ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 36,No. 1,Jan. 2021,pp. 63－75
DOI:10. 16337/j. 1004-9037. 2021. 01. 006

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

基于卷积神经网络的光流估计模型

丰　艳，刘　帅，王传旭

（青岛科技大学信息科学技术学院，青岛 266100）

摘　要：光流信息是图像像素的运动表示，现有光流估计方法在应对图像遮挡、大位移和细节呈现等复杂情况时难以保证高精度。为了克服这些难点问题，本文建立一种新型的卷积神经网络模型，通过改进卷积形式和特征融合的方式来提高估计精度。首先，加入调整优化能力更强的可形变卷积，以便于提取相邻帧图像的大位移和细节等空间特征；然后利用基于注意力机制生成特征关联层，将相邻两帧的特征进行融合，以其作为由反卷积和上采样构成的解码部分的输入，旨在克服基于特征匹配等估计光流传统方法精度低的缺点；最后将得到的估计光流通过多网络堆栈的循环优化模型实现最终的光流估计。实验表明，本文网络模型在处理遮挡、大位移和细节呈现等方面的表现优于现有方法。

关键词：光流估计；可形变卷积；卷积神经网络；注意力机制

中图分类号：TP391　　**文献标志码：**A

Optical Flow Estimation Model Based on Convolutional Neural Network

FENG Yan, LIU Shuai, WANG Chuanxu

(School of Information Science and Technology, Qingdao University of Science and Technology, Qingdao 266100, China)

Abstract: The optical flow information is the motion representation of the image pixels. The existing optical flow estimation methods are difficult to ensure high precision in dealing with complex situations, such as occlusion, large displacement and detailed presentation. In order to overcome these difficult problems, a new convolutional neural network is proposed. The model improves the estimation accuracy by improving the convolution form and feature fusion. Firstly, the deformable convolution with stronger adjustment and optimization ability is added to extract the spatial features such as large displacement and details of adjacent frame images. Then, the feature correlation layer is generated by using the attention-based mechanism to carry out the feature fusion of the two adjacent frames, which is used as the input of the decoding part composed of deconvolution and upsampling and aims to overcome the disadvantage of low accuracy for the traditional methods of estimating optical flow based on feature matching. And finally the above estimated optical flow is optimized with a set of network stack. Experiments show that the proposed network model performs better than existing methods in dealing with occlusion, large displacement and detail presentation.

Key words: optical flow estimation; deformable convolution; convolutional neural networks; attention mechanism

---

基金项目：国家自然科学基金(61672305)资助项目；国家青年科学基金(61702295)资助项目。
收稿日期：2019-08-08；**修订日期：**2020-03-21

引　言

从Gibson等在1951年首先提出光流的概念[1]开始，图像光流的估计算法已经取得了非常好的成绩[2-5]。由于光流包含着空间运动物体在观察成像平面上像素运动的瞬时速度和位移矢量等信息，且可以通过颜色域和空间域结合的形式表示图像物体运动状态，因此，它在目标检测[6]、图像分割[7]及人类行为识别[8]等领域有了广泛的应用，是计算机视觉研究中的常用方法和重要问题。在光流估计中，遮挡、大位移和图像细节表现等一直都是亟待解决的难点问题，最近的研究也大都致力于解决此类问题。文献[9]基于无监督的方法设计了一种新型Warp，希望解决大位移和遮挡问题，但细节呈现方面结果不如人意。文献[10]提出一种多假设性网络，结合多种损失计算方法，再对估计结果进行多次局部评估，以此来优化光流的细节表现。文献[11]设计了专门评估遮挡情况的子网络，将其加到网络级联架构中，并通过优化训练策略的方式提高网络性能，但是由于网络结构过于复杂庞大，内部有着海量的权重参数，导致网络过于臃肿。

以上是前人对解决光流估计难点问题作出的尝试，其中鲜有针对网络特征提取和融合部分的针对性优化，而本文认为对于光流估计方法的关键，就是寻找两幅图像之间的变化和联系，相邻帧图像特征的提取和特征的融合方式很大程度上决定了结果的优劣。因此为了加强网络的学习调整能力，针对性地解决遮挡、大位移和细节呈现等问题，本文做出的主要贡献是：(1)加入可形变卷积[12]到光流估计网络模型的卷积层，旨在提取相邻帧的空间细节特征和大位移特征捕获；(2)利用基于注意力机制[13]的关联层，将相邻帧的特征融合重构，克服了现有融合方法调整适应能力不足的缺点，旨在学习两帧图像间的相关性，并对融合特征进行反卷积得出光流；(3)通过本文设置的图像Warp和网络级联对光流进行优化，得出最终的光流估计。实验结果表明本文方法在处理遮挡和大位移等问题上，显现出了较高的估计精度和鲁棒性，更在图像细节保护方面有显著的效果。

1　相关工作

目前光流估计方法可分为传统方法和深度学习两大类。在传统方法中，变分法最为主流且国内研究最多。文献[14]基于全变分L_1变分法，通过提取置信度较高的图像相互结构区域构造新型的全局目标函数，并采用金字塔分层细化的策略优化光流估计结果。文献[15]先进行最邻近场匹配得出粗略光流，再通过融合算法对估计光流进行大位移补偿。文献[16]将输入图像视作马尔科夫随机场，在超像素和像素两个层面上执行置信度传播，可以在减少计算量的同时得到高精度光流。而随着深度学习的快速发展，出现了许多基于深度学习的光流估计方法。Dosovitskiy等[17]最先将卷积神经网络(Convolutional neural network，CNN)的方法应用在光流估计上，使用类似于U型网的网络结构对光流进行估计。后来Ranjan等[18]将空间金字塔结构与卷积神经网络结合，通过多个尺度逐步放大和图像Warping的方式优化光流估计结果。Gadot等[19]则使用Siamese-CNN来分别计算两帧图像的特征描述符，再结合近似最邻近模板匹配算法，最后通过学习训练新型的损失函数实现光流的估计。其中FlowNet是将深度学习应用到光流估计算法中的开山之作，后面的研究多是对该网络进行改进。Tran等[20]使用3D卷积代替传统卷积，来实现对图片像素级的估计。文献[21-22]使用无监督形式训练网络，其损失函数使用经典的光流约束方程，实现了在缺乏有效Ground truth情况下基于卷积神经网络的光流估计。Teney等[23]提出具有图像不变性结构的网络模型，网络可针对图像的规则形变进行自适应调整。Thewlis等[24]将卷积神经网络与深度匹配相结合，在保留深度匹配方法高精度的优点下实现了深度网络的端到端训练。后来基于深度学习的光流估计算法又有了大的突破，Eddy等[25]通过对原有FlowNet做改进，将具有不同结构适用于解决不同图像问题的“变形FlowNet”级联起来形成一个大型网络来解决复杂光流估计问题，自此基于深度学习的方法与最先进的传统方法在精度上已经持平，甚至在某些

方面已经超越传统算法。近期文献[26]又提出了一个轻量型的光流估计网络,通过金字塔网络提取多尺度的空间特征,再依次在各个尺度下进行一系列估计优化并最终级联到一起输出光流,在精度牺牲不大的前提下减小了网络体积并提高了运行速度。

这些基于卷积神经网络的深度学习光流估计方法,对简单条件下的图像效果尚可,但是应对诸如遮挡、大位移和细节呈现等难点问题时,还有很大的优化空间。本文为了更好地解决此类难题,提出一种基于卷积神经网络的新型光流估计模型,通过对卷积形式和特征融合部分做针对性优化,让网络可以更好地学习邻帧图像间的变化和联系,其在处理遮挡、大位移和细节呈现方面的实验结果优于上述已有方法。

2　可形变卷积和注意力机制的网络实现

光流估计需要利用图像序列中运动像素的变化以及相邻帧之间的关联性来找到帧间存在的相互关系,从而计算出相邻帧之间物体的运动信息(得到光流图)。因此,本文从相邻帧特征提取和特征融合两方面入手,提出一种基于卷积神经网络的新型光流估计模型。

2.1　总体网络架构

网络具体流程如图1所示。首先,将相邻的两帧RGB图像I_1和I_2分别输入包含可形变卷积的卷积层,提取包含空间信息的特征图;然后将特征图逐通道叠加输入到基于注意力机制的特征关联层中,分析两帧特征的相关性后进行融合重构,然后对重构的特征进行一系列反卷积和上采样操作估计出光流,并且每层反卷积都会有对光流的初步估计并作为参考输入到下一层反卷积,本文称作单一网络(Deformable and attention-single, DANet-S);最后将估计的光流通过Warp计算出估计损失,再利用不同的网络结构进行级联优化后,输出最终的光流估计结果。网络的重点在于可形变卷积和注意力关联层部分,二者的内部具体细节将在后面作详细阐述。

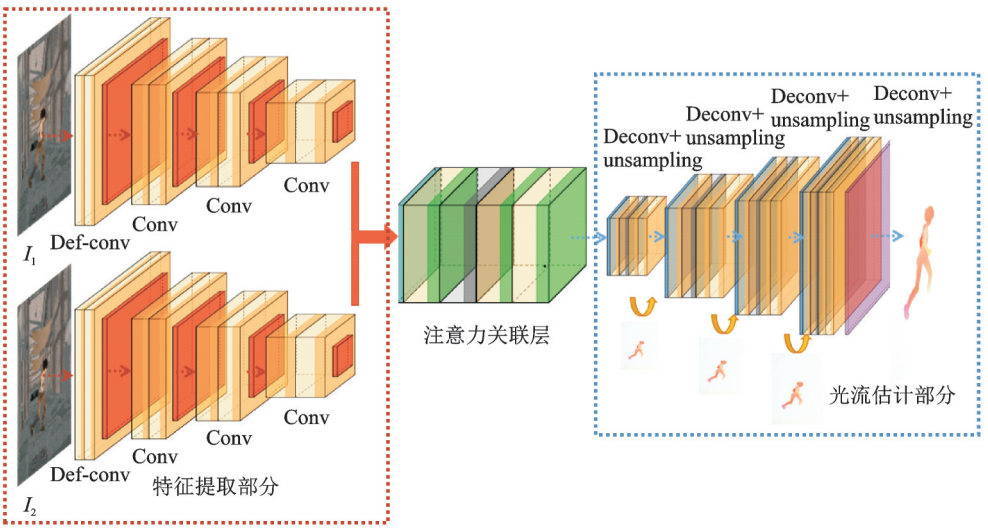

图1　DANet-S结构

Fig.1　DANet-S structure

2.2　可形变卷积特征提取

卷积作为卷积神经网络的核心,其操作是在局部感受野上,将空间上和特征维度上的信息进行聚合的信息聚合体。以往基于深度学习的光流估计方法使用的卷积核多为方形卷积核,其对输入映射的

固定位置进行采样，即在一个卷积层中所有的激活单元感受野都是一样的，这种卷积形式限制了网络在图像自适应优化的空间，不利于捕捉运动轮廓的细节，并且对帧之间像素大小位移的适应性差，尤其是大位移情况下需要更大的感受野来捕捉像素的运动。因此本文对网络中的卷积进行了改进，将网络的第1层卷积改为了适应调整能力更强的可形变卷积，希望通过可形变卷积可以更好地捕捉图像物体运动的细节和大位移。

2.2.1 可形变卷积捕获细节和大位移特征

如图2所示，可形变卷积与传统方形卷积的区别在于，其在各个卷积采样点的位置都增加了一个偏移量。通过这些偏移变量，卷积核就可以在当前位置附近随意地采样，而不再局限于之前的规则格点。

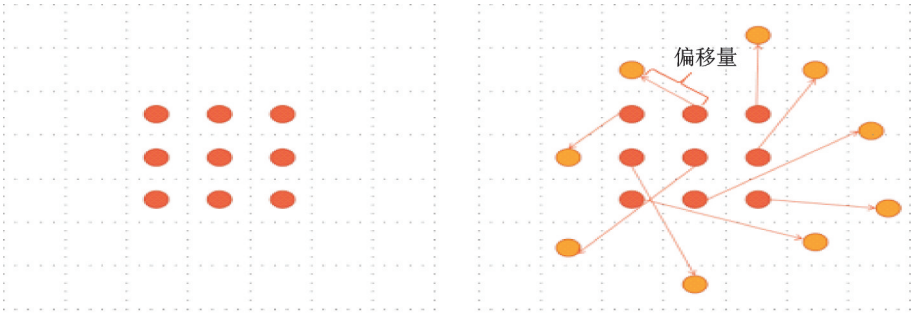

图2 卷积结构图

Fig.2 Convolution structure

对于卷积层所输出特征映射 $\boldsymbol{Y}$ 的一点 $\boldsymbol{P}_0$，传统的方块卷积核操作原理可以用公式表示为

$$y(\boldsymbol{P}_0)=\sum_{\boldsymbol{P}_n\in R}\boldsymbol{w}(\boldsymbol{P}_n)\cdot\boldsymbol{x}(\boldsymbol{P}_0+\boldsymbol{P}_n) \tag{1}$$

式中：$\boldsymbol{x}$ 代表该层的输入特征映射或者原始图像；R 为卷积核所覆盖在 $\boldsymbol{x}$ 的区域；$\boldsymbol{w}$ 为采样的权重值；$\boldsymbol{P}_n$ 则为 R 在 $\boldsymbol{x}$ 所覆盖区域中的遍历。

针对可形变卷积，其增加的偏移量是卷积网络的一部分，可以通过另外一个平行的标准卷积计算得到，进而也可以通过反向传播进行学习，具体过程如图3所示，也可通过公式表示为

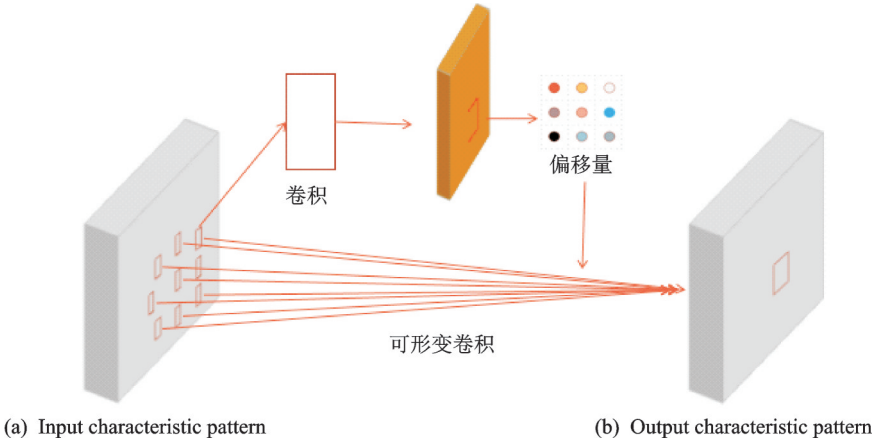

图3 可形变卷积操作

Fig.3 Deformable convolution operation

$$y(\boldsymbol{P}_0)=\sum_{\boldsymbol{P}_n\in R}\boldsymbol{w}(\boldsymbol{P}_n)\cdot\boldsymbol{x}(\boldsymbol{P}_0+\boldsymbol{P}_n+\Delta\boldsymbol{P}_n) \tag{2}$$

式中$\Delta\boldsymbol{P}_n$即为采样点的偏移量,偏移量作为网络参数的一部分,可以通过网络训练自适应调整得到。其获得方式如下:使用一个与原始卷积平行其大小相同的卷积,其卷积核每个采样点的内部权重参数,作为原始卷积核的对应采样点的偏移量$\Delta\boldsymbol{P}_n$,经过网络训练后得到最优化的采样点偏移量,然后通过该偏移量调整原始卷积采样位置。通过加入这一变量,卷积核采样的位置就从固定的规则格点变成了可通过训练调整的随机位置。

由于学习到的偏移量$\Delta\boldsymbol{P}_n$通常为小数,而小数坐标显然无法在图像上操作,因此采用双线性插值将采样点坐标转换为整数。假设原始采样点坐标为(4,7),偏移量为0.5,对应的坐标点为(4.5,7.5),那么寻找距其最近的4个像素点为(4,7),(4,8),(5,7)和(5,8),对这4点的值进行双线性插值作为(4.5,7.5)的数值。可用公式表示为

$$\boldsymbol{x}(\boldsymbol{p})=g(G(\boldsymbol{q})) \tag{3}$$

式中:$G(\boldsymbol{q})$表示距采样点$\boldsymbol{q}$最近的4个像素点,$\boldsymbol{x}(\boldsymbol{q})$为输出特征图坐标;$g$代表双线性插值过程。

加上该偏移量的学习之后,可变形卷积核的大小和采样点的位置可以根据当前图像的特征进行自适应的调整,这样就可以更好地获取图片中不同物体的细节,并满足不同大小的位移所需要的感受野。

2.2.2　可形变池化

卷积往往是一个特征升维的过程,特征维度高不仅计算耗时,而且容易导致过拟合,所以卷积后要进行降维。在卷积神经网络中通常使用池化的操作来对特征进行降维,由于前面特征提取部分采用的是可形变卷积,因此这里采用相对应的可形变池化。和可形变卷积相同,可形变池化在其中加入了一个偏移量,偏移量$\Delta\boldsymbol{P}_{ij}$的产生过程与可形变卷积操作过程相同,其作为网络参数的一部分,可以通过网络训练自适应调整得到。偏移量获得方式如下:使用一个与原始池化操作平行独立的卷积,其卷积核每个采样点的内部权重参数,作为原始池化操作的每一个采样点的偏移量$\Delta\boldsymbol{P}_{ij}$,经过网络训练后得到最优化的采样点偏移量,然后通过该偏移量调整原始池化采样点坐标。具体过程如图4所示,可用公式表示为

$$y(i,j)=\sum_{\boldsymbol{P}\in\mathrm{bin}(i,j)}x(\boldsymbol{P}_0+\boldsymbol{P}+\Delta\boldsymbol{P}_{ij})/n_{ij} \tag{4}$$

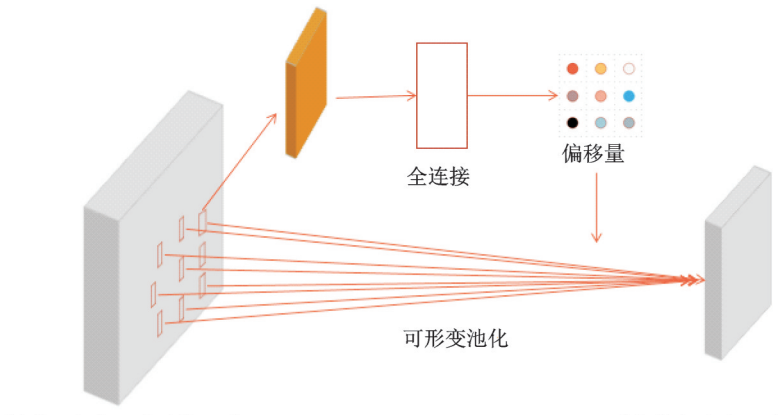

图4　可形变池化操作

Fig.4　Deformable pooling operation

2.3　基于注意力机制的帧间特征关联层

光流估计要寻找相邻帧图像像素的运动状态,所以有了通过前面卷积层操作提取的两个相邻帧的

独立特征，还需要将两部分特征融合起来计算出两者的关联性，才能进行反卷积得到高精度的光流图像，但是简单的特征叠加或匹配难以凸显特征间的关联性。文献[27]表明引入注意力机制能够提高网络对图像的理解和处理能力，因此本文建立基于注意力机制的特征关联层，将相邻两帧的特征逐通道叠加后进行重构，最大限度保留有用图像空间特征的同时计算两部分特征的相关性，以便后续的反卷积操作能够更好地估计出精确、清晰的光流。

将前面所提两帧特征图叠加，得到一个特征通道数为 C，宽和高为 W、H 的融合特征 $\boldsymbol{U}$。其每个通道都代表着原始帧的一部分相关信息，但是由于像素点移动的区域和运动状态的不同，各个通道包含的信息利用价值也不同，因此就需要通过网络学习训练找到“重要”的通道，并让这些通道内的信息扮演更重要的角色，与此同时抑制次要特征，避免学习过多无关信息导致光流图像失真。具体流程如图5所示，首先将特征 $\boldsymbol{U}$ 进行空间维度上的降维，然后通过全连接层学习训练，得出长度为通道数的权重 $\boldsymbol{S}$，最后使用权重参数对原特征 U 进行乘法加权，输出重构后的特征 $\boldsymbol{U}'$。

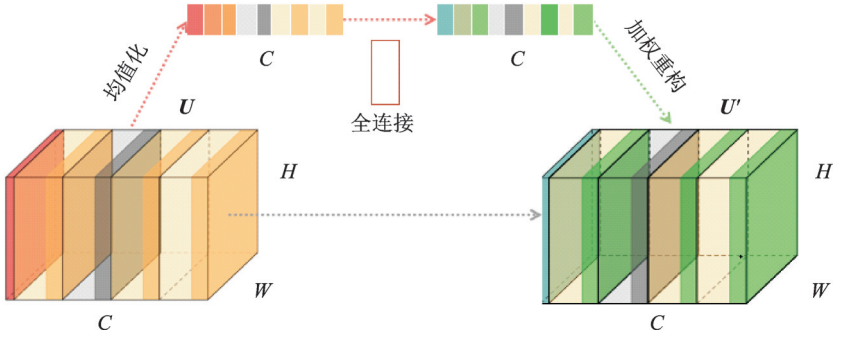

图5 关联层操作

Fig.5 Association layer operation

首先通过全局平均池化将融合的特征 $\boldsymbol{U}$ 在空间维度上进行压缩，全局平均池化可以将 $\boldsymbol{U}$ 中各个通道上的空间信息转化为一个数值，而这个实数具有全局的感受野，表示特征在通道上相应的全局数值分布情况，也可以一定程度上代表该通道的特征属性，这样一个多通道的特征就被转化为长度为通道数的一维向量。具体操作可用公式表示为

$$z_c = F_c(\boldsymbol{u}_c) = \frac{1}{W \times H}\sum_{i=1}^{W}\sum_{j=1}^{H}\boldsymbol{u}_c(i,j) \tag{5}$$

式中 $\boldsymbol{u}_c$ 代表融合特征 $\boldsymbol{U}$ 中通道为 c 的二维特征，其在全区域上累加取平均数，最终每个通道得到一个标量 Z_c，C 个通道组合成一个长度为 C 的一维向量 $\boldsymbol{z}$。该向量后面通过全连接层和激活函数进行学习训练来表示对应通道的重要程度(权重)，然后将特征 $\boldsymbol{U}$ 的每个通道用对应的权重进行加权，即对应通道特征中每个元素与权重分别相乘，然后得到重新标定权重的特征，即

$$\boldsymbol{s} = F_e(\boldsymbol{z}, \boldsymbol{W}) = \sigma(\boldsymbol{W}_2\delta(\boldsymbol{W}_1\boldsymbol{z})) \tag{6}$$

式中：向量 $\boldsymbol{z}$ 首先与一个维度为 $C/r \times C$ 的参数矩阵 $\boldsymbol{W}_1$ 相乘，即一步全连接层操作，其中 r 为一个固定值，起到减少通道数从而降低计算量的作用；然后再经过一个 ReLU(Rectified linear unit)线性整流层 δ，输出维度变化结果；后面再进行一次全连接层操作，与一个维度为 $C \times C/r$ 的参数矩阵 $\boldsymbol{W}_2$ 相乘，这里输出的向量长度就变回 C；最后再经过非线性激活函数 σ，得到长度为 C 的权重向量 $\boldsymbol{s}$。

得到了代表 $\boldsymbol{U}$ 各通道重要性的权重向量 $\boldsymbol{s}$ 之后，将其逐通道加权到先前的融合特征 $\boldsymbol{U}$ 上，就完成了对特征的融合重构，即

$$\boldsymbol{u}_c' = F_{\text{scale}}(\boldsymbol{u}_c, \boldsymbol{s}_c) = \boldsymbol{s}_c \cdot \boldsymbol{u}_c \tag{7}$$

式中:$\boldsymbol{u}_c$为$\boldsymbol{U}$的第c个通道特征;$\boldsymbol{s}_c$为权重向量$\boldsymbol{s}$的第c维权重;$\boldsymbol{u}_c'$为重标定后的第c维融合特征,最终融合特征表示为$\boldsymbol{U}'$。

加入基于通道注意力机制的特征融合重构部分,将图像的高维映射特征重新分配权重,增强了网络的调整适应能力,有利于捕捉相邻帧间的相关性,使网络能够更好地解决大位移与遮挡问题。

2.4　光流估计

将上面融合后的低分辨率的高维特征转化为光流,所采用的方法是上采样和反卷积。反卷积操作是卷积操作的逆运算,主要功能是放大特征映射,提高图像分辨率,能够让网络更好地学习输入输出关系。上采样则是负责将估计出的小分辨率光流逐步放大,使最终光流分辨率达到要求的精度。这里使用FlowNet[17]的优化方法,在反卷积步骤中加入上一层反卷积的估计F(粗略光流)作为下一层反卷积的参考,这种方法在光流估计领域被广泛应用。通过这种方式,既保留了上层的特征图所传递的高级信息,也保留了底层特征图中提供的精细局部信息。在这一部分共包括4层反卷积和上采样,每层反卷积后包括一层ReLU激活函数,每次反卷积操作都将Feature map放大两倍,最后得到的光流图通过上采样(双线性插值法)恢复到原始图片级别的清晰度。

2.5　网络级联光流优化

由于可形变卷积和关联层运算比较复杂,且单一的网络结构难以应对复杂多样的图像运动。为了优化最终结果、提升网络性能,本文在后续加入了Warp和网络堆栈的部分,其主要原理是将几个结构、特点不同的网络级联在一起形成一个网络堆栈,使各个子网络输出的光流经过多个网络循环优化再组合在一起,以起到提高光流估计精度的效果。

2.5.1　Warp计算损失量

Warp操作是基于光流信息的原理,计算出子网络光流所描述的运动场与实际运动场之间的差距(损失量),并将这个损失量输入到下级子网络,使下级子网络能够专注于学习这个差距。具体描述为

$$\boldsymbol{I}'_2 = (\boldsymbol{I}_1, \boldsymbol{F}) \tag{8}$$

$$\boldsymbol{C} = \left\| \boldsymbol{I}_2 - \boldsymbol{I}'_2 \right\| \tag{9}$$

设$\boldsymbol{I}_1$、$\boldsymbol{I}_2$分别为视频相邻帧,式(8)中$\boldsymbol{F}$为上层子网络所估计光流,由于光流的定义为图像的运动场,则可知$\boldsymbol{I}_1$结合$\boldsymbol{F}$可得近似的相邻帧$\boldsymbol{I}'_2$,为描述$\boldsymbol{I}_2$与$\boldsymbol{I}'_2$之间的差距,式(9)定义了损失量$\boldsymbol{C}$。

2.5.2　级联网络优化光流

如图6所示,本文中的网络堆栈设置了3种结构和内部模块不同的子网络。网络1即DANet-S,采用可形变卷积(第1层卷积)和基于注意力机制的关联层(如图2所示);网络2使用可形变卷积并去除基于注意力机制的关联层,将特征直接叠加;网络3使用了传统卷积加基于注意力机制的关联层。实验表明,若堆栈中加入可形变卷积并去除基于注意力机制的关联层的子网络,则整个网络无法收敛,因此在此不做该组合的网络设定。其中每个网络中设置6级卷积层,每次卷积移动步长为2,其中第1层卷积核大小为7×7,第2层5×5,后面4层大小皆为3×3,每层卷积后都跟着一个ReLu整流函数。后面光流估计部分设置4级反卷积层,每层反卷积后同样都加1个ReLu整流函数。每级反卷积层的操作如下:首先将该层的融合特征和上层反卷积层所估计的光流(通过上采样将尺度恢复到与该层特征相同尺度)进行叠加;然后再通过反卷积估计出光流,这样每一层输出的光流不仅来自于融合特征,还利用了上级反卷积层估计出的光流,可以有效提高光流精度,如此传递到子网络的最后一级反卷积层输出光流估计结果。而最后的合成模块将相邻帧和子网络估计的光流合并再进行两次缩小和放大(卷积和反卷积)操作,目的是要把前面几个子网络的估计结果进行融合。合成模块本质上是一个小型卷积网络,

如前面的网络2的结构，但是网络内部层数较少，只有两级卷积和反卷积层，且输入为前面子网络估计出的光流图和损失量，这样融合模块就可以结合不同子网络的优点和特性，输出最终的优化结果。通过实验测试，按照图6结构进行子网络堆栈组合效果最优，本文将这个大型网络称为DANet-C(DANet-cascade)。

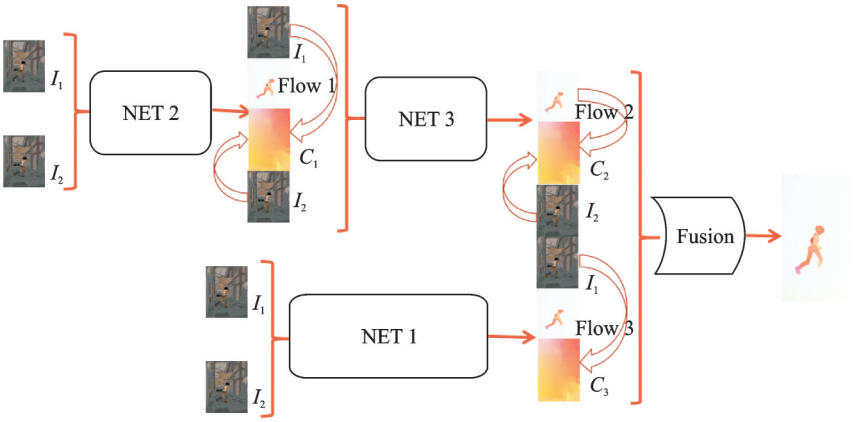

图6 DANet-C结构

Fig.6 DANet-C structure diagram

3 网络训练

本文模型在Pytorch 0.4.1框架下构建和训练，Ubuntu版本为18.04，Cuda版本为9.0，显卡型号为NVIDIA GTX1080Ti。网络以监督学习的形式学习训练，平均终点误差作为网络损失函数，来自数据集中Ground truch与估计光流插值后图像两者的对比，因此训练所采用的数据集以及策略将会很大程度上影响网络性能，而且多个网络级联虽然可以有效提高估计精度，但又有以下几个缺点：网络结构复杂庞大，训练速度慢且容易出现过拟合或不收敛的情况；多层级多支流的子网络下行分享信息，不仅导致误差的传递，还会引起损失计算混乱的问题；堆栈网络需要大量的计算成本，在内存较小的设备上容易导致空间不足。

分步训练是一种级联网络常用的训练策略，其思想是通过分批次训练大型网络分割出的各个子网络。首先训练子网络收敛并达到一定的精度要求，然后再将子网络连接后对融合模块(即级联网络最后一层)进行训练微调。由于小型网络的参数量小、所需计算成本低、反向传播和参数更新速度快，因此可以提高网络训练效率，并防止不收敛和过拟合现象发生；训练好的子网络级联后微调，可以固定已训练好的网络层，不破坏相应权重参数，所以有减少误差传递的效果。采用这种策略可以从网络训练的角度优化最终估计结果，因此本文采用分步训练的策略。下面是本文分步训练的具体细节：将数据集分为训练集和测试集，分别对几个子网络进行训练，然后连接在一起对合成模块进行微调。对于网络2，训练的初始学习率设置为$\lambda=10^{-4}$，然后进行3×10^5次迭代，之后每进行1×10^5次迭代就将学习率减小一半；对于网络3，初始学习率同样设置为$\lambda=10^{-4}$，然后进行4×10^5次迭代，后面的策略与网络2相同。由于DANet-S(网络1)的内部结构最为复杂，导致采用类似策略训练时出现了梯度爆炸的情况，为了解决这个问题，本文将网络1的初始学习率设为$\lambda=10^{-6}$，在经过10^4次迭代后，再增加到10^{-4}，然后再依照网络3的策略进行后续训练；而对于合成模块则设置初始学习率为10^{-5}，2×10^5次迭代后就可以减小学习率。

4　实验与分析

为了验证本文网络采用的可形变卷积和基于注意力机制的关联层是否合理有效,本文分别在Flying Chairs和Mpi Sintel两个数据集上分别进行了测试。

4.1　数据集介绍

本文在两个光流数据集上进行试验。第1个是Flying Chairs数据集,它是一个合成数据集,由22 872个图像对和相应的光流图像组成。图像内容是3D椅子模型在随机的背景前做无规则运动,但椅子和背景只在平面上运动,缺少复杂的运动形式,是光流估计中的基础数据集。使用该数据集是为了检验本文方法应对简单运动时是否有着较高的精度和适应性。第2个是Mpi Sintel数据集,它是由1 064个从电影中采集的动画图像对组成的开源数据集,并分别对图像施加不同的处理效果,分为Clean和Final两部分,其中每对原始图像都有相对应的Ground truth。图像包含运动模糊、多帧分析和非刚性运动等多个光流估计的常见问题,是光流估计领域中最常用的数据集之一。其中包括本文方法所期望解决的遮挡、大位移和细节呈现等问题,因此使用Mpi Sintel数据集来验证本文方法解决上述复杂问题的能力。

4.2　基于Flying Chairs数据集的算法测试和精度比较

4.2.1　对比方法

本文选取了4个对比网络,包括FlowNetC[17]、SPyNet[18]、LiteFlowNet[26]和FlowNet2[25]。其中,FlowNetC[17]基于卷积神经网络,使用匹配形式的特征融合模块估计光流,但由于网络结构较为简单,对复杂图像的适应性较差;SPyNet[18]将空间金字塔与深度学习相结合,从多个尺度估计光流,实现由粗到细的优化,并减少了网络参数量,但其缺少对遮挡问题的针对性优化,容易出现大面积模糊;FlowNet2[25]在原有FlowNet[17]的基础上使用网络堆栈的形式进行循环优化,并改进了训练策略,缺点在于网络结构庞大、训练困难,且存在边缘模糊、图像细节显示不清晰的问题;LiteFlowNet[26]是通过金字塔特征逐级正则化,并结合中值滤波建立的轻量型光流估计网络,由于追求网络轻量化损失了部分性能,整体精度较低,在应对遮挡、大位移和图像细节呈现上都不够理想。

4.2.2　数据分析对比

表1给出了上述网络与本文方法在Flying Chairs数据集上的估计结果,使用网络输出光流和数据集中Ground truth之间的平均终点误差(Average endpoint error,AEE)评估光流估计精度。从表1中可知,SPyNet[18]由于其多尺度估计机制,导致光流图像容易出现大范围模糊,因此整体精度较差。而FlowNetC[17]和LiteFlowNet[26]在特征融合部分所做的优化,使其在解决Flying Chairs数据集中的简单运动问题,效果反而要比SPyNet[18]更为突出,但结果仍不理想。FlowNet2[25]的级联结构可以很好地帮助网络适应各种运动情况,因此整体精度较高。而本文级联网络DANet-C的AEE误差为1.75,达到最小值;这表明本文DANet-C网络在应对平面和小位移等简单运动上,同样具有更好的鲁棒性和光流估计精度,显示出本文所采用的可形变卷积、基于注意力机制的关联层和网络级联结构的优越性。

表1　不同的深度学习网络在Flying Chairs上的表现

Table 1　Performance of different deep learning networks on Flying Chairs dataset

Model	AEE
FlowNetC[17]	2.19
SPyNet[18]	2.63
LiteFlowNet[26]	2.25
FlowNet2[25]	1.81
DANet-S	1.93
DANet-C	**1.75**

4.3 基于Mpi Sintel数据集的算法测试和精度比较

由于Flying Chairs数据集中的图像物体运动比较简单,为了检测本文网络面对大位移、细节呈现、遮挡等复杂情况下的表现。本文又使用Mpi Sintel这一包含复杂运动的数据集对网络进行了训练和测试。

4.3.1 对比方法

为了全面验证本文方法的优越性,不仅优于深度学习方法,更比传统变分方法优越。本文在Mpi Sintel数据集上,除上述4种深度学习方法外,又添加了当前具有代表性的变分方法,LDOF[28]、DeepFlow[29]和PCA-Layers[30],与本文方法进行量化对比和结果分析。其中LDOF[28]采用的是基于特征匹配的光流计算模型,由于采用匹配机制,导致在光流图像的边缘存在大区域的过度平滑;DeepFlow[29]提出一种针对大位移光流的描述符匹配算法,采用能量最小化策略提高估计精度,其所采用手工设计的描述符适应性差,难以应对包含遮挡和细节呈现问题的光流估计;PCA-Layers[30]通过提取光流场的主要成分,并以不同权重叠加稀疏特征进行匹配,估算光流,但提取主成分的同时也损失了很多图像细节,导致整体精度较低,存在大范围的边缘模糊和噪声。

4.3.2 精度分析对比

从表2中各方法的终点误差(End point error,EPE)可知,采用基于特征匹配的LDOF[28]与DeepFlow[29]方法难以应对Mpi Sintel数据集中存在的复杂运动,存在的大范围过度平滑和模糊也降低了整体估计精度。采用主成分分析和稀疏特征的PCA-Layers[30],也由于无法完整保留原始图像的细节和边缘,导致误差较大。同时FlowNetC[17]、SPyNet[18]和LiteFlowNet[26]这3种深度网络由于自身的单一结构导致逼近能力差,难以拟合复杂原始图像与光流图像的输入输出关系,因此在Mpi Sintel数据集上整体误差高。而本文级联的DANet-C在Mpi Sintel数据集Clean部分上的误差降低到了3.830和1.285,明显优于其他的估计算法,且一定程度上优于同样为级联网络FlowNet2[25]的3.959和1.468,说明本文采用的模型针对图像中的复杂运动也能起到比较好的提升效果,也证明了可形变卷积+注意力关联层的组合能够更好地提取相邻帧的空间细节特征,提高网络调整适应和学习相邻帧的相关性的能力,最终提高整体估计精度。由于Mpi Sintel数据集Final部分的图像数据存在较多的模糊现象,而本文方法并没有针对模糊问题做相应优化,因此在该部分的精度表现没有达到最优,误差为5.500和2.978,高于采用了多种后续细化方法的LiteFlowNet[26],但精度上仍优于其他对比方法。表明本文提出的可形变卷积和注意力关联层网络模型在应对模糊问题时,仍然有一定的优化效果。

表2 不同方法在Mpi Sintel的表现

Table 2 Performance of different methods on Mpi Sintel dataset

Method	Sintel Clean		Sintel Final	
	EPE all	EPE matched	EPE all	EPE matched
LDOF[28]	7.563	3.432	9.116	5.037
DeepFlow[29]	5.377	1.771	7.212	3.336
PCA-Layers[30]	5.730	2.455	7.886	4.256
SPyNet[18]	6.640	3.013	8.360	4.512
SPyNet-tf[18]	6.689	3.020	8.431	4.549
FlowNetS[17]	6.158	2.800	7.218	3.752
FlowNetC[17]	6.081	2.576	7.883	4.132
FlowNet2[25]	3.959	1.468	6.016	2.977
LiteFlowNet[26]	4.539	1.630	5.381	2.419
LiteFlowNetX[26]	4.664	1.540	5.417	2.549
DANet-S	7.527	2.681	7.851	3.855
DANet-C	**3.830**	**1.285**	**5.500**	**2.978**

4.4 光流估计的鲁棒性分析

为了更直观地评估本文模型对具有复杂运动图像的光流估计效果,本文从Mpi Sintel数据集中选取了几个具有典型复杂运动特征的图像对进行展示。图7给出了原始RGB图像、真实光流和各个方法输出的光流估计结果。

4.4.1 遮挡问题的鲁棒性分析

图7第1行中存在遮挡现象,前景的

人物遮挡了后面物体的一部分。可以看出SPyNet[18]所呈现出的效果最差,边缘模糊不清,存在过度平滑现象,且在物体的某些部分出现了明显的错误估计;LiteFlowNet[26]在图像遮挡部分也存在大量噪声,背景与物体的边缘保护表现差;FlowNet2[25]在物体分割表现良好,且边缘呈现清晰,但损失了部分图像细节;而本文模型DANet-C则完整显现了图像交叠部分,局部遮挡位置的轮廓线条也比较清晰,和真实图像相似性更高。

4.4.2　大位移问题的鲁棒性分析

图7第2行的相邻帧像素运动范围大,属于大位移现象,比较适合检验光流估计应对大位移的能力。从图中可以看出SPyNet[18]的不同尺度提取图像特征的方式,虽然处理速度较快,但难以捕捉相邻帧像素的大范围位移,图像模糊不清,独立整体被分割,部分细节完全丢失。LiteFlowNet[26]则在位移较大的部分模糊了前景与背景,部分出现大面积错误,损失了物体边缘。FlowNet2[25]整体表现良好,物体轮廓存在过度平滑的现象,区分度较差。而本文所提出的光流估计网络DANet-C在完整估计出大位移物体的同时,又保留了较多的图像细节,物体边缘轮廓也比较清晰。

4.4.3　细节图像的光流估计鲁棒性分析

为了进一步展示本文方法在细节呈现上效果,图8分别截取了图7中框定的局部图像并进行放大。从图中可以看出LiteFlowNet[26]由于采用轻量级网络,损失了部分逼近能力,图像细节损失较多,存在大片模糊现象;第2行中SPyNet[18]所显示的尾部被分离,损失了物体的整体性;FlowNet2[25]在第1行

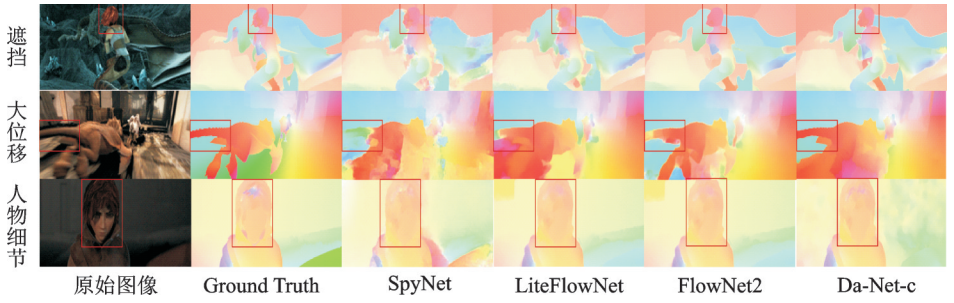

图7　图像光流估计结果对比

Fig.7　Comparison of image optical flow estimation results

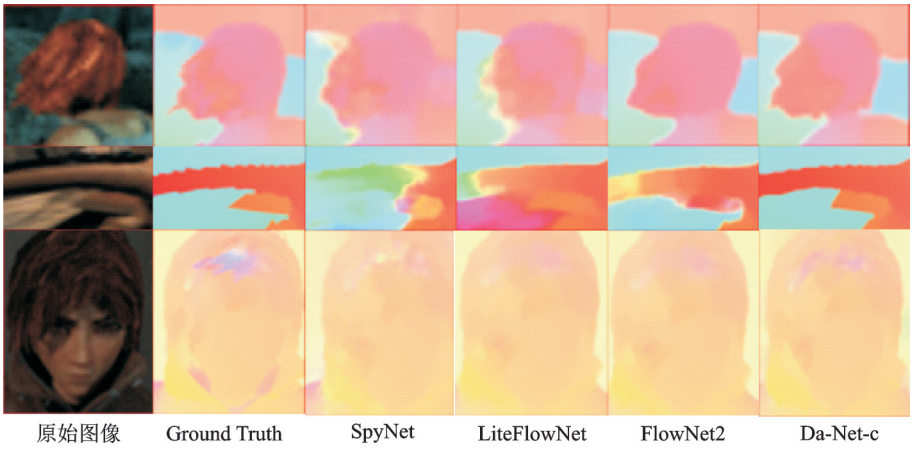

图8　光流图像局部放大对比

Fig.8　Magnification contrast of optical flow image local

中,人物头部边缘估计粗糙,在第3行中也存在丢失细节现象;而从第1行看出,本文方法在保留大部分细节的同时,边缘也比较清晰,基本没有模糊,说明本文对细小边缘有着较好的保护效果。同时在第3行中可看出,本文方法的估计结果更接近于真实值,说明本文方法针对图像中亮度近似部分的细节估计同样有着良好的适用性。

5 结束语

本文构建了一个基于可形变卷积和注意力机制的光流估计模型,其通过包括可形变卷积的特征提取部分提取相邻帧的图像空间特征,然后利用基于注意力机制的特征关联层将特征融合重构,再对特征进行反卷积来估计光流,最后通过多网络堆栈对光流循环优化,实现最终的光流估计输出。通过对比本文模型与其他方法的实验结果,显示出本文方法有明显优势,主要表现在面对遮挡、大位移和图像细节呈现等复杂问题上有更高的精度和鲁棒性;证明了可形变卷积和基于注意力机制的关联层在解决此类问题中的重要作用,同时也证明了本文模型的合理性和有效性。未来的工作计划进一步优化网络模型,提升在背景分离等方面的不足,使得该模型能够适用于更多的图像运动。

参考文献:

[1] MALCOLM N, GIBSON J J. The perception of the visual world[J]. Philosophical Review, 1951, 60(4): 594.

[2] LAI H Y, TSAI Y H, CHIU W C. Bridging stereo matching and optical flow via spatio-temporal correspondence[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Angeles, USA: IEEE, 2019: 1890-1899.

[3] REN Z, GALLO O, SUN D, et al. A fusion approach for multi-frame optical flow estima-tion[C]//Proceedings of 2019 IEEE Winter Conference on Applications of Computer Vision (WACV). [S.l.]: IEEE, 2019: 2077-2086.

[4] LIU P, KING I, LYU M R, et al. Ddflow: Learning optical flow with unlabeled data di-stillation[C]//Proceedings of the AAAI Conference on Artificial Intelligence. [S.l.]: AAAI, 2019, 33: 8770-8777.

[5] LIU X, QI C R, GUIBAS L J. Flownet3D: Learning scene flow in 3D point clouds[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2019: 529-537.

[6] 储林臻, 闫钧华, 杭谊青, 等. 基于改进光流法的旋转运动背景下对地运动目标实时检测[J]. 数据采集与处理, 2015, 30(6): 1325-1331.
CHU Linzhen, YAN Junhua, HANG Yiqing, et al. Real time ground moving object detection in rotational[J].Journal of Data Acquisition and Processing, 2015, 30(6): 1325-1331.

[7] 魏本征, 尹义龙. 基于局部特征约束的 TEM 图像分割算法[J]. 数据采集与处理, 2018, 33(3): 400-408.
WEI Benzheng, YIN Yilong. Local feature-constraint information based TEM image segmentation algorithm[J].Journal of Data Acquisition and Processing, 2018, 33(3): 400-408 .

[8] 刘赏, 董林芳. 人群运动中的视觉显著性研究[J]. 数据采集与处理, 2017, 32(5): 890-897.
LIU Shang, DONG Linfang. Research on visual saliency of crowd movement[J].Journal of Data Acquisition and Processing, 2017, 32(5): 890-897.

[9] WANG Y, YANG Y, YANG Z, et al. Occlusion aware unsupervised learning of optical flow[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2018: 4884-4893.

[10] ILG E, CICEK O, GALESSO S, et al. Uncertainty estimates and multi-hypotheses networks for optical flow[C]// Proceedings of the European Conference on Computer Vision (ECCV). Munich, Germany: Springer, 2018: 652-667.

[11] ILG E, SAIKIA T, KEUPER M, et al. Occlusions, motion and depth boundaries with a generic network for disparity, optical flow or scene flow estimation[C]//Proceedings of the European Conference on Computer Vision (ECCV). Munich, Germany: Springer, 2018: 614-630.

[12] DAI J, QI H, XIONG Y, et al. Deformable convolutional networks[C]//Proceedings of the IEEE International Conference On Computer Vision. Venice, Italy: IEEE, 2017: 764-773.

[13] HU J, SHEN L, SUN G. Squeeze-and-excitation networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA: IEEE, 2018: 7132-7141.

[14] 葛利跃, 张聪炫, 陈震, 等. 相互结构引导滤波 TV-L1 变分光流估计[J]. 电子学报, 2019, 47(3): 707-713.
GE Liyue, ZHANG Congxuan, CHEN Zhen, et al. Mutual-structure guided filtering based TV-L1 optical flow estimation[J]. Acta Electronica Sinica, 2019, 47(3): 707-713.
[15] 张聪炫, 陈震, 熊帆, 等. 非刚性稠密匹配大位移运动光流估计[J]. 电子学报, 2019, 47(6): 1316-1323.
ZHANG Congxuan, CHEN Zhen, XIONG Fan, et al. Large displacement motion optical flow estimation with non-rigid dense patch matching [J].Acta Electronica Sinica, 2019, 47(6): 1316-1323.
[16] 张子星, 文颖. 基于分层置信度传播的光流估计方法[J]. 计算机系统应用, 2018, 27(9): 25-32.
ZHANG Zixing, WEN Ying. Hierarchical belief propagation for optical flow estimation[J]. Computer Systems and Applications, 2018, 27(9): 25-32.
[17] DOSOVITSKIY A, FISCHER P, ILG E, et al. Flownet: Learning optical flow with convolutio-nal networks[C]//Proceedings of the IEEE International Conference on Computer Vision. Santiago, Chile: IEEE, 2015: 2758-2766.
[18] RANJAN A, BLACK M J.Opticalflow estimation using a spatial pyramid network[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI, USA: IEEE, 2017: 4161-4170.
[19] GADOT D, WOLF L. PatchBatch: A batch augmented loss for optical flow[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA: IEEE, 2016: 4236-4245.
[20] TRAN D, BOURDEV L, FERGUS R, et al. Deep end2end voxel2voxel prediction[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. [S.l.]: IEEE, 2016: 17-24.
[21] AHMADI A, PATRAS I. Unsupervised convolutional neural networks for motion estimation[C]//Proceedings of 2016 IEEE International Conference on Image Processing (ICIP). [S.l.]: IEEE, 2016: 1629-1633.
[22] JASON J Y, HARLEY A W, DERPANIS K G. Back to basics: Unsupervised learning of optical flow via brightness constancy and motion smoothness[C]//Proceedings of European Conference on Computer Vision. Cham: Springer, 2016: 3-10.
[23] TENEY D, HEBERT M. Learning to extract motion from videos in convolutional neural networks[C]//Proceedings of Asian Conference on Computer Vision. Cham: Springer, 2016: 412-428.
[24] THEWLIS J, ZHENG S, TORR P H S, et al. Fully-trainable deep matching[C]//Proceedings of British Machine Vision Conference (BMVC). [S.l.]: BMVC, 2016: 145.1-145.12.
[25] EDDY I L G, MAYER N, SAIKIA T, et al. Flownet 2.0: Evolution of optical flow estimation with deep networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2017: 2462-2470.
[26] HUI T W, TANG X, CHANGE L C. Liteflownet: A lightweight convolutional neural network for optical flow estimation [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA: IEEE, 2018: 8981-8989.
[27] 张文, 谭晓阳. 基于 Attention 的弱监督多标号图像分类[J]. 数据采集与处理, 2018, 33(5): 801-808.
ZHANG Wen, TAN Xiaoyang.Weakly supervised multi-label classification based attention mechanism[J]. Journal of Data Acquisition and Processing, 2018, 33(5): 801-808.
[28] BROX T, MALIK J. Large displacement optical flow: Descriptor matching in variational motion estimation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2010, 33(3): 500-513.
[29] WEINZAEPFEL P, REVAUD J, HARCHAOUI Z, et al. DeepFlow: Large displacement optical flow with deep matching [C]//Proceedings of the IEEE International Conference on Computer Vision. [S.l.]: IEEE, 2013: 1385-1392.
[30] WULFF J, BLACK M J. Efficient sparse-to-dense optical flow estimation using a learned basis and layers[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2015: 120-130.

作者简介:

丰艳(1984-),女,博士,副教授,研究方向:图像处理、计算机视觉和虚拟现实,E-mail: fywmh@163.com。

刘帅(1994-),男,硕士研究生,研究方向:图像处理、计算机视觉和虚拟现实。

王传旭(1968-),**通信作者**,男,博士,教授,研究方向:计算机视觉,E-mail:wangchuanxu_qd@163.com。

(编辑:刘彦东)