自动化学报
ACTA AUTOMATICA SINICA

联合深度超参数卷积和交叉关联注意力的大位移光流估计

王梓歌 葛利跃 陈震 张聪炫 王子旭 舒铭奕

Large Displacement Optical Flow Estimation Jointing Depthwise Over-parameterized Convolution and Cross Correlation Attention

WANG Zi-Ge, GE Li-Yue, CHEN Zhen, ZHANG Cong-Xuan, WANG Zi-Xu, SHU Ming-Yi

在线阅读 View online: https://doi.org/10.16383/j.aas.c230049

联合深度超参数卷积和交叉关联注意力的大位移光流估计

王梓歌 [1, 2] 葛利跃 [1, 3] 陈 震 [1, 2, 4] 张聪炫 [1, 2, 4] 王子旭 [1, 2] 舒铭奕 [1, 2]

摘 要 针对现有深度学习光流估计模型在大位移场景下的准确性和鲁棒性问题, 提出了一种联合深度超参数卷积和交叉关联注意力的图像序列光流估计方法. 首先, 通过联合深层卷积和标准卷积构建深度超参数卷积以替代普通卷积, 提取更多特征并加快光流估计网络训练的收敛速度, 在不增加网络推理量的前提下提高光流估计的准确性; 然后, 设计基于交叉关联注意力的特征提取编码网络, 通过叠加注意力层数获得更大的感受野, 以提取多尺度长距离上下文特征信息, 增强大位移场景下光流估计的鲁棒性; 最后, 采用金字塔残差迭代模型构建联合深度超参数卷积和交叉关联注意力的光流估计网络, 提升光流估计的整体性能. 分别采用 MPI-Sintel 和 KITTI 测试图像集对本文方法和现有代表性光流估计方法进行综合对比分析, 实验结果表明本文方法取得了较好的光流估计性能, 尤其在大位移场景下具有更好的估计准确性与鲁棒性.

关键词 光流, 大位移, 注意力, 深度超参数卷积, 深度学习

引用格式 王梓歌, 葛利跃, 陈震, 张聪炫, 王子旭, 舒铭奕. 联合深度超参数卷积和交叉关联注意力的大位移光流估计. 自动化学报, 2024, **50**(8): 1−15

DOI 10.16383/j.aas.c230049

Large Displacement Optical Flow Estimation Jointing Depthwise Over-parameterized Convolution and Cross Correlation Attention

WANG Zi-Ge[1, 2] GE Li-Yue[1, 3] CHEN Zhen[1, 2, 4] ZHANG Cong-Xuan[1, 2, 4] WANG Zi-Xu[1, 2] SHU Ming-Yi[1, 2]

Abstract To improve the computation accuracy and robustness of deep-learning based optical flow models under large displacement scenes, we propose an optical flow estimation method jointing depthwise over-parameterized convolution and cross correlation attention. First, we construct a depthwise over-parameterized convolution model by combining the common convolution and depthwise convolution, which extracts more features and accelerates the convergence speed of optical flow network. This improves the optical flow accuracy without increasing computation complexity. Second, we exploit a feature extraction encoder based on cross correlation attention network, which extracts multi-scale long distance context feature information by stack the attention layers to obtain a larger receptive field. This improves the robustness of optical flow estimation under large displacement scenes. Finally, a pyramid residual iteration network by combing cross correlation attention and depthwise over-parameterized convolution is presented to improve the overall performance of optical flow estimation. We compare our method with the existing representative approaches by using the MPI-Sintel and KITTI datasets. The experimental results demonstrate that the proposed method achieves better computation accuracy and robustness, especially under large displacement areas.

Key words Optical flow, large displacement, cross correlation attention, cross correlation attention, deep learning

Citation Wang Zi-Ge, Ge Li-Yue, Chen Zhen, Zhang Cong-Xuan, Wang Zi-Xu, Shu Ming-Yi. Large displacement optical flow estimation jointing depthwise over-parameterized convolution and cross correlation attention. *Acta Automatica Sinica*, 2024, **50**(8): 1−15

收稿日期 2023-02-10 录用日期 2023-08-29

Manuscript received February 10, 2023; accepted August 29, 2023

国家自然科学基金 (62222206, 62272209), 江西省重大科技研发专项 (20232ACC01007), 江西省重点研发计划重点专项 (20232BBE50006), 江西省技术创新引导类计划项目 (2021AEI91005), 江西省教育厅科学技术项目 (GJJ210910), 江西省图像处理与模式识别重点实验室开放基金 (ET202104413) 资助

Supported by National Natural Science Foundation of China (62222206, 62272209), National Science and Technology Major Project of Jiangxi Province (20232ACC01007), Key Research and Development Program of Jiangxi Province (20232BBE50006), the Technological Innovation Guidance Program of Jiangxi Province (2021AEI91005), Science and Technology Program of Education Department of Jiangxi Province (GJJ210910), and the Open Fund of Jiangxi Key Laboratory for Image Processing and Pattern Recognition (ET202104413)

本文责任编委 桑农

Recommended by Associate Editor SANG Nong

1. 南昌航空大学江西省图像处理与模式识别重点实验室 南昌 330063 2. 南昌航空大学测试与光电工程学院 南昌 330063 3. 北京航空航天大学仪器科学与光电工程学院 北京 100083 4. 南昌航空大学无损检测技术教育部重点实验室 南昌 330063

1. Key Laboratory of Jiangxi Province for Image Processing and Pattern Recognition, Nanchang Hangkong University, Nanchang 330063 2. School of Measuring and Optical Engineering, Nanchang Hangkong University, Nanchang 330063 3. School of Instrumentation and Optoelectronic Engineering, Beihang University, Beijing 100083 4. Key Laboratory of Nondestructive Testing, Ministry of Education, Nanchang Hangkong University, Nanchang 330063

光流是指图像序列中运动目标或场景表面像素点的二维运动矢量, 其不仅提供了图像中运动目标和场景的运动参数, 还携带了丰富的结构信息. 因此, 图像序列光流估计技术研究是图像处理与计算机视觉领域的研究热点, 研究成果被广泛应用于人体姿态估计[1]、目标跟踪[2]、三维重建[3]、动作识别[4]和表情识别[5−6] 等更高级的视觉任务.

光流估计的目的是找到同一像素点在连续两帧图像序列变化过程中的对应匹配关系, 从而估计出该像素点的运动大小和方向. 自 Horn 和 Schunck[7] 开创性地将光流估计表述为能量最小化问题以来, 出现了许多有效的方法[8−10] 来提高光流估计的性能. 传统方法将光流估计视作一对图像间稠密位移场空间上的手工优化问题. 首先利用视觉相似图像区域对齐的数据项和对运动合理性施加先验正则的平滑项构建经典能量函数, 然后通过最小化能量方程以获取光流估计最优解. 虽然这种方法可以有效提升光流估计的准确性, 但是由于难以设计出对各种情况都具有鲁棒性的优化目标, 因而制约了其进一步发展与在工程领域的应用.

得益于深度学习理论和技术的突破性发展, 目前基于深度学习的光流估计方法已经在估计精度、鲁棒性以及推理时间方面全面超越经典的传统方法. 在模型结构方面, 基于深度学习的光流估计方法主要由特征提取模块、成本量模块和光流估计子网络模块组成. 其中, 特征提取模块是模型实现光流估计的关键, 其特征提取的质量严重影响后续成本量模块和光流估计子网络的工作性能. 在深度学习光流估计早期, 直接采用 U-Net 模式从连续两帧图像序列中提取图像特征用于光流估计[11], 但该方法获取的特征过于粗糙. 此后, 相关研究将图像金字塔引入光流估计网络用于捕获不同运动幅度的图像特征[12], 但分辨率的变化导致图像特征存在较为严重的信息丢失. 后续, 基于由粗到细特征金字塔编码结构的深度学习光流估计模型被证明可以有效处理大位移运动问题. 然而, 在金字塔采样过程中, 由于目标像素损失致使目标在传递过程中存在特征稀释问题, 从而造成大位移运动光流估计存在局部信息丢失, 特别是位移较大的小目标. 同时, 受标准卷积核可学习参数量的内在因素限制, 当前仅依赖标准卷积构建的特征提取网络在特征提取内容丰富度与置信度方面仍存在较大不足. 针对该问题, 传统的光流估计方法[13] 通过将随机搜索策略与由粗到细的方案相结合, 以提高大位移运动光流估计精度. 但由于光流估计网络需要大量迭代估计, 导致模型精度与估计效率不能较好地平衡. 为此, Hur 等[14] 提出一种权值共享编码网络并使用迭代残差优化方案进一步细化光流结果, 在减少模型参数的同时提高了光流估计准确性. 但该模型仅集中于对光流估计的后置处理, 因此, 对特征稀释造成的大位移运动局部信息损失问题, 仍无法妥善解决.

为解决上述问题, 本文提出一种联合深度超参数卷积和交叉关联注意力的大位移光流估计方法. 首先, 针对光流估计模型特征提取置信度与丰富度较低问题, 构建基于深度超参数卷积的光流估计网络, 通过将深层卷积与标准卷积耦合提升卷积特征学习的丰富度, 从而捕获置信度更高的图像特征. 其次, 针对基于由粗到细策略的金字塔模型引起大位移运动局部信息丢失问题, 设计基于交叉关联注意力的特征提取编码网络进行局部到全局的特征编码模型, 通过改变不同尺度下的特征提取感受野增强长距离目标上下文信息建模能力, 从而提高大位移场景下光流估计的准确性与鲁棒性. 本文的主要贡献总结如下:

1) 首次将深层卷积引入光流估计任务, 并与标准卷积耦合构建基于深度超参数卷积的光流估计网络, 通过提高模型特征提取的置信度与内容丰富度, 不仅加快模型训练收敛速度还有效提升了光流估计的可靠性;

2) 提出一种交叉关联注意力的特征提取编码网络, 通过建立局部到全局的注意力感受野变化策略, 实现了不同尺度目标长距离上下文特征关联, 进一步提高了大位移运动光流估计的准确性与鲁棒性;

3) 采用 MPI-Sintel 与 KITTI 等权威测试数据集对本文方法和现有代表性深度学习方法进行综合实验对比分析. 结果表明, 本文方法在大多数测量指标上均取得了最优结果, 尤其在大位移运动区域.

本文内容安排如下: 第 1 节介绍了光流估计方法的相关工作; 第 2 节详述了所提出的联合深度超参数卷积和交叉关联注意力的光流估计方法; 第 3 节给出了本文方法模型损失函数与训练策略; 第 4 节详细叙述了实验结果与分析; 第 5 节是对全文的总结.

1 相关工作

在深度学习技术兴起之前, 基于变分框架的光流估计方法一直占据着传统光流估计研究的主导. 变分框架的光流计算方法将光流估计视为一个在数据项和平滑项寻求平衡的能量最小化问题, 并通过计算图像梯度来估计稠密的光流场[15]. 由于其假设光流在整个图像上的变化是平滑均匀的, 即速度变化率趋近于零, 而实际场景难以满足该理想假设, 因此这导致该类方法普遍存在鲁棒性较差的问题.

针对基于图像亮度守恒假设的数据项难以处理光照变化的问题, 基于图像梯度等高阶数据守恒假设成为后续变分光流数据项的必要补充[16]. 针对一致性平滑策略易导致光流边缘模糊的问题, 基于图像与光流驱动的平滑项扩散策略成为解决该问题的有效手段[17−19]. 针对图像序列存在运动不连续的问题, 基于 L_1 范数的全变分 TV-L_1 模型在实现运动不连续光流估计的同时有效地抑制了异常值[20]. 针对大位移光流估计问题, 基于图像金字塔分层迭代优化方案被证明是解决该问题的有效方法[21]. 虽然经过数十年的研究发展, 基于变分框架的光流估计方法在估计精度和鲁棒性等方面已取得显著提升, 但由于该类方法需要执行大量迭代优化过程, 导致其光流估计时间消耗巨大, 因此难以满足实时任务的要求.

近年来, 随着深度学习理论与技术的快速发展, 卷积神经网络模型被广泛应用于光流估计技术研究, 该类方法[22−23] 由于具有估计速度快、稳定性高等显著优点, 逐渐成为光流估计研究领域的热点. Dosovitskiy 等[11] 使用卷积神经网络搭建了基于有监督学习的光流估计模型 FlowNet, 首次证明了利用通用 CNN (Convolutional neural network) 架构直接估计光流的可行性. 然而, 尽管 FlowNet 模型在运算时间上可以达到实时估计, 但光流精度仍低于传统变分光流估计方法. 后续, Ilg[24] 将 FlowNetS 和 FlowNetC 网络进行堆叠, 通过增加网络深度显著提升了深度学习光流估计的精度与鲁棒性. 但是网络堆叠导致模型结构过于复杂、参数过多, 大幅增加了模型训练的难度. 此外, 为了更好地处理大位移运动, Ranjan 等[12] 将经典的空间金字塔方法与深度学习结合, 通过将大位移运动交由金字塔处理, 在一定程度上提升了大位移运动光流估计的准确性, 但该方法获取的图像特征并不鲁棒. 为此, Sun 等[25] 提出一种紧凑的 PWC-Net 光流估计模型, 该模型通过构建可学习的特征金字塔并将变形技术引入深度学习模型估计大位移, 在实现高精度光流估计的同时在模型尺寸与性能之间取得了最佳的平衡. 虽然上述方法显示出良好的估计精度, 但由粗到细的金字塔估计过程往往因像素点的丢失而产生一定的信息损失. 针对该问题, Wang 等[26] 提出一种混合特征提取模块弥补特征提取初期丢失的空间信息. Teed 等[27] 提出 RAFT 模型, 通过采用循环递归方案以恢复采样过程中丢失的小目标运动信息. 受到 Transformer[28] 视觉检测任务中成功应用的影响, 文献 [29] 基于一种广义的注意力变体构建全局运动聚合模块并将其附加于 RAFT 框架, 显著改善了遮挡与闭塞区域的光流估计精度. GMFlow 方法更是直接基于 Transformer 构建了一种定制化的数据增强方案并用于全局匹配相关计算, 进一步提升了大位移光流估计的准确性[30].

然而, 现有深度学习光流估计方法大多关注于如何通过后置处理来弥补前期的信息损失, 而忽视了标准卷积可学特征参数的内在限制以及目标长距离上下文信息关联对模型光流估计性能的影响. 针对以上问题, 本文提出一种联合深度超参数卷积和交叉关联注意力的大位移光流估计方法, 实验结果表明本文方法相对现有代表性深度学习方法可以获得更好的光流估计效果.

2 联合深度超参数卷积和交叉关联注意力的光流估计模型

2.1 网络模型整体架构

虽然传统特征金字塔提取模型可以让不同尺寸目标在相应尺度下拥有合适的特征表示, 但降比例采样引起的目标语义代沟问题, 使得模型难以准确捕获置信度较高的目标特征. 针对该问题, 本文提出基于交叉关联注意力的特征编码结构, 通过增强长距离像素之间关联度, 以捕捉目标的长距离上下文信息, 进而提高模型特征提取的置信度. 此外, 为了加快训练并收敛到最佳的参数组合, 本文将交叉关联注意力特征提取编码网络与深度超参数卷积耦合, 构造联合深度超参数卷积和交叉关联注意力的光流估计网络, 以实现高精度的光流估计.

图 1 展示了本文所提方法模型示意图, 从图中可以看出, 网络主要由特征提取编码网络和光流解码网络组成, 其中特征提取编码网络由交叉关联注意力模块和卷积层组成, 解码器采用 IRR-PWC 中的联合遮挡检测残差迭代优化光流估计策略. 在光流估计过程中, 首先将连续两帧图像输入到特征编码网络 (图中仅用其中一帧图像作为示例), 用卷积核为 3×3 的深度超参数卷积进行下采样至原图像 1/4 分辨率, 以减小特征编码网络的估计量. 然后将 1/4 分辨率的特征图输入到交叉关联注意力模块. 整个交叉关联注意力模块主要分为四个阶段, 每个阶段操作相似, 两个阶段之间卷积层的作用是使通道数量加倍, 以扩大感受野并增加特征维度. 为了避免对全局信息估计量过大, 本文使用多头注意力将自注意力的 4 个头部均分成两组, 一组在水平上做横向自注意力, 另外一组在垂直上做纵向自注意力, 两组自注意力并行对图像进行处理, 最后将输出进行拼接. 随着每个阶段的加深, 多头注意力可以关联更多的区域, 这使得网络在降低估计复

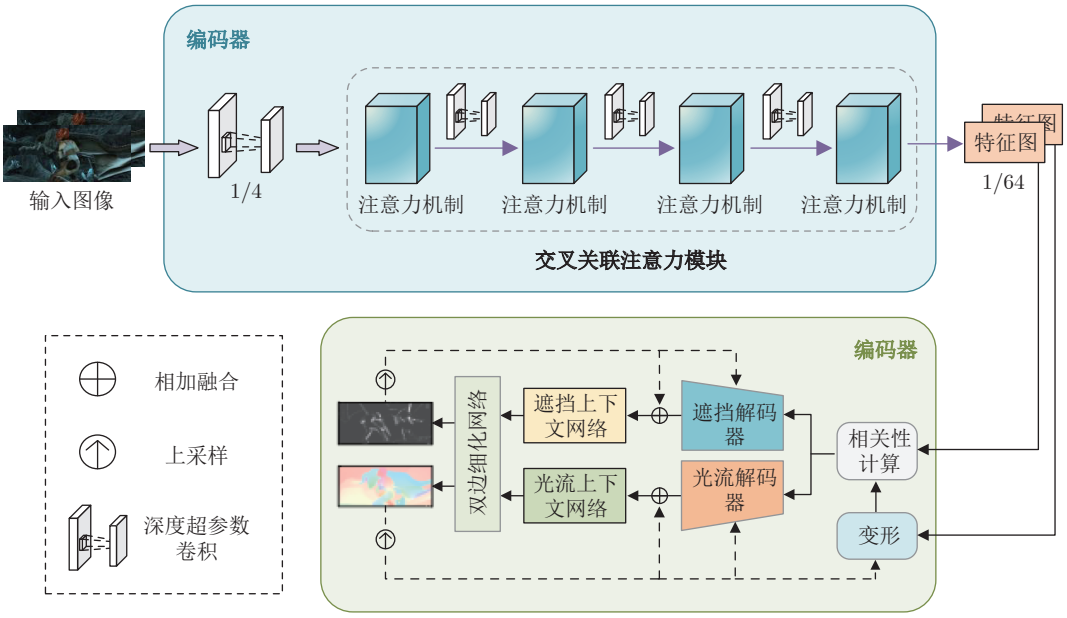


图 1 基于深度超参数卷积和交叉关联注意力的大位移光流估计网络示意图

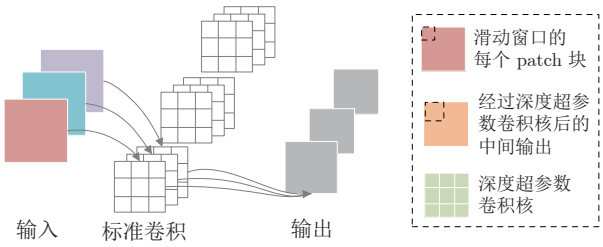
Fig. 1 Structure diagram of large displacement optical flow estimation based on depthwise over-parameterized convolution and cross correlation attention

杂度的同时增大图像感受野以便对图像全局信息进行提取. 最后, 通过使用双向估计、双边细化和遮挡上采样层联合迭代估计光流和遮挡. 本文方法通过构建联合深度超参数卷积和交叉关联注意力的光流估计模型, 显著提升了大位移运动区域光流估计的准确性和鲁棒性.

2.2 基于深度超参数卷积的光流估计方法

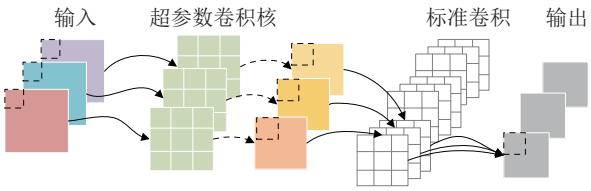
现有基于深度学习的光流估计网络模型普遍采用标准卷积对所有样本都采用相同的卷积核参数, 但由于训练参数“卷积核”是提取到的图像特征且光流属于稠密估计任务, 因此, 为了提升模型的拟合能力, 需要大量参数来进行学习. 为此, 一般做法是通过增加线性层——非线性层的数量来加深网络的深度, 从而提高模型的特征抽取能力. 但只增加线性层会造成过拟合问题.

针对该问题, 本文从卷积核可学习参数出发, 为了学习到更多参数同时加快训练速度, 避免过拟合, 在标准卷积中可附加一个深度卷积, 构成深度超参数卷积[31]. 图 2 展示了标准卷积和深度超参数卷积工作过程对比示意图, 图 2(a) 为标准卷积过程, 图 2(b) 为深度超参数卷积过程. 从图中可以看出, 标准卷积的卷积核是对 3 个通道同时做卷积. 深度超参数卷积是深度卷积和标准卷积的组合, 其估计过程主要分为两部分: 首先把深度卷积核的参数和标准卷积核的参数相乘, 然后再与原图的输



(a) 标准卷积

(a) Conventional convolution



(b) 深度超参数卷积

(b) Depthwise over-parameterized convolution

图 2 深度超参数卷积和标准卷积示意图

Fig. 2 The structure diagram of conventional convolution and depthwise over-parameterized convolution

入进行一次标准卷积. 给定一个输入特征图, 卷积层以滑动窗口的方式对其进行处理, 在每个窗口位置上, 将一组卷积内核应用于相应大小的块 $\boldsymbol{P} \in \mathbf{R}^{(M\times N)\times C_{\text{in}}}$ 上, $\otimes$ 为深度超参数卷积层输出, 其卷积算子可公式化如下:

$$\boldsymbol{O} = (\boldsymbol{D}, \boldsymbol{W}) \otimes \boldsymbol{P} = (\boldsymbol{D}^{\mathrm{T}} \circ \boldsymbol{W}) * \boldsymbol{P} \tag{1}$$

式 (1) 中, $\boldsymbol{D}^{\mathrm{T}} \in \mathbf{R}^{D_{\text{mul}}\times(M\times N)\times C_{\text{in}}}$ 是 $\boldsymbol{D} \in \mathbf{R}^{(M\times N)\times D_{\text{mul}}\times C_{\text{in}}}$

在第一维和第二维的转置操作, 一个深度卷积算子首先使用可训练的核 $\boldsymbol{D}^{\mathrm{T}}$ 来变换 $\boldsymbol{W}$, 然后在 $\boldsymbol{W}'$ 和 $\boldsymbol{P}$ 之间应用一个常规的卷积算子 $*$ 生成最后的结果.

图 3 展示了深度超参数卷积的结构, 深度超参数卷积由深度卷积的卷积核 $\boldsymbol{D} \in \mathbf{R}^{(M\times N)\times D_{\mathrm{mul}}\times C_{\mathrm{in}}}$ 和标准卷积的卷积核 $\boldsymbol{W} \in \mathbf{R}^{C_{\mathrm{out}}\times D_{\mathrm{mul}}\times C_{\mathrm{in}}}$ 组成, 其中, M 和 N 是卷积核在两个空间方向上的大小, 例如使用 3×3 卷积核, 则 $M=3$、$N=3$. C_{in} 和 C_{out} 分别代表输入的维度数和输出的维度数. 对于某个特定通道的输入特征, 有 D_{mul} 个卷积核作用于窗口 $M\times N$, 输出 D_{mul} 个特征, 本文 $D_{\mathrm{mul}}=M\times N$. 图中深度卷积操作首先使用可训练的卷积核 $\boldsymbol{D}^{\mathrm{T}}$ 来变换 $\boldsymbol{W}$ (图中红色部分), 生成 $\boldsymbol{W}'$. 然后在 $\boldsymbol{W}'$ 和 $\boldsymbol{P}$ 之间应用一个标准的卷积操作得到最终结果 $\boldsymbol{O}$.

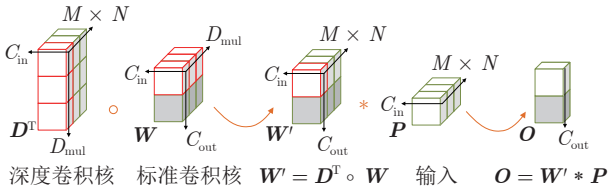


图 3　深度超参数卷积操作

Fig. 3　The operation of depthwise over-parameterized convolution

在训练过程中, 在标准卷积上增加一次深度卷积, 可以显著增加可学习训练参数量, 相较于标准卷积可以学习到更多特征, 为后续稠密的光流估计任务提供可靠的先验特征. 图 4 展示了不同类型卷积在 1/4 分辨率下的特征提取结果可视化对比, 从图中可以看出, 融合深度超参数卷积的光流估计网络与使用标准卷积的光流网络相比, 捕捉到了更为丰富且置信度更高的图像特征. 例如更加清晰的轮廓和纹理, 因此, 本文所提出的融合深度超参数卷积光流估计网络能够有效提升网络性能.

2.3　网络模型整体架构

现有光流估计模型通常采用由粗到细的金字塔策略来应对大位移问题, 该策略在每个金字塔级别上通过对图像进行相同比例降采样来减少源图像与目标图像之间的距离, 从而优化大位移运动区域光流估计结果. 但由于降采样导致图像分辨率减小, 引起目标特征稀释, 进而丢失重要的位置信息, 导致已有方法针对大位移区域的光流估计性能仍存在较大不足. 因此, 为了提高大位移区域光流估计的鲁棒性和准确性, 本文从图像特征金字塔编码部分入手, 设计基于交叉关联注意力的特征提取模块. 该模块主要目的是提高网络长距离上下文建模能力, 以获取图像全局上下文信息, 从而捕获更为准确的大位移运动目标特征.

(a) 参考帧 (a) Reference frame　(b) Baseline 特征图 (b) The feature map of Baseline　(c) 深度超参数卷积特征图 (c) The feature map of depthwise over-parameterized convolution

图 4　不同光流模型特征图对比

Fig. 4　Comparison of feature maps of different optical flow models

尽管可学习的特征金字塔已被多种基于深度学习的视觉任务证明可以有效处理大位移. 但由金字塔引起的特征稀释问题, 常会丢失部分运动信息, 且该信息在后续光流迭代过程中难以被恢复. 例如, 基于金字塔迭代的经典光流方法 IRR-PWC 所采用的特征编码部分是由多层卷积滤波器对 L 层的特征进行连续下采样操作. 但由于 CNN 只有固定有限的感受野, 对全局上下文信息捕获能力较弱, 导致基于金字塔的光流估计方法虽能有效改善大位移问题, 但容易丢失局部位置信息, 特别是自身尺寸相对较小的小目标运动特征.

针对该问题, 本文提出一种基于交叉关联注意力的光流特征编码网络模型. 该模型在高分辨率下采用并行的水平和垂直窗口获取局部特征的位置信息, 这有利于纠正在金字塔前期因目标特征稀释而导致的初始光流估计错误问题. 之后, 通过特征金字塔每个阶段对特征图的降采样, 注意区域逐步扩大从而实现全局特征信息的提取. 因此, 基于交叉关联注意力的全局特征提取编码网络在保证全局上下文信息获取的同时可以有效改善目标局部特征丢失问题, 图 5 展示了交叉关联注意力机制的基本结构.

大位移运动的产生是由于像素点位移变化相对剧烈, 而导致连续帧图像间部分区域像素点运动的不连续. 基于这一特性, 以较少的估计成本和内存成本建模远程的像素依赖关系, 让模型聚集图像所有像素的上下文信息, 有望提高模型大位移光流估计性能. 受到 CSWin transformer[32] 和 CCNet[33] 局部注意力窗口的启发, 本文设计了一种层级的注意力光流特征提取编码器. 该模型主要由两部分不同尺度下的卷积层和注意力机制模块组成, 输入图像

首先经过两层卷积进行特征提取后, 将特征图输入到后续的注意力模块得到包含丰富上下文信息的特征图. 如图 6 所示, 基于交叉关联注意力的光流特征提取模块主要有两个串行部分, 第一部分包含两层卷积核为 3 × 3 的卷积层, 用来提取 1/2 分辨率和 1/4 分辨率下图像的特征, 再将输出特征图经过一个标准卷积层后进入分层注意力模块形成最终的特征编码网络. 其中每个注意力模块之间的卷积层是卷积核为 3 × 3 的标准卷积, 用于改变通道数量. 通过将注意力机制进行层级叠加, 模型最终可以得到聚合图像全局上下文信息的特征图, 生成大位移运动线索, 从而恢复被稀释的大位移运动局部特征与小目标运动特征.

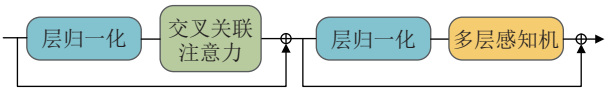


图 5 交叉关联注意力机制

Fig. 5 The cross correlation attention block

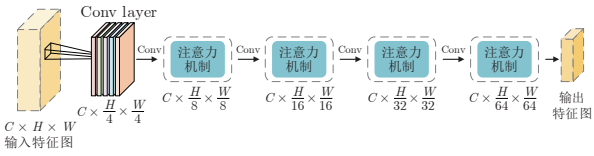


图 6 基于交叉关联注意力的光流特征编码网络示意图

Fig. 6 Structure diagram of optical flow feature encoder network based on cross correlation attention

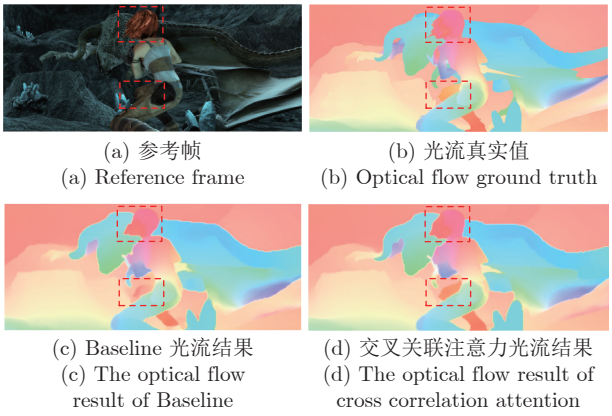
每一个注意力机制模块都包含相同操作, 如图 5 所示, 令 F^{l-1} 代表注意力机制模块在第 l 层的输入, 每一层注意力机制的输入都是上一层注意力机制的输出再经过一个标准卷积层, 最后输出 1/64 分辨率的特征图, 一个注意力机制模块的输出估计公式可以表示为:

$$\begin{cases} \widetilde{F}^{l} = CSwin\text{-}Attention(LN(F^{l-1}) + F^{l-1}) \\ F^{l} = MLP(LN(\widetilde{F}^{l})) + \widetilde{F}^{l} \end{cases} \tag{2}$$

式 (2) 中, F^{l-1} 代表了第 1-1 层卷积层的输出, 运算符号 LN 代表了层归一化, MLP 代表多层感知机, $CSwin\text{-}Attention$ 是交叉关联注意力模块, F^{l} 是第 l 层注意力机制最后的输出, 基于交叉关联注意力的光流特征提取编码网络最终的输出可以表示为:

$$F_{\text{out}} = CS(conv(conv_2(conv_1(F_{\text{in}})))) \tag{3}$$

式 (3) 中, F_{in} 代表输入的图片, $conv_1$、$conv_2$ 均代表三个卷积核为 3 × 3 的卷积层, $CS(conv(\cdot))$ 是四个注意力机制模块的操作. 如图 7 所示, 本文提出的基于交叉关联注意力的光流特征提取网络相较于



(a) 参考帧
(a) Reference frame

(b) 光流真实值
(b) Optical flow ground truth

(c) Baseline 光流结果
(c) The optical flow result of Baseline

(d) 交叉关联注意力光流结果
(d) The optical flow result of cross correlation attention

图 7 不同光流模型估计结果对比

Fig. 7 Comparison of results of different optical flow models

传统的基于 CNN 的特征编码网络, 可以从输入信息中聚合全局长距离上下文信息, 受益于注意力机制对远程依赖关系的捕获能力和对全局信息的关注, 网络可以在保持原有遮挡检测精度的前提下, 提高对大位移区域特征提取的能力, 从而改善在大位移区域光流估计的性能.

3 模型损失函数与训练策略

3.1 模型损失函数

本文损失函数的采用了 IRR-PWC[21] 的损失函数, 最终的损失由光流损失和遮挡损失的加权和组成, 本文使用 $L_{2,1}$ 范数来监督前向流和后向流, 其中光流损失函数 l^{i}_{flow} 如下:

$$l^{i}_{\text{flow}} = \frac{1}{2}\sum(||f^{i}_{fw} - f_{fw,GT}||_2 + ||f^{i}_{bw} - f_{bw,GT}||_2) \tag{4}$$

式 (4) 中, f^{i}_{fw}、f^{i}_{bw} 分别表示预测的正向光流和反向光流图, $f_{fw,GT}$、$f_{bw,GT}$ 则代表正向光流和反向光流图的真实值, N 为迭代的总步数, $i = 1, \cdots, N$. 本文使用一个加权的二元交叉熵来作为遮挡图的监督:

$$\begin{aligned} l^{i}_{\text{occ}} = \frac{1}{2}\sum(&w^{i}_{1}o^{i}_{1}\log(o_{1,GT}) + \\ &\overline{w}^{i}_{1}(1-o^{i}_{1})\log(1-o_{1,GT}) + \\ &w^{i}_{2}o^{i}_{2}\log(o_{2,GT}) + \overline{w}^{i}_{2}(1-o^{i}_{2})\log(1-o_{2,GT})) \end{aligned} \tag{5}$$

式 (5) 中, o^{i}_{1} 和 o^{i}_{2} 分别代表在第 i 步迭代时在第一张图片和第二张图片的遮挡图, $o_{1,GT}$ 和 $o_{2,GT}$ 则表示第一张和第二张遮挡图的真实值. 权重 w^{i}_{1} 和 $\overline{w}^{i}_{1}$ 为 $w^{i}_{1} = H \cdot W/(\sum o^{i}_{1} + \sum o_{1,GT})$ $\overline{w}^{i}_{1} = H \cdot W/(\sum(1-o^{i}_{1}) + \sum(1-o_{1,GT}))$. 因此最终的损失函数为:

$$I_{\text{full}} = \frac{1}{N}\sum_{i=1}^{N}\alpha_i(l^i_{\text{flow}} + \lambda \cdot l^i_{\text{occ}}) \tag{6}$$

式 (6) 中, λ 根据遮挡损失对光流进行加权. 在每次迭代中, 本文估计使光流损失和遮挡相等的 λ.

3.2　模型训练策略

为了提高模型的通用性, 本文首先将模型在包括双向光流地面真实值和每个图像遮挡图的 FlyingChairsOcc[14] 数据集上进行训练, 使模型学习到简单的平移、缩放等二维运动特征; 然后在具有三维运动和更大位移的数据集 FlyingThings3D_subset[24] 上微调, 使得模型可以学习到更加丰富的三维运动信息. 在前两个阶段的预训练后, 模型已经可以学习到充足的二维和三维运动特征. 最后, 在此基础上, 进一步使用 MPI-Sintel[34] 和 KITTI[35] 数据集进行微调, 从而增强模型的泛化性.

4　实验结果与分析

4.1　光流评价指标

实验采用光流估计研究领域权威的 MPI-Sintel 与 KITTI2015 测试图像数据集进行算法性能测试. 分别采用平均端点误差 (Average end-point error, EPE) 和异常值百分比 (Percentage of flow error, *Fl-all*) 两种评价指标对本文方法和对比方法的光流估计结果进行量化评估. 其中 MPI-Sintel 数据集采用端点误差指标来衡量估计光流与真实光流之间的几何距离误差. 其计算公式如下:

$$EPE = \frac{1}{N}\sqrt{(\boldsymbol{u}_{\text{gt}} - \boldsymbol{u})^2 + (\boldsymbol{v}_{\text{gt}} - \boldsymbol{v})^2} \tag{7}$$

式 (7) 中, N 表示整幅图像像素点个数, $\boldsymbol{u}_{\text{gt}}$ 是水平方向上的真实光流值, $\boldsymbol{u}$ 是水平方向上的估计光流值, $\boldsymbol{v}_{\text{gt}}$ 是垂直方向上的真实光流值, $\boldsymbol{v}$ 是垂直方向上的估计光流值.

KITTI2015 常用异常值百分比来对光流估计结果进行评价, 异常值百分比表示异常值像素占整幅图像像素点的比重. 其计算公式如下:

$$Fl\text{-}all = \frac{\sum \boldsymbol{P}(EPE > \tau)}{N} \times 100\% \tag{8}$$

式 (8) 中, N 表示整幅图像像素点个数, *Fl-all* 表示光流异常值百分比, $\boldsymbol{P}(EPE > \tau)$ 表示光流端点误差大于 τ 的像素点, 其中 $\tau = 3$.

4.2　MPI-Sintel 数据集实验对比

MPI-Sintel 数据集是一个用于算法光流估计性能评估的合成数据集, 由 Clean 和 Final 两个子数据集组成, 包含了非刚性运动、大位移运动和遮挡等困难场景. 其中, Final 数据集相较于 Clean 数据集包含更多运动模糊、大气效果、图像噪声和镜面反射等特效, 光流估计难度较大.

首先, 为了更为客观地分析本文方法光流估计性能, 实验选取 8 种具有代表性的深度学习光流估计方法与本文方法进行综合对比, 并使用 MPI-Sintel 测试数据集对各方法进行性能测试. 此外, 各对比方法遵循文献 [25] 的训练策略, 先在 FlyingChairs 或 FlyingChairsOcc 数据集进行预训练, 之后在 FlyingThings3D 或 FlyingThings3D_subset 数据集中进行微调, 最后在 MPI-Sintel 测试集上再次微调训练. 定量实验结果如表 1 所示, 其中, EPE 代表所有像素点的平均端点误差, Matched 代表在相邻帧中仍然可见区域上的端点误差, Unmatched 代表仅在两个相邻帧之一中可见区域的端点误差. 从表 1 中可以看出, 本文方法除在 Matched 指标上取得了次优结果外, 在其他指标均获得了最优的光流估计精度, 说明本文方法相对其他对比方法具有更加准确的整体光流估计精度. 这主要得益于本文所提出的全局特征提取编码网络, 能够有效捕获多尺度长距离上下文特征信息, 在提高特征匹配的准确率的同时显著提升了模型整体光流估计精度. 值得注意的是, HMFlow 方法在 Clean 数据集取得了次优的光流估计结果, 而 IOFPL-ft 方法在 Final 数据集仅低于本文方法. 原因在于 HMFlow 方法主要通过提高由粗到细过程中粗级信息的匹配精度来改善光流估计, 但是当面对具有运动模糊的 Final 数据集时由于难以获取高精度的图像粗级特征因此光流估计性能呈现下降趋势. IOFPL-ft 方法主要通过改进成本量构建过程避免了伪影对光流估计的影响, 保留了更多细节光流信息.

为了更加详细地展示本文方法与对比方法在不同区域的光流估计性能差异, 表 2 统计了各方法在运动边缘和大位移运动区域的光流估计量化结果. 其中指标 $d_{0\text{-}10}$、$d_{10\text{-}60}$、$d_{60\text{-}140}$ 分别表示距离遮挡边界 0-10、10-60、60-140 个像素区域的端点误差, $s_{0\text{-}10}$、$s_{10\text{-}40}$、s_{40+} 表示不同位移速度区域的端点误差. 从表 2 可以看出, 本文方法在包含较多大位移运动且临近运动边界的区域光流估计结果表现较好, 如在 Clean 数据集的 $d_{0\text{-}10}$、$d_{10\text{-}60}$、$d_{60\text{-}140}$ 指标上本文方法取得了最佳的估计精度. 在大位移运动指标 s_{40+} 方面, 本文方法相比于次优算法也分别取得了 19%、6.8% 的精度提升, 原因在于本文方法使用的局部到全局的感受野策略有效增强了网络的

表 1 MPI-Sintel 数据集图像序列光流估计结果 (pixels)

Table 1 Optical flow calculation results of image sequences in MPI-Sintel dataset (pixels)

对比方法	Clean			Final		
	All	Matched	Unmatched	All	Matched	Unmatched
IRR-PWC[14]	3.844	1.472	23.220	4.579	2.154	24.355
PPAC-HD3[36]	4.589	1.507	29.751	4.599	2.116	24.852
LiteFlowNet2[37]	3.483	1.383	20.637	4.686	2.248	24.571
IOFPL-ft[38]	4.394	1.611	27.128	4.224	**1.956**	22.704
PWC-Net[25]	4.386	1.719	26.166	5.042	2.445	26.221
HMFlow[39]	3.206	1.122	20.210	5.038	2.404	26.535
SegFlow153[40]	4.151	1.246	27.855	6.191	2.940	32.682
SAMFL[41]	4.477	1.763	26.643	4.765	2.282	25.008
本文方法	**2.763**	**1.062**	**16.656**	**4.202**	2.056	**21.696**

表 2 数据集运动边缘与大位移指标对比结果 (pixels)

Table 2 Comparison results of motion edge and large displacement index in MPI-Sintel dataset (pixels)

对比方法	Clean						Final					
	$d_{0\text{-}10}$	$d_{10\text{-}60}$	$d_{60\text{-}140}$	$s_{0\text{-}10}$	$s_{10\text{-}40}$	s_{40+}	$d_{0\text{-}10}$	$d_{10\text{-}60}$	$d_{60\text{-}140}$	$s_{0\text{-}10}$	$s_{10\text{-}40}$	s_{40+}
IRR-PWC[14]	3.509	1.296	0.721	0.535	1.724	25.430	4.165	1.843	**1.292**	0.709	2.423	28.998
PPAC-HD3[36]	2.788	1.340	1.068	**0.355**	**1.289**	33.624	3.521	1.702	1.637	**0.617**	2.083	30.457
LiteFlowNet2[37]	3.293	1.263	0.629	0.597	1.772	21.976	4.048	1.899	1.473	0.811	2.433	29.375
IOFPL-ft[38]	3.059	1.421	0.943	0.391	1.292	31.812	**3.288**	**1.479**	1.419	0.646	**1.897**	27.596
PWC-Net[25]	4.282	1.657	0.674	0.606	2.070	28.793	4.636	2.087	1.475	0.799	2.986	31.070
HMFlow[39]	2.786	0.957	0.584	0.467	1.693	20.470	4.582	2.213	1.465	0.926	3.170	29.974
SegFlow153[40]	3.072	1.143	0.656	0.486	2.000	27.563	4.969	2.492	2.119	1.201	3.865	36.570
SAMFL[41]	3.946	1.623	0.811	0.618	1.860	29.995	4.208	1.846	1.449	0.893	2.587	29.232
本文方法	**2.772**	**0.854**	**0.443**	0.541	1.621	**16.575**	3.884	1.660	**1.292**	0.753	2.381	**25.715**

长距离上下文建模能力. 在挑战性较大的 Final 数据集, 虽然本文方法在 $d_{0\text{-}10}$、$d_{10\text{-}60}$ 指标并未取得最佳的光流估计结果, 但整体精度估计也仅略低于 IOFPL-ft. 原因在于图像边缘被运动模糊、大气变化和噪声所破坏, 导致对图像边缘特征信息的提取存在一定误差[29]. 如图 8 所示, 分别列举了两个有代表性 bamboo_3 序列和 PERTURBED_shaman_1 序列在 Clean (不包含噪声、运动模糊等干扰) 和 Final 数据集 (包含运动模糊、大气变化和噪声等干扰) 中特征图可视化结果. 其中图 8(a) 和 8(c) 是 Clean 数据集中的两个图像序列, 图 8(b) 和 8(d) 是 Final 数据集中的两个图像序列, 由红框区域及其放大图部分可以明显看出由于运动模糊等原因对原图像的影响, 导致在 Final 数据集中对于目标边缘特征信息的提取存在一定误差. 其次, 在 $s_{0\text{-}10}$、$s_{10\text{-}40}$ 指标本文方法低于对比方法的主要原因是, 由于位移较小的区域目标的运动信息不足, 致使网络对于该区域在连续下采样过程中易丢失重要的结构信息, 从而影响网络对于位移较小区域的光流估计结果[27]. 图 9 为不同金字塔层数下不同大小的输入目标的特征可视化及光流可视化结果 (本文只列举了金字塔第一层到第三层的特征可视化结果, 因为金字塔高层的特征极其抽象不易辨认). 其中 A 标签区域为位移较大的目标区域, B 标签区域为位移较小的目标区域. 由图 9 可知在运动信息较为丰富的 A 标签区域, 虽经过了连续的下采样, 但在不同层的特征图中仍清晰可见, 但在运动信息不足的 B 标签区域, 经过了相同的降采样比例后, 特征信息发生了明显的丢失, 从而导致对于包含较小位移区域的光流估计结果出现偏差.

为了定性分析本文方法光流估计效果, 图 10 以 MPI-Sintel 测试数据集中包含多个运动目标、运动场景相对较为复杂的 Temple 1 图像序列为例, 分别展示了本文方法和 IOFPL-ft、IRR-PWC、SegFlow153 三种代表性对比方法的光流估计可视化结果对比. 图中第一行为整体光流估计效果, 为了便于观察对比, 本文以 A, B, C 为标签标记了三个不同尺寸的运动目标, 第二列为同对比方法对应

(a) Clean 数据集中的 bamboo_3 序列 (a) The bamboo_3 sequences of Clean dataset

(b) Final 数据集中的 bamboo_3 序列 (b) The bamboo_3 sequences of Final dataset

(c) Clean 数据集中的 PERTURBED_shaman_1 序列 (c) The PERTURBED_shaman_1 sequences of Clean dataset

(d) Final 数据集中的 PERTURBED_shaman_1 序列 (d) The PERTURBED_shaman_1 sequences of Final dataset

(e) Clean 数据集中的 bamboo_3 序列特征图 (e) The feature map of bamboo_3 sequences of Clean dataset

(f) Final 数据集中的 bamboo_3 序列特征图 (f) The feature map of bamboo_3 sequences of Final dataset

(g) Clean 数据集中的 PERTURBED_shaman_1 序列特征图 (g) The feature map of PERTURBED_shaman_1 sequences of Clean dataset

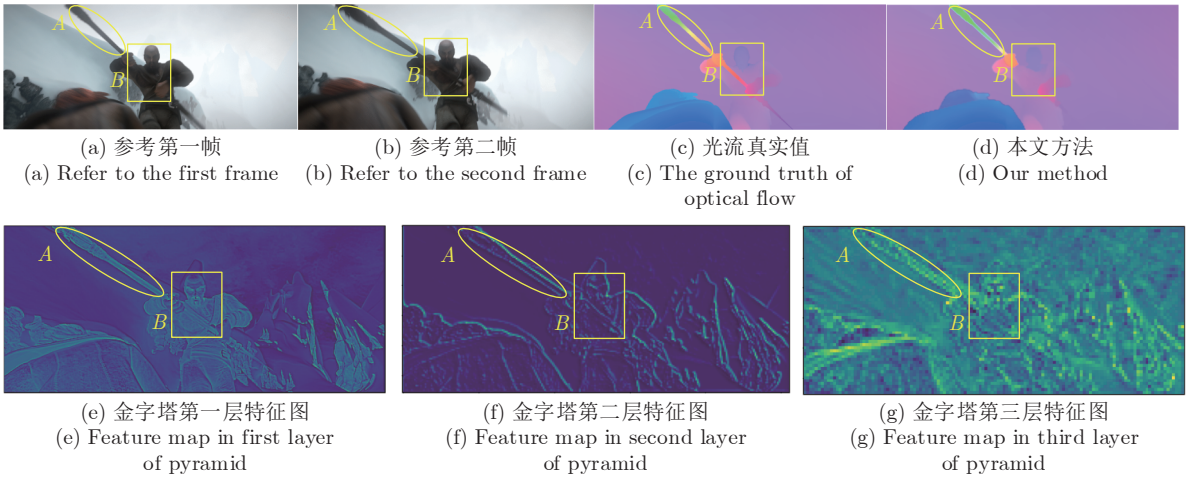
(h) Final 数据集中的 PERTURBED_shaman_1 序列特征图 (h) The feature map of PERTURBED_shaman_1 sequences of Final dataset

(i)　(ii)　(iii)　(iv)

(i) 标签区域局部放大对比图 (i) Partial enlargement and comparison of the label area

图 8　Clean 和 Final 数据集不同序列特征图可视化 (其中红框区域内为存在明显区别的边缘特征信息结果)

Fig. 8　Visualization of feature maps of different sequence in Clean and Final datasets (The red bounding box contains edge feature information results with significant differences)



(a) 参考第一帧 (a) Refer to the first frame

(b) 参考第二帧 (b) Refer to the second frame

(c) 光流真实值 (c) The ground truth of optical flow

(d) 本文方法 (d) Our method

(e) 金字塔第一层特征图 (e) Feature map in first layer of pyramid

(f) 金字塔第二层特征图 (f) Feature map in second layer of pyramid

(g) 金字塔第三层特征图 (g) Feature map in third layer of pyramid

图 9　金字塔不同层数下不同尺度目标特征可视化

Fig. 9　Visualization of Feature maps at different scales under different layers of pyramid

标签区域的光流估计结果可视化对比. 从图中可以看出, 在相对目标尺寸较大的 A 区域, 所有方法均可以较为完整地估计出目标区域光流信息, 在细节上本文方法受 Final 数据集模糊效果的影响, 在运动边缘光流估计效果略低于 IOFPL-ft 方法. 在相对最小的尺寸 B 区域, 本文方法获得了最佳的光流估计效果, IRR-PWC 和 SegFlow 存在明显的目标光流信息丢失, 而 IOFPL-ft 存在间断现象. 最后, 在中等尺寸 C 区域, 本文方法仍然取得了最佳的光流估计效果, 相对其他方法唯一捕捉到了飞行龙的脚部光流运动信息. 因此, 本文方法能够在不同尺寸目标运动估计效果方面获得最佳的结果.

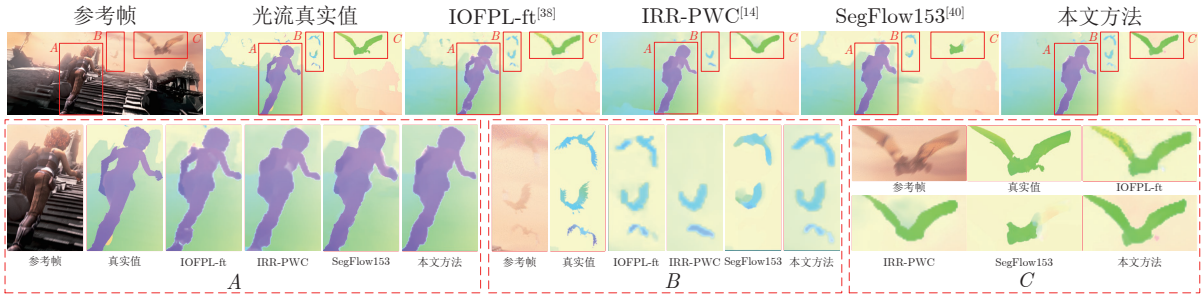


图 10 MPI-Sintel 测试集图像序列对比方法光流估计可视化结果

Fig. 10 Flow field results of the comparable methods evaluated on some MPI-Sintel test datasets

4.3 KITTI 数据集实验对比

KITTI 测试图像数据集由包含大量真实道路场景图像序列构成, 其主要用于测试算法针对真实场景任务时光流估计的准确性与鲁棒性. 为进一步验证本文方法在真实场景下的光流估计准确性与鲁棒性, 本文采用包含更多光照变化、大位移运动、遮挡以及运动模糊等场景的 KITTI2015 测试图像集进行测试对比分析, 实验采用与 MPI-Sintel 数据集实验相同的训练策略. 表 3 展示了本文方法和对比方法针对 KITTI2015 数据集图像序列光流估计异常值百分比结果对比, 表中 *Fl-bg* 代表图像中背景区域平均的异常值百分比, *Fl-fg* 代表图像中前景区域平均的异常值百分比, *Fl-all* 代表图像中平均光流异常值百分比. 由于 HMFlow 方法并未在 KITTI2015 上进行测评, 所以未展示其对比结果. 从表 3 可以看出, 虽然本文方法在 *Fl-all* 指标未取得最佳的结果, 但是在 *Fl-fg* 和 *Fl-bg* 指标分别获取了次优和最优结果. SegFlow153 方法通过使用基于分割的 PatchMatch 框架在最佳级别上进行稠密匹配以应对大位移光流估计, 但在包含真实道路场景的数据集上表现较差. 相较于本文方法, PPAC-HD3 方法光流估计精度在前景区域背景区域和整体区域效果较好, 产生该现象的原因是 PPAC-HD3 引入概率像素自适应卷积 (PPACs) 对密集预测网络进行细化, 在允许图像自适应平滑的同时可以将高置信度的像素传播到置信度较低区域, 从而有效抑制了异常值. 与 PPAC-HD3 相比, 本文方法在包含像素点位移变化不明显的弱纹理背景区域, 光流估计效果提升不显著, 这是由于背景区域所包含的运动信息不足, 致使本文方法中所使用的交叉关联注意力机制无法在像素点之间建立正确的匹配关系. 但本文方法在前景区域误差结果最小, 这说明本文方法对于前景中发生较明显位移变化的区域有显著的提升效果.

图 11 展示了本文方法与三种精度相近对比方法: PPAC-HD3、LiteFlowNet2、IRR-PWC 在 KITTI2015 测试集 000004 和 000019 图像序列的光流估计误差可视化对比, 图中颜色越接近红色表示光流误差越大, 颜色越接近蓝色表示光流误差越小. 为了更好地观察图中各对比方法在大位移运动区域光流估计效果, 本文对标签区域进行了局部放大. 从图中可以看出, 在光照不足且背景复杂的 000004 序列中, 本文方法取得了最佳的光流估计效果, 在小目标的车头周围, 光流估计异常值占比最小. 在 000019 序列, 针对大位移运动的汽车目标, 虽然其尺寸占比较小, 本文方法仍然取得了较为优越的光流估计效果, 异常值占比仅略低于 PPAC-HD3, 说明针对复杂场景的小目标大位移运动具有较高的光流估计效果.

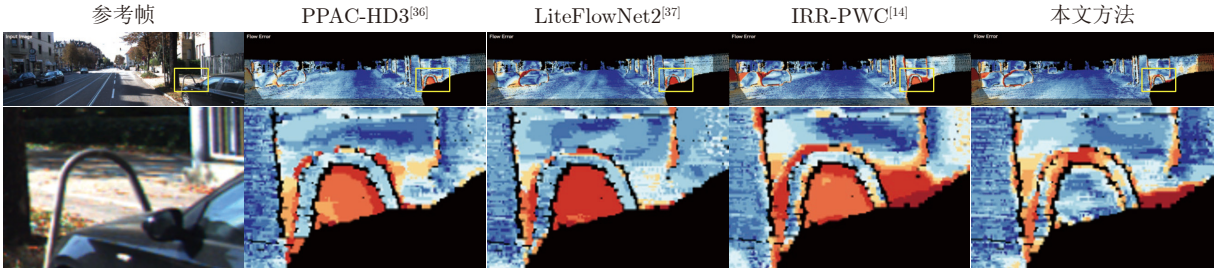
表 3 KITTI2015 数据集计算结果 (%)

Table 3 Calculation results in KITTI2015 dataset (%)

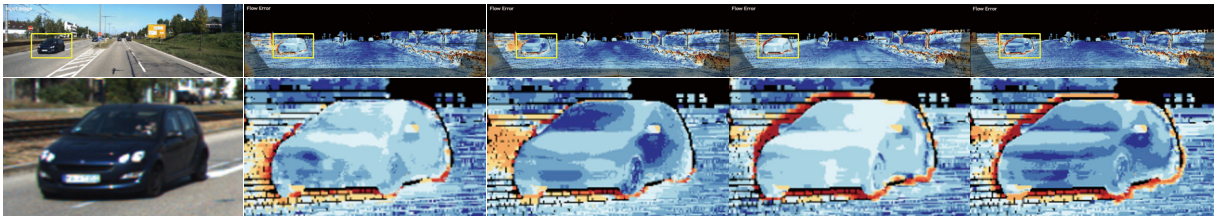
对比方法	*Fl-bg*	*Fl-fg*	*Fl-all*
IRR-PWC[14]	7.68	7.52	7.65
PPAC-HD3[36]	**5.78**	7.48	**6.06**
LiteFlowNet2[37]	7.62	7.64	7.62
IOFPL-ft[38]	—	—	6.52
PWC-Net[25]	9.66	9.31	9.60
SegFlow153[40]	22.21	23.72	22.46
SAMFL[41]	7.72	7.43	7.68
本文方法	7.43	**6.65**	7.30

4.4 消融实验

为了分析所提方法各模块对模型光流估计性能提升的作用, 本文进行了独立且完整的消融实验. 实验采用 MPI-Sintel 数据集中的 Clean 数据集和 KITTI2015 测试图像集对本文方法进行消融实验对比分析, 此外, 为了保证客观公正, 本文采取的训练步骤与策略均相同. 各消融实验模型数据如表 4 和表 5 所示, 其中, Baseline 为基准模型, Baseline_CS 为基准模型加基于交叉关联注意力的全局特征提取



(a) 000004 序列光流误差结果极其大位移运动区域局部放大图
(a) Flow error maps of 000004 sequence and close-up views within the yellow squares of large displacement region



(b) 000019 序列光流误差结果极其大位移运动区域局部放大图
(b) Flow error maps of 000019 sequence and close-up views within the yellow squares of large displacement region

图 11　KITTI2015 测试集图像序列对比方法光流估计误差可视化结果
Fig. 11　Flow error maps of the comparable methods tested on KITTI2015 datasets

表 4　Clean 数据集上消融实验结果对比 (pixels)
Table 4　Comparison of ablation experiment results in Clean dataset (pixels)

消融模型	All	Matched	Unmatched	$s_{10\text{-}40}$	s_{40+}
Baseline	3.844	1.472	23.220	1.724	25.430
Baseline_CS	2.892	1.070	17.765	1.662	17.460
Baseline_deconv	3.621	1.461	21.272	1.659	23.482
Full model	**2.763**	**1.062**	**16.656**	**1.621**	**16.575**

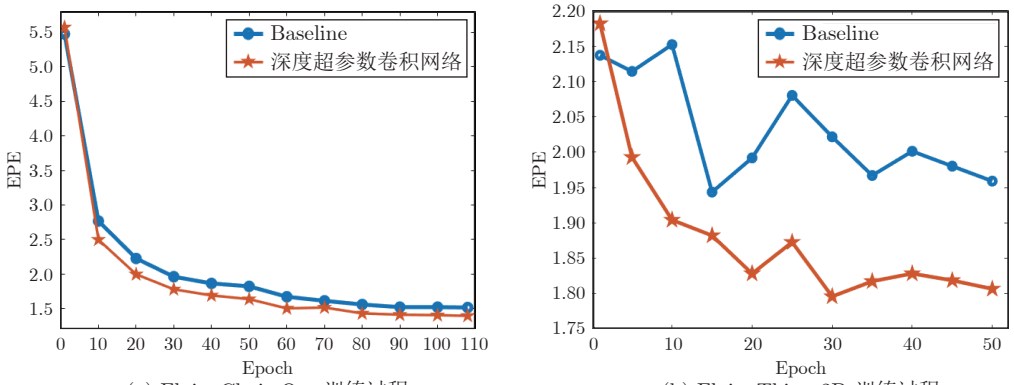
表 5　KITTI2015 数据集上消融实验结果对比
Table 5　Comparison of ablation experiment results in KITTI2015 dataset

消融模型	*Fl-bg* (%)	*Fl-fg* (%)	*Fl-all* (%)	训练时间 (min)
Baseline	7.68	7.52	7.65	621
Baseline_CS	7.74	7.58	7.71	690
Baseline_deconv	**7.28**	7.30	**7.29**	632
Full model	7.43	**6.65**	7.30	**616**

网络, Baseline_deconv 为基准模型加深度超参数卷积, Full model 是基准模型耦合交叉关联注意力的全局特征提取网络和深度超参数卷积后的全模型. 由表 4 可以看出, 相比于单独去除交叉关联注意力的全局特征提取网络的 Baseline_deconv 模型和单独去除深度超参数卷积的 Baseline_CS 模型, 全模型在 Clean 数据集的所有指标上取得了最好的光流估计效果. 并且从 $s_{10\text{-}40}$, s_{40+} 指标可以看出, 通过分别增加基于交叉关联注意力的全局特征提取网络和深度超参数卷积, 可以有效提升大位移运动光流估计的精度. 当基于交叉关联注意力的全局特征提取网络和深度超参数卷积共同作用时, 可以显著提高网络光流估计精度, 二者的协同作用进一步提升了 $s_{10\text{-}40}$, s_{40+} 指标精度. 说明所提方法各模块可以有效提高网络的光流估计精度, 特别是对大位移运动区域. 从表 5 可以看出, 在加入 CS 后, 在 KITTI 数据集上却未实现理想的光流估计精度, 这是因为 KITTI 数据集包含的可用训练数据较少, 致使交叉关联注意力机制无法较好关注邻近的标记, 在早期阶段聚合局部信息所包含的物体表征能力相对较弱, 这也是注意力机制中一个长期存在的问题[42]. 引入 deconv 后, 在 KITTI 数据集中的背景、前景和整体区域分别提升了 5%、3% 和 5%, 表明深度超参数卷积的引入能够增加网络的表达能力, 从而提高光流估计网络的性能. 当 CS 与 deconv 相结合时, 完整的模型在具有明显位移的前景区域提升了 11.6%, 然而在背景和全部区域却取得了次优的结果, 这是由于基于交叉关联注意力的全局特征编码网络主要关注于前景区域具有明显位移的目标物体. 虽然相较于 Baseline_deconv 模型, 全模型的精度有轻微下降, 但不同消融模型的运行时间对比表明, 本文方法 (即 Full model) 可以在光流估计精度和推理速度间取得较好的平衡.

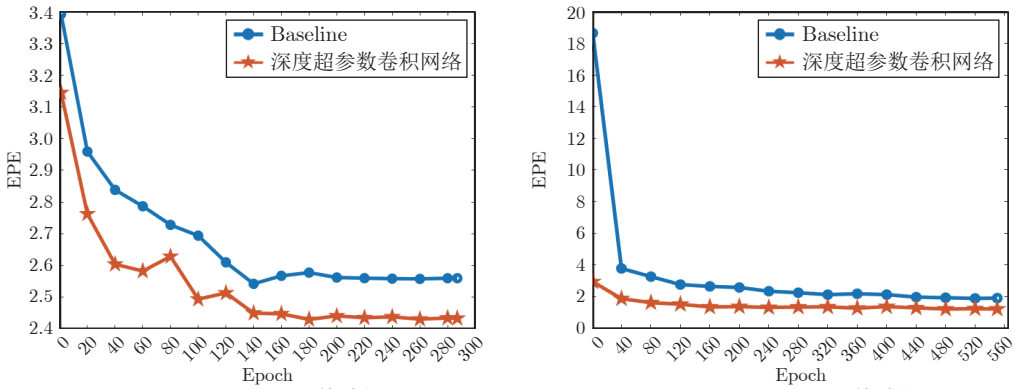
1) 深度超参数卷积的影响

由表 4 和表 5 可以看出, 在去除深度超参数卷积后, Baseline 模型在 MPI-Sintel 数据集上的性能较差, 特别是在 Unmatched 和 s_{40+} 指标上. 在 KITTI2015 测试图像集上, 相较于 Baseline 模型,



(a) FlyingChairsOcc 训练过程
(a) The training process of FlyingChairsOcc

(b) FlyingThings3D 训练过程
(b) The training process of FlyingThings3D



(c) MPI-Sintel 训练过程
(c) The training process of MPI-Sintel

(d) KITTI2015 训练过程
(d) The training process of KITTI2015

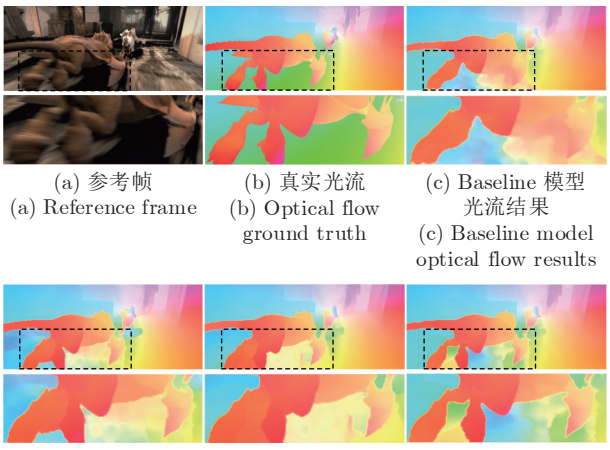
图 12 Baseline_deconv 在各数据集训练过程

Fig. 12 The training process of Baseline_deconv on each dataset

Baseline_deconv 模型在图像整体区域、前景背景区域的光流异常值百分比明显降低, 并在背景区域和整体区域精度略高于全模型. 这表明所提出的深度超参数卷积因其训练时权重参数的增加, 对图像特征提取的有效性提升, 从而更有利于光流整体性能的提高. 为了验证深度超参数卷积对训练过程的影响, 本文使用了四个数据集进行对比实验, 由图 12 可以看出, 使用深度超参数卷积后的网络不仅在相同迭代次数的情况下更快收敛, 并且能够达到更低的端点误差. 此外, 图 13 也定性地展示出加入深度超参数卷积后, 模型可以更加准确地估计出大位移运动光流结构信息.

2) 基于交叉关联注意力的特征提取模块的影响

由表 4 和图 13 可以看出, 去除交叉关联注意力的特征提取模块后, Baseline 方法在 MPI-Sintel 数据集的整体估计性能较低, 特别是在 s_{40+} 指标上, 而 Baseline_CS 相比于 Baseline 模型对于复杂大位移场景下光流估计效果较好, 例如能够较为准确地估计出黑色标签区域中龙的身体区域光流信息. 这表明所提出的交叉关联注意力可以更加有效



(a) 参考帧
(a) Reference frame

(b) 真实光流
(b) Optical flow ground truth

(c) Baseline 模型光流结果
(c) Baseline model optical flow results

(d) Baseline_deconv 模型光流结果
(d) Baseline_deconv model optical flow results

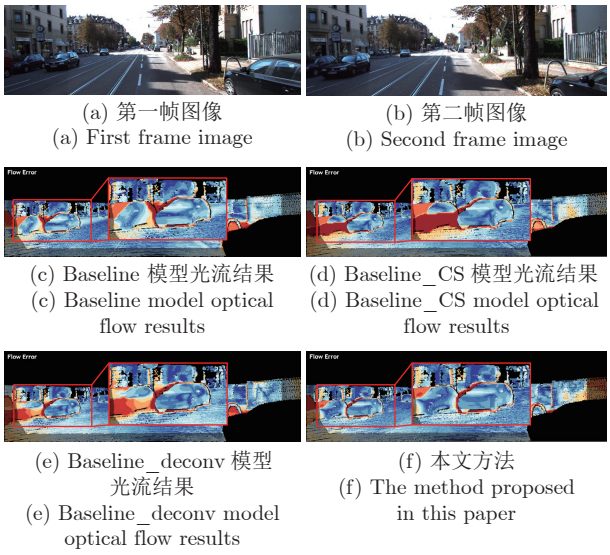
(e) Baseline_CS 模型光流结果
(e) Baseline_CS model optical flow results

(f) Full 模型光流结果
(f) Full model optical flow results

图 13 消融模型光流估计结果在 MPI-Sintel 测试数据集可视化对比

Fig. 13 Comparison of visualization results of each ablation model on MPI-Sintel test datasets

地扩大感受野从而获取全局信息. 然而, 由表 5 和图 14 可以看出, 在 KITTI 测试图像集上 Baseline_CS 的异常值百分比却略高于 Baseline 模型, 但全模型的异常值百分比却显著低于单独去除基于交叉关联注意力的特征提取模块的 Baseline_deconv 模型. 说明所提出的基于交叉关联注意力的特征提取模块对光流估计模型中对全局信息的提取起重要作用, 在运动物体大位移运动区域具有更好的准确性与鲁棒性.



(a) 第一帧图像
(a) First frame image

(b) 第二帧图像
(b) Second frame image

(c) Baseline 模型光流结果
(c) Baseline model optical flow results

(d) Baseline_CS 模型光流结果
(d) Baseline_CS model optical flow results

(e) Baseline_deconv 模型光流结果
(e) Baseline_deconv model optical flow results

(f) 本文方法
(f) The method proposed in this paper

图 14　消融模型光流估计结果在 KITTI2015 测试数据集可视化对比

Fig. 14　Comparison of visualization results of each ablation model on KITTI2015 datasets

5　结束语

本文提出了一种联合深度超参数卷积和交叉关联注意力的大位移光流估计模型, 针对深度学习光流估计模型特征提取信息不足问题, 通过构建基于深度超参数卷积的特征提取网络, 在有效提升特征置信度的同时加速模型训练收敛速度. 针对大位移运动区域光流准确性较低问题, 本文设计了一个基于交叉关联注意力的全局特征提取编码网络, 通过扩大感受野并增强长距离上下文建模能力, 提高了大位移区域光流估计的准确性. 实验分别采用 MPI-Sintel 和 KITTI 数据集对本文方法和现有代表性深度学习光流估计方法进行了综合实验对比. 实验结果表明, 本文方法对于大位移运动区域具有较高的精度和鲁棒性, 尤其在包含较多位移运动的前景区域具有更显著的优势. 但在运动信息不足的背景区域本文方法仍存在一定的局限, 未来将结合像素级对象分割方法来提取目标运动信息, 并通过后处理优化的手段对背景丢失的光流信息进行恢复, 从而提高对背景区域光流估计的有效性.

References

1 Zhang Jiao-Yang, Cong Shuang, Kuang Sen. Real-time state estimation and feedback control for n-qubit stochastic quantum systems. *Acta Automatica Sinica*, 2024, **50**(1): 42−53
(张骄阳, 丛爽, 匡森. n 比特随机量子系统实时状态估计及其反馈控制. 自动化学报, 2024, **50**(1): 42−53)

2 Zhang Wei, Huang Wei-Min. Multi-strategy adaptive multi-objective particle swarm optimization algorithm based on swarm partition. *Acta Automatica Sinica*, 2022, **48**(10): 2585−2599
(张伟, 黄卫民. 基于种群分区的多策略自适应多目标粒子群算法. 自动化学报, 2022, **48**(10): 2585−2599)

3 Zhang Fang, Zhao Dong-Xu, Xiao Zhi-Tao, Geng Lei, Wu Jun, Liu Yan-Bei. Research progress of single image super-resolution reconstruction technology. *Acta Automatica Sinica*, 2022, **48**(11): 2634−2654
(张芳, 赵东旭, 肖志涛, 耿磊, 吴骏, 刘彦北. 单幅图像超分辨率重建技术研究进展. 自动化学报, 2022, **48**(11): 2634−2654)

4 Yang Tian-Jin, Hou Zhen-Jie, Li Xing, Liang Jiu-Zhen, Huan Juan, Zheng Ji-Xiang. Recognizing action using multi-center subspace learning-based spatial-temporal information fusion. *Acta Automatica Sinica*, 2022, **48**(11): 2823−2835
(杨天金, 侯振杰, 李兴, 梁久祯, 宦娟, 郑纪翔. 多聚点子空间下的时空信息融合及其在行为识别中的应用. 自动化学报, 2022, **48**(11): 2823−2835)

5 Yan Meng-Kai, Qian Jian-Jun, Yang Jian. Weakly aligned cross-spectral face detection. *Acta Automatica Sinica*, 2023, **49**(1): 135−147
(闫梦凯, 钱建军, 杨健. 弱对齐的跨光谱人脸检测. 自动化学报, 2023, **49**(1): 135−147)

6 Guo Ying-Chun, Feng Fang, Yan Gang, Hao Xiao-Ke. Cross-domain person re-identification on adaptive fusion network. *Acta Automatica Sinica*, 2022, **48**(11): 2744−2756
(郭迎春, 冯放, 阎刚, 郝小可. 基于自适应融合网络的跨域行人重识别方法. 自动化学报, 2022, **48**(11): 2744−2756)

7 Horn B K P, Schunck B G. Determining optical flow. *Artificial Intelligence*, 1981, **17**(1−3): 185−203

8 Sun D Q, Roth S, Black M J. Secrets of optical flow estimation and their principles. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR). San Francisco, USA: IEEE, 2010. 2432−2439

9 Menze M, Heipke C, Geiger A. Discrete optimization for optical flow. In: Proceedings of the 37th German Conference Pattern Recognition (GCPR). Aachen, Germany: Springer, 2015. 16−28

10 Chen Q F, Koltun V. Full flow: Optical flow estimation by global optimization over regular grids. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, USA: IEEE, 2016. 4706−4714

11 Dosovitskiy A, Fischer P, Ilg E, Häusser P, Hazirbas C, Golkov V. FlowNet: Learning optical flow with convolutional networks. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). Santiago, Chile: IEEE, 2015. 2758−2766

12 Ranjan A, Black M J. Optical flow estimation using a spatial pyramid network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, USA: IEEE, 2017. 2720−2729

13 Amiaz T, Lubetzky E, Kiryati N. Coarse to over-fine optical flow estimation. *Pattern Recognition*, 2007, **40**(9): 2496−2503

14 Hur J, Roth S. Iterative residual refinement for joint optical flow and occlusion estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA: IEEE, 2019. 5754−5763

15 Tu Z G, Xie W, Zhang D J, Poppe R, Veltkamp R C, Li B X, et al. A survey of variational and CNN-based optical flow techniques. *Signal Processing: Image Communication*, 2019, **72**: 9−24

16 Zhang C X, Ge L Y, Chen Z, Li M, Liu W, Chen H. Refined TV-

L_1 optical flow estimation using joint filtering. *IEEE Transactions on Multimedia*, 2020, **22**(2): 349−364

17 Dalca A V, Rakic M, Guttag J, Sabuncu M R. Learning conditional deformable templates with convolutional networks. In: Proceedings of the 33rd International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc., 2019. Article No. 32

18 Chen J, Lai J H, Cai Z M, Xie X H, Pan Z G. Optical flow estimation based on the frequency-domain regularization. *IEEE Transactions on Circuits and Systems for Video Technology*, 2021, **31**(1): 217−230

19 Zhai M L, Xiang X Z, Lv N, Kong X D. Optical flow and scene flow estimation: A survey. *Pattern Recognition*, 2021, **114**: Article No. 107861

20 Zach C, Pock T, Bischof H. A duality based approach for real-time TV-L_1 optical flow. In: Proceedings of the 29th DAGM Symposium on Pattern Recognition. Heidelberg, Germany: Springer, 2007. 214−223

21 Zhao S Y, Zhao L, Zhang Z X, Zhou E Y, Metaxas D. Global matching with overlapping attention for optical flow estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE, 2022. 17571−17580

22 Li Z W, Liu F, Yang W J, Peng S H, Zhou J. A survey of convolutional neural networks: Analysis, applications, and prospects. *IEEE Transactions on Neural Networks and Learning Systems*, 2022, **33**(12): 6999−7019

23 Han J W, Yao X W, Cheng G, Feng X X, Xu D. P-CNN: Part-based convolutional neural networks for fine-grained visual categorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022, **44**(2): 579−590

24 Ilg E, Mayer N, Saikia T, Keuper M, Dosovitskiy A, Brox T. FlowNet 2.0: Evolution of optical flow estimation with deep networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, USA: IEEE, 2017. 1647−1655

25 Sun D Q, Yang X D, Liu M Y, Kautz J. PWC-Net: CNNs for optical flow using pyramid, warping, and cost volume. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Salt Lake City, USA: IEEE, 2018. 8934−8943

26 Wang Z G, Chen Z, Zhang C X, Zhou Z K, Chen H. LCIF-Net: Local criss-cross attention based optical flow method using multi-scale image features and feature pyramid. *Signal Processing: Image Communication*, 2023, **112**: Article No. 116921

27 Teed Z, Deng J. RAFT: Recurrent all-pairs field transforms for optical flow. In: Proceedings of the 16th European Conference on Computer Vision (ECCV). Glasgow, UK: Springer, 2020. 402−419

28 Han K, Xiao A, Wu E H, Guo J Y, Xu C J, Wang Y H. Transformer in transformer. In: Proceedings of the 35th International Conference on Neural Information Processing Systems. Montreal, Canada: NIPS, 2021.15908−15919

29 Jiang S H, Campbell D, Lu Y, Li H D, Hartley R. Learning to estimate hidden motions with global motion aggregation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, USA: Canada, 2021. 9752−9761

30 Xu H F, Zhang J, Cai J F, Rezatofighi H, Tao D C. GMFlow: Learning optical flow via global matching. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE, 2022. 8111−8120

31 Cao J M, Li Y Y, Sun M C, Chen Y, Lischinski D, Cohen-Or D, et al. DO-Conv: Depthwise over-parameterized convolutional layer. *IEEE Transactions on Image Processing*, 2022, **31**: 3726−3736

32 Dong X Y, Bao J M, Chen D D, Zhang W M, Yu N H, Yuan L, et al. CSWin transformer: A general vision transformer backbone with cross-shaped windows. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE, 2022. 12114−12124

33 Huang Z L, Wang X G, Huang L C, Huang C, Wei Y C, Liu W Y. CCNet: Criss-cross attention for semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea (South): IEEE, 2019. 603−612

34 Butler D J, Wulff J, Stanley G B, Black M J. A naturalistic open source movie for optical flow evaluation. In: Proceedings of the 12th European Conference on Computer Vision (ECCV). Florence, Italy: Springer, 2012. 611−625

35 Menze M, Geiger A. Object scene flow for autonomous vehicles. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, USA: IEEE, 2015. 3061−3070

36 Wannenwetsch A S, Roth S. Probabilistic pixel-adaptive refinement networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE, 2020. 11639−11648

37 Hui T W, Tang X O, Loy C C. A lightweight optical flow CNN——revisiting data fidelity and regularization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021, **43**(8): 2555−2569

38 Hofinger M, Bulò S R, Porzi L, Knapitsch A, Pock T, Kontschieder P. Improving optical flow on a pyramid level. In: Proceedings of the 16th European Conference on Computer Vision (ECCV). Glasgow, UK: Springer, 2020. 770−786

39 Yu S H J, Zhang Y M, Wang C, Bai X, Zhang L, Hancock E R. HMFlow: Hybrid matching optical flow network for small and fast-moving objects. In: Proceedings of the 25th International Conference on Pattern Recognition (ICPR). Milan, Italy: IEEE, 2021. 1197−1204

40 Chen J, Cai Z M, Lai J H, Xie X H. Efficient segmentation-based PatchMatch for large displacement optical flow estimation. *IEEE Transactions on Circuits and Systems for Video Technology*, 2019, **29**(12): 3595−3607

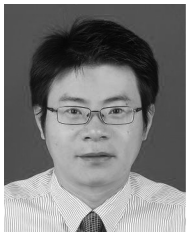
41 Zhang C X, Zhou Z K, Chen Z, Hu W M, Li M, Jiang S F. Self-attention-based multiscale feature learning optical flow with occlusion feature map prediction. *IEEE Transactions on Multimedia*, 2022, **24**: 3340−3354

42 Lu Z H, Xie H T, Liu C B, Zhang Y D. Bridging the gap between vision transformers and convolutional neural networks on small datasets. In: Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans, USA, 2022: 14663−14677

王梓歌 南昌航空大学测试与光电工程学院硕士研究生. 主要研究方向为计算机视觉.
E-mail: Wangzggg@163.com
(**WANG Zi-Ge** Master student at the School of Measuring and Optical Engineering, Nanchang Hangkong University. Her main research interest is computer vision.)

葛利跃 南昌航空大学助理实验师. 北京航空航天大学仪器科学与光电工程学院博士研究生. 主要研究方向为图像检测与智能识别.
E-mail: lygeah@163.com
(**GE Li-Yue** Assistant experimenter at Nanchang Hangkong Uni-

versity. Ph.D. candidate at the School of Instrumentation and Optoelectronic Engineering, Beihang University. His research interest covers image detection and intelligent recognition.)



陈 震 南昌航空大学测试与光电工程学院教授. 2003 年获得西北工业大学博士学位. 主要研究方向为图像处理与计算机视觉.
E-mail: dr_chenzhen@163.com
(**CHEN Zhen** Professor at the School of Measuring and Optical Engineering, Nanchang Hangkong University. He received his Ph.D. degree from Northwestern Polytechnical University in 2003. His research interest covers image processing and computer vision.)

张聪炫 南昌航空大学测试与光电工程学院教授. 2014 年获得南京航空航天大学博士学位. 主要研究方向为图像处理与计算机视觉. 本文通信作者.
E-mail: zcxdsg@163.com
(**ZHANG Cong-Xuan** Professor at the School of Measuring and Optical Engineering, Nanchang Hangkong University. He received his Ph.D. degree from Nanjing University of Aeronautics and Astronautics in 2014. His research interest covers image processing and computer vision. Corresponding author of this paper.)

王子旭 南昌航空大学测试与光电工程学院硕士研究生. 主要研究方向为计算机视觉.
E-mail: wangzixu0827@163.com
(**WANG Zi-Xu** Master student at the School of Measuring and Optical Engineering, Nanchang Hangkong University. His main research interest is computer vision.)

舒铭奕 南昌航空大学测试与光电工程学院硕士研究生. 主要研究方向为计算机视觉.
E-mail: shumingyi1997@163.com
(**SHU Ming-Yi** Master student at the School of Measuring and Optical Engineering, Nanchang Hangkong University. His main research interest is computer vision.)