硕士学位论文

基于可变形卷积网络的视频去模糊算法

DEFORMABLE CONVOLUTIONAL NETWORKS FOR VIDEO DEBLURRING

段风志

哈尔滨工业大学

2022 年 6 月

国内图书分类号：TP391.4　　学校代码：10213
国际图书分类号：004.8　　密级：公开

工程硕士学位论文

基于可变形卷积网络的视频去模糊算法

硕士研究生：段风志
导　　师：姚鸿勋教授
申请学位：工程硕士
学　　科：电子信息（软件工程）
所在单位：计算学部
答辩日期：2022 年 6 月
授予学位单位：哈尔滨工业大学

Classified Index: TP391.4
U.D.C: 004.8

Dissertation for the Master's Degree in Engineering

DEFORMABLE CONVOLUTIONAL NETWORKS FOR VIDEO DEBLURRING

Candidate: Fengzhi Duan
Supervisor: Prof. Hongxun Yao
Academic Degree Applied for: Master of Engineering
Speciality: Electronic Information(Software Engineering)
Affiliation: Faculty OF Computing
Date of Defence: June, 2022
Degree-Conferring-Institution: Harbin Institute of Technology

摘 要

相机抖动、目标运动和景深变化等原因会导致设备拍摄的视频中存在模糊。而模糊的视频会影响人类视觉感官质量和高层次的视觉任务。视频去模糊是一项具有挑战性的任务，其通常分为四个阶段：特征提取、相邻视频帧对齐、特征融合以及特征重建。本文的研究思路是利用相邻帧中的清晰像素融合参考帧中的对应像素，充分挖掘输入视频序列中的时空信息。实现该思路，需要解决两个最关键的问题：相邻视频帧精确对齐和自适应时序特征融合。针对这些问题，本文提出了两个递进式的解决方案。

第一个工作，我们提出了基于多尺度可变形卷积网络的视频去模糊算法，提升了相邻帧对齐的准确性。它在相邻帧特征层面执行隐式对齐，解决了传统光流估计对齐不准确、计算量大的问题。我们首先通过实验验证了特征对齐比图像对齐具有更好的结果。然后提出了基于多尺度可变形卷积网络的隐式特征对齐模块，它利用可变形卷积的几何形变建模能力结合多尺度策略实现了从粗略到精细的特征对齐。对齐后的相邻帧特征和参考帧特征，在相同空间位置处具有对应的时序关系。所以特征融合阶段中使用1×1卷积网络进行通道维度的时序特征融合，然后利用卷积层和残差块组成的网络进一步融合特征。

第二个工作，我们提出了基于自适应时空卷积网络的视频去模糊算法，进一步提升了相邻帧对齐的准确性并且极大地提高了时序特征融合的效率。我们改进了可变形卷积，提出一种新的卷积计算方式 Dcn Align。它将光流作为基础的位置偏移，并通过卷积网络生成残差偏移，解决了第一个工作中对齐网络训练不稳定、位置偏移值溢出的问题。特征融合阶段，基于动态滤波网络提出了动态局部滤波层，它具有特定于输入和空间位置进行局部特征变换的能力，实现了利用相邻帧中的清晰特征对参考帧中的模糊特征做自适应地像素级融合。

我们将这两个算法在公开数据集上进行了定量评估和定性测试，实验结果表明算法具有非常高的准确率，可以有效去除动态场景存在中的非均匀模糊。

关键词：视频去模糊；可变形卷积；动态滤波网络；多尺度策略

Abstract

Causes such as camera shake, object motion, and depth-of-field changes can cause blur in video captured by the device. And blurry video affects the quality of human visual senses and high-level visual tasks. Video deblurring is a challenging task, which is usually divided into four stages: feature extraction, adjacent video frame alignment, feature fusion, and feature reconstruction. The research idea of this paper is to use the clear pixels in adjacent frames to fuse the corresponding pixels in the reference frame to fully mine the spatiotemporal information in the input video sequence. To realize this idea, two most critical problems need to be solved: accurate alignment of adjacent video frames and adaptive timing feature fusion. Aiming at these problems, this paper proposes two progressive solutions.

In the first work, we propose a video deblurring algorithm based on multi-scale deformable convolutional networks, which improves the accuracy of adjacent frame alignment. It performs implicit alignment at the feature level of adjacent frames, which solves the problems of inaccurate alignment and heavy computation in traditional optical flow estimation. We first experimentally verify that feature alignment has better results than image alignment. Then, an implicit feature alignment module based on multi-scale deformable convolutional network is proposed, which utilizes the geometric deformation modeling ability of deformable convolution combined with multi-scale strategy to achieve feature alignment from coarse to fine. The aligned adjacent frame features and reference frame features have a corresponding temporal relationship at the same spatial position. Therefore, in the feature fusion stage, the convolutional network is used to fuse the time series features of the channel dimension, and then the network composed of the convolutional layer and the residual block is used to further fuse the features.

In the second work, we propose a video deblurring algorithm based on adaptive

spatio-temporal convolutional network, which further improves the accuracy of adjacent frame alignment and greatly improves the efficiency of temporal feature fusion. We improve the deformable convolution and propose a new convolution calculation method Dcn Align. It uses optical flow as the basic position offset and generates residual offset through the convolutional network, which solves the problems of unstable alignment network training and overflow of position offset values in the first work. In the feature fusion stage, a dynamic local filter layer is proposed based on the dynamic filter network, which has the ability to transform local features specific to the input and spatial position, and realizes the use of the clear features in adjacent frames to adapt to the fuzzy features in the reference frame on pixel-level.

We quantitatively evaluate and qualitatively test these two algorithms on public datasets. The experimental results show that the algorithms have very high accuracy and can effectively remove non-uniform blur in the presence of dynamic scenes.

Keywords: video deblurring; deformable convolution; dynamic filtering network; multi-scale strategy

目 录

摘 要.....I
ABSTRACT.....I
第 1 章 绪 论.....1
1.1 课题研究背景和意义.....1
1.1.1 课题研究背景.....1
1.1.2 课题研究意义.....1
1.2 图像和视频去模糊的研究现状分析.....2
1.2.1 图像去模糊.....2
1.2.2 视频去模糊.....4
1.3 存在的问题与挑战.....6
1.3.1 相邻视频帧的精确对齐.....6
1.3.2 真实模糊数据集的获取.....6
1.4 本文的研究内容和组织结构.....7
1.4.1 本文的研究内容.....7
1.4.2 文章的组织结构.....8
第 2 章 卷积神经网络理论基础.....10
2.1 引言.....10
2.2 普通卷积神经网络.....12
2.3 可变形卷积网络.....16
2.3.1 基础可变形卷积网络.....17
2.3.2 增强的可变形卷积网络.....19
2.4 动态滤波网络.....20
2.4.1 滤波器生成网络.....21
2.4.2 动态滤波层.....22
2.5 本章小结.....23
第 3 章 基于多尺度可变形卷积的视频去模糊算法.....24
3.1 引言.....24
3.2 算法原理概述和整体网络结构.....25
3.3 相邻视频帧对齐.....26

3.3.1 基于光流估计的图像对齐和特征对齐 26
3.3.2 基于多尺度可变形卷积网络的隐式特征对齐 28
3.4 特征融合和特征重建部分 30
3.4.1 特征融合 30
3.4.2 特征重建 31
3.5 实验设置 31
3.5.1 数据集和数据增强 31
3.5.2 损失函数 32
3.5.3 模型实现和训练细节 32
3.6 实验结果和分析 32
3.6.1 算法定量评估结果 33
3.6.2 算法定性评估结果 33
3.7 本章小结 36
第 4 章 基于自适应时空卷积网络的视频去模糊算法 37
4.1 引言 37
4.2 原理概述和整体网络结构 37
4.3 基于增强可变形卷积网络的相邻帧对齐模块 39
4.4 基于动态滤波网络的自适应时序特征融合模块 42
4.5 基于自适应时空卷积网络的视频去模糊算法 46
4.6 实验结果和分析 46
4.6.1 算法定量评估结果 46
4.6.2 算法定性评估结果 47
4.7 消融实验 51
4.7.1 相邻帧对齐模块 51
4.7.2 时序特征融合模块 52
4.8 本章小结 52
结论 54
参考文献 55
攻读硕士学位期间发表的论文及其它成果 59
哈尔滨工业大学学位论文原创性声明和使用权限 60
致 谢 61

第1章 绪 论

1.1 课题研究背景和意义

1.1.1 课题研究背景

图像去模糊一直以来是计算机视觉和图像处理领域一项重要的任务。给定一幅由于相机抖动、目标运动或者失焦等原因从而引入模糊的图像，图像去模糊的目标是根据图像必要的边缘信息和隐藏的细节恢复出锐利清晰的图像。根据输入的信息源不同，图像去模糊分为单幅图像去模糊和视频去模糊。

单幅图像去模糊是高病态的。传统的方法将各种约束条件应用于模糊的模型特征，以及利用了不同的自然图像先验来规范求解空间。这些方法大多数涉及启发式参数调整，并伴随着昂贵的计算代价。此外，关于模糊的简化假设模型，通常在真实模糊图像上失效。因为真实世界中的模糊图像的退化过程要比假设模型复杂的多。这也导致模型泛化性能较差。之后提出了基于学习的方法进行去模糊。早期的方法用学习的参数替换了传统框架中的一些模块或步骤，以利用外部数据。最近的工作开始使用端到端的可训练神经网络对图像进行去模糊。

视频去模糊其旨在从模糊视频序列中恢复清晰帧，相较于单幅图像去模糊，区别在于可以有效利用输入视频帧与帧之间的冗余信息。近年来，随着短视频的井喷式爆发和手持以及机载视频捕获设备的广泛普及，这一问题得到相关专业人员的积极关注和研究。视频中的模糊通常是由目标运动、相机抖动和深度变化等原因引起的，这种模糊是一种空间变化模糊，即视频序列中的不同视频帧和同一视频帧的不同区域具有不同程度的模糊。所以使用全局统一的模糊核很难消除这种空间变化模糊。

1.1.2 课题研究意义

模糊的图像和视频不仅会导致人类视觉感官体验的质量下降，还会妨碍一

些高级计算机视觉任务，例如目标检测、视觉跟踪和 SLAM 等。去模糊算法可以从模糊的图像中恢复出更多的细节和高频信息，能够有效提高视觉感官体验质量和下游视觉任务的性能。在我们实际生活中有很多去模糊算法的应用，比如老旧图像修复、经典老电影修复等。这些图像和视频，经过去模糊算法处理后变得清晰。而计算机视觉任务的输入基本都是图像或者视频，去模糊算法通常作为这些任务的数据预处理手段。因为图像中存在的模糊，会严重影响下游计算机视觉任务的性能。比如目标检测。如果被检测的物体由于运动、相机抖动、失焦等原因，造成大片的模糊，会大大降低目标检测算法的准确率。但是经过去模糊算法的处理，目标检测算法的输入就是清晰的图像，不会影响到算法性能。因此，研究一个有效的图像和视频去模糊算法是非常有意义的。

1.2 图像和视频去模糊的研究现状分析

1.2.1 图像去模糊

图像去模糊的任务是对于输入的模糊图像，输出恢复出的清晰图像。但由于现实世界中的图像模糊产生的原因不同，模糊类型也不同，所以很难使用一个统一的模型或算法对所有的模糊进行处理。国外的研究者试图处理动态场景中由于相机抖动、目标高速运动和深度变化引起的空间变化模糊。这些方法大多数基于以下图像退化模型：

$$B = L \otimes K + N \tag{1-1}$$

其中 B、L、N 分别代表模糊图像、清晰图像和额外的噪声，K 是模糊核，⊗代表卷积运算。为每个像素找到模糊核是一个高病态问题。传统的方法试图通过对模糊源进行先验假设来对模糊模型进行建模。在文献[1,2]中，他们假设模糊仅由三维相机运动引起。但是，在动态场景中，由于存在多个运动对象以及摄像机运动，因此核估计更具挑战性。文献[3]提出了一种动态场景去模糊方法。但是，这些模糊核的估计仍然不准确，尤其是在物体突然运动和相机严重抖动的情况下。这种错误的模糊核估计直接影响潜像的质量，从而导致在恢复的图像中产生振铃伪像。

近几年，卷积神经网络已应用于图像处理问题，并表现出很好的效果[4-7]。由于没有现实世界中真实的模糊图像和其对应标签的清晰图像可用于监督学习，因此相关研究人员通常使用由模糊内核卷积清晰图像来生成对应的模糊图像。在文献[4,5,7]中使用具有统一模糊核合成的模糊图像进行训练。从图像去模糊数据集[8,9]问世以来，通过直接估计去模糊的输出而无需内核估计，从而提出了端到端学习方法[8,10,11]。这些端到端的方法，只需要输入模糊的图像，经过模型处理之后便可以得到最终去模糊处理后的清晰图像。为了获得用于处理大模糊的大感受野，在文献[8,10]中使用了多尺度策略。为了处理动态场景的非均匀模糊，张等人使用空间变体 RNN[12]在特征空间通过神经网络生成的 RNN 权重消除模糊。为了生成具有更多细节的清晰图像，在文献[8,13]中使用了对抗损失来训练网络，使得恢复的图像更符合人类视觉感受。

最近另一类神经网络架构 Transformers 在自然语言和高级视觉任务上表现出了显着的性能提升。因此，也有许多学者开始研究能否将 Transformer 架构用于底层图像处理任务（去噪、去模糊、超分辨率等），并已提出一些新的模型。采用 Transformer 进行图像复原主要有两个问题:局部上下文信息弱和模型计算量大。局部上下文信息对于图像复原非常重要。对于图像去模糊任务，可以利用图像像素的局部自相似性信息来去模糊。Transformer 庞大的计算量导致其不适用于高分辨率图像。针对上述两个问题，文献[16]提出了一种有效的基于 Transformer 的架构 Uformer 用于图像复原。它采用 Transformer 模块构建了一种分层编解码网络。Uformer 的两个核心设计使其适用于图像复原任务：（1）跳过连接机制。通过跳过连接机制搭建编码器与解码器之间的信息桥接。（2）局部增强窗口 Transformer 模块。采用非重叠的基于窗口的自注意力降低计算量，同时在前向网络中采用深度卷积进一步改善其捕获局部上下文的能力。受益于上述两种设计，Uformer 具有为图像复原捕获有用依赖关系的能力。近来还有许多其他基于 Transformer 架构的图像去模糊工作涌现出来，取得了不错的效果。这也是最近图像去模糊领域中的一个研究热点和趋势。

1.2.2 视频去模糊

近年来，基于深度神经网络的视频去模糊方法得到相关人员的广泛研究。早期对视频去模糊的研究仅将其看作图像去模糊的简单拓展，视频帧与帧之间的冗余信息没能得到充分利用。这些方法通常会使用视频序列中的多个连续帧作为模型输入，以各种方式利用输入帧之间的时序关系来恢复其中间视频帧。DBN [9]在通道维度中堆叠 5 个连续的帧，卷积神经网络在相邻帧之间聚合时空信息。KIM[19]等人通过深度循环神经网络将多帧特征连接起来以恢复当前图像。

最近的研究通过更复杂的流程将视频去模糊分为特征提取、相邻帧对齐、特征融合以及特征重建四部分，充分地利用了视频的帧间空域信息。国内研究者提出的方法由于使用了来自相邻帧的较清晰区域[12,13]或来自连续帧的光流[14,15]对中间视频帧进行质量补偿，因此取得了重大进展。但是，直接利用相邻帧的清晰区域来补偿中间帧对应的区域，通常会产生明显的伪影，因为相邻视频帧没有完全对齐。大多数现有方法[16-18]都是通过显式估计参考帧及其相邻帧之间的光流场来执行对齐的。但光流估计计算量很大而且很难做到准确，往往使得参考帧产生伪影。

对于视频去模糊任务，最大的难题就是相邻视频帧的精确对齐。只有对齐了输入的视频帧，才可以使用相邻帧中的冗余空域信息补偿中间帧。2018 年动态滤波器这一架构的提出，让研究者们探索出可以解决相邻视频帧对齐难题的方法，例如 STFAN[22]和 EDVR[27]。STFAN[22]提出了滤波器自适应卷积层，将对齐和去模糊这两个过程视为特征域中的两个滤波器自适应卷积。STFAN[22]对需要恢复的视频帧应用逐元素卷积核，根据输入自适应地对视频帧进行像素级别的特征变换。STFAN[22]将相邻帧对齐和去模糊集成到同一个框架中，而无需显示的运动估计，这解决了传统的光流估计计算量大的难题。EDVR[27]借助可变形卷积网络，提出了多尺度的相邻视频帧对齐模块。该模块使用一种金字塔结构以从粗到细的方式在特征层将每个相邻帧与参考帧执行对齐，以处理大而复杂的运动。对齐过程利用了可变形卷积的几何形变建模能力，使得模型可以处理不同程度的帧间抖动。但是由于可变形卷积网络中位置偏移的生成过程是

没有监督信号的，其网络的解释性差、训练过程不稳定，模型结果也难复现。

在其他视频增强领域同样涌现了很多优秀的工作，可以为视频去模糊任务提供一些新的思路和灵感。MFQE [26]提出了一种针对压缩视频的质量增强方法。其核心思想是通过质量好（清晰）的帧补偿质量差（模糊）的帧：首先 MFQE [26]使用一个分类器，将视频中的质量好帧找出来，对于每一个差帧，借助其相邻的两个好帧进行质量增强。在质量增强之前，相邻的两个好帧要进行运动补偿，对齐到差帧所处时刻的状态。该方法实现了较好的性能，提升了视频序列中质量较差的帧的恢复效果。但同时存在两个问题：（1）MFQE [26]依赖于精确的运动补偿（光流估计）。如果运动补偿不准确，后续的质量增强方法便会失效。（2）不同的相邻帧、同一视频帧的不同区域都具有不同程度的模糊，所以 MFQE [26]在利用质量好的帧弥补质量差的帧的同时，可能也引入了质量好的帧中模糊的区域。因此应该在像素级别利用相邻帧中质量好的区域补偿当前帧中对应的质量差的区域。

对于视频去模糊任务，影响其性能的除了算法本身，最重要的就是模糊数据集了。目前去模糊领域两个常用的数据集 GOPRO 数据集和 REDS 数据集是 Nah 提出的，都是人工合成的，并不是真实世界中的模糊图像。因为去模糊领域里，从传统方法到深度学习方法，实际上并没有真的需要被去模糊的图片。人工合成数据集的方法是用高速相机照出清晰的图像，然后在这个基础上处理成模糊的图像，原图像拿来做真值。这就出现一个问题，人工合成的模糊图像可能不符合现实世界的模糊图像。所以，人工处理的过程是没办法完全建模现实世界中真实图像的退化过程，这也是现阶段去模糊算法泛化能力差的最大原因。解决这个问题，现阶段有两种思路：一是学习生成更符合现实世界的模糊图像，二是直接通过相机拍摄出现实世界的模糊图像。对于第一种思路，优图实验室的工作[31]方法是通过两个 GAN，一个来生成更符合真实世界的模糊图像，一个用来学习去模糊，相比 DeblurGan 有了一定的进步。文献[32]通过两台相机同时拍摄，其中一个相机通过低快门速度捕获模糊图像，另一个通过高快门速度捕获 GT 图像，然后再通过他们的后处理方法来生成高质量的真实 GT 图

像。这个工作提出第一个用于学习图像去模糊的大规模真实世界模糊数据集RealBlur Dataset。通过实验表明，RealBlur Dataset 可以极大地提高基于深度学习的去模糊方法在相机抖动和移动对象上对真实世界模糊图像的性能。这个工作对解决真实世界的模糊数据集难以获取这一难题，提出了可行的解决方案。

1.3 存在的问题与挑战

1.3.1 相邻视频帧的精确对齐

目前视频去模糊工作中相邻视频帧对齐的方法主要有两种：基于运动估计（例如光流）的显示对齐和基于可变形卷积网络的隐式对齐。基于光流的显示对齐方法，存在的最大问题是对于模糊的图像，计算的光流是不准确的。不准确的光流，会导致视频帧在根据光流变形之后引入伪影，影响视频帧质量。所以 2019 年之前大部分基于光流的视频去模糊算法，相邻视频帧对齐的效果并不理想。另一种对齐方法是基于可变形卷积的。可变形卷积网络根据输入通过一系列卷积层生成采样位置偏移，在之后的卷积过程中根据采样位置偏移实现隐式对齐。但是采样位置偏移的生成，是没有监督信号的。其网络的解释性较差，而且训练过程非常不稳定，导致模型结果难以复现。综上，现阶段针对视频增强任务（视频去模糊、视频去噪、视频超分辨率重建等）中的相邻视频帧对齐部分，还没有非常好的解决方案，这是该领域非常值得研究的一大难点。

1.3.2 真实模糊数据集的获取

目前大多数图像和视频去模糊工作中，使用的数据集都是人工合成的，并不是真实世界中的模糊图像。因为去模糊领域里，从传统方法到深度学习方法，实际上并没有真的需要被去模糊的图片。如果用真实的图片作为模糊图片去训练，没有可供参考的真值，导致无法计算损失。人工合成的模糊图像无法完全建模现实世界中真实图像的退化过程，导致现阶段去模糊算法泛化能力差。而获取真实模糊数据集，存在着非常大的困难：在存在模糊的情况下，需要将模糊图像的内容同其真值清晰图像进行几何对齐。这意味着两个图像应该在同一

相机位置拍摄，由于必须摇晃相机才能拍摄模糊的图像，这极大增加了拍摄难度。此外，用于图像去模糊的真实世界模糊数据集应满足以下要求。首先，数据集应涵盖相机抖动的最常见情况，即运动模糊最频繁发生的弱光环境。其次，真值清晰图像应该具有尽可能小的噪声。最后，模糊和真值清晰图像应进行光度对准。这些要求无疑对构建真实世界的模糊数据集提出了严峻的挑战。

1.4 本文的研究内容和组织结构

1.4.1 本文的研究内容

本文的研究目标是自主设计和实现可以有效去除动态场景中非均匀模糊的视频去模糊算法。通过调研近几年视频去模糊领域的发展现状和趋势，了解到目前该领域存在的尚待解决的问题：相邻视频帧精确对齐和自适应时序特征融合，形成了本文的研究内容。

（1）相邻视频帧精确对齐。　　因为视频序列中的相邻视频帧之间存在一定程度的抖动，要利用相邻帧中清晰的区域补偿参考帧中对应的模糊区域，需要在时序特征融合之前对相邻帧进行对齐，将其对齐到参考帧所处时刻的状态。传统的对齐方法是基于运动估计的，例如光流估计。这类方法计算量大，而且光流估计很难做到准确，尤其对于相邻视频帧之间存在较大抖动的情况下。不准确的运动估计可能会在恢复出的视频帧中引入伪像，严重影响网络的性能。另一类对齐方法是利用卷积神经网络的特征提取能力，进行隐式对齐。但是视频序列中，不同的相邻帧和参考帧对之间存在不同程度的抖动，而卷积神经网络由于其固定的卷积核配置使其在几何变换建模方面具有局限性，所以无法使用普通卷积神经网络来处理这种不同程度的抖动。文献[20]提出了一种可变形的卷积运算，增强了普通卷积神经网络对几何变换的自适应建模能力。其后文献[21]将可变形卷积网络应用到视频超分辨率重建任务中，在特征层面进行相邻帧对齐。但是由于其网络分支中采样位置偏移的生成是没有监督信号的，其网络解释性较差，而且训练过程非常不稳定，导致模型结果难以复现。本文的研究内容之一就是研究如何解决相邻视频帧精确对齐的问题。我们将基于可变

形卷积网络，结合多尺度的策略尝试解决这个问题。并且我们会对比将相邻视频帧在图像层面对齐和特征层面对齐的差异。进一步，我们将改进可变形卷积网络的计算方式，引入光流估计对采样位置偏移的生成过程进行约束，尝试解决可变形卷积网络训练不稳定的问题。

（2）自适应时序特征融合。　　在相邻视频帧对齐后，如何高效地融合它们的时序特征是本文的另一个研究重点。相邻帧和参考帧对齐之后，它们同一空间位置处的特征具有对应的时序关系。我们需要在所有的空间位置处，分别融合这些具有时序关系的特征。由于不同相邻视频帧像素模糊程度不同，时序融合时不同视频帧所占权重就应不同；而同一个视频帧不同空间位置处像素模糊程度也不同，它们的融合权重也应该不同。所以对齐后的时空特征，在时序特征融合过程中应该具有像素级聚合权重（特定于时序位置和空间位置）。

1.4.2 文章的组织结构

如图 1-4 所示，这是我们文章的组织层次和结构。首先是我们的研究目标：设计和实现可以有效去除视频中非均匀模糊的算法。第 1 章我们介绍了去模糊任务的研究背景和研究意义，梳理了图像去模糊和视频去模糊的研究现状以及该领域尚未解决的问题。第 2 章是普通卷积网络、可变形卷积网络和动态滤波网络的理论基础。我们的研究思路是利用相邻视频帧中的清晰像素融合参考帧的模糊像素，充分挖掘输入视频序列中的时空信息。要实现这个想法，需要解决两个问题：相邻视频帧精确对齐和自适应时序特征融合，这也是我们论文的主要研究内容。第 3 章我们提出了第一个研究方案：基于多尺度可变形卷积网络的视频去模糊算法。该方案分别利用多尺度可变形卷积网络和 1x1 卷积网络解决上述两个问题。第 4 章我们提出了改进的方案：基于自适应时空卷积网络的视频去模糊算法。在对齐部分，我们改进了可变形卷积网络，提出了光流约束的增强可变形卷积。其次，我们构建了动态滤波网络实现自适应的像素级时序特征融合。在第 4 章，针对我们提出的相邻帧对齐模块和自适应时序特征融合模块，做了消融实验并对实验结果进行了分析和讨论。实验结果表明，我们的第二个研究方案相较于第一个研究方案取得了很大的性能提升。

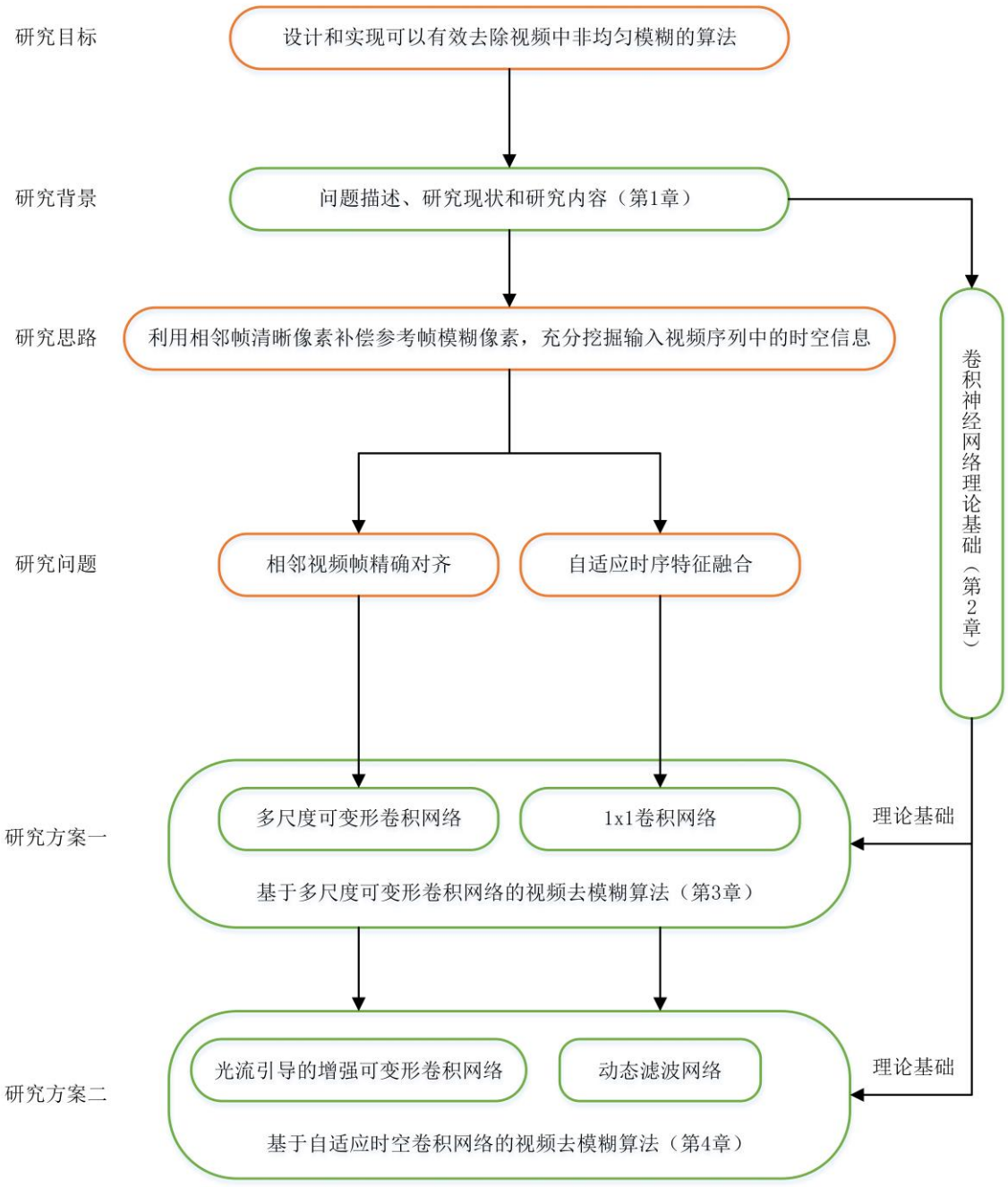


图 1-4　文章的组织结构

第 2 章 卷积神经网络理论基础

2.1 引言

卷积神经网络这一术语来自于对生物学中大脑视觉皮层的研究，该层中视觉神经细胞对特定视觉区域很敏感。一些特别的神经细胞只会对特定方向的边缘做出响应。比如，某些神经元会对垂直边缘做出响应，而其他的则会对水平或者斜边缘做出反应。在生物大脑中的存在一种局部敏感和方向选择的神经元网络结构，这种结构可以有效降低神经网络的复杂程度，这也就是卷积神经网络的生物理论基础。卷积神经网络就是受生物启发的一种多层人工神经网络，而人工神经网络经历了 M-P 神经元、单层感知机和多层前馈神经网络等阶段的发展。1957 年，罗森布拉特从生物神经系统的角度，一步步抽象并构建起感知机模型，发明了单层感知机，其结构如图 2-1 所示。

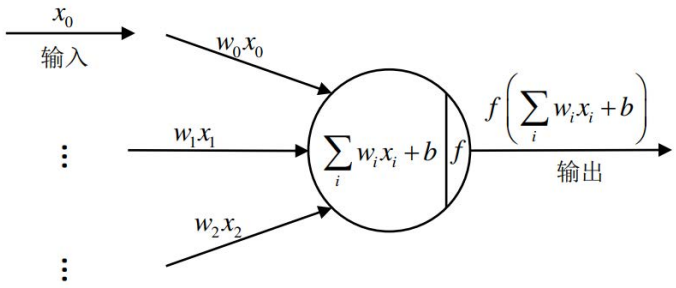


图 2-1　单层感知机示意图

单层感知机的计算过程可以用公式（2-1）表示：

$$Y = f(\sum_i w_i x_i + b) \tag{2-1}$$

其中 x_i 表示感知机的外部输入数据，w_i 表示 x_i 对应的权重，b 代表偏置项，f 是激活函数。单层感知器属于单层前向网络，即除输入层和输出层之外，只拥有一层神经元节点。输入数据从输入层经过隐藏层向输出层逐层传播，相邻两层的神经元之间相互连接，同一层的神经元之间没有连接。单层感知机的局限性

在于只能解决线性可分的问题，但是它不能解决类似异或这种线性不可分的问题。为了解决这个问题，多层感知机即人工神经网络被提出来。多层感知机在输入输出层中间加了多个隐层。最简单的多层感知机只含一个隐层，即三层的结构，如图 2-2 所示。多层感知机通过对线性分类器的组合叠加，具有拟合非线性函数的能力，从而可以解决线性不可分的问题。多层感知机成为了后来名声大噪的卷积神经网络的实现基础。

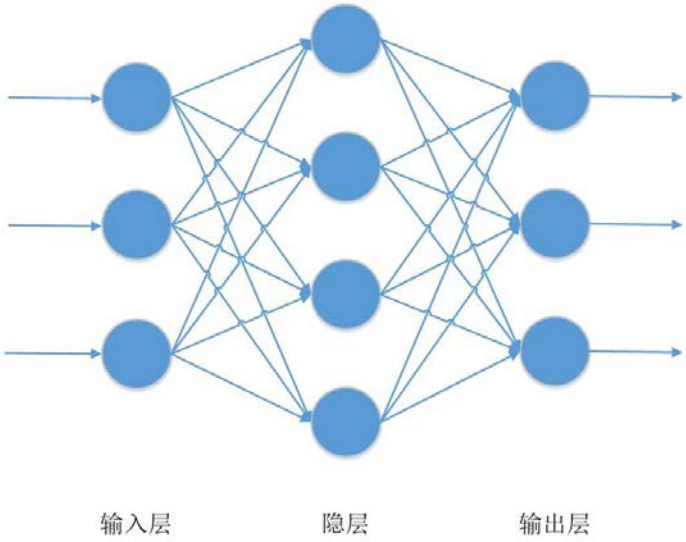


图 2-2　多层感知机示意图

卷积神经网络自从出现之后，就被应用到各种计算机视觉任务中并取得了很大的成功。在普通卷积神经网络提出之后，又有许多其他优秀的网络变体不断被提出，并且在某些方面拥有超越普通卷积神经网络的能力。本章介绍普通卷积神经网络及其变体网络的理论基础，变体网络包括本文使用到的可变形卷积网络和动态滤波网络。所以本章 2.2 节、2.3 节和 2.4 节将分别介绍普通卷积神经网络、可变形卷积网络和动态滤波网络。其中可变形卷积网络又分为基础的可变形卷积网络和增强可变形卷积网络，动态滤波网络根据滤波器生成网络的不同可以被实例化成动态卷积层和动态局部滤波层两种类型。

2.2 普通卷积神经网络

普通卷积神经网络和上一节讲的人工神经网络非常相似：它是由包含权重和偏置项的神经元组成。所有的神经元将其输入数据进行加权求和运算，得到的值再输入到激活函数中进行计算。但是卷积神经网络的结构基于一个假设：输入数据是图像。基于该假设，卷积神经网络向结构中添加了一些特有的性质。这些特有属性使得前向传播函数实现起来更高效，并且大幅度降低了网络中参数的数量。与人工神经网络不同，卷积神经网络各层中的神经元是 3 维排列的：宽度、高度和深度（这里的深度指的是激活数据体的第三个维度，而不是整个网络的深度，整个网络的深度指的是网络的层数）。如图 2-3 所示，是人工神经网络和卷积神经网络的结构示意图。我们可以看到，卷积神经网络中的神经元采取局部连接的方式而不是全连接方式。

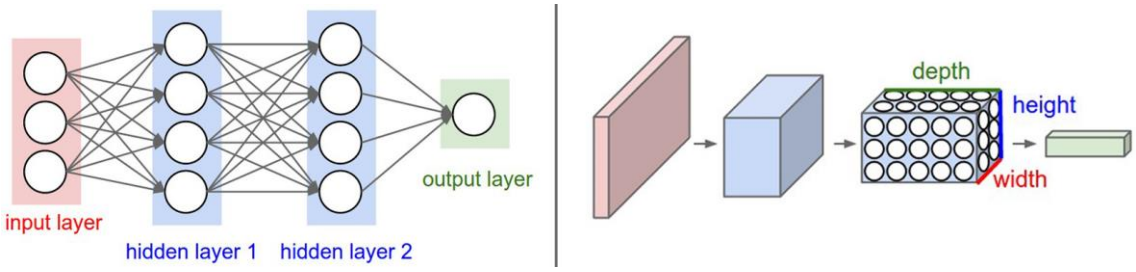


图 2-3　人工神经网络和卷积神经网络结构示意图

卷积神经网络通常由以下几部分构成：卷积层、池化层、全连接层和激活函数。

（1）卷积层　　卷积层产生了卷积神经网络中大部分的计算量。卷积层通过卷积核对输入数据进行卷积运算。如图 2-4 所示是卷积计算过程示意图，我们以二维卷积计算为例进行说明。卷积计算将各个位置上滤波器的元素和输入的对应元素相乘，然后再求和。最后将这个结果保存到输出的对应位置。将这个卷积过程在所有位置都进行一遍，就可以得到卷积运算的输出。

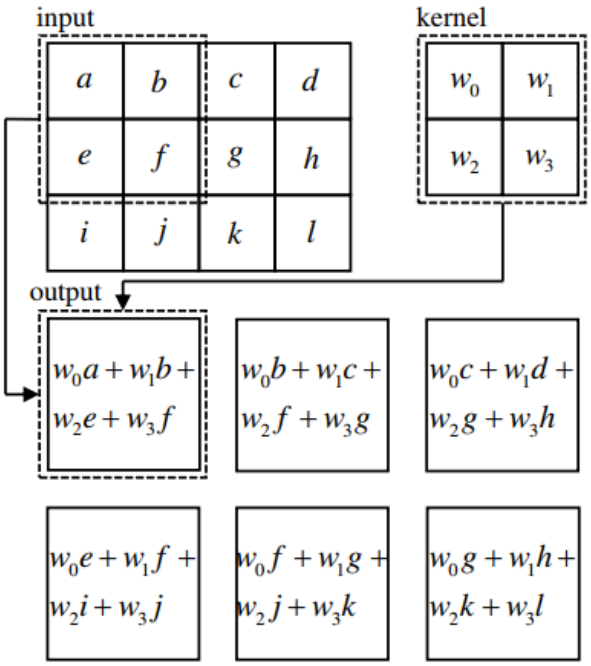


图 2-4　卷积计算过程示意图

在进行卷积层的处理之前，有时要向输入数据的周围填入固定的数据（比如 0 等），这称为填充。使用填充主要是为了调整输出的大小。卷积核在输入图像上滑动的距离，则称为步长。卷积核每个采样点之间的距离，称为间隔。间隔大于 1 的卷积，叫做空洞卷积。空洞卷积通过在卷积核采样点之间插 0 的方式，获得比普通卷积更大的感受野。除了普通卷积和空洞卷积之外，还有许多其他卷积方式涌现出来，在很多任务上取得了优秀的表现。比如深度可分离卷积、可变形卷积、动态卷积等。

（2）池化层　　卷积层通过卷积运算对输入图像进行特征提取和下采样，降低了图像维数。但降维后的图像维数还是很高。而池化层可以更好地对图像进行降维，用更高层次的特征表示图像。池化层还可以降低信息冗余、提升模型的尺度不变性、旋转不变性以及防止网络过拟合。常见的池化层包括最大池化层、全局最大池化层、平均池化层、全局平均池化层、随机池化层等。

如图 2-5，展示了一个二维特征图的最大值池化过程。最大池化层选图像区域的最大值作为该区域池化后的值。

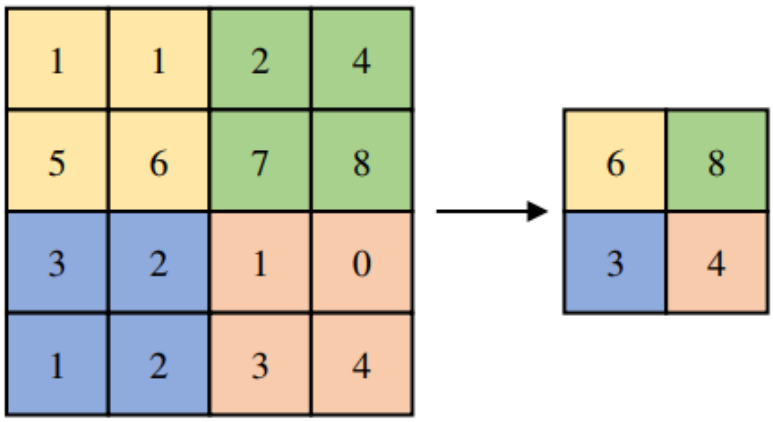


图 2-5　最大池化计算过程示意图

（3）全连接层　　如图 2-6 所示，全连接层中的每个神经元都会和前一层的所有神经元相连。全连接层可以对输入数据进行特征空间变换，再加上激活函数的非线性映射，多层全连接层理论上可以模拟任何非线性变换。但其缺点也很明显：无法保持空间结构，而且网络参数量很大。

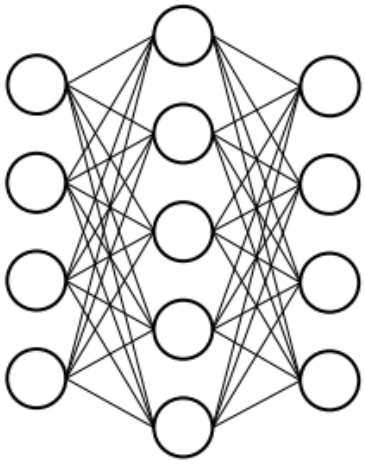


图 2-6　全连接层示意图

（4）激活函数　　激活函数是在卷积神经网络层间输入与输出之间的一种函数变换，目的是为了加入非线性因素，增强模型的表达能力。通过引入激活函数，卷积神经网络就可以拟合各种曲线。激活函数主要分为饱和激活函数和非饱和激活函数。常见的饱和激活函数有 Sigmoid 函数和 Tanh 函数；常见的非饱和激活函数有 ReLU 函数以及其变种。非饱和激活函数相较于饱和激活函数主要有如下优势：可以解决梯度消失问题、可以加速网络收敛。如图 2-7 所示，是这三种常见激活函数的函数曲线。

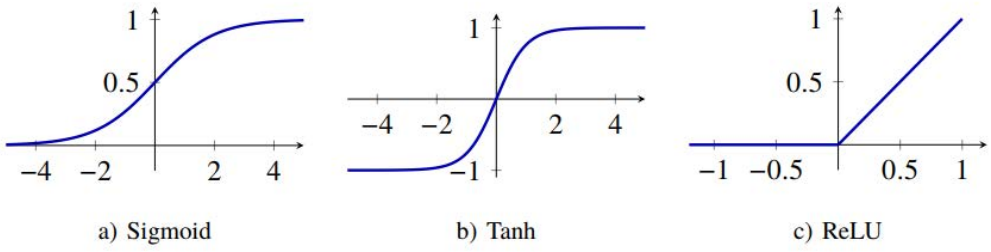


图 2-7　三种激活函数曲线

Sigmoid 激活函数的计算公式如（2-5）所示，它将输入值映射到（0,1）范围内。从图 2-7（a）可以看出，Sigmoid 激活函数连续、光滑且严格单调，以(0,0.5)中心对称，是一个非常良好的阈值函数。

$$\mathrm{Sigmoid}(x) = \frac{1}{1 + e^{-x}} \tag{2-5}$$

然而，Sigmoid 激活函数也有其自身的缺陷，最明显的就是饱和性：其两侧导数逐渐趋近于 0。Sigmoid 激活函数极容易导致梯度消失问题。饱和神经元会使得梯度消失问题雪上加霜，假设神经元输入 Sigmoid 的值特别大或特别小，对应的梯度约等于 0，即使从上一步传导来的梯度较大，该神经元权重和偏置的梯度也会趋近于 0，导致参数无法得到有效更新。

Tanh 激活函数把输入值映射到 $[-1, 1]$ 之间，其计算公式如公式（2-6）所示。Tanh 激活函数解决了 Sigmoid 激活函数收敛变慢的问题，相对于 Sigmoid 激活函数提高了网络收敛速度。但是它和 Sigmoid 激活函数一样，存在梯度消

失问题。因为 Tanh 激活函数两边的饱和性使得网络梯度消失，导致模型难以训练。

$$\mathrm{Tanh}(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}} \tag{2-6}$$

ReLU 激活函数的提出就是为了解决 Sigmoid 激活函数和 Tanh 激活函数存在的梯度消失问题。ReLU 的梯度只可以取两个值：0 或 1。当输入小于 0 时，梯度为 0；当输入大于 0 时，梯度为 1。好处就是 ReLU 梯度的连乘不会收敛到 0，连乘的结果也只可以取两个值：0 或 1 。如果值为 1，梯度保持值不变进行前向传播；如果值为 0 ,梯度从该位置停止前向传播。Sigmoid 函数是双侧饱和的，即朝着正负两个方向函数值都会饱和；但 ReLU 函数是单侧饱和的，即只有朝着负方向，函数值才会饱和。假设神经元为检测某种特定特征的开关，高层神经元负责检测抽象的高级特征，其有着更丰富的语义信息，例如眼睛或者轮胎；低层神经元负责检测具象的低级特征(曲线或者边缘)。当开关处于开启状态，说明在输入范围内检测到了对应的特征，且正值越大代表特征越明显。加入某个神经元负责检测边缘，则正值越大代表边缘区分越明显。假设一个负责检测边缘的神经元，激活值为 1 相对于激活值为 0.5 来说，检测到的边缘区分地更明显；但激活值-1 相对于-0.5 来说就没有意义了，因为低于 0 的激活值都代表没有检测到边缘。所以用一个常量值 0 来表示检测不到特征是更为合理的，像 ReLU 这样单侧饱和的神经元就满足要求。使用 ReLU 激活函数在计算上也是高效的。相对于 Sigmoid 函数梯度的计算，ReLU 函数梯度取值只有 0 或 1。且 ReLU 函数将负值截断为 0 ，为网络引入了稀疏性，进一步提升了计算高效性。但是这样做也可能阻碍训练过程。如果对于所有的样本，该激活函数的输入都是负的，那么该神经元再也无法学习，这就是神经元”死亡“问题。Leaky ReLU 激活函数的提出就是为了解决神经元”死亡“问题。Leaky ReLU 函数在输入大于 0 的部分与 ReLU 函数一样，仅在输入小于 0 的部分有差别。ReLU 输入小于 0 的部分值都为 0，而 LeakyReLU 输入小于 0 的部分值为负，且有微小的梯度。这就可以解决 ReLU 函数存在的神经元“死亡”问题。

2.3 可变形卷积网络

2.3.1 基础可变形卷积网络

普通卷积神经网络因为它的卷积核具有固定的几何结构，所以其几何变换能力有限。普通卷积神经网络的卷积核，针对不同的网络输入具有相同的权重和形状，这对于很多计算机视觉任务是不合理的，比如不同尺寸物体的目标检测、动态场景中的非均匀模糊去除。为增强普通卷积网络的自适应几何变换建模能力，文献[29]提出了可变形卷积：由输入特征学习得到的偏移量来改变标准卷积的采样位置，更好的适应视觉识别任务中物体的几何变换（大小、姿态、角度等）。可变形卷积网络的输入不同，其卷积核的采样位置就不同。决定卷积核采样位置的参数根据输入自适应的生成。

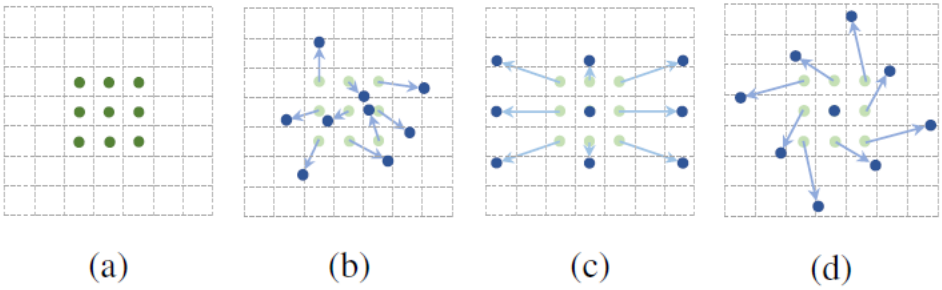


图 2-8　常规卷积和可变形卷积采样位置示意图

如图 2-8 中所示，是常规卷积神经网络和可变形卷积网络的卷积核采样位置示意图。图 2-8（a）是标准3×3卷积的常规采样网格；图 2-8（b）则是可变形卷积中,原采样点增加偏移量后的变形采样位置；图 2-8（c）和图 2-8（d）是图 2-8 (b)的特例，分别是采样尺寸放大和采样尺寸放大加旋转的情况。在卷积操作中，我们以坐标化的形式来表示接受域 R，其定义了卷积操作中感受野的大小。此处以3×3卷积为例，卷积核元素间隔设置为 1。则 R 如公式（2-9）所示：

$$R = \{(-1,-1),(-1,0),\ldots,(0,1),(1,1)\} \tag{2-9}$$

常规 2D 卷积运算包括两个步骤：（1）在输入特征图 x 上使用规则网格 R 进行采样（2）使用权重参数 w 对所有采样点进行加权求和。对于每个位置 p_0 在输出特征图 y 上的值如公式（2-10）所示：

$$y(p_0) = \sum_{p_n \in R} w(p_n) \bullet x(p_0 + p_n) \tag{2-10}$$

其中 p_n 枚举 R 中的所有采样位置。在可变形卷积中，规则采样位置的集合 R 中的每一个采样点，额外学习一个偏移作为最终的采样位置，此时 R 不再是规则的正方形区域。可变形卷积的运算过程可表示为公式（2-11）：

$$y(p_0) = \sum_{p_n \in R} w(p_n) \bullet x(p_0 + p_n + \Delta p_n) \tag{2-11}$$

其中 Δp_n 表示每个特征点的偏移量，其值通常为分数，在卷积神经网络中通过双线性插值使每个特征点的偏移量的学习是可微的。

如图 2-9 所示，是可变形卷积网络的架构。可变形卷积包括两个分支：（1）上面的分支根据输入的特征图生成参数化的滤波器，这里生成的参数是输入特征图中所有特征点的偏移量。因为每一个特征点都有 x 和 y 两个方向的偏移，所以对于通道数为 N 的输入特征图，会在该分支生成通道数为 2N 的偏移特征图。（2）下面的分支将生成的所有特征点的偏移值作用于输入的特征图，进行可变形卷积运算。

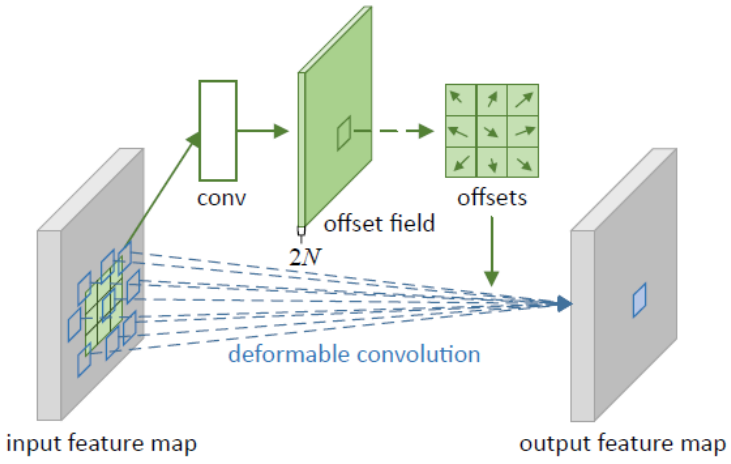


图 2-9　可变形卷积示意图

2.3.2 增强的可变形卷积网络

可变形卷积网络的优越性能源于其适应物体几何变化的能力。虽然其神经特征的空间支持比常规卷积神经网络更接近对象结构，但这种支持可能远远超出感兴趣区域，导致特征受到不相关图像内容的影响，恶化目标任务的性能。为了解决这个问题，文献[30]提出了增强的可变形卷积网络，通过增加建模能力和加强训练来提高其关注相关图像区域的能力。通过在网络中更全面地集成可变形卷积，并通过引入扩展变形建模范围的调制机制来增强网络建模能力。为了进一步增强可变形卷积网络的空间区域支持能力，改进的可变形卷积网络引入了一种调制机制。有了这种机制，新的可变形卷积网络不仅可以调整特征采样点的位置偏移量，还可以调制来自不同空间位置的特征点采样权重。在特殊情况下，新的可变形卷积网络可以通过将其特征点采样权重设置为零来决定不感知来自特定位置的特征或信号。因此，来自相应空间位置的图像内容将不会影响模块的输出。所以调制机制为网络模块提供了另一个维度来调整其空间支持区域。

给定具有 K 个采样位置的卷积核，让 w_k 和 p_k 分别表示第 k 个采样位置的卷积核权重和位置偏移量。例如 K 的值为 9，则 p_k 可以表示为公式（2-12）

$$p_k = \{(-1,-1),(-1,0),..,(0,1),(1,1)\} \tag{2-12}$$

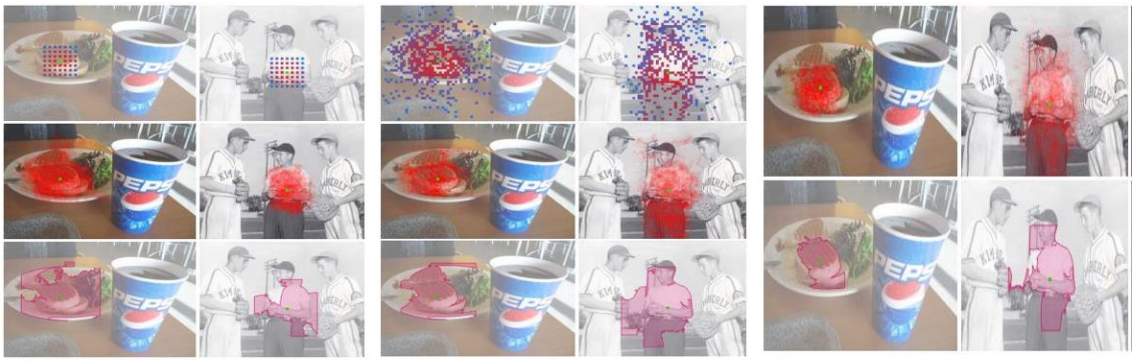
其定义了一个 3 x 3 的卷积核采样位置。增强的可变形卷积网络的计算过程可以表示为公式（2-13）：

$$y(p) = \sum_{k=1}^{K} w_k \bullet x(p + p_k + \Delta p_k) \bullet \Delta m_k \tag{2-13}$$

其中 $x(p)$表示输入特征图 x 在位置 p 处的特征值，$y(p)$表示输出特征图 y 在位置 p 处的特征值。Δp_k 和 Δm_k 分别表示第 k 个特征采样点可学习的位置偏移量和调制权重。调制权重 Δm_k 的值位于[0, 1]范围内，而 Δp_k 是范围不受限制的实数。Δp_k 决定了卷积计算过程中的采样位置，调制权重 Δm_k 决定了来自不同特征采样点的权重。$p + p_k + \Delta p_k$ 通常是小数，所以通过双线性插值来计算

$x(p + p_k + \Delta p_k)$位置处的特征值。Δp_k和Δm_k都是通过应用在输入特征图 x 上的单独卷积层 layer 获得的。卷积层 layer 对输入特征图 x 做卷积计算后，得到的输出结果和输入特征图 x 具有相同的空间分辨率，但通道数是 $3K$。其中前 $2K$ 通道的特征值对应偏移量 Δp_k，剩下的 K 个通道的特征值进一步输入到 Sigmoid 激活函数中，获得调制权重 Δm_k。增强的可变形卷积网络通过增加了 Δp_k和Δm_k这两个自由度，使其具有更强的几何形变建模能力和自适应特征提取能力。

如图 2-10 所示，是常规卷积网络、基础可变形卷积网络和增强的可变形卷积网络的空间支持示意图。在每个子图中，从上到下的每行分别显示了有效采样位置、有效感受野和有误差的显着性区域。2-10（c）中省略了有效采样位置，因为它们与 2-10（b）中的相似。可以看出，增强的可变形卷积网络有着比常规卷积网络和基础可变形卷积网络更大的有效感受野、更小的有误差显著性区域。



(a) regular conv　　(b) deformable conv　　(c) modulated deformable conv

图 2-10　不同卷积网络的空间支持示意图

2.4 动态滤波网络

常规卷积网络中，卷积核的参数在网络训练完成后，对任意输入都是固定的。而动态滤波网络中，滤波器的参数是根据输入自适应生成的，其随着输入的变化而变化。输入不同，生成的滤波器参数就随之不同。动态卷积网络最早

是在文献 [28]中提出的。如图 2-11 所示，其网络结构包括两个分支：（1）滤波器生成网络：以输入为条件产生滤波器（2）动态滤波层：将生成的滤波器应用于另一个输入。

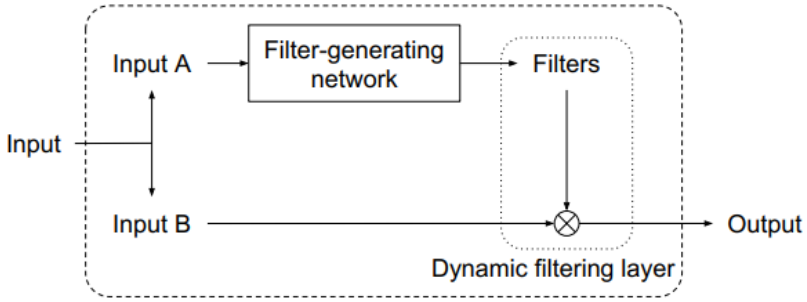


图 2-11　动态滤波网络框架图

这种架构由于其自适应特性而具有更强的灵活性，而且滤波器是根据输入动态生成的，没有需要学习的参数，所以模型参数的数量没有过多增加。通过这种方式可以学习各种各样的滤波操作，包括局部空间变换，这种变换是特定于输入、特定于位置的，所以具有自适应性，更加适用于需要网络具有几何形变建模能力的任务，例如动态场景的非均匀模糊去除、不同尺寸的物体检测等。而常规卷积网络由于其固定的几何结构，不能有效地对几何形变过程进行建模。

2.4.1 滤波器生成网络

滤波器生成网络的输入是 I_A，如公式（2-14）所示。其中 h、w、c_A 分别是输入 A 的高度、宽度和通道数。

$$I_A \in \mathrm{R}^{h\times w\times c_A} \tag{2-14}$$

滤波器生成网络的输出是被参数 θ 参数化的滤波器 F_θ，θ 如公式（2-15）所示：

$$\theta \in \mathrm{R}^{s\times s\times c_B\times n\times d} \tag{2-15}$$

其中 s 是滤波器大小，决定感受野，可以根据目标任务进行选择。感受野的大小也可以通过堆叠多个动态滤波器模块来增加。这在可能涉及较大局部位移的

应用中很有用。c_B是输入 B 的通道数，n 是滤波器数目，d 视动态滤波层的类型而定：当动态滤波层是动态卷积时，$d = 1$；当动态滤波层是动态局部滤波时，$d = h \times w$。将生成的滤波器F 应用于输入I_B，I_B如公式（2-16）所示：

$$I_B \in \mathrm{R}^{h \times w \times c_B} \tag{2-16}$$

得到最终的输出$G \in \mathrm{R}^{h \times w \times n}$如公式（2-17）所示：

$$G = F_\theta(I_B) \tag{2-17}$$

滤波器生成网络根据动态滤波层的不同，生成了不同的滤波器F。动态滤波层可以被实例化为动态卷积层和动态局部滤波层两种类型。如图 2-12 所示，是它们的网络结构。当动态滤波层是动态卷积层时，滤波器生成网络生成了单个滤波器在输入I_B上进行卷积计算；当动态滤波层是动态局部滤波层时，输入I_B每个空间位置周围的特征值都使用特定于位置的动态生成的滤波器进行卷积计算。

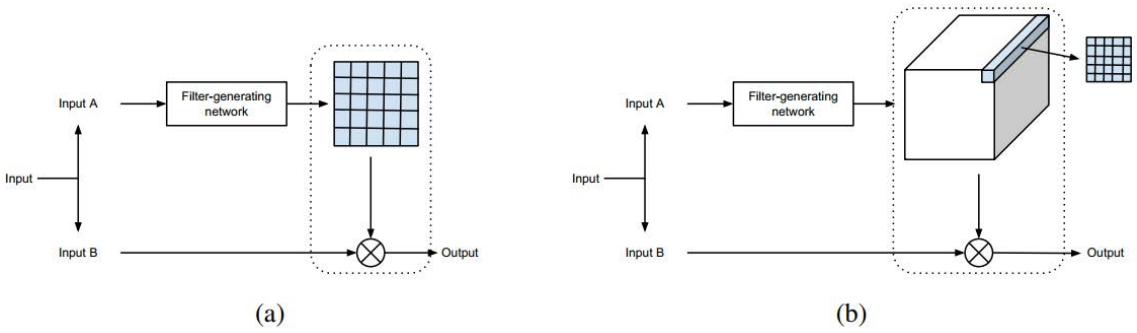


图 2-12　动态卷积层（左）和动态局部滤波层（右）示意图

2.4.2 动态滤波层

动态滤波层将图像或特征图I_B作为输入，并输出滤波后的结果G。动态滤波层可以被实例化为动态卷积层和动态局部滤波层两种类型。

动态卷积层类似于传统的卷积层，它们在输入I_B的每个空间位置都应用了同样的滤波器。但传统的卷积层中卷积核权重是模型参数，而在动态卷积层中，

滤波器参数θ是中间的计算结果，其由滤波器生成网络动态生成。动态卷积层中的滤波器是特定于样本的，并以滤波器生成网络的输入为条件。其滤波运算如公式（2-18）所示：

$$G(i,\ j)\ =\ F_{\theta}(I_B(i,\ j)) \tag{2-18}$$

在动态局部滤波层中，对于输入I_B的每一个位置$(i,\ j)$都有一个特定的局部滤波器应用于以(i,j)为中心、s为边界长度的正方形区域$I_B(i,\ j)$。其运算过程如公式（2-19）所示：

$$G(i,\ j)\ =\ F_{\theta}^{(i,\ j)}(I_B(i,\ j)) \tag{2-19}$$

动态局部滤波层中的滤波操作，不仅是特定于输入的，而且是特定于位置的。

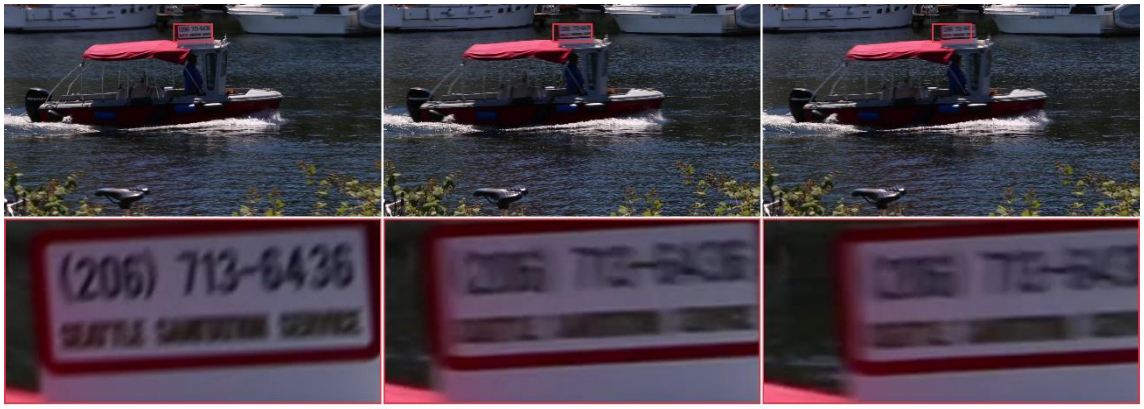
2.5 本章小结

本章我们讲解了卷积神经网络的理论基础，包括普通卷积神经网络、可变形卷积网络和动态滤波网络。普通卷积神经网络小节，详细介绍了其卷积层、池化层、全连接层、激活函数和归一化层五部分组成以及其局部连接、权值共享的特点。可变形卷积网络小节，分别介绍了基础可变形卷积网络和增强可变形卷积网络。它们相较于普通卷积网络，拥有更强的几何形变建模能力和自适应特征提取能力，所以适用于空间变化的任务，例如动态场景的非均匀去模糊。增强可变形卷积网络在基础可变形卷积网络之上，引入了调制权重，使网络不仅可以调整感知输入特征的偏移量，还可以调制来自不同空间位置的输入特征幅度。最后一节介绍了动态滤波网络，其组成包括滤波器生成网络和动态滤波层两部分。其中动态滤波层又可以被实例化为动态卷积层和动态局部滤波层两种类型。动态滤波网络在没有增加模型参数的前提下，增强了网络的自适应特性而具有更强的灵活性，更加适用于需要网络具有几何形变建模能力的任务。由于我们的目标任务是视频中动态场景的非均匀模糊去除，它具有空间变化的特点，即视频帧中不同空间位置处的像素，模糊程度不同。所以可变形卷积网络和动态滤波网络就可以应用于我们的任务。在下面章节中，我们利用了基础的可变形卷积网络进行视频去模糊中多尺度的特征提取和相邻视频帧对齐。

第 3 章 基于多尺度可变形卷积的视频去模糊算法

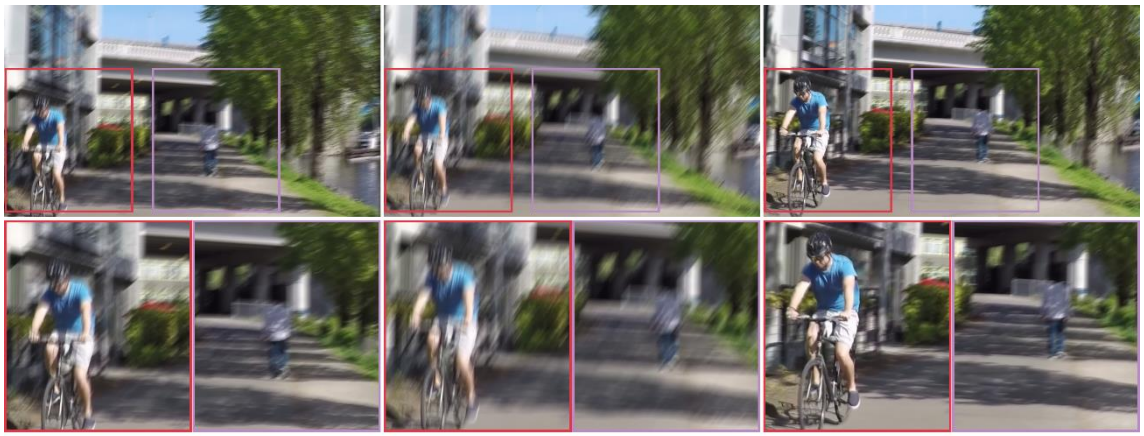
3.1 引言

通过观察 DVD[9]数据集中的视频序列，我们发现这样一个现象：对于同一个视频序列中的连续视频帧，它们对应像素的模糊程度并不一样。由于失焦、目标运动、相机抖动等原因，造成在某一视频帧的一些像素非常模糊，而在其相邻视频帧对应的像素却是清晰的，如图 3-1 和图 3-2 所示。



(a)T-1 (b)T (c)T+1

图 3-1 DVD[9]数据集中的 Boat 视频序列



(a)T-1 (b)T (c)T+1

图 3-2 DVD[9]数据集中的 Street 视频序列

基于这样的观察，我们构建了如下研究思路：可以利用相邻视频帧中的清晰像素补偿参考帧中的模糊像素，充分利用输入连续视频帧中的时空信息，以提升视频去模糊算法的性能。算法的目标就是利用相邻帧中冗余的空域信息，对参考帧中的特征做融合。由于相机抖动或者目标运动等原因，相邻帧和参考帧之间存在抖动，所以在特征融合之前需要先执行相邻视频帧的对齐。我们利用可变形卷积网络的几何形变建模能力，在特征层面实现相邻视频帧的隐式对齐。对齐后的相邻帧和参考帧特征，在同一空间位置处具有对应的时序关系。此时我们利用 1x1 卷积网络进行时序特征融合，相邻帧中的像素质量越高，融合权重就应该越大。

3.2 算法原理概述和整体网络结构

为了充分地利用视频帧间的空域信息，我们将视频去模糊分为特征提取、相邻帧对齐、特征融合以及特征重建四部分。算法的输入是视频序列中的连续 N 帧，输出是去模糊后的中间帧。中间帧称为参考帧，其他帧称为相邻帧。算法的结构如图 3-3 所示（以连续 3 帧输入为例进行说明），整体是一个残差的结构。由于输入的相邻帧 B_{t+i} 和参考帧 B_t 之间存在抖动，需要将其输入到多尺度可变形卷积对齐模块中，将相邻帧对齐到参考帧所处的时刻。对齐模块输出对齐后的时空特征 Aligned Feature，然后先经过 1x1 卷积网络降低特征维度，再输入到卷积层和残差块组成的网络中进行特征融合，得到融合后的特征 C_t。融合的特征最后经过特征重建网络，获得重建的特征。参考帧 B_t 和重建的特征相加求和，得到清晰的参考帧 R_t。

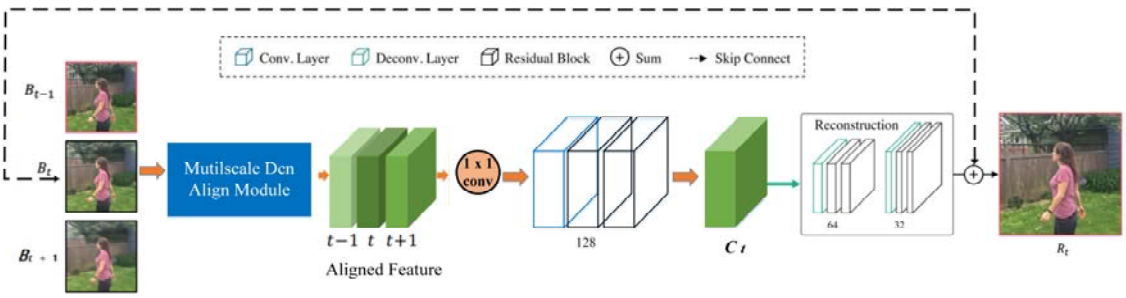


图 3-3　算法整体框架图

3.3 相邻视频帧对齐

3.3.1 基于光流估计的图像对齐和特征对齐

这一小节我们使用光流估计的方法，分别在图像层面和特征层面显示执行相邻视频帧的对齐。通过比较它们对齐结果的差异性，我们得到了在不同图像和特征尺度进行相邻帧对齐的一些思考，这些思考启示了我们利用多尺度策略结合可变形卷积网络的几何形变建模能力，实现相邻帧的隐式对齐。我们使用PWC-Net[39]作为光流估计网络，计算相邻帧和参考帧之间的光流。给定一个视频序列中的第 T 帧 F_t、第 T+I 帧 F_{t+i}，要将 F_{t+i} 对齐到 F_t 所处时刻，需要利用PWC-Net[39]网络计算出 F_{t+i} 到 F_t 之间的光流 $U_{t+i->t}$，如公式（3-1）所示：

$$U_{t+i->t} = N(F_t, F_{t+i}) \tag{3-1}$$

其中 N 是光流估计网络，即 PWC-Net[39]。我们将基于光流分别在图像层面和特征层面执行相邻视频帧的对齐。

（1）图像对齐。　　如图 3-4 所示，是我们从 DVD[9]数据集中某个视频序列里得到的第 T-2 帧、第 T 帧以及计算出的它们之间的光流 $U_{t-2->t}$。

(a)frame T-2　　(b)frame T　　(c)optical flow

图 3-4　第 T-2 帧、T 帧和它们的光流

如公式（3-2）所示，我们利用图像间的光流，对视频帧 F_{t-2} 进行文献[41, 42]中的 Warp 变形操作，先将其对齐到参考帧 F_t 所处时刻的状态；

$$\hat{F}_{t-2} = W(F_{t-2}, U_{t-2->t}) \tag{3-2}$$

公式（3-3）再将对齐后的图像利用卷积层进行特征提取和下采样，得到对齐后的相邻帧特征，如图 3-6（a）所示。

$$\hat{I}_{t-2} = C(\hat{F}_{t-2}) \tag{3-3}$$

其中 $\hat{F}_{t-2}$ 代表对齐后的第 T-2 帧，W 代表 Warp 变形操作，C 代表卷积操作。

（2）特征对齐。　　如公式（3-4）和公式（3-5）所示，我们对 F_{t-2} 和 F_t 分别执行卷积操作进行特征提取和下采样，得到对应的特征图 I_{t-2} 和 I_t；然后对光流 $U_{t-2->t}$ 进行下采样操作，如公式（3-6）所示得到特征图 I_{t-2} 和 I_t 之间的光流 $V_{t-2->t}$；

$$I_{t-2} = C(F_{t-2}) \tag{3-4}$$

$$I_t = C(F_t) \tag{3-5}$$

$$V_{t-2->t} = D(U_{t-2->t}) \tag{3-6}$$

其中 C 代表卷积操作，D 代表下采样。得到的特征图 I_{t-2} 和 I_t 和它们之间的光流 $V_{t-2->t}$ 如图 3-5 所示。

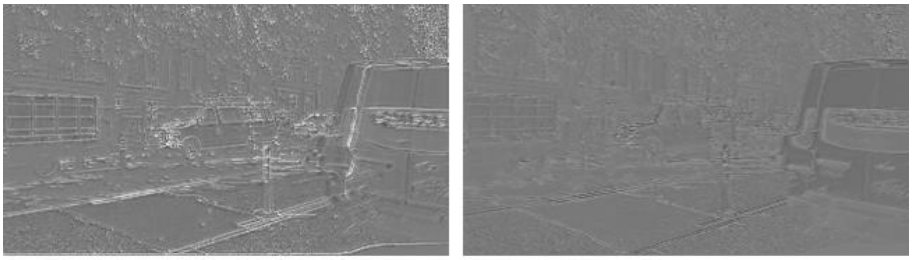
(a)feature T-2　　(b)feature T　　(c)optical flow

图 3-5　第 T-2 帧特征、T 帧特征和它们的光流

最后对 I_{t-2} 进行 Warp 变形操作，如公式（3-7）所示，将其对齐到 I_t 所处时刻的状态，得到对齐后的特征图如图 3-6（b）所示。

$$\hat{I}_{t-2} = W(I_{t-2}, V_{t-2->t}) \tag{3-7}$$

通过对比图 3-6（a）图像对齐和图 3-6（b）特征对齐的结果，我们发现：特征对齐相较于图像对齐保留了更多原始图像的纹理信息和高频信息，而图像对齐的结果有更多的噪点。通过模型最终去模糊的结果来看，采用特征对齐的方法比采用图像对齐的方法有更高的性能。



(a)Image Warping (b)Feature Warping

图 3-6 图像对齐结果(a)和特征对齐结果(b)

这启发了我们可以在不同尺度的特征层面对相邻视频帧进行对齐，融合不同尺度对齐的结果，采用从粗略到精细对齐的方式。在下一小节我们就采用了这种多尺度的策略，确定了视频去模糊算法最终采用的对齐方法。

3.3.2 基于多尺度可变形卷积网络的隐式特征对齐

在上一小节，我们探究了利用光流在图像层面和特征层面执行相邻视频帧对齐的差异性，发现在特征层面对齐的效果更好。然而对于模糊的图像，计算的光流是不准确的。不准确的光流，会导致视频帧在根据光流对齐之后引入伪影，影响视频帧质量。为了解决这一问题，我们借助可变形卷积网络的几何形变建模能力，实现了相邻视频帧的隐式特征对齐，避免了显示的运动估计。更进一步，我们受特征金字塔的启发，在对齐模块中采用了多尺度的策略，融合不同尺度特征对齐的结果。由于不同尺度的特征图，所处网络深度不同，其感受野和分辨率就不同。所以不同尺度的特征对齐结果，具有不同的语义信息表征能力和几何信息表征能力。而采用多尺度的策略，可以充分利用不同尺度特征的语义信息和几何信息，使得最终的对齐结果更精确。

如图 3-7 所示，是我们提出的基于多尺度可变形卷积网络的特征对齐模块。对于输入的任意相邻帧 $B_{t+i} = \mathrm{H} \times \mathrm{W} \times 3$，图中的实线代表卷积操作，不断的对输入视频帧进行特征提取和下采样，蓝色长方体代表卷积操作输出的未对齐的特征图 F_{t+i}，绿色长方体代表 Dcn Align 和卷积操作输出的对齐后的特征图。底部的数字是特征图的通道数。在 L1、L2、L3 提取的特征图的尺寸分别是

$16 \times H \times W$ 、 $32 \times \frac{1}{2} H \times \frac{1}{2} W$ 和 $64 \times \frac{1}{4} H \times \frac{1}{4} W$ 。

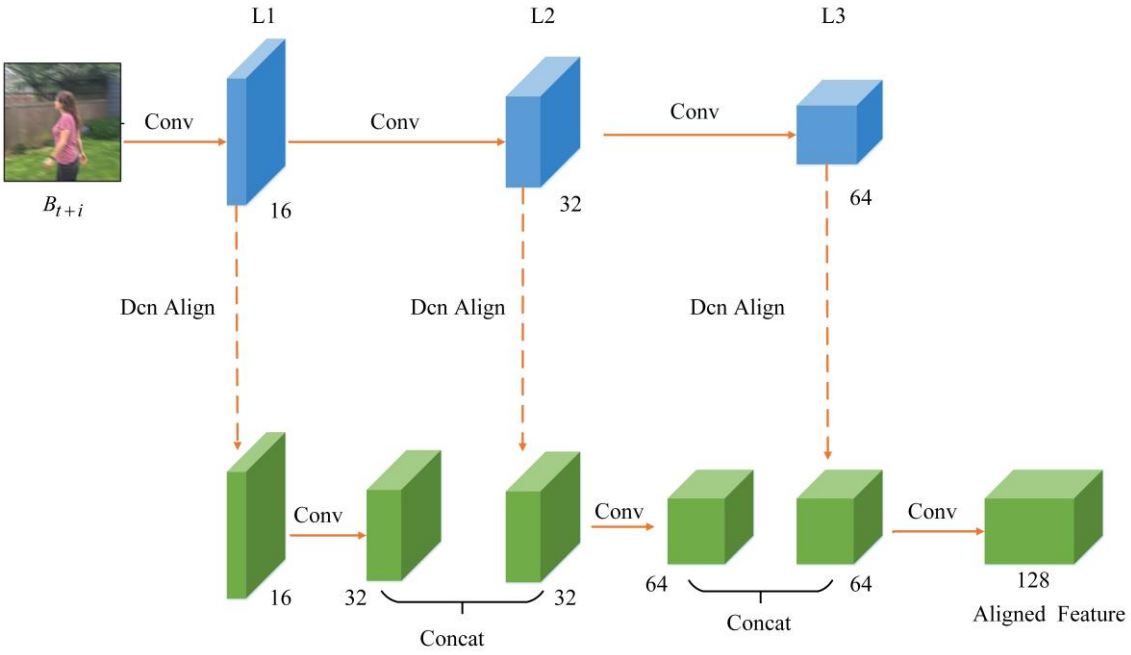


图 3-7 基于多尺度可变形卷积网络的特征对齐模块

在每一尺度具体对齐过程如下：在每个层次，我们利用可变形卷积[29]执行 Dcn Align，其计算过程如公式（3-8）所示：

$$\hat{F}_{t+i}(p) = \sum_{k=1}^{K} w_k \bullet F_{t+i}(p + p_k + offset_k) \tag{3-8}$$

其中 K 是卷积核采样数目， w_k 和 p_k 分别表示第 k 个特征采样点的卷积核权重位置偏移量。例如 K 的值为 9，则 p_k 可以表示为公式（3-9）

$$p_k = \{(-1,-1),(-1,0),\ldots,(0,1),(1,1)\} \tag{3-9}$$

其定义了一个 3 x 3 的卷积核采样位置。$offset_k$ 代表 F_{t+i} 中第 k 个采样位置处的特征点和参考帧对应特征点的位置偏移量，其生成过程如公式（3-10）所示：

$$offset = C([F_{t+i}, F_t]) \tag{3-10}$$

其中[,]代表连接操作，C 是卷积操作。公式（3-8）和可变形卷积[29]的不同之处

在于，*offset* 的生成方式以及它的物理意义。在公式（3-10）中，首先将同尺度的相邻帧特征 F_{t+i} 和参考帧特征 F_t 在通道维度连接起来，然后输入到一系列卷积层中去学习生成 *offset*，它的物理意义是相邻帧每个特征点和参考帧对应特征点的位置偏移量。所以公式（3-8），实现了在可变形卷积计算的过程中，根据位置偏移量 *offset* 去自适应采样与参考帧对应的特征点，实现了隐式的对齐。

在 L1 尺度对齐的特征图，通过卷积层提取特征后，传递到 L2 尺度并且和 L2 尺度对齐的特征图在通道维度拼接起来，通过卷积层实现不同尺度对齐结果的融合，再传递到 L3 尺度，后序流程依次类推。由于卷积神经网络通过逐层抽象的方式来提取图像特征，处于不同层次的特征图具有不同大小的感受野和分辨率。浅层网络的感受野小、语义信息表征能力弱，但分辨率高、几何细节信息表征能力强；而深层网络感受野比较大，语义信息表征能力强，但是特征图的分辨率低，几何信息的表征能力弱。所以对齐模块中我们设计了具有 L1、L2 和 L3 三个尺度的特征金字塔，在每个尺度都将相邻帧特征 F_{t+i} 对齐到参考帧特征 F_t 所处时刻的状态，并且在 L2 尺度融合 L1 尺度对齐的结果、L3 尺度融合 L2 尺度对齐的结果。采用这种从粗到细的多尺度对齐方式，不仅使网络最终输出的对齐特征具有良好的语义信息和空间信息，而且对齐结果更精细。

3.4 特征融合和特征重建部分

3.4.1 特征融合

视频去模糊任务中，相邻帧对齐和特征融合是最为关键的两个部分。如何高效地融合对齐后的相邻帧时空特征是该部分的核心。

由于在对齐模块中，执行了相邻视频帧的对齐。所以特征融合模块的输入，是在不同空间位置具有时序对应关系的相邻帧特征和参考帧特征。我们首先使用简单的 1x1 卷积网络，在通道维度实现初步的时序特征融合。融合过程中，卷积网络会将相邻帧清晰像素的特征和参考帧对应像素的特征加权求和。但由于普通卷积网络的卷积核参数在特征图不同空间位置处权值共享，所以在不同空间位置处其时序特征融合的权重是一样的。然而同一视频帧内，不同空间位

置处的像素模糊程度是不同的。所以它们的权重也应该不同。但为了减少模型参数量，我们这里就使用简单的 1x1 卷积网络实现融合。初步时序信息融合后的特征图，再输入到一个卷积层和两个残差块组成的网络中进一步融合特征，得到融合后的特征C_t。随着网络深度的增加，卷积神经网络的感受野变大，可以处理视频帧中存在的更大模糊。

3.4.2 特征重建

特征重建网络负责将融合后的特征C_t重建成和参考帧相同分辨率的残差图像，最终将该重建的残差图像和参考帧相加得到恢复出的清晰帧。重建网络由两个反卷积模块组成，每个反卷积模块由一个反卷积层和两个残差块组成。反卷积层对输入的特征进行上采样，残差模块可以很好地防止网络因为深度过深导致梯度消失。特征提取网络和特征重建网络整体是一个 Encoder-Decoder 结构，分别负责对输入进行特征降维下采样和特征升维上采样。特征提取会过滤掉图像中存在的噪声，所以 Encoder-Decoder 结构本身就利于图像去模糊等图像增强任务。

3.5 实验设置

我们的实验设置包括使用的训练数据集和测试数据集、数据增强方案、损失函数和网络训练方案等。

3.5.1 数据集和数据增强

在实验中，我们使用 DVD[9]数据集来训练该算法。该数据集包含 71 个视频（6708 个模糊清晰图像对），分为 61 个训练视频（5708 对模糊-清晰图像）和 10 个测试视频（1000 个模糊-清晰图像）。我们执行一些数据增强操作以进行网络训练。首先在训练和测试过程中，数据加载器每次返回由 5 个连续模糊帧组成的视频序列和该视频序列的中间帧对应的清晰图像。为了将运动多样性添加到训练数据中，随机反转视频序列的顺序。对于每个序列，执行相同的图像变换：（1）从[0.8，1.2]中均匀采样的亮度，对比度和饱和度等色度转换（2）图

像随机水平和垂直翻转（3）将图像随机裁剪为 256×256 像素的色块。最后为了使网络具有更好的泛化能力，将 N（0，0.01）的高斯随机噪声添加到所有输入图像中。

3.5.2 损失函数

为了有效训练该网络，我们使用均方误差（MSE）损失函数，如公式（3-11）所示，它衡量恢复的视频帧 R 及其对应标签的清晰视频帧 S 之间的差异：

$$L_{mse} = \frac{1}{CHW}\left\|R - S\right\|^2 \tag{3-11}$$

其中 C、H、W 分别代表图像的通道数、高度和宽度。

3.5.3 模型实现和训练细节

我们使用 Xavier[42]初始化方法对网络进行初始化，并使用 Adam[43]优化器对其进行训练，其中 β1= 0.9，β2= 0.999。由于可变形卷积网络在学习特征偏移的过程中不稳定，所以对齐模块的初始学习率设为较小的 1e-6，其余模块初始学习率设为 1e-4，整个网络一起进行端到端的训练。所有模块的学习率每 400k 次迭代就衰减 0.1 倍，网络最终在 850k 次迭代后收敛。我们使用 Pytorch 框架实现该算法，并在 NVIDIA GeForce 2080Ti 上进行训练和测试。

由于对齐模块中可变形卷积网络生成特征位置偏移量的过程不可解释，导致网络训练过程不稳定，极端情况下会出现位置偏移值溢出的情况。可变形卷积的特征采样点位置偏移值溢出，导致采样点的特征值为 0，此时视频去模糊退化成单幅图像去模糊。相邻帧中存在的空域信息无法利用，网络性能受到影响。我们采取的训练策略是在每一轮 epoch 都保存完整的模型检查点，如果在某一轮 epoch 训练过程中出现上述情况，则模型退回到上一轮的训练结果重新开始训练。

3.6 实验结果和分析

我们在 DVD[9]数据集和 GOPRO[8]数据集上对该算法进行了定量评估和定

性评估。

3.6.1 算法定量评估结果

为了评估算法的性能，我们将其与近几年的视频去模糊算法在 DVD[9]数据集和 GOPRO[8]数据集上进行了定量比较，如表 3-1 和表 3-2 所示。实验中使用 PSNR 和 SSIM 作为评估指标，它们反映了每个算法的准确率。

表 3-1　各算法在 DVD[9]数据集上的定量评估结果

Method	Tao[10]	Su[9]	STFAN[22]	Xiang[35]	TSP[30]	Suin [36]	Ours
PSNR	29.98	30.01	31.15	31.68	32.13	32.53	32.15
SSIM	0.8842	0.8877	0.9049	0.9157	0.9268	0.9468	0.9302

表 3-2　各算法在 GOPRO[8]数据集上的定量评估结果

Method	Tao[10]	Su[9]	STFAN[22]	Nah[37]	TSP[30]	Suin [36]	Ours
PSNR	30.29	27.31	28.59	29.97	31.67	32.10	31.72
SSIM	0.9014	0.8255	0.8608	0.8947	0.9279	0.9600	0.9283

3.6.2 算法定性评估结果

为了验证算法的泛化能力，我们对其在 DVD[9]测试数据集上进行了定性测试。如图 3-6、3-7、3-8 和 3-9 所示,是各算法在 DVD[9]测试数据集中的去模糊结果。以图 3-6 为例，图 3-6（a）是模糊的参考帧，图 3-6（b）是对应的真值，图 3-6（c）、（d）和（e）分别是 EDVR[27]、PVDNet[37]和 TSP[30]对参考帧的处理结果，图 3-6（f）是我们算法的处理结果。可以看出，相较于其他算法的处理结果，我们的算法恢复出了更多的模糊图像中的细节信息，可以有效地处理动态场景中的非均匀模糊。

图 3-8　DVD[9]测试数据集 IMG_0021 视频序列的去模糊结果

图 3-9　DVD[9]测试数据集 IMG_0030 视频序列的去模糊结果

图 3-10　DVD[9]测试数据集 IMG_0037 视频序列的去模糊结果

图 3-11　DVD[9]测试数据集 IMG_0039 视频序列的去模糊结果

3.7 本章小结

本章介绍了我们的第一个工作：基于多尺度可变形卷积网络的视频去模糊算法。该算法采用多尺度的策略结合可变形卷积网络的几何形变建模能力，在特征层面将相邻帧特征和参考帧特征进行对齐。算法之后使用简单的 1x1 卷积网络对对齐后的特征进行时序融合。为了一进步扩大卷积网络的感受野，使之可以处理输入视频帧中可能存在的更大的模糊，我们使用一个卷积层和两个残差块组成的网络继续融合特征。最后融合的特征输入到重建网络中进行特征重建，并和参考帧相加得到恢复出的清晰帧。我们在 DVD[9]和 GOPRO[8]数据集上对该算法进行了定量评估和定性评估，结果表明相较于近几年的视频去模糊算法，我们的算法具有较高的性能，而且由于我们使用了更简单高效的特征融合网络，模型参数量更少。

第 4 章 基于自适应时空卷积网络的视频去模糊算法

4.1 引言

视频去模糊任务的两个关键子问题分别是相邻视频帧的精确对齐和自适应时序特征融合。第 3 章的工作，我们利用多尺度可变形卷积网络和 1x1 卷积网络初步解决了这两个问题，但是该方案还可以进一步得到优化和改进。在本章，我们提出了基于自适应时空卷积网络的视频去模糊算法，它在执行相邻帧精确对齐的同时实现帧内的空域信息融合，并且基于动态滤波网络实现了自适应地像素级时序特征融合。该算法更好的解决了动态场景中非均匀模糊去除的问题，相较于第一个算法有了很大的性能提升。

4.2 原理概述和整体网络结构

同第 3 章一样，算法采用连续 5 帧作为输入来恢复出清晰的中间帧。连续 5 帧的中间帧表示为参考帧，其余帧表示为相邻帧。算法的研究思路是利用相邻视频帧中的清晰像素，融合参考帧中的模糊像素，充分挖掘输入视频序列中的时空信息。但由于相机抖动或者目标运动等原因，相邻帧和参考帧之间存在抖动，需要先执行对齐才能进一步利用相邻帧中冗余的空域信息。相邻帧和参考帧对齐之后，它们同一空间位置处的特征具有对应的时序关系。我们需要在所有的空间位置处，分别融合这些具有时序关系的特征。由于这些特征所对应的视频帧像素模糊程度不同，时序融合时不同视频帧所占权重就应不同；而同一个视频帧不同空间位置处像素模糊程度也不同，它们的融合权重也应该不同。所以对齐后的时空特征，在时序特征融合过程中应该具有像素级聚合权重（特定于时序位置和空间位置）。针对不同的输入视频帧，需要自适应的生成这个像素级聚合权重。

我们的算法分为四个阶段：特征提取网络、可变形卷积对齐模块、时序特征融合模块和特征重建网络。其整体结构如图 4-1 所示（以连续 3 帧输入为例进行说明）。为了保留参考帧中的信息，网络整体采用残差的结构。参考帧和特

征重建网络输出的重建特征相加，得到去模糊后的清晰图像。由于对齐模块中我们使用可变形卷积网络在执行相邻帧时序对齐的同时实现帧内空域信息融合，融合模块中使用了动态滤波网络实现了自适应的像素级时序特征融合，所以我们称该算法为基于自适应时空卷积网络的视频去模糊算法。

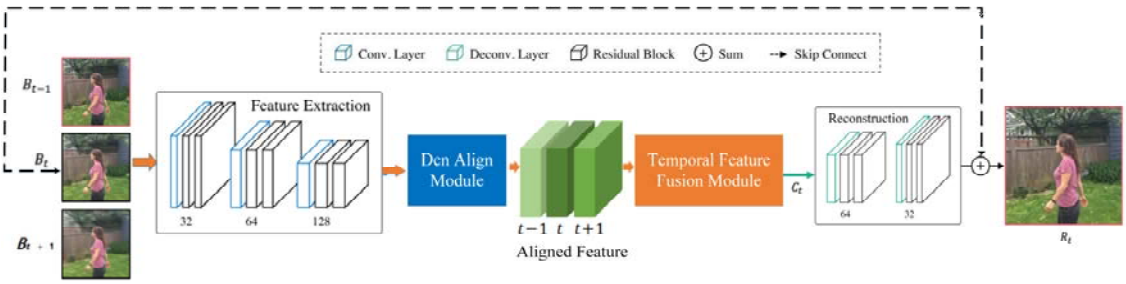


图 4-1　算法整体框架图

输入的连续视频帧，分别经过 Feature Extraction 网络进行下采样和特征提取，将其转换为较低分辨率的特征图，以在后续过程中节省内存和计算资源的消耗。不同的视频帧，共享特征提取网络的参数。Feature Extraction 网络由三个卷积模块构成，每个卷积模块包含一个步长为 2 的卷积层和两个残差模块。带步长的卷积层会在提取特征的同时降低其分辨率。而残差模块可以很好地解决模型训练过程中由于网络过深导致的梯度消失问题，使得整体网络可以设计的更深，从而增大网络的感受野。这使得深度神经网络可以处理图像中更大的模糊。Feature Extraction 网络的输出是输入视频帧三维的时空特征，这些特征是未对齐的。Dcn Align Module 将在特征层面执行对齐。对齐后的特征经过 Temporal Feature Fusion Module 进行像素级别的时序特征融合，相邻帧中清晰像素的特征会提升参考帧中对应模糊像素特征的质量，达到去模糊的效果。融合后的特征，经过特征重建网络，重建成和参考帧具有相同分辨率的残差图像，最终将该重建的残差图像和参考帧相加得到恢复出的清晰帧。重建网络由两个反卷积模块组成，每个反卷积模块由一个反卷积层和两个残差块组成。反卷积层对输入的特征进行上采样，残差模块可以很好地防止网络因为深度过深导致梯度消

失。特征提取网络和特征重建网络整体是一个 Encoder-Decoder 结构，可以有效去除图像中的噪声。

4.3 基于增强可变形卷积网络的相邻帧对齐模块

在 3.3 章节，我们探究了特征对齐相较于图像对齐具有的优越性，并且提出了基于多尺度可变形卷积网络的隐式特征对齐模块，替代了传统的光流估计的方法，实现了隐式的对齐。但由于对齐模块可变形卷积网络中，相邻帧和参考帧对应特征点的位置偏移量 *offset* 的生成过程是没有监督信号的。导致网络训练不稳定，极端情况下会出现 *offset* 值溢出的情况。此时，可变形卷积根据 *offset* 采样到的特征值是零，视频去模糊便退化成单幅图像的去模糊。我们无法再利用相邻帧中冗余的空域信息。为了解决这一问题，我们改进了增强的可变形卷积网络，使其在执行相邻视频帧的特征精确对齐的同时实现帧内的空域信息融合。如图 4-2 所示，是我们提出的基于增强可变形卷积网络的相邻帧对齐模块。

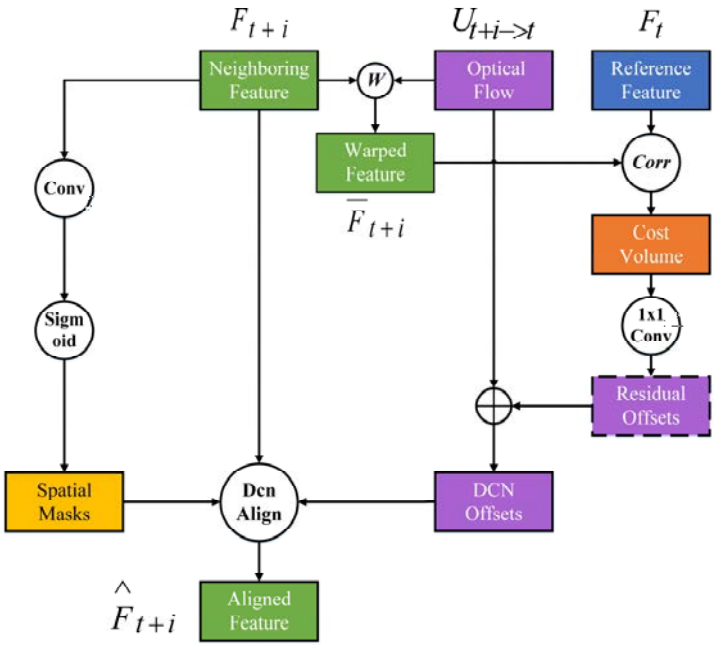


图 4-2　基于增强可变形卷积网络的相邻帧对齐模块

我们改进了增强的可变形卷积网络，提出了一种新的卷积计算网络：Dcn Align。它的结构包括两个部分：（1）Dcn Align 的卷积参数生成网络：以输入为条件产生卷积核参数，包括卷积核的采样位置偏移 DCN Offsets 和采样权重 Spatial Masks。（2）Dcn Align 卷积层：将生成的卷积核应用于相邻帧特征，执行对齐的同时实现帧内空域信息融合。DCN Align 的计算过程如公式（4-1）所示：

$$\hat{F}_{t+i}(p) = \sum_{k=1}^{K} w_k \bullet F_{t+i}(p + p_k + offset_k) \bullet mask_k \qquad (4\text{-}1)$$

其中 F_{t+i} 代表相邻帧特征，$\hat{F}_{t+i}$ 代表对齐后的相邻帧特征，它们的尺寸为 $128 \times \mathrm{H} \times \mathrm{W}$。$K$ 是卷积核采样数目，w_k 和 p_k 分别表示第 k 个特征采样点的卷积核权重以及位置偏移量。例如 K 的值为 9，则 p_k 可以表示为公式（4-2）：

$$p_k = \{(-1,-1),(-1,0),\ldots,(0,1),(1,1)\} \qquad (4\text{-}2)$$

$offset = 2\mathrm{K} \times \mathrm{H} \times \mathrm{W}$，代表卷积核的采样位置偏移；$mask = 1 \times \mathrm{H} \times \mathrm{W}$，代表对应采样点的采样权重。*offset* 和 *mask* 分别对应图 4-2 中的 DCN Offsets 和 Spatial Masks。$offset_k$ 代表第 k 个采样点的位置偏移，$mask_k$ 代表第 k 个采样点所处空间位置处的权重。由于偏移后的采样点位置一般是小数，所以对应的采样特征值和采样权重通过双线性插值计算得到。

采样位置偏移 DCN Offsets 对应的物理意义是参考帧中的特征点和相邻帧对应特征点的位置偏移。所以 DCN Offsets 的生成过程，就是相邻帧和参考帧特征对齐的过程。这部分使用残差学习的思想，如公式（4-3）所示，它由两部分构成：光流作为基础的位置偏移，通过光流网络计算得到；由于模糊的图像计算的光流并不准确，所以我们在此基础上继续学习一个残差的位置偏移，实现精确的运动估计。残差偏移通过卷积层学习得到。

$$\text{DCN Offsets} = \text{Optical Flow} + \text{Residual Offsets} \qquad (4\text{-}3)$$

首先通过光流网络 PWC-Net[39]计算出相邻帧特征 F_{t+i} 和参考帧特征 F_t 之间的光流 $U_{t+i->t}$，如公式（4-4）所示：

$$U_{t+i->t} = N(F_t, F_{t+i}) \qquad (4\text{-}4)$$

然后用光流对相邻帧特征执行 Warp 变形操作，得到和参考帧特征 F_t 初步对齐的相邻帧特征 $\overline{F}_{t+i}$，如公式（4-5）所示：

$$\overline{F}_{t+i} = W(F_{t+i}, U_{t+i->t}) \tag{4-5}$$

接下来将初步对齐的相邻帧特征 $\overline{F}_{t+i}$ 和参考帧特征 F_t 进行互相关运算，得到匹配代价容量 Cost Volume。匹配代价容量 Cost Volume 表示执行互相关运算的两个特征图之间的相似性，即初步对齐的相邻帧特征 $\overline{F}_{t+i}$ 和参考帧特征 F_t 之间的匹配误差。将其再输入到卷积网络中学习得到残差的位置偏移 Residual Offsets。上述计算过程如公式（4-6）和公式（4-7）所示：

$$\text{Cost Volume} = Corr(F_t, \overline{F}_{t+i}) \tag{4-6}$$

$$\text{Resudual Offsets} = C(\text{Cost Volume}) \tag{4-7}$$

其中 *Corr* 代表互相关运算，*C* 代表卷积运算。

采样权重 Spatial Masks 大小为 $1 \times H \times W$，它表示相邻视频帧中不同空间位置处特征点的清晰程度。它的值位于（0,1）之间，像素越模糊，其值越小。采样权重 Spatial Masks 通过对相邻视频帧特征执行一系列的卷积操作，最终应用 Sigmoid 激活函数得到，其计算过程如公式（4-8）所示：

$$\text{Spatial Masks} = Sigmoid(C(F_{t+i})) \tag{4-8}$$

其中 *Sigmoid* 代表 Sigmoid 激活函数运算，*C* 代表卷积运算。

所以公式（4-1）相较于普通的卷积计算，增加了两个额外的自由度 *offset* 和 *mask*，来控制卷积过程中特征点的采样位置和采样权重。因为 *offset* 代表了相邻帧和参考帧对应特征点的位置偏移，所以 Dcn Align 在特征层面实现了相邻视频帧的对齐。又因为 *mask* 反映了相邻视频帧特征的清晰程度，在卷积计算时，越清晰的特征点所占权重越大，对最终加权求和的结果影响就越大。这其实是一种注意力机制，实现了视频帧内空域的信息融合，使得清晰的像素对卷积的结果贡献更大。而 Dcn Align 相较于增强的可变形卷积，有两点不同：（1）其 *offset* 的生成过程完全不同，Dcn Align 中的 *offset* 具有实际的物理意义。我们将其应用到了视频去模糊任务中的相邻帧对齐问题上，使得 *offset* 的生成基于光流的引导，网络具有可解释性，解决了可变形卷积训练过程中网络不稳定、

offset 值溢出的情况（2）Dcn Align 中的 *mask* 同样具有实际的物理意义，它反映的是视频帧特征点的清晰程度，其尺寸是$1 \times H \times W$。而增强的可变形卷积中 Δm 代表调制权重，其尺寸是$K \times H \times W$（K 是卷积核采样点数目）负责调制来自不同空间位置的输入特征幅度。所以 Dcn Align 和增强的可变形卷积一个明显的区别是：它们的卷积核在特征图上以固定步长滑动时，前后两次卷积中在同一空间位置处的采样点，Dcn Align 具有相同的 *mask* 值，其代表的就是该位置处特征点的清晰程度；而增强的可变形卷积的调制权重 Δm 和卷积核采样空间位置不是一对一的关系，前后两次卷积在同一空间位置处的采样点可能具有不同的调制权重，这显然是不合理的。

4.4 基于动态滤波网络的自适应时序特征融合模块

在视频去模糊任务中，相邻帧对齐和特征融合是两个最关键的部分。如何高效融合对齐后的连续视频帧特征是这部分的核心。相邻帧和参考帧特征对齐之后，它们在同一空间位置处的特征具有对应的时序关系。我们需要在所有的空间位置处，分别融合这些具有时序关系的特征。由于这些特征所对应的视频帧像素模糊程度不同，时序融合时不同视频帧所占权重就应不同；而同一个视频帧不同空间位置处像素模糊程度也不同，它们的融合权重也应该不同。所以对齐后的时空特征，在时序特征融合过程中应该具有像素级聚合权重（特定于时序位置和空间位置）。针对不同的输入视频帧，需要自适应的生成这个像素级聚合权重。

为了解决这个问题，我们利用了动态滤波网络的自适应性。常规卷积网络中，卷积核的参数在网络训练完成后，对任意输入都是固定的。而动态滤波网络中，滤波器的参数是根据输入自适应生成的，其随着输入的变化而变化。输入不同，生成的滤波器参数就随之不同。这种架构由于其自适应特性而具有更强的灵活性，而且滤波器是根据输入动态生成的，没有需要学习的参数，所以模型参数的数量没有过多增加。动态滤波网络可以学习各种各样的滤波操作，包括局部特征变换，这种变换是特定于输入、特定于位置的，所以具有自适应性。而我们需要做的时序特征融合，就是一种特定于输入和空间位置的局部特

征变换。时序特征融合的权重，也是特定于输入和位置的，是一个像素级聚合权重。

受 KPN[28]的启发，我们提出了基于动态滤波网络的自适应时序特征融合模块，如图 4-3 所示。该模块将生成的逐元素卷积滤波器应用于对齐后的特征，真正实现了像素级别的特征融合。

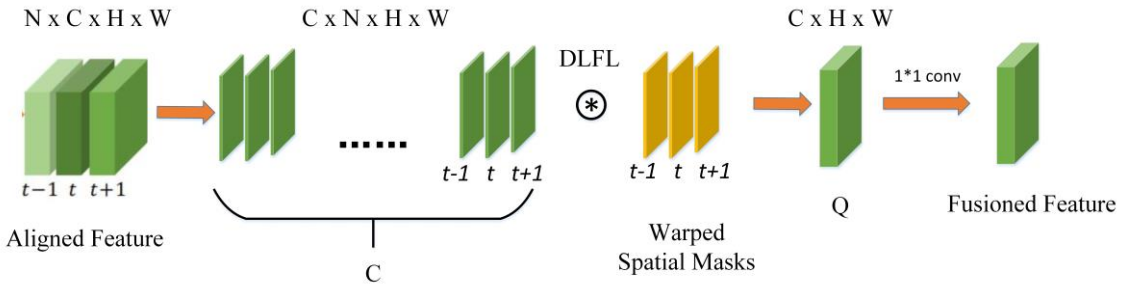


图 4-3　基于动态滤波网络的自适应时序特征融合模块

对齐后的特征图尺寸为 $N \times C \times H \times W$，N 是输入的连续视频帧的数目，C 是特征通道数。先将其形状重塑为 $C \times N \times H \times W$，得到 C 组 $N \times H \times W$ 的特征图。每一组这样的特征图，都具有完整的输入连续视频帧的时空特征，不同组的特征图是特征在通道维度的不同响应。我们先对每组特征图做时序特征融合，并将每组的融合结果在通道维度连接起来，得到特征图 Q。再利用 1×1 对 Q 在通道维度做特征融合，得到最终融合后的特征。

在相邻视频帧对齐模块生成的 Spatial Masks，反应了不同视频帧中不同空间位置处像素的清晰程度。我们将它的值作为时序特征融合的权重，这样就能实现用相邻帧中更清晰的像素融合参考帧中模糊的像素。不同的相邻视频帧中对应像素越清晰，时序特征融合过程中所占权重就越大。我们用对齐模块中计算得到的相邻帧和参考帧之间的光流，对 Spatial Masks 做变形操作获得 Warped Spatial Mask。将 Warped Spatial Mask 和对齐后的相邻视频帧特征的每个组应用到我们构建的动态局部滤波层（Dynamic Local Filtering Layer）中，其计算过程示意图如图 4-4 所示。

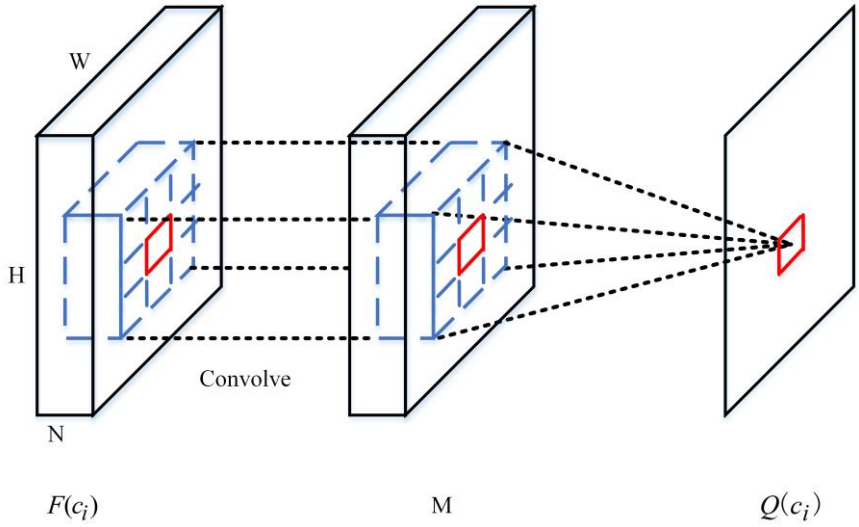


图 4-4　基于动态滤波网络的自适应时序特征融合模块

动态局部滤波层 DLFL 计算过程如公式（4-9）所示：

$$Q(c_i, x, y) = \sum_{d=1}^{N} \sum_{n=-r}^{r} \sum_{m=-r}^{r} \hat{F}(c_i, d, x+n, y+m) \times M(d, x+n, y+m) \quad (4\text{-}9)$$

其中，N 是输入连续视频帧的数目，$r = \frac{k-1}{2}$，k 是动态局部滤波层卷积边长，M 是 Warped Spatial Mask。这里的卷积核没有需要学习的参数，而是中间的计算结果。模型参数大大减少，推理速度加快。通过动态局部滤波层 DLFL，我们实现了对对齐后的时空特征，进行像素级别的特征融合。由于特征融合的权重 Warped Spatial Mask，是根据输入动态生成的。不同输入的视频帧，具有不同的权重；同一视频帧内不同空间位置处，也具有不同的权重。所以利用 Warped Spatial Mask 对对齐的时空特征做动态局部滤波操作，实现了特定于输入和位置的局部特征变换，即特征融合。通过这种像素级的时序特征融合，实现了我们一开始的研究思路：利用相邻帧中清晰的像素补偿参考帧中模糊的像素，充分挖掘输入视频序列中的时空信息。

4.5 基于自适应时空卷积网络的视频去模糊算法

通过特征提取网络、基于增强可变形卷积网络的相邻帧对齐模块、基于动

态滤波网络的自适应时序特征融合模块和特征重建网络，我们实现了该算法：

算法 1 基于自适应时空卷积网络的视频去模糊算法

Input: 连续 N 个模糊视频帧 $\{B_{t-n},\ldots,B_t,\ldots,B_{t+n}\}$，其中 $N=2n+1$

Output: 清晰的参考帧 R_t

从训练完成的模型中加载网络参数

将参考帧 B_t 输入到 Feature Extraction 网络，得到提取的特征 F_t

For i = -n *to* n **Do**

    将相邻帧 B_{t+i} 输入到 Feature Extraction 网络，得到提取的特征 F_{t+i}

    根据公式（4-4），计算相邻帧特征 F_{t+i} 和参考帧特征 F_t 之间的光流 $U_{t+i->t}$

    根据公式（4-5）、（4-6）、（4-7），计算残差的位置偏移 Residual Offsets

    根据公式（4-3），计算采用位置偏移 DCN Offsets

    根据公式（4-8），计算采样权重 Spatial Masks

    根据公式（4-1），将相邻帧特征 F_{t+i} 对齐，得到对齐后的特征 $\hat{F}_{t+i}$

End

将对齐后的相邻帧特征和参考帧特征，在通道维度连接，得到形状为 N × C × H × W 的特征：Aligned Feature

将 Aligned Feature 进行 Reshape 操作，得到形如 C × N × H × W 的特征 F

For i = 1 *to* C **Do**

    根据公式（4-9），将每组形如 N × H × W 特征 $F(c_i)$ 和对应的 N 个融合权重 Spatial Masks 输入到动态局部滤波层中进行 DLFL 计算，得到像素级融合的时序特征 $Q(c_i)$

End

将 C 个特征 $Q(c_i)$ 在通道维度连接，再经过 1x1 卷积得到最终融合的特征 C_t

将 C_t 输入到特征重建网络中输出重建的特征，和参考帧 B_t 相加得到 R_t

Return R_t

4.6 实验结果和分析

我们在 DVD[9]和 GOPRO[8]数据集上对该算法进行了定量和定性评估。该算法的实验设置同 3.5 节一致。

4.6.1 算法定量评估结果

为了评估算法的性能，我们将其与近几年的视频去模糊算法在 DVD[9]数据集和 GOPRO[8]数据集上进行了定量比较，如表 4-1 和表 4-2 所示。

实验中使用 PSNR 和 SSIM 作为评估指标，它们反映了每个算法的准确率。从实验结果可以看出，我们的算法相较于近几年最新的视频去模糊算法，达到了较高的性能和准确率。由于 GOPRO[8]数据集中的视频序列，相邻帧和参考帧之间的抖动更大，所以更加难以对齐。各算法在 GOPRO[8]数据集上的表现结果相较于在 DVD[9]数据集上都有不同程度的下降。但我们算法中的对齐模块由于采用了从粗略到精细的对齐方式，使得网络具有更好的鲁棒性，所以性能下降较少。

表 4-1　各算法在 DVD[9]数据集上的定量评估结果

Method	Tao[10]	Su[9]	STFAN[22]	Xiang[35]	TSP[30]	Suin [36]	Ours
PSNR	29.98	30.01	31.15	31.68	32.13	32.53	32.54
SSIM	0.8842	0.8877	0.9049	0.9157	0.9268	0.9468	0.9410

表 4-2　各算法在 GOPRO[8]数据集上的定量评估结果

Method	Tao[10]	Su[9]	STFAN[22]	Nah[37]	TSP[30]	Suin [36]	Ours
PSNR	30.29	27.31	28.59	29.97	31.67	32.10	32.18
SSIM	0.9014	0.8255	0.8608	0.8947	0.9279	0.9600	0.9316

4.6.2 算法定性评估结果

为了进一步验证算法的泛化能力，我们对其在 DVD[9]测试数据集和真实模糊数据集上进行了定性测试。在测试阶段使用视频序列中的连续 5 张视频帧作为算法输入，这和训练阶段的设置保持一致。

如图 4-4、4-5、4-6 和 4-7 所示,是各算法在 DVD[9]测试数据集中的去模糊结果。以图 4-4 为例，图 4-4（a）是模糊的参考帧，图 4-4（b）是对应的真值，图 4-4（c）、4-4（d）和 4-4（e）分别是 EDVR[27]、PVDNet[37]和 TSP[30]算法对参考帧的处理结果，图 4-4（f）是我们算法的处理结果。从定性评估结果可以看出，我们的算法相较于其他算法，恢复出了模糊图像中更多的细节信息，可以有效地处理动态场景中的非均匀模糊。图 4-8 和图 4-9 则是在 DVD[9]真实模糊数据集上的测试结果，该数据集中只有模糊的图像没有与之对应的真值。从图 4-8（d）和图 4-9（d）可以看出，虽然没有与模糊图像对应的真值，我们的算法依然有效地去除了真实模糊图像中存在的由于相机抖动和目标运动产生的模糊。所以算法具备较强的泛化能力。

图 4-4　DVD[9]测试数据集 IMG_0021 视频序列的去模糊结果

图 4-5　DVD[9]测试数据集 IMG_0030 视频序列的去模糊结果

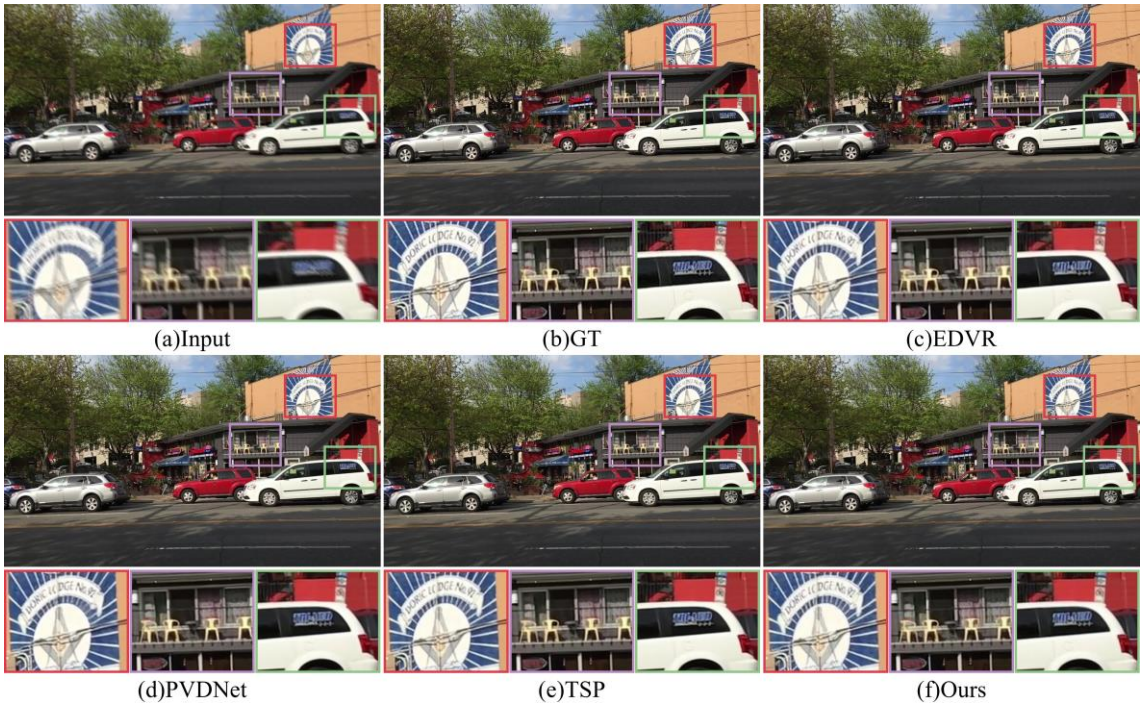


图 4-6　DVD[9]测试数据集 IMG_0037 视频序列的去模糊结果

图 4-7　DVD[9]测试数据集 IMG_0039 视频序列的去模糊结果

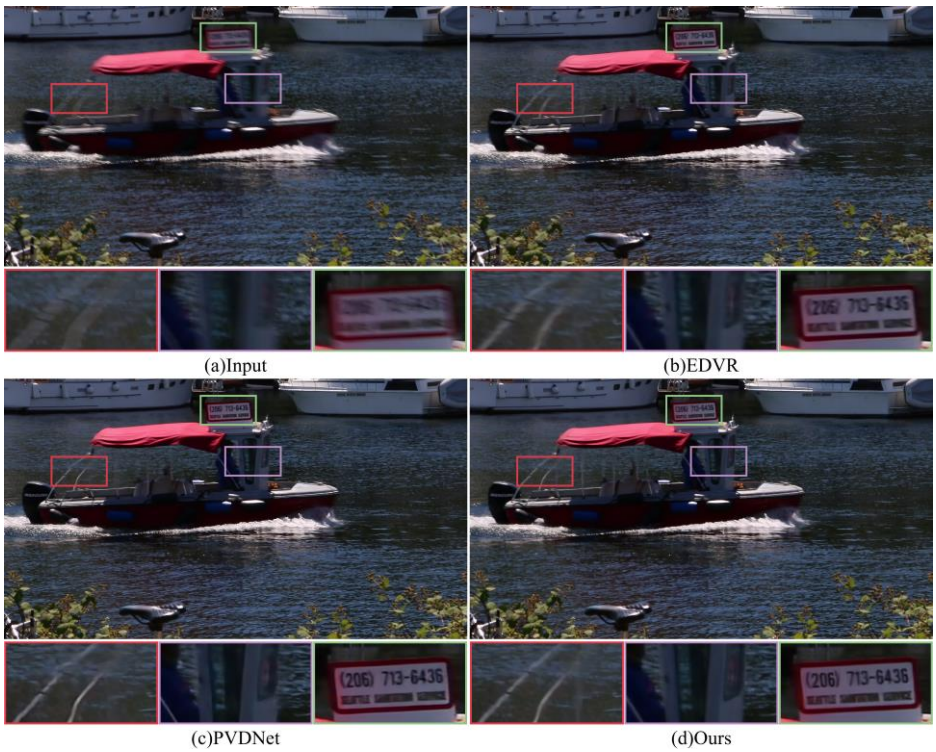


图 4-8　DVD[9]测试数据集 Boat 视频序列的去模糊结果

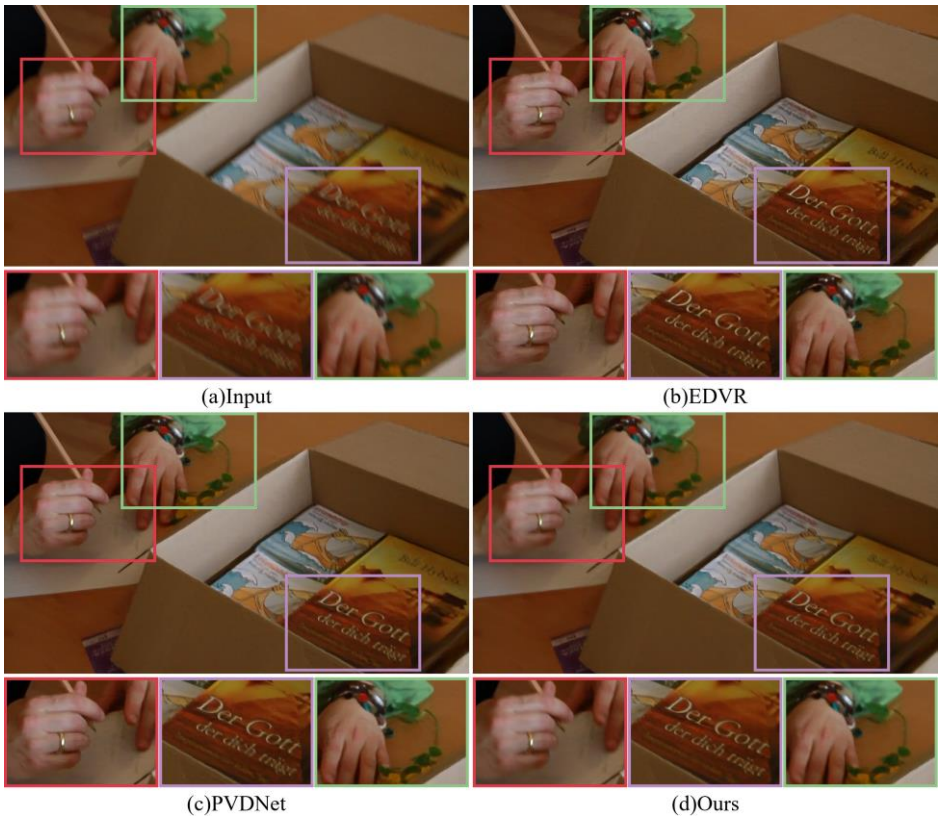


图 4-9　DVD[9]测试数据集 Books 视频序列的去模糊结果

4.7 消融实验

该算法最重要的两个模块就是我们提出的基于增强可变形卷积网络的相邻帧对齐模块和基于动态滤波网络的自适应时序特征融合模块。为了验证这两个模块的有效性，我们分别做了以下消融实验，并针对实验结果做了分析和讨论。

4.7.1 相邻帧对齐模块

基于增强可变形卷积网络的相邻帧对齐模块中，相邻帧和参考帧对应特征点的位置偏移 *offset* 由基础光流和残差偏移组成。为了探究相邻帧对齐模块对去模糊结果的影响，我们在 DVD[9]数据集上做了三个消融实验。如表 4-3 所示，w/o Align、w/o Flow 和 w/o Residual 分别表示网络结构中去除整个对齐模块、去除对齐模块中的光流和去除对齐模块中的残差偏移。

表 4-3　对齐模块在 DVD[9]数据集上的消融实验结果

Method	w/o Align	w/o Residual	w/o Flow	Ours
PSNR	29.26	31.38	31.62	32.54
SSIM	0.8730	0.9113	0.9148	0.9410

从实验结果可以看出，对齐模块对最终的去模糊结果有着至关重要的影响。而在对齐模块中，相邻帧和参考帧对应特征的位置偏移的两部分组成：基础光流和残差偏移，都对对齐有着一定的影响。去除基础光流后，算法 PSNR 指标下降了 0.92db；去除残差偏移后，算法 PSNR 指标下降了 1.16db；去除整个对齐模块，算法 PSNR 指标下降了 3.28db。所以在 Dcn Align 卷积计算时，引入光流作为基础的位置偏移解决了只使用卷积操作生成位置偏移存在的 *offset* 值溢出导致网络性能下降的问题；而引入残差偏移作为基础光流的精细化补偿，解决了模糊的图像计算的光流不精确的问题。基础光流和残差偏移一起组成了相邻帧和参考帧对应特征的位置偏移，以一种从粗到细的方式进行准确的运动估计。

4.7.2 时序特征融合模块

在基于动态滤波网络的自适应时序特征融合模块中，我们探究了动态局部滤波层的滤波器尺寸对算法性能的影响。如表 4-4 所示，随着滤波器尺寸的不断增大，算法性能也不断提升。但是当滤波器尺寸 $k=3$ 后，随着 k 的增加，算法性能提升很少，随之带来的却是计算量的大幅增加。所以为了平衡性能和效率，我们论文中将 k 设置为 3。

表 4-4　滤波器尺寸在 DVD[9]数据集上对算法性能的影响

Filter Size	$k=1$	$k=3$	$k=5$	$k=7$
PSNR	32.26	32.54	32.58	32.61
SSIM	0.9347	0.9410	0.9418	0.9420

4.8 本章小结

本章介绍了我们的第二个工作：基于自适应时空卷积网络的视频去模糊算法。该算法的研究思路是利用相邻帧中的清晰像素融合参考帧中的模糊像素，充分挖掘输入视频序列中的时空信息。算法分为四个阶段：特征提取网络、基于增强可变形卷积网络的相邻帧对齐模块、基于动态滤波网络的自适应时序特征融合模块和特征重建网络。特征提取网络和特征重建网络整体是一个 Encoder-Decoder 结构，分别负责特征提取下采样和特征重建上采样。算法中最重要的两个部分就是对齐模块和特征融合模块，它们分别解决了视频去模糊任务中最关键的两个问题：相邻视频帧精确对齐和高效的特征融合。

在对齐模块中我们改进了增强可变形卷积，提出了一种新的卷积计算方式 Dcn Align，其在执行相邻帧精确对齐的同时实现了帧内空域信息融合。在相邻帧和参考帧对应特征点位置偏移生成过程中，我们使用从粗略到精细的方式进行准确的运动估计。首先计算相邻帧和参考帧特征之间的光流，作为基础的偏移；然后利用卷积网络学习利用光流初步对齐的相邻帧特征和参考帧特征之间

的残差偏移，进行精确的偏移补偿。采取这样的位置偏移生成策略，解决了第3章3.3.2节对齐模块中仅仅使用卷积网络学习生成偏移可能存在的*offset*值溢出、网络训练不稳定的问题。对齐模块中同样通过卷积网络学习生成了视频帧特征质量分布图 Spatial Mask，其反应了不同空间位置处的像素清晰程度。在Dcn Align卷积计算过程中，通过Spatial Mask实现了更清晰的特征点对最终的卷积结果贡献更大。这便是帧内的空域信息融合。

在时序特征融合模块中，我们不仅需要实现基于时序位置和空间位置的像素级特征融合，而且需要根据输入视频序列的不同实现自适应的融合。自适应的含义是根据输入视频序列的不同，生成与之对应的像素级聚合权重。由于常规卷积网络的卷积核权重在特征图不同空间位置处参数共享，所以不能使用全局固定的卷积核进行特征融合。在第3章3.4.1节，我们使用简单的1×1卷积网络进行时序特征融合，它的局限性就在于：常规卷积网络在模型训练结束后，针对同一视频帧不同空间位置处的特征和不同的视频帧，在时序特征融合时具有相同的卷积核权重。这是不合理的，因为同一视频帧不同空间位置处的像素模糊程度并不同，不同视频序列中的视频帧像素模糊程度更不同。我们借助动态滤波网络具有的能够特定于输入和位置进行局部空间变换的能力，构建了自己的动态局部滤波层。利用在对齐模块中生成的 Spatial Mask 作为聚合权重，通过动态局部滤波层实现了利用相邻帧中的清晰特征对参考帧中的对应特征做像素级别的融合。

该算法在DVD[9]数据集和GOPRO[8]数据集上进行了定量评估和定性测试。结果表明算法具有较高的性能，可以有效处理动态场景中的非均匀模糊，包括真实世界中没有对应真值的模糊图像。最后我们针对相邻帧对齐模块和时序特征融合模块分别做了消融实验，验证了它们的有效性。

结　论

本文对视频去模糊任务进行了深入的理论研究和实验，重点对其中的两个子问题：相邻帧对齐和时序特征融合展开了分析和研究。本文的研究思路是利用相邻帧中的清晰像素融合参考帧中的对应像素，充分挖掘输入视频序列中的时空信息。

本文的主要贡献和创新性工作如下：

（1）通过实验验证了相邻帧在特征层面对齐比在图像层面对齐具有更好的结果。

（2）改进了增强可变形卷积网络，提出了一种新的卷积计算方式 Dcn Align。它提高了相邻帧对齐的准确性，同时实现了帧内的空域信息融合。

（3）利用动态局部滤波层，极大地提高了特征融合的效率和视频序列中的时空信息的利用率，实现了自适应的像素级别时序特征融合。

其中我们提出的 Dcn Align 对齐操作，不仅克服了传统光流估计进行对齐存在的光流计算不准确问题，而且解决了只使用卷积网络学习相邻帧和参考帧对应特征点的位置偏移存在的网络训练不稳定、偏移值溢出的问题。更一进步，Dcn Align 对齐操作不仅适用于视频去模糊任务，也可以用于其他视频任务中存在的对齐问题，例如视频超分辨率重建任务中的相邻视频帧对齐。其次我们构建的动态局部滤波层，提供了一种特定于输入和位置进行局部空间变换的能力。在我们的工作中，这种局部空间变换就是时序特征融合，利用相邻帧中的清晰特征对参考帧中的对应特征进行像素级别的融合。这种局部空间变换也可以应用到其他任务中，执行特定于输入和空间位置的特征变换。

本文的工作可以进一步改进和优化的点如下：

（1）在对齐模块中，通过卷积网络学习生成 Spatial Mask 来表示视频帧中不同空间位置处像素的清晰程度。这个学习生成的过程可以引入图像像素的先验知识，以一种更具有解释性的方式生成。

（2）继续探究 Dcn Align 在其他视频任务中存在的对齐问题上的表现。

参考文献

[1] Gupta A , Joshi N , Zitnick C L , et al. Single image deblurring using motion density functions[C]. ECCV , 2010, 68(1):562-573

[2] Oliver Whyte, Josef Sivic, Andrew Zisserman, and Jean Ponce. Non-uniform deblurring for shaken images[C]. IJCV, 2012, 98(2):168–186.

[3] Kim T H , Ahn B , Lee K M . Dynamic Scene Deblurring[C]. ICCV, 2013, 53(2):065-074.

[4] L. Xu, J. S. Ren, C. Liu, and J. Jia. Deep convolutional neural network for image deconvolution[C]. Advances in Neural Information Processing Systems, 2014, 46(1):1790–1798.

[5] C. J. Schuler, M. Hirsch, S. Harmeling, and B. Sch¨olkopf. Learning to deblur[C]. IEEE transactions on pattern analysis and machine intelligence, 2016, 38(7):1439–1451.

[6] J. Sun, W. Cao, Z. Xu, and J. Ponce. Learning a convolutional neural network for non-uniform motion blur removal[C]. IEEE Conference on Computer Vision and Pattern Recognition, 2015, 63(1): 769–777.

[7] Leibe B , Matas J , Sebe N , et al. A Neural Approach to Blind Motion Deblurring[J]. 2016, 10.1007/978-3-319-46487-9(Chapter 14):221-235.

[8] Nah S , Kim T H , Lee K M . Deep Multi-scale Convolutional Neural Network for Dynamic Scene Deblurring[C]. IEEE Conference on Computer Vision and Pattern Recognition, 2017,13(1):623-634.

[9] Su S , Delbracio M , Wang J , et al. Deep Video Deblurring for Hand-Held Cameras[C]. IEEE Conference on Computer Vision and Pattern Recognition, 2017, 86(7):130-139.

[10] Tao X , Gao H , Wang Y , et al. Scale-recurrent Network for Deep Image Deblurring[J]. 2018, 20(4):231-240.

[11] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution[C]. The IEEE Conference on Computer Vision and Pattern Recognition, 2018, 35(8):302-312.

[12] Yasuyuki Matsushita, Eyal Ofek, Weina Ge, Xiaoou Tang, and Heung-Yeung Shum. Full-frame video stabilization with motion inpainting[J]. TPAMI, 2006,

28(7):1150–1163.

[13] Sunghyun Cho, Jue Wang, and Seungyong Lee. Video deblurring for hand-held cameras using patch-based synthesis[J]. TOG, 2012, 31(4):64.

[14] Tae Hyun Kim and Kyoung Mu Lee. Generalized video deblurring for dynamic scenes[C]. IEEE Conference on Computer Vision and Pattern Recognition, 2012,16(1):323-334.

[15] Wenqi Ren, Jinshan Pan, Xiaochun Cao, and Ming-Hsuan Yang. Video deblurring via semantic segmentation and pixelwise non-linear kernel[C]. ICCV, 2017, 12(3):32-41.

[16] Jose Caballero, Christian Ledig, Aitken Andrew, Acosta Alejandro, Zehan Wang, and Wenzhe Shi. Realtime video super-resolution with spatio-temporal networks and motion compensation[C]. The IEEE Conference on Computer Vision and Pattern Recognition, 2018, 27(2):372-382.

[17] Xue T , Chen B , Wu J , et al. Video Enhancement with Task-Oriented Flow[J]. International Journal of Computer Vision, 2017, 10(3):23-32.

[18] Tae Hyun Kim, Mehdi S M Sajjadi, Michael Hirsch, and Bernhard Scholkopf. Spatio-temporal transformer network for video restoration[C]. ECCV, 2018, 12(1):230-231.

[19] Tae Hyun Kim, Kyoung Mu Lee, Bernhard Scholkopf, and Michael Hirsch. Online video deblurring via dynamic temporal blending network[C]. CVPR, 2017, 45(2):120-129.

[20] Jo Y , Oh S W , Kang J , et al. Deep Video Super-Resolution Network Using Dynamic Upsampling Filters Without Explicit Motion Compensation[C]. IEEE, 2018, 13(1):13-21.

[21] Tian Y , Zhang Y , Fu Y , et al. TDAN: Temporally Deformable Alignment Network for Video Super-Resolution[J]. 2018, 13(1):23-33.

[22] Zhou S ,Zhang J , Pan J , et al. Spatio-Temporal Filter Adaptive Network for Video Deblurring[J]. 2019, 35(1):36-47.

[23] Jiawei Zhang, Jinshan Pan, Jimmy Ren, Yibing Song, Linchao Bao, Rynson WH Lau, and Ming-Hsuan Yang. Dynamic scene deblurring using spatially variant recurrent neural networks[C]. CVPR, 2018, 15(1):78-89.

[24] Orest Kupyn, Volodymyr Budzan, Dmytro Mishkin, and Jiri Matas. Deblurgan:

Blind motion deblurring using conditional adversarial networks[C]. CVPR, 2018, 25(2):78-89.

[25] Sifei Liu, Jinshan Pan, and Ming-Hsuan Yang. Learning recursive filters for low-level vision via a hybrid neural network[C]. ECCV, 2016, 13(5):897-903.

[26] Yang R , Xu M , Wang Z , et al. Multi-frame Quality Enhancement for Compressed Video[C] IEEE, 2018, 19(2):765-773.

[27] Wang X , Chan K C K , Yu K , et al. EDVR: Video Restoration with Enhanced Deformable Convolutional Networks[J]. 2019, 21(5):5563-5575.

[28] De Brabandere B , Jia X . Dynamic Filter Networks[J]. 2016, 23(2):980-993.

[29] Dai J , Qi H , Xiong Y , et al. Deformable Convolutional Networks[J]. 2017,27(2):213-224.

[30] Xizhou Zhu, Han Hu, Stephen Lin, and Jifeng Dai. Deformable convnets v2: More deformable, better results.cvpr 2018.11(3)14:22.

[31] Pan J , Bai H , Tang J . Cascaded Deep Video Deblurring Using Temporal Sharpness Prior[J]. 2020, 18(2):1534-1542.

[32] Kaihao Z, Wenhan L, Yiran Z. Deblurring by Realistic Blurring. CVPR 2020.

[33] Jaesung R ,Haeyun L , Jucheol W. Real-World Blur Dataset for Learning and Benchmarking Deblurring Algorithms[C]. ECCV 2020. 11(2) 16:23.

[34] K. C. K. Chan, X. Wang, K. Yu, C. Dong. Understanding Deformable Alignment in Video Super-Resolution. AAAI 2021.

[35] Dongxu Li, Chenchen Xu, Chenchen Xu. ARVo: Learning All-Range Volumetric Correspondence for Video Deblurring. CVPR 2021.

[36] Xinguang Xiang, Hao Wei, and Jinshan Pan. Deep video deblurring using sharpness features from exemplars. *IEEE Transactions on Image Processing*, 29:8976–8987, 2020.

[37] Maitreya Suin, A. N. Rajagopalan Gated Spatio-Temporal Attention-Guided Video Deblurring. CVPR 2021.

[38] Seungjun Nah, Sanghyun Son, and Kyoung Mu Lee. Recurrent neural networks with intra-frame iterations for video deblurring. In *CVPR*, pages 8102–8111, 2019. 5, 7

[39] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz.PWC-Net: CNNs for optical flow using pyramid, warping,and cost volume. In CVPR, pages 8934–8943, 2018. 3, 4, 5,9

[40] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox. FlowNet 2.0: Evolution of optical flow estimation with deep networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1, 3, 4, 5, 7,8

[41] M. Jaderberg, K. Simonyan, A. Zisserman, et al. Spatial transformer networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2015. 4

[42] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun.Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *ICCV*, 2015.

[43] Diederik P Kingma and Jimmy Ba. Adam: A method forstochastic optimization. In *ICLR*, 2015.

[44] Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions[C] Proceedings of the IEEE conference on computer vision and pattern recognition. 2015 : 1-9.

[45] Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition[C] 3rd International Conference on Learning Representations. 2015.

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

[47] Hu J, Shen L, Sun G. Squeeze-and-excitation networks[C] Proceedings of the IEEE conference on computer vision and pattern recognition. 2018 : 7132-7141.

[48] Deng J, Dong W, Socher R, et al. Imagenet: A large-scale hierarchical image database[C] 2009 IEEE conference on computer vision and pattern recognition.2009 : 248-255.

攻读硕士学位期间发表的论文及其它成果

（一）发表的学术论文

[1] Fengzhi Duan, Hongxun Yao. Adaptive Spatio-Temporal Convolutional Network for Video Deblurring [C]. International Conference on Image and Graphics (ICIG),2021. (published) Lecture Notes in Computer Science(), vol 12890. Springer, Cham.

（二）申请及已获得的专利

[1] 姚鸿勋，段风志，张慧琮，陶胤旭，韩国权，李佳忆. 一种基于卷积神经网络的游行和暴恐识别方法.（已申请）

哈尔滨工业大学学位论文原创性声明和使用权限

学位论文原创性声明

本人郑重声明：此处所提交的学位论文《基于可变形卷积网络的视频去模糊算法》，是本人在导师指导下，在哈尔滨工业大学攻读学位期间独立进行研究工作所取得的成果，且学位论文中除已标注引用文献的部分外不包含他人完成或已发表的研究成果。对本学位论文的研究工作做出重要贡献的个人和集体，均已在文中以明确方式注明。

作者签名：段风志　　日期：2022 年 6 月 15 日

学位论文使用权限

学位论文是研究生在哈尔滨工业大学攻读学位期间完成的成果，知识产权归属哈尔滨工业大学。学位论文的使用权限如下：

（1）学校可以采用影印、缩印或其他复制手段保存研究生上交的学位论文，并向国家图书馆报送学位论文；（2）学校可以将学位论文部分或全部内容编入有关数据库进行检索和提供相应阅览服务；（3）研究生毕业后发表与此学位论文研究成果相关的学术论文和其他成果时，应征得导师同意，且第一署名单位为哈尔滨工业大学。

保密论文在保密期内遵守有关保密规定，解密后适用于此使用权限规定。

本人知悉学位论文的使用权限，并将遵守有关规定。

作者签名：段风志　　日期：2022 年 6 月 15 日

导师签名：　　日期：2022 年 6 月 15 日

致　谢

时光匆匆而逝，两年的研究生生涯即将画上句号。我也将踏入社会、奔赴职场开启下一段旅程。我们在每一次人生阶段交替中成长。回顾这两年，我遇到了很多良师益友，他们给予了我帮助和指导，让我受益匪浅。

首先要感谢我的导师姚鸿勋教授，她不仅是我课题工作上的指路人，也是生活中给予我关心、照顾的长辈。记得第一次跟姚老师汇报工作，作为刚刚接触学术研究的我，其实还不具备很高的学术素养。但第一次工作汇报后，姚老师给予了我很大的鼓励，支持对我的想法。这让我对学术研究产生了信心和兴趣，从而在之后的研究道路上坚持下来、有所沉淀。读研期间的每次组会汇报，姚老师都对我悉心指导，帮我把控整体的研究方向。在姚老师的指导下，研二上学期我投稿的 ICIG 论文被顺利接受。我们一起去海南参会期间，姚老师帮我拍照、带我吃粤式早茶，我们之间的关系又拉进了一步。姚老师跟我做到了真正的亦师亦友，感谢您对我的照顾和帮助。

其次要感谢我们实验室的所有师兄师姐给我的帮助和指导。感谢金声师兄、浩哲师兄、慧琮师兄、璐璐师姐、雅思师姐、浩然师兄、雨欣师姐、帝麟师兄、王波师兄、张啸师兄。和金声师兄相处的日子，非常的快乐。平时大家会互相开玩笑，谈天论地。希望他一切顺利，取得更高的成就。浩哲师兄是我来这个实验室的介绍人，也是我的本科校友。在大学期间就听说了关于浩哲师兄的诸多传说，是一个开发能力巨强的大佬。也是因为浩哲师兄的一次回校演讲，让我对计算机视觉领域产生了兴趣。现在我也即将追随浩哲师兄的脚步，离开校园前往腾讯工作。希望和他在 Teg 有更多交流合作的机会。认识慧琮师兄是在北京大学信息技术研究院实习期间。当时我还没有确定具体的研究课题，是慧琮师兄带我学习图像去模糊任务，从而确定了现在的研究方向。璐璐师姐是我们实验室的开心果，和师姐相处非常的轻松愉快。跟师姐回她母校那天，是我非常非常开心的一天。

感谢实验室的其他小伙伴们这几年的陪伴，包括浩森、陶胤旭、陈斌、力

凝、陈希、志伟、兆攀、李明慧。跟你们相处的时光，十分快乐。感谢我的大学同学也是研究生同学凌基发和陈姗姗。感谢我的室友谷博文。除此之外，还要感谢在阿里实习期间带我的师兄白福裕，他真的教会了我很多。还要感谢我的大学室友贾定坤，对我的职业发展给了很多指导。

最后要感谢我的家人，他们在背后默默支持着我，让我可以专注于自己的学习和工作。平时跟父母交流，他们会提醒我不要太累了。累了就休息休息，打打球，锻炼好身体。

回顾这两年，收获了很多也成长了很多。但我相信未来的几年，我会收获更多、成长更多。凡是过往，皆为序章。告别自己的学生时代，人生的大幕，正式拉开！