处理模块，通过对齐输入帧与插值帧，计算出纹理残差，对特征抽取过程中可能缺失的一些细节进行增强，在前一章工作的基础上，为了进一步减少计算量，我对后处理模块进行了简化。具体而言，在经过特征抽取以及特征聚合获得粗糙的插帧结果$\tilde{I}_t$后,将原始输入帧I_{t-1}和I_{t+1}和$\tilde{I}_t$一起送入金字塔级联网络中，实现对齐，再将对齐后得到的特征图送入一个 2D 卷积神经网络当中生成残差，最后将得到的残差与$\tilde{I}_t$相加即可获得最终的插帧结果$\hat{I}_t$。其流程在图 4.右侧可以清晰看到，其中后处理模块有一个金字塔级联网络和三个 2D 卷积层构成。

4.3 复杂度分析

相比于传统的卷积方法和 Transformer 方法，本章方法所用的网络结构在时间复杂度上更优，在本节中将详细分析本章方法的时间复杂度，并且与卷积方法以及 Transformer 方法进行比较。

首先对于传统的卷积方法，一层卷积层，假设输入特征的尺寸为$H \times W \times C$，输入输出通道数相同均为 C，则该层卷积操作的计算过程可以描述为对每个输入通道 C 中全部的$H \times W$个像素点都通过一个$K \times K$的卷积核得到一个输出通道的一个值，输出通道数为 C，

所以一共进行了$H \times W \times C^2 \times K^2$次计算，最终的计算复杂度为$O(HWC^2K^2)$。

对于多头自注意力方法，主要的计算量集中在对于 Q、K、V 三个矩阵的计算，同样假设输入为$H \times W \times C$，得到 Q、K、V 所需要的计算次数均为$H \times W \times C^2$，根据 Q、K 计算注意力图需要对它们做矩阵乘法，矩阵乘法的复杂度为$(H \times W)^2 \times C$，并通过注意力图进行加权运算，即用注意力图与矩阵 V 做矩阵乘法，该矩阵乘法的复杂度同样为$(H \times W)^2 \times C$，考虑到是多头注意力机制，最后需要对不同的特征图进行拼接融合，这个映射层是一个线性层，计算复杂度与计算 Q、K、V 时一致，也为$H \times W \times C^2$，所以 Transformer 方法的计算复杂度为$O(4HWC^2 + 2H^2W^2C)$。

本章中所用到的窗口注意力方法包括三个注意力模块，一个是窗口注意力模块，一个是移动窗口注意力模块，还有一个通道注意力模块。通道注意力模块的时间复杂度为$O(C^2)$，相比于窗口注意力机制，复杂度量级很小，为便于观察，将其省略。移动窗口注意力模块经过分割补全后计算过程与窗口注意力模块是一样的，它们的计算过程基本与 Transformer 方法相同，区别在于，长和宽不再是$H \times W$，而是变成计算窗口，假定窗口大小为$M \times M$，则一共有$H/M \times W/M$个窗口需要计算，而每个窗口的计算复杂度与上述多头自注意力方法一致，所以窗口注意力模块的计算复杂度为$(H/M \times W/M) \times (4M^2C^2 + 2M^4C) = 4HWC^2 + 2HWM^2C$。

对上述计算的时间复杂度进行比较，本章方法相对于卷积操作，对计算复杂度进行求差，可以得到不等式$CK^2 \geq 4C + 2M^2$，假设窗口尺寸与卷积核大小相等，即$K = M$，通过不等式可以得到当$K \geq 3$且$C \geq 4$时（注意 K 和 C 均为整数），本章方法的计算复杂度均优于卷积方法。

本章方法相对于多头自注意力方法，同样地，对两种方法的计算复杂度求差，可以得到不等式$M^2 \leq HW$时，本章方法将优于多头自注意力方法，注意，M 代表的是窗口的尺寸，对于一个$H \times W$的输入，$M \leq \min(H, W)$，可以推得，本章方法在计算复杂度上优于基于多头注意力的方法。当然，由于引入了额外的通道注意力模块，相对于纯粹的窗口注意力方法，本章方法的计算复杂度有一些增加。

4.4 实验细节及参数设置

4.4.1 实验细节

本章模型的损失函数包括两个部分：一部分是用来计算第一阶段生成结果$\tilde{I}_t$与真实帧I_{gt}之间的像素级损失，记为表示损失L_p，另一部分是用来计算经过纹理增强之后的插值帧$\hat{I}_t$与真实帧I_{gt}之间的像素级损失，记为纹理损失L_t。最终训练所用的损失函数可以表示为：

$$L_{total} = \rho(\tilde{I}_t - I_{gt}) + \sigma * \rho(\hat{I}_t - I_{gt}) \tag{4.1}$$

其中，σ取值为 0.1，$\rho(\cdot)$代表求 1 范数的函数，如公式(3.2)所示。

$$\rho(\hat{I}_t - I_{gt}) = \| \hat{I}_t - I_{gt} \|_1 \tag{4.2}$$

之所以采用了 1 范数，是因为在帧插值任务当中，2 范数往往会导致模糊的结果。本章方法在训练过程中采用的是 Adam 优化器，其中β_1和β_2分别取 0.9 和 0.999，初始学习率设置为 0.001，衰减率设为 0.05，此外，本章的训练数据集使用的是 Viemo90k 数据集，并将其分为训练集和验证集。其中，训练集选取了包含 45000 个三元组，验证集包含 6313 个三元组，每帧的分辨率为 448 × 256。三元组的中间帧作为真实值，其余两帧作为输入数据。为了增强数据，还将输入序列的时间顺序颠倒。在训练过程中，批量大小设置为 4，并在 RTX 3080 上进行实验。经过大约 200 个 epoch 的训练后，训练过程收敛。

4.4.2 数据集

此外，本文还在以下数据集上对本章方法进行了测试，以验证本章方法的有效性以及鲁棒性。

1. Vimeo-90K 数据集由约 90000 个视频片段组成，每个片段包含 3 帧，分辨率为 448×256。数据集包含了从 vimeo.com 下载的约 90000 个视频片段，涵盖了多种场景和动作。本文中以三元组序列的前后两帧作为输入，以中间帧作为真实值进行训练。

2. UCF101[63]是一个由 YouTube 收集的真实动作视频的动作识别数据集，是一个用于视频动作识别的数据集。数据集由 101 种动作类别组成，每种类别包含 100 多个视频片段，总共有 13320 个视频片段，时长约 27 小时。它囊括了非常多的动作类别，而这些不同的动作类别可以帮助本文验证本章方法的泛化性。

3. DAVIS480p[64]是一个用于视频目标分割的数据集，由 50 个高质量视频序列组成，每个序列包含 30 到 100 帧，分辨率为 480p。这个数据集是 DAVIS2016 数据集的扩展，增加了更多的目标类别和更难的场景，并覆盖了许多常见的视频对象分割挑战，如遮挡、运动模糊、外观变化和多个对象。其中遮挡与运动模糊正是帧插值问题目前所面临的难点和痛点，以此作为测试集，可以有效展现出不同方法在极端情况下的性能表现。

4.4.3 评价指标

1. PSNR 是一种客观的图像质量评估指标，它反映了原始图像与失真图像之间的相似度。PSNR 的定义是基于均方误差（MSE），即原始图像与失真图像之间每个像素值的差的平方的平均值。MSE 越小，说明两幅图像越相似，PSNR 越高，说明图像质量越好。PSNR 通常用分贝（dB）表示，以便处理图像的动态范围较大的情况。采用 PSNR 作为度量可以比较直观的展现出原始图像与失真图像像素级的相似度。

2. SSIM 是一种完全参考度量，这意味着它需要一个未压缩或无失真的图像作为参考。SSIM 指数由三个部分组成：亮度、对比度和结构部分。它们分别评估图像亮度、对比度和结构变化的视觉影响。如公式(4.3)所示，第一项为图像亮度，通过计算亮度均值得到，第二项为对比度，通过计算亮度值的标准差得到，第三项为结构相似性，通过计算亮度值的协方差得到。结合了这三个部分，SSIM 作为度量可以反映图像结构之间的相似性。

$$S(x,y)) = l(x,y)^{\alpha} \cdot c(x,y)^{\beta} \cdot s(x,y)^{\gamma} \tag{4.3}$$

4.5 实验结果

表 4.3 展示了本章方法与最近提出的一些先进的方法在 Vimeo90K、UCF101、UCF101 数据集上的定量比较，其中包括基于光流与核估计结合的方法如 DAIN、MEMC，基于特征表示的方法 CAIN、FLAVR，基于 Transformer 的方法 VFIformer。通过 PSNR 和 SSIM 两个指标对这些方法进行了评估，红色字体标注的是最优的结果，蓝色字体标注的是次优的结果，其中 TCVFI 是在第三章中所提出的方法。我发现本章所提出的方法除了在一些数据集上略逊于 VFIformer 以外，要优于其他所有最先进的算法，而之所以会略逊于 VFIformer，可能是因为 VFIformer 中利用了光流信息，但是该模型作为一种基于直接特征表示的方法，所提取的特征是可以应用于下游任务当中的。总体上来说，该方法取得了非常具有竞争力的效果。

表 4.3 本章方法在 Vimeo90K、UCF101、UCF101 数据集上的定量比较

方法	Vimeo90K		UCF101		UCF101	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
DAIN	34.04	0.958	35.26	0.963	27.31	0.870
CAIN	33.93	0.964	32.28	0.965	26.46	0.856
FLAVR	36.19	0.975	33.29	0.971	27.41	0.874
VFIformer	36.50	0.981	35.43	0.970	28.31	0.891
MEMC	34.20	0.959	35.16	0.963	27.27	0.865
TCVFI	36.32	0.979	34.21	0.976	27.60	0.877
本章方法	36.48	0.979	35.48	0.978	28.14	0.883

为了进一步分析所提出的方法，我进行了充分地消融实验对本文的网络结构进行分析。我将特征聚合模块分别替换成了 2D 转置卷积、双线性插值，对于 2D 转置卷积，我增加了特征聚合模块的层数，以使得两者参数量保持基本一致。我用完全一样的设置对重组后的模块进行训练，并在多个数据集上进行了测试。定量测试的结果如所示。

表 4.4 本章方法在 Vimeo90K、UCF101、UCF101 数据集上的消融研究

方法	Vimeo90K		UCF101		UCF101	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
T+P+2D	36.07	0.975	33.40	0.971	27.45	0.874
T+P+B	35.60	0.966	34.08	0.963	26.57	0.866
T+P+3D	36.48	0.979	35.48	0.978	28.14	0.883

其中 T 表示基于 Transformer 的特征提取模块，2D 代表 2D 转置卷积，B 代表双线性插

值，3D 代表 3D 转置卷积，P 代表后处理模块，可以直接观察到第二行仅仅只是进行双线性插值效果并不好，它并没有进行进一步的完善的特征聚合。而观察第一行与第三行可以看到 3D 转置卷积比 2D 转置卷积更适合作上采样操作，可以生成具有更高质量的图像。

4.6 实验可视化分析

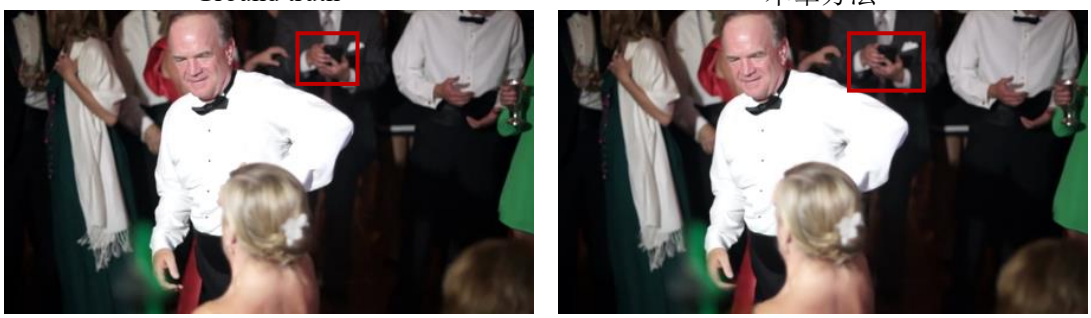
本节选取了一些复杂场景下的插帧结果进行展示并分析。

Ground truth　　本章方法

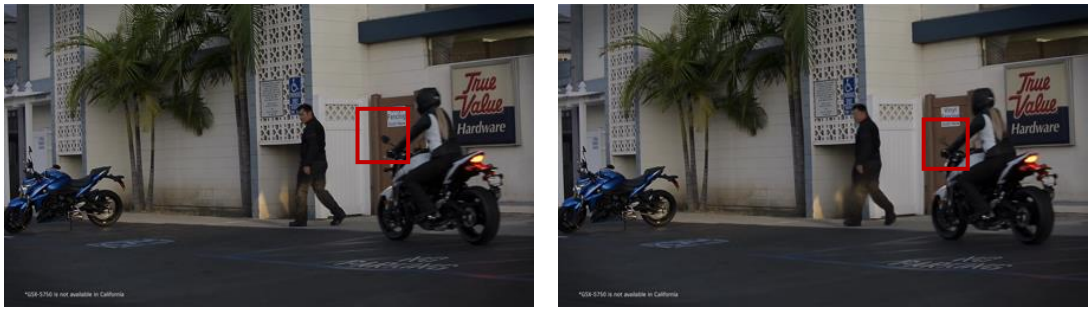
图 4.5 复杂运动场景下插帧结果

如图 4.5 所示，绿框标注出了插帧结果出现模糊的区域，结果显示，在复杂运动场景下，当视频中运动对象较多时，本章方法在一些亮度对比度不强的区域插帧的结果会出现模糊。分析认为在面对一些复杂运动场景时，网络模型对于运动特征的抽取能力受限，原因在于本章方法结构虽然简单，但是仅提取一个特征流同时建模时序上的运动信息和空间上的表观信息是存在一些困难的，这导致最终的插帧结果在一些边缘运动处发生了模糊。

Ground truth　　本章方法



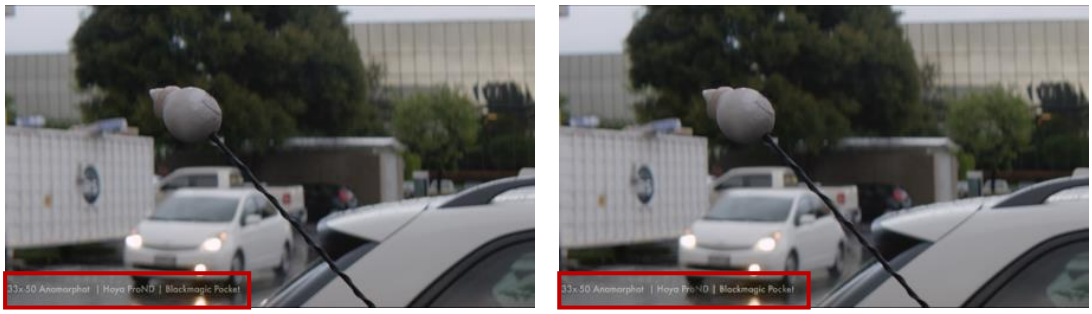
Ground truth　　本章方法



Ground truth　　本章方法

Ground truth　　本章方法

Ground truth　　本章方法



Ground truth　　本章方法

图 4.6 高对比度运动场景下插帧结果

如图 4.6 所示，在运动目标较少或者视频具有较高的对比度的情况下，本章所提出的方法可以很好的提取出运动信息，并生成非常清晰的纹理，这展现了本章所提出的基于 Transformer 的特征提取模块具备对运动进行隐式估计的能力。

以往的依赖于光流信息的帧插值方法往往无法有效地处理遮挡问题。本文分析认为依赖于光流的方法都需要满足比较强的先验假设，对于大运动的情况，无法满足光流法中的微小移动假设，这使得基于光流的方法在大运动的情况下无法取得好的效果，而对于遮挡的情况，由于一些背景像素点被遮挡，也就无法求偏导去获得运动信息，一般采取遮挡推理的方式对被遮挡的像素点进行近似，但这样依然会损失精度。本文通过基于 Transformer

的方法直接估计运动信息和表观信息的混合特征，而不依赖于光流信息，可以有效地避免大运动以及遮挡情况下的插值帧失真、模糊、伪影等问题。

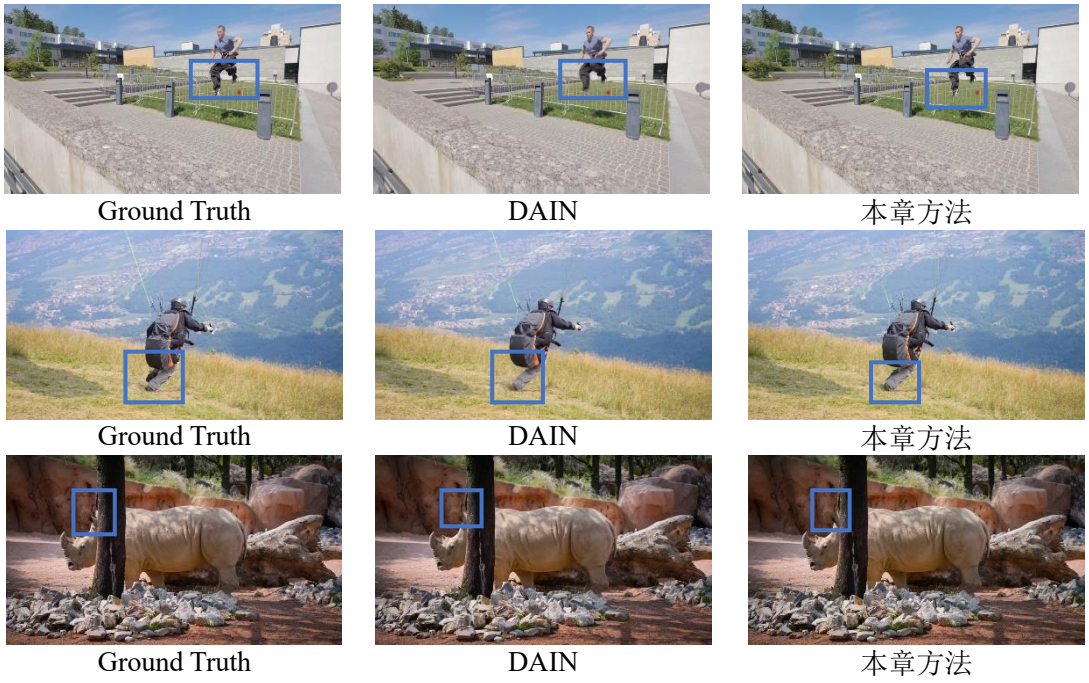


图 4.7 大运动及遮挡场景下插帧结果比较

如图 4.7 所示，本章方法与基于光流的先进方法 DAIN 在大运动以及有遮挡情况下进行了可视化对比，在遮挡区域以及大运动区域，DAIN 均出现了伪影以及模糊的情况，而在上述情况中本章方法呈现了更加清晰。纹理分明的可视化结果，这展现了本章所提出的方法在困难情境下的插帧性能，反映了本章方法相较于对比方法具有更好的鲁棒性。

4.7 本章小结

本章提出了一种新的帧插值方法，由于卷积神经网络受限于模型容量，容易出现过拟合的情况，而基于光流的方法又过于依赖光流估计网络的精度，本文基于 Transformer 设计新的基于直接特征表示的帧插值方法，在网络结构上，结合 Swin Transformer 所提出的窗口注意力以及移动窗口注意力并加入了通道注意力实现特征提取，获得低分辨率高通道数的特征图。考虑到特征图中蕴含的丰富的时序信息，选择 3D 转置卷积实现上采样操作，这样可以充分利用特征图中的丰富信息。最后我借鉴了基于光流的方法的处理过程，在模型最后引入了后处理模块，对插帧结果进行一些细节上的补充，从而获得更好的视觉效果。本章我还在三个数据集上进行了测试，结果显示，本章所提出的方法取得了有竞争力的性能结果。通过消融实验，也充分验证了所提出模块的有效性。

第五章 总结与展望

5.1 总结

随着多媒体设备的高速发展，人们对帧率和分辨率的要求都与日俱增，但是高帧率高分辨率的硬件设备造价昂贵，如果每个视频都使用这种硬件，那会大大抑制多媒体内容的产出，同时许多视频处理应用都需要用到帧插值作为核心，像是视频的编解码，动画电影的高帧率转换以及体育赛事的慢动作回放等等，这都体现出了帧插值在应用上的巨大潜力，也促使我们去开发更加优秀的帧插值算法。

现有的帧插值算法往往基于光流，但是这样会导致光流估计中的误差向后传播，从而影响插值帧的质量，并且由于基于光流的算法需要在估计出光流后，利用光流对输入帧进行翘曲操作，而这个操作是非常消耗计算资源的，为了解决这些问题，从先前的工作受到启发，我们设计了两种基于特征表示的直接帧插值方法。

具体来说，在本文中我们主要做出了以下两点贡献：

1. 本文提出了一种基于纹理增强的帧插值方法。模型先通过一个以 3D 卷积神经网络和通道注意力为基本块的 U-net 骨干网络，实现对视频特征的下采样和上采样过程，同时受到一些基于光流方法的启发，在不破坏特征流的情况下加入了一个纹理增强的后处理模块，增强视频插值帧的视觉效果，同时又不会引入新的误差。最后考虑到通道间关系在帧插值中的重要性，引入了 Pixel Shuffle 在第一步和最后一步分别对输入和输出进行下采样和上采样，这样既没有引入额外的参数，同时又可以实现通道级别的特征重塑。
2. 受到 Swin Transformer 的启发，本文考虑使用 Transformer 架构来实现特征的提取。因为卷积神经网络固有的局部性和平移不变性在帧插值问题中很容易导致无法提取出大运动的信息，因此，本文提出了一种基于 Transformer 的特征提取模块，但同时考虑到 Transformer 在图像生成过程中无法捕获到 Patch 内部联系，所以引入了 3D 转置卷积来进行上采样，并且沿用了之前后处理模块的思想，最后加入了纹理增强的后处理模块，在不影响特征流提取的情况下，增强插值帧的纹理效果。

5.2 展望

本文针对帧插值任务，分别设计了两种基于特征表示的直接帧插值方法，一种基于 3D 卷积神经网络进行特征提取，另一种则基于 Transformer 结构进行特征提取，但这些工作仍然有需要完善的部分，因此我们认为可以继续考虑的研究重点如下：

首先，不论是 3D 卷积神经网络还是 Transformer 架构，直接进行特征表示，会导致空域特征与时域特征等混合在一起，这无疑会加大训练难度，但假如我们可以利用网络结构实现同时提取出两个特征，再将两个特征进行一些融合，特征的融合同样可以利用类似注意力机制的模块来实现，或许会减少训练难度，并且所得到的特征表示也具有很强的启发性。

其次基于特征表示的直接方法面临的另一个问题是不利于指定在任意时刻的帧插值，对于光流法，在获取双向光流后，我们可以通过不同的时刻，以不同权重对输入帧进行翘曲，这样我们只需要训练一个模型就可以实现在任意时刻插帧，而基于特征表示的方法在这方面仍没有很好的解决办法，实现在任意时刻的帧插值或许是我们未来可以完善的一个问题。

参考文献

[1] S Baker, D Scharstein, J. P. Lewis, S Roth, MJ Black, R Szeliski. A Database and Evaluation Methodology for Optical Flow[J]. Int. J. Comput. Vision, 2011.92(1): 1-31.ISSN 0920-5691.

[2] S C Tai, Y R Chen, Z B Huang, C C Wang. A multi-pass true motion estimation scheme with motion vector propagation for frame rate up-conversion applications[J]. Journal of display technology, 2008. 4(2):188-197.

[3] B D Choi, J W Han, C S Kim, S J Ko. Motion-compensated frame interpolation using bilateral motion estimation and adaptive overlapped block motion compensation[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2007. 17(4):407-416.

[4] Y Dar, A M Bruckstein. Motion-Compensated Coding and Frame Rate Up-Conversion: Models and Analysis[J]. IEEE Transactions on Image Processing, 2015. 24(7):2051-2066.

[5] Q Lu, N Xu, X Fang. Motion-compensated frame interpolation with multiframe-based occlusion handling[J]. IEEE Journal of Display Technology, 2016. 12(1):45-54.

[6] D Choi, W Song, H Choi, T Kim. MAP-Based Motion Refinement Algorithm for Block-Based Motion-Compensated Frame Interpolation[J].IEEE Transactions on Circuits and Systerns for Vdeo Technology, 2016. 26(10):1789-1804.

[7] T H Tsai, A T Shi, K T Huang. Accurate frame rate up-conversion for advanced visual quality[J]. IEEE Transactions on Broadcasting, 2016. 62(2):426-435.

[8] H Y Wu, M Rubinstein, E Shih, J Guttag, F Durand, W Freeman. Eulerian video magniflca-tion for revealing subtle changes in the world[J]. 2012.

[9] S Meyer, O Wang, H Zimmer, M Grosse, A Sorkine-Hornung. Phase-based frame interpolation for video[C]//Proceedings of the IEEE Conference on Computer Vsion and Pattern Recognition. 2015:1410-1418.

[10] S Meyer, A Djelouah, B Mc Williams, A Sorkine-Hornung, M Gross, C Schroers. PhaseNet for video frame interpolation[C]I/Proeeedings of the IEEE Conference on Computer Vsion and Pattern Recognition. 2018:498-507.

[11] S L Pintea, J C van Gemert.Making a Case for Learning Motion Representations with Phase[C]//ECCV Workshops, Lecture Notes in Computer Science. volume 9915. Cham:Springer, 2016.

[12] S et al. Meyer. Phase-Based modification transfer for video[C]//Proceedings of Computer Vision (ECCV). volume 9907. 2016:633-648.

[13] L Zhou, R Sun, X Tian, Y Chen. Phase-based frame rate up-conversion for depth video[J].Journal of Electronic Imaging, 2018. 27(4):043036.

[14] S Niklaus, L Mai, F Liu. Video Frame Interpolation via Adaptive Separable Convolution[C]//2017 IEEE International Conference on Computer Vision (ICCV). IEEE, oct 2017.

[15] S Niklaus, L Mai, F Liu. Video frame interpolation via adaptive convolution[C]//IEEE Conference on Computer Vision and Pattern Recognition. volume 1. 2017:3.

[16] A Dosovitskiy, P Fischer, E Ilg, P Hausser, C Hazirbas, V Golkov, P Van Der Smagt, D Cremers, T Brox. Flownet: Learning optical flow with convolutional networks[C]//Proceedings of the IEEE International Conference on Computer Vision. 2015:2758-2766.

[17] E Ilg, N Mayer, T Saikia, M Keuper, A Dosovitskiy, T Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks[C]//IEEE conference on computer vision and pattern recognition (CVPR). volume 2. 2017:6.

[18] A Ranjan,M J Black.Optical flow estimation using a spatial pyramid net-work[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.2017:4161-4170.

[19] D Sun, X Yang, M Y Liu, Jan K. PWC-Net: CNNs for optical flow using pyramid, warping, and cost volume[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2018.

[20] T W Hui, X Tang, C Change Loy. Liteflownet: A lightweight convolutional neural network far optical flow estimation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018:8981-8989.

[21] T Zhou, S Tulsiani, W Sun, J Malik, A A Efros.View synthesis by appearance flow[CJ//European conference on computer vision. Springer, 2016:286-301.

[22] Z Liu, R A Yeh, X Tang, Y Liu, A Agarwala. Video Frame Synthesis Using Deep Voxel Flow.[C]//ICCV 2017:4473-4481.

[23] T Xue, B Chen, J Wu, D Wei, W T Freeman. Video enhancement with task-oriented flow[J].International Journal of Computer Vision, 2019. 127(8):1106-1125.

[24] G Lu, X Zhang, L Chen, Z Gao. Novel integration of frame rate up conversion and hevc coding based on rate-distortion optimization[J].IEEE Transactions on Image Processing, 2017. 27(2):678-691.

[25] Y Dar, A M Bruckstein. Motion-Compensated Coding and Frame Rate Up-Conversion: Models and Analysis[J]. IEEE Transactions on Image Processing, 2015. 24(7):2051-2066.

[26] G Lu, X Zhang, Z Gao.A novel framework of frame rate up conversion integrated within HEVC coding[C]//IEEE International Conference on Image Processing.2016:pp. 4240-4244.

[27] T Xue, J Wu, K Bouman, B Freeman. Visual dynamics: Probabilistic future frame synthesis via cross convolutional networks[C]//Advances in neural information processing systems. 2016:91-99.

[28] T C Wang, M Y Liu, J Y Zhu, A Tao, J Kautz, B Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018:8798-8807.

[29] S E Chen, L Williams. View interpolation for image synthesis[C]//Proceedings of the 20th annual conference on Computer graphics and interactive techniques. ACM, 1993:279-288.

[30] X Artigas, L Torres. Iterative generation of motion-compensated side information for distributed video coding[C]//IEEE International Conference on Image Processing 2005. volume 1. IEEE, 2005:I-833.

[31] B D Choi, J W Han, C S Kim, S J Ko. Motion-compensated frame interpolation using bilateral motion estimation and adaptive overlapped block motion compensation[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2007. 17(4):407-416.

[32] Q Lu, N Xu, X Fang. Motion-compensated frame interpolation with multiframe-based occlusion handling[J]. IEEE Journal of Display Technology, 2016. 12(1):45-54.

[33] D Wang, A Vincent, P Blanchfield, R Klepko. Motion-compensated frame rate up-conversion—part ii: New algorithms for frame interpolation[J]. IEEE Transactions on Broadcasting, 2010. 56(2):142-149.

[34] W Song, P Heo, G Choi, S R Oh, H Park. Motion compensated frame interpolation of occlusion and motion ambiguity regions using color-plus-depth information[C]//2018 25th IEEE International Conference on Image Processing (ICIP). IEEE, 2018:1478-1482.

[35] C Zhang, Z Chen, M Wang, M Li, S Jiang. Robust Non-Local TV-L^1 Optical Flow Estimation With Occlusion Detection[J]. IEEE Transactions on Image Processing, 2017. 26(8):4055-4067.

[36] X Yang, J Liu, et al. Depth-assisted frame rate up-conversion for stereoscopic video[J]. IEEE Signal Processing Letters, 2014. 21(4):pp. 423-427.

[37] W Bao, W Lai, X Zhang, Z Gao, M-H Yang. Memc-net: Motion estimation and motion compensation driven neural network for video interpolation and enhancement[J]. IEEE transactions on pattern analysis and machine intelligence, 2019.

[38] Y Guo, L Chen, Z Gao, X Zhang. Frame rate up-conversion method for video processing applications[J]. IEEE Transactions on Broadcasting, 2014. 60(4):659-669.

[39] K Lee, J Jeong. Bilateral frame rate up-conversion algorithm based on the comparison of texture complexity[J]. Electronics Letters, 2016. 52(5):354-355.

[40] Y Guo, Z Gao, L Chen, X Zhang. Occlusion handling frame rate up-conversion[C]//2013 IEEE International Conference on Acoustics,Speech and Signal Processing.IEEE, 2013:1424-1428.

[41] Y L Liu, Y T Liao, Y Y Lin, Y Y Chuang. Deep video frame interpolation using cyclic frame generation[C]//AAAI. 2019.

[42] D Mahajan, F C Huang, W Matusik, R Ramamoorthi, P Belhumeur. Moving gradients:a path-based method for plausible image interpolation[C]//ACM Transactions on Graphics (TOG). volume 28. ACM, 2009:42.

[43] N Wadhwa, M Rubinstein, F Durand, W T Freeman. Phase-based video motion processing[C]//ACM Transactions on Graphics. volume 32. 2013:80.

[44] E P Simoncelli, W T Freeman. The steerable pyramid: a flexible architecture for multi-scale derivative computation[C]//IEEE International Conference on Image Processing. 1995:pp.444-447.

[45] Ian Goodfellow, J Pouget-Abadie, M Mirza, B Xu, D Warde-Farley, S Ozair, A Courville, Y Bengio. Generative adversarial nets[C]//Advances in neural information processing sys- terns. 2014:2672-2680.

[46] Ho J , Jain A , Abbeel P . Denoising Diffusion Probabilistic Models[J]. 2020.

[47] J van Amersfoort, W Shi, A Acosta, F Massa, J Totz, Z Wang, J CabaIlero. Frame interpolation with multi-scale deep loss functions and generative adversarial networks[J].arXiv preprint arXiv:1711.06045, 2017.

[48] X Liang, L Lee, W Dai, E P Xing. Dual motion gan for future-flow embedded video prediction[C]//Proceedings of the IEEE International Conference on Computer Vision. 2017:1744-1752.

[49] C Vondrick,H Pirsiavash,A Torralba.Generating videos with scene dynamics[C]//Advances In Neural Information Processing Systems. 2016:613-621.

[50] T Xue, J Wu, K Bouman, B Freeman. Visual dynamics: Probabilistic future frame synthesis via cross convolutional networks[C]//Advances in neural information processing systems. 2016:91-99.

[51] H Jiang, D Sun, V Jampani, M H Yang, E Learned-Miller, J Kautz. Super SloMo: High quality estimation of multiple intermediate frames for video interpolation[J]. jun 2018.

[52] J Johnson, A Alahi, F Li, F. Perceptual losses for real-time style transfer and super-resolution[C]//European conference on computer vision. Springer, 2016:694-711.

[53] S Nilclaus, F Liu. Context-aware synthesis for video frame interpolation[J]. arXiv preprint arXiv:1803.10967, 2018.

[54] Bao W, Lai W S, Ma C, et al. Depth-aware video frame interpolation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 3703-3712.

[55] Choi M, Kim H, Han B, et al. Channel attention is all you need for video frame interpolation[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2020, 34(07): 10663-10671.

[56] Kalluri T, Pathak D, Chandraker M, et al. Flavr: Flow-agnostic video representations for fast frame interpolation[C]//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2023: 2071-2082.

[57] 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.

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

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

[60] Lu L, Wu R, Lin H, et al. Video frame interpolation with transformer[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 3532-3542.

[61] Danier D, Zhang F, Bull D. Enhancing deformable convolution based video frame interpolation with coarse-to-fine 3d cnn[C]//2022 IEEE International Conference on Image Processing (ICIP). IEEE, 2022: 1396-1400.

[62] Xue T, Chen B, Wu J, et al. Video enhancement with task-oriented flow[J]. International Journal of Computer Vision, 2019, 127: 1106-1125.

[63] Soomro K, Zamir A R, Shah M. UCF101: A dataset of 101 human actions classes from videos in the wild[J]. arXiv preprint arXiv:1212.0402, 2012.

[64] Perazzi F, Pont-Tuset J, McWilliams B, et al. A benchmark dataset and evaluation methodology for video object segmentation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 724-732.

致 谢

三年的硕士生涯倏忽而过，回首这三年，酸甜苦辣千般滋味涌上心头，往日回忆走马灯般在脑海中不停闪烁，想停却停不住。在此期间，我经历了许多，不敢言所谓进步，所谓成长，但想来总是走过了一段与以往想法截然不同的分岔路，感谢在这段时光陪我一起走过的所有人，陪伴的时间或长或短，相互的了解或深或浅，但是这一路上你们都不可或缺。

首先，我要感谢我的导师。他是一位对学术有着极高热忱的老师，始终保持着对新知识的好奇，以及对学术的严谨态度，这种态度深深地影响了我，让我意识到了做科研不是浅尝辄止，张冠李戴就可以了，它需要深入地了解领域研究现状，要明白研究的核心问题是什么，解决问题的思想是什么，而后才是百花齐放的解决方案，方案只是工具，只有理解了问题，然后善用工具，我们才可能在科研的路上前进一小步，此中乐，不足为外人道也。

其次，我要感谢我的女朋友，人生不如意事十之八九；与你相识，幸事；没有更早遇见你，憾事。我实在是一个徘徊不定的人，不敢选择亦不敢尝试，有她的陪伴和督促，让我重新有了勇气与方向，此中滋味，难以言说，百般心绪，亦无法言表。过去的这段日子我一度迷茫、抑郁，我可以在所有人面前表现如常，唯独在她面前不可以。幸亏有她，曾经我或许是只在沙漠中迷路的骆驼，很幸运，遇上了她这样一片绿洲。

我还要感谢我的父母，是他们帮我挡住了太多风浪，我才能省去很多烦恼，虽然在科研上他们也知之甚少，但是他们始终支持着我，鼓励着我，在我许多次低迷，落寞的时候，每当想起他们，我都还有继续加油的勇气，人生无离别，怎知恩爱重，求学至今，这句话也让我愈发感同身受，若非是一次次的离别，又怎会意识到家的意义。

研究生的求学过程中，我也结交认识了许多好朋友好球友，他们个性各异，但是都非常可爱，我们获得了许多美好的回忆，也见证了每一个人遇到困难的迷茫彷徨，但最终都将战胜困难的过往，此刻，他们的身影一个个在我脑海里浮现，挥之不去，愿我们友情常在，愿大家前程似锦。

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

1. Jin Wang, Xiaoyang Tan. Texture-Based Multi-Frame Interpolation System. The 10th International Academic Conference for Graduates of NUAA.2022:273-276.