联合深度可分离残差与多尺度双通道注意力的全局匹配优化光流估计方法

王子旭[1,2],陈弘烨[1],葛利跃[3,4*],张聪炫[1],陈 震[1,2],王梓歌[1]
(1. 南昌航空大学仪器科学与光电工程学院,江西南昌 330063;2. 西北工业大学计算机学院,陕西西安 710129;
3. 南昌航空大学信息工程学院,江西南昌 330063;4. 北京航空航天大学仪器科学与光电工程学院,北京 100019)

摘 要: 随着深度学习理论与技术的快速发展,基于深度学习的光流估计方法在计算精度与鲁棒性方面取得显著提升. 然而,受标准卷积感受野局部属性和现有匹配代价体积策略容易产生匹配歧义的限制,当前方法在大位移运动和弱纹理区域普遍存在光流估计精度较低,运动模糊现象较严重的问题. 针对上述问题,本文提出一种联合深度可分离残差与多尺度双通道注意力的全局匹配优化光流估计方法. 首先,构建联合深度可分离残差块与多尺度双通道注意力的编码模块,在平衡参数量与运算速度的同时获取连续帧间更准确的深度特征. 然后,设计基于可学习的全局匹配优化光流估计策略,通过排除遮挡并高效利用全局匹配信息,有效缓解因匹配歧义引起的运动模糊. 最后,为了提高模型的训练稳定性与泛化性,本文提出联合全局与局部的光流损失函数,约束模型训练. 实验分别采用MPI-Sintel、KITTI-2015 和 Middlebury 测试数据集对本文方法和现有代表性方法进行综合对比分析. 结果表明,本文方法在所有对比方法中取得了最优的光流估计精度,尤其在大位移和弱纹理区域具有更好的准确性和鲁棒性.

关键词: 光流;代价体积;匹配歧义;大位移与弱纹理;运动模糊

基金项目: 国家自然科学基金(No.62222206,No.62272209);江西省重大科技研发专项(No.20232ACC01007);江西省自然科学基金(No.20242BAB20048)

中图分类号: TP391 **文献标识码:** A **文章编号:** 0372-2112(2025)05-1622-15

电子学报 URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20240818

A Global Matching Optimization Approach for Optical Flow Estimation Using Joint Depth-Separable Residual Blocks and Multi-Scale Dual-Channel Attention

WANG Zi-xu[1,2], CHEN Hong-ye[1], GE Li-yue[3,4*], ZHANG Cong-xuan[1], CHEN Zhen[1,2], WANG Zi-ge[1]
(1. *School of Instrument Science and Optoelectronic Engineering*, *Nanchang Hangkong University*, *Nanchang*, *Jiangxi* 330063, *China*;
2. *School of Computer Science*, *Northwestern Polytechnical University*, *Xi'an*, *Shaanxi* 710129, *China*;
3. *School of Information Engineering*, *Nanchang Hangkong University*, *Nanchang*, *Jiangxi* 330063, *China*;
4. *School of Instrumentation and Optoelectronic Engineering*, *Beihang University*, *Beijing* 100019, *China*)

Abstract: With the rapid development of deep learning theory and technology, deep learning-based optical flow estimation methods have significantly improved in computational accuracy and robustness. However, due to the limitations of standard convolution's local receptive field and existing matching cost volume strategies that can lead to matching ambiguities, current methods often suffer from low accuracy in optical flow estimation and severe motion blur, particularly in large displacement motions and weak-texture regions. To address these issues, this paper proposes a global matching optimization optical flow estimation method combining deep separable residuals with multi-scale dual-channel attention. First, an encoding module is constructed that integrates deep separable residual blocks with multi-scale dual-channel attention, achieving more accurate depth features between consecutive frames while balancing parameter count and computational speed. Then, a learnable global matching optimization strategy for optical flow estimation is designed, which alleviates motion blur caused by matching ambiguities by excluding occlusions and efficiently utilizing global matching information. Finally, to

收稿日期:2024-09-03;修回日期:2025-02-23;责任编辑:朱梅玉
*通讯作者:葛利跃

enhance the model's training stability and generalization, a combined global and local optical flow loss function is proposed to constrain model training. Experiments conducted on the MPI-Sintel, KITTI-2015 and Middlebury test datasets demonstrate that the proposed method achieves the best optical flow estimation accuracy among all compared methods, especially showing better accuracy and robustness in large displacement and weak-texture regions.

Key words: optical flow; cost volume; matching ambiguity; large displacement and weak texture; motion blur

Foundation Item(s): National Natural Science Foundation of China (No.62222206, No.62272209); Major Research and Development Project of Jiangxi Province (No.20232ACC01007); Natural Science Foundation of Jiangxi Province (No.20242BAB20048)

1　引言

光流估计是图像处理和计算机视觉研究领域重点方向之一,旨在从连续图像序列中计算每个像素点的二维运动矢量. 由于光流包含了物体与场景的运动参数和相关结构信息,因此通常作为高级视觉任务的关键基础,广泛应用于行为分析[1]、目标检测[2]、视频压缩感知[3]、图像超分辨率[4]等领域.

自 Horn 和 Schunck[5]首次提出光流基本守恒假设和计算模型后,针对图像序列的光流估计方法研究得到快速发展. 其中,基于变分理论的光流估计模型[6,7]由于其既可以获取稠密的光流场又有灵活的延展性,成为传统光流估计方法研究的主流解决方案. 然而,该类方法需要设计严格的手工特征约束条件,在针对包含大位移、运动遮挡等复杂场景运动光流估计时,往往存在光流估计精度低、鲁棒性差的问题[8]. 此外,由于需要大量的迭代步骤最小化能量函数,该类方法时间成本消耗过大,严重限制了光流估计技术在实际任务场景的应用.

随着深度学习理论与技术的快速发展,得益于其强大的参数学习与网络拟合能力,使光流估计技术在计算精度与时效性方面得到突破性提升[9,10]. 通常,基于深度学习的光流估计方法基本流程[11,12]可描述如下:首先基于卷积神经网络提取图像连续帧间运动特征,然后基于运动特征构建匹配代价获取连续图像帧间对应像素点匹配关系,最后,通过构建光流编-解码模块从匹配代价体积中回归出光流. 因此,准确地捕获场景运动特征和构建最佳匹配模型,是实现高精度深度学习光流估计方法的关键. 然而,随着任务场景的不断复杂化,当目标场景包含大位移运动和弱纹理区域时,由于像素位移过大使当前普遍较小卷积核的标准卷积或窗口注意力构建运动编码器的深度学习方法[12,13],难以准确捕捉大位移运动场景中像素点间的长程依赖关系. 同时,由于缺乏丰富的纹理模式和大位移运动引起的局部遮挡现象,使现有方法在构建匹配代价体积时容易产生对应像素点的匹配歧义[13,14],使光流估计的准确性和鲁棒性出现明显下降.

针对上述问题,本文提出一种联合深度可分离残差与多尺度双通道注意力的全局匹配优化光流估计方法. 首先,利用深度可分离残差块与多尺度双通道注意力机制,构建深度特征编码器,在平衡参数量与运算速度的同时获取连续帧间更准确的图像特征. 然后,针对现有匹配代价体积策略存在局部匹配歧义问题,设计了一种基于可学习的全局匹配优化策略,通过排除遮挡并高效利用全局匹配信息,有效缓解匹配歧义引起的运动模糊现象. 最后,为了提高模型训练的稳定性与泛化能力,本文提出一种联合全局与局部的光流损失函数,约束模型训练. 实验结果显示,本文方法可以显著提升大位移运动和弱纹理区域的光流估计精度和鲁棒性.

2　相关工作

近年来,得益于基于卷积神经网络的深度学习方法在图像处理与计算机视觉领域取得的巨大成功[15,16],将深度学习技术引入光流估计任务成为研究人员关注的重点与热点. FlowNet[11](learning otical Flow with cnvolutional Networks)是第1个使用端到端网络的深度学习光流计算模型,尽管其计算精度低于传统变分方法,但它突破了实时光流估计难题并为后续深度学习光流估计任务带来了启发. 例如FlowNet2[17](evolution of optical Flow with deep Networks2)以堆叠多个FlowNet网络的形式,通过增加网络深度与尺度提高模型光流预测性能,但显著增加了模型的参数量和运算时间. 为了处理大位移运动光流估计准确性的问题,SpyNet[18](Spatial pyramid Network)基于图像金字塔和变形操作构建了一个紧凑的光流估计模型,缓解大位移对光流估计的影响. Sun等人[12]借鉴传统方法由粗到细的计算思想,开发了一个由堆叠图像金字塔、图像变形操作和匹配代价体积组成的特征金字塔光流计算网络模型PWC-Net(Pyramid Warpingand Cost volume Network),进一步提升大位移光流计算的准确性与鲁棒性. 由于其出色的性能,使其成为后续诸多光流估计方法的基础模型. 例如,Hui等人[19]采用类似思想并引入正则化约束构建了一个轻量化的光流估计网络LiteFlowNet(lightweight convolutional neural Network for optical Flow estimation),在保持计算精度的同时显著提升模型的计算效率. IRR-PWC[20](Iterative Residual Refinement version of PWC)引入了迭代细化方案,通过在

不同金字塔级别重复使用相同的光流解码器模块，缓解了由于下采样层数过深导致小目标丢失的问题．体积对应网络（Volumetric Correspondence Networks，VCN）[21]通过构建4D匹配代价体积获取更精确的匹配关系，提高大位移区域的光流估计准确度．针对遮挡场景光流估计问题，Zhao等人[22]联合遮挡约束和运动约束对光流估计任务进行建模，有效提升了遮挡区域的光流估计精度．

不同于上述光流估计策略，Teed等人[13]通过构建4D匹配代价体积设计一种基于GRU（Gated Recurrent Unit，GRU）的循环迭代方案，提出了一种新颖的RAFT（Recurrent All-pairs Field Transforms for ptical flow）光流估计模型，既避免了模型对像素点间匹配特征的多分辨率搜索，又有效平衡了模型参数与估计精度．此后，许多方法都以RAFT为基线模型进行优化与改进．例如，SeparableFlow[23]利用非局部聚合来优化匹配代价体积中存在的错误匹配信息，增强了模型的鲁棒性．Jiang等人[14]利用上下文信息构造全局聚合模块，通过上下文信息与运动信息的交互，提升遮挡区域的光流估计精度．文献[24]在其解码器中通过引入大核卷积提升解码过程的感受野，进一步提升了遮挡区域的光流估计精度．此外，GMFlow[25]（larning otical Flow via Global Matching）方法通过先使用Transformer进行特征增强，再由匹配代价直接获得初始光流场，通过自注意力改善不匹配问题，显著提升了大位移场景下的光流估计精度．虽然，当前基于深度学习的光流估计方法在计算精度和鲁棒性方面取得了明显进步，但当目标场景包含弱纹理和大位移运动情况时，仍存在运动模糊和准确性较低的问题．

3 本文方法

3.1 模型框架

基于深度学习的光流估计模型通常采用特征编码-匹配代价体积-光流解码的3段式结构．其中特征编码主要用于提取输入图像序列特征，为后续的匹配代价体积提供可靠的数据支持．目前，主流的特征编码结构是采用PWC-Net[12]中的特征金字塔编码架构，通过逐层下采样关注不同尺度的图像特征．但是，随着下采样深度的增加，像素点的过度损失导致模型光流估计的可靠性显著降低．此外，现有匹配代价体积策略的局部性也使模型在大位移运动和弱纹理区域产生匹配歧义，从而产生较为严重的运动模糊．针对上述问题，本文提出一种联合深度可分离残差与多尺度双通道注意力的全局匹配优化光流估计网络，在提取更准确的图像深度特征同时采用基于可学习的全局匹配优化策略降低匹配歧义，显著提升大位移和弱纹理区域的光流估计准确性和鲁棒性．

图1为本文所提方法的网络模型示意图．模型主要由深度特征编码器、基于可学习的全局匹配优化模块、上下文编码器和光流解码器、联合全局与局部信息的光流损失函数构成．其中，深度编码器主要由深度可分离残差与多尺度双通道注意力模块组成，用于提取深度图像特征．基于可学习的全局匹配优化模块，从匹配代价体积中获得粗级光流和遮挡掩码．上下文特征编码器主要由3层连续的卷积残差块组成，获取图像上下文信息．光流解码器主要由匹配代价搜索、运动特征编码与聚合和GRU循环迭代优化部分组成，回归出更为精细的光流．

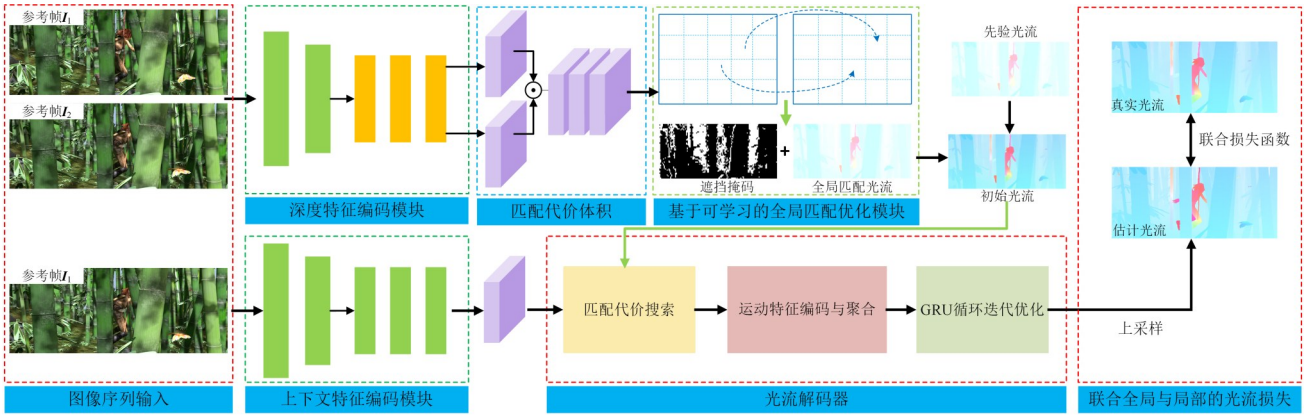


图1 联合深度可分离残差与多尺度双通道注意力的全局匹配优化光流估计模型示意图

具体来说，如图1所示，首先将连续2帧参考图像序列I_1、I_2输入深度特征编码器中获取图像序列特征，同时将第1帧图像I_1单独输入上下文特征编码用于捕获上下文特征．接着，得到的图像序列深度特征通过点积相似度计算得到匹配代价体积，即储存2帧图像序列像素之间的位置关系．然后，通过可学习的全局匹配优化模块获取匹配粗级光流场和遮挡掩码，结合上一帧输出的先验光流获取初始光流．最后，本文将初始光流

与上下文特征结合输入光流解码模块回归出更精细的输出光流. 需要注意的是,为了获得与输入图像序列相同分辨率的光流结果,本文通过视觉相似引导光流上采样策略[26]将解码器输出光流上采样至原始图像分辨率,便于使用联合全局与局部信息的光流损失函数优化.

3.2　深度特征编码器

基于深度学习的光流估计方法通常采用基于标准卷积的U-Net[11](U Network)或ResNet[13](Residual Network)方案来构建特征编码模块,通过引入窗口注意力进行增强特征,提升模块性能. 但标准卷积和窗口注意力的内在局部属性和远距离像素建模能力的不足,使特征提取模块难以捕捉像素位移较大或纹理模式缺失的区域图像特征. 为此,本文提出了联合深度可分离残差与多尺度双通道注意力的深度特征编码模块,在平衡参数量与运算速度的同时获取连续帧间更准确且全面的图像特征. 图2展示了所提深度特征编码模块的网络结构,从图2(*a*)可以看出,深度可分离残差模块和多尺度双通道注意力模块以串联方式组合构成深度编码器. 下面将详细介绍各组成部分的网络结构与工作原理.

3.2.1　深度可分离残差模块

如图2(*b*)所示,本文使用深度可分离卷积和逐点卷积构建深度可分离残差块,然后通过组合深度可分离残差块构建初始特征编码器. 从图2(*b*)可以看出,整个模块包含2个残差结构,在输出端均使用⊕加法操作进行特征融合,允许网络在不同层之间直接传递信息,在一定程度缓解特征传递过程中产生的特征稀释问题,提高信息传递效率增强模型的性能. 具体来说,本文首先通过分离的下采样层对每层特征进行下采样,深度卷积提取空间特征,通过2个连续的逐点卷积对通道特征进行扩张融合,提取更加充分的初始特征. 过程如下:

$$\boldsymbol{F}^{n}_{\mathrm{LM}}=\mathrm{DSR}(\boldsymbol{F}^{n}),n\in 1,2,\cdots,N \tag{1}$$

式中,n表示图像特征采样层数,$\boldsymbol{F}^{n}$表示第n层提取的图像特征,例如,当n=1时,$\boldsymbol{F}^{1}$表示输入特征为原始图像;$\boldsymbol{F}^{n}_{\mathrm{LM}}$表示第$n$层输出图像特征经DSR(Depth-Sparable Residual module)处理后得到的低级运动特征(Low-level Motion features,LM),其中DSR表示深度可分离残差块,结构表示为

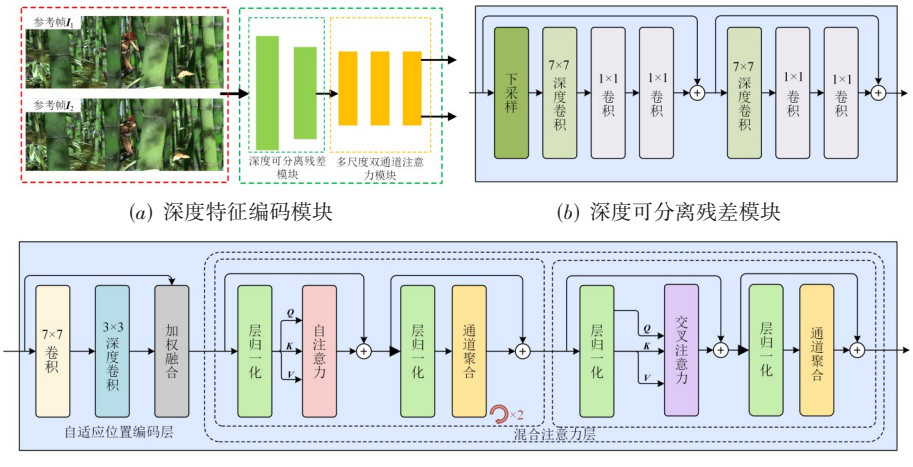
$$\mathrm{DSR}=\mathrm{Conv}_{1\times1}\Big(\mathrm{GELU}\Big(\mathrm{Conv}_{1\times1}\Big(\mathrm{norm}\Big(\mathrm{Dwconv}_{7\times7}(\boldsymbol{F})\Big)\Big)\Big)\Big) \tag{2}$$

式中,$\boldsymbol{F}$表示输入图像特征;$\mathrm{Conv}_{1\times1}(\cdot)$表示逐点卷积;$\mathrm{Dwconv}_{7\times7}(\cdot)$为卷积核尺寸为7×7的深度卷积;norm(·)表示归一化层;GELU(Gaussian Error Linear Unit)为激活函数.

该模块相比于传统仅使用3×3标准卷积的特征提取方法,可以获得内容更为丰富的初始特征,为后续深度特征提取提供支持.

3.2.2　多尺度双通道注意力模块

通过深度可分离残差块提取初始图像运动特征,为了提取深度特征并建模远距离像素间的长程依赖关系,本文构建了基于Transformer的多尺度双通道注意力的深度特征提取模块. 如图2(*c*)所示,该模块分为2个部分,第1部分是自适应位置编码层,第2部分是混合注意力层.



(*a*) 深度特征编码模块　　(*b*) 深度可分离残差模块

(*c*) 多尺度双通道注意力模块

图2　联合深度可分离残差与多尺度双通道注意力的深度特征编码模块

下面介绍具体实现步骤:

(1)自适应位置编码层. 通过增强图像帧本身内部像素间的位置信息,提取更密集的图像特征,有助于注意力模块感知像素点的上下文位置关系. 首先,使用卷积核尺寸为7×7标准卷积对输入的初始图像特征进行下采样,使用深度卷积进行局部性质学习. 然后,通过激活函数对特征进行变换,最后将原始特征与使用位置编码后的特征进行点积并做残差连接进行加权融

合，实现自适应位置编码，该过程如下：

$$\boldsymbol{F}_{\mathrm{M}}^{\mathrm{P}}=\mathrm{Fusion}(\boldsymbol{F}_{\mathrm{LM}}) \tag{3}$$

式中，$\boldsymbol{F}_{\mathrm{LM}}$表示输入的初始图像运动特征；Fusion(·)表示加权融合操作；$\boldsymbol{F}_{\mathrm{M}}^{\mathrm{P}}$表示经自适应位置编码的图像运动特征，P表示位编码(Position encoding，P)，M表示运动特征(Motion features，M).

(2)混合注意力层. 混合注意力层如图2(*c*)所示，其运算过程分为2步：第1步使用自注意力层加强单一图像内的特征，第2步使用交叉注意力层建立连续帧图像之间的联系.

首先，将经过位置编码的特征输入自注意力层，通过学习不同位置特征之间的依赖关系，获取全局特征和上下文关系. 接着，对经自注意力加权的特征执行通道聚合，捕获更复杂的特征模式和表达能力. 通道聚合操作，本质是一个多层感知机(MultiLayer Perceptron，MLP)，包含多个全连接层和3×3深度卷积，每个全连接层对输入特征进行非线性映射，以引入更高阶的特征交互与表示能力. 3×3深度卷积的作用是利用图像中的空间上下文信息，建立相邻像素之间的空间依赖关系，更好地捕获局部信息. 因此，自注意力层公式为

$$\boldsymbol{q}_{\mathrm{SA}}=\mathrm{Linear}(\boldsymbol{F}_{\mathrm{M}}^{\mathrm{P}}),\ (\boldsymbol{k}_{\mathrm{SA}},\boldsymbol{v}_{\mathrm{SA}})=\mathrm{Linear}\left(\mathrm{MS}(\boldsymbol{F}_{\mathrm{M}}^{\mathrm{P}})\right) \tag{4}$$

$$\boldsymbol{F}_{\mathrm{M}}^{\mathrm{SA}}=\mathrm{MLP}\left(\mathrm{Softmax}\left(\frac{\boldsymbol{q}_{\mathrm{SA}}\boldsymbol{k}_{\mathrm{SA}}^{\mathrm{T}}}{\sqrt{d}}\right)\boldsymbol{v}_{\mathrm{SA}}\right) \tag{5}$$

式中，$\boldsymbol{q}_{\mathrm{SA}}$、$\boldsymbol{k}_{\mathrm{SA}}$、$\boldsymbol{v}_{\mathrm{SA}}$分别表示自注意力使用的$\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$矩阵，均来源于相同的一组归一化层输出特征；Linear(·)表示线性映射；$\boldsymbol{F}_{\mathrm{M}}^{\mathrm{SA}}$表示通过自注意力层后获得的输出特征.

由于自注意力仅对单帧图像进行建模，而运动信息是通过连续帧图像像素点之间的位置关系来确定的. 因此，本文又使用交叉注意力层对2帧图像像素间位置关系进行建模，即构建2帧图像相互间的依赖关系. 图3展示了交叉注意力层结构示意图，首先，本文对自注意力层输出特征$\boldsymbol{F}_{\mathrm{M}}^{\mathrm{SA}}$进行拆分并对应线性投影生成$\boldsymbol{K}$、$\boldsymbol{V}$矩阵. 这里本文将注意力头分为2个部分，一部分通过多尺度空间采样模块MS(·)进行注意力操作，另一部分则保留原始分辨率的特征. 然后，将不同注意力头产生的特征输入交叉注意力进行特征学习建模. 最后，通过通道聚合和多层感知机MLP对不同层次的特征进行特征变换和聚合，获取最终的深度特征$\boldsymbol{F}_{\mathrm{M}}^{\mathrm{CA}}$. 完整交叉注意力层过程如下：

$$\begin{aligned}&\boldsymbol{q}_{\mathrm{CA}}=\mathrm{Linear}(\boldsymbol{F}_{\mathrm{M}}^{\mathrm{SA}'}),\\&(\boldsymbol{k}_{\mathrm{CA}},\boldsymbol{v}_{\mathrm{CA}})=\mathrm{Linear}\left(\mathrm{MS}(\boldsymbol{F}_{\mathrm{M}}^{\mathrm{SA}'})\right)\end{aligned} \tag{6}$$

$$\boldsymbol{F}_{\mathrm{M}}^{\mathrm{CA}}=\mathrm{MLP}\left(\mathrm{Softmax}\left(\frac{\boldsymbol{q}_{\mathrm{CA}}\boldsymbol{k}_{\mathrm{CA}}^{\mathrm{T}}}{\sqrt{d}}\right)\boldsymbol{v}_{\mathrm{CA}}\right) \tag{7}$$

式中，$\boldsymbol{q}_{\mathrm{CA}}$、$\boldsymbol{k}_{\mathrm{CA}}$、$\boldsymbol{v}_{\mathrm{CA}}$分别表示交叉注意力使用的$\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$矩阵，$\boldsymbol{Q}$来源于第1帧生成的特征，$\boldsymbol{K}$、$\boldsymbol{V}$来源于第2帧生成的特征；$\boldsymbol{F}_{\mathrm{M}}^{\mathrm{SA}'}$表示将运动特征分为2个相同通道数的前后2帧特征，作用是产生交叉注意力所需的$\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$矩阵；$\boldsymbol{F}_{\mathrm{M}}^{\mathrm{CA}}$表示通过交叉注意力层后获得的最终输出的深度特征.

综上，本文采用的混合注意力模块表述如下：

$$\boldsymbol{F}_{\mathrm{M}}=\mathrm{CA}\left(\mathrm{SA}\left(\mathrm{SA}(\boldsymbol{F}_{\mathrm{M}}^{\mathrm{P}})\right)\right) \tag{8}$$

式中，CA(·)表示交叉注意力层；SA(·)表示自注意力层；$\boldsymbol{F}_{\mathrm{M}}$表示输出的深度特征.

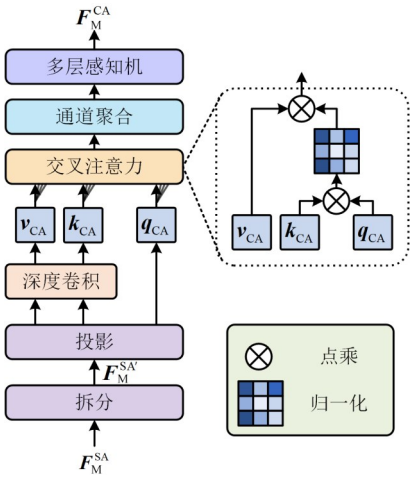


图3 交叉注意力层结构

图4以MPI-Sintel(Max Planck Institute Sintel optical flow dataset)数据集[27]中bamboo_2和ambush_1为例展示了所提联合深度可分离残差与多尺度双通道注意力深度特征编码模块的有效性. 其中第2行和第3行分别展示了对比方法GMA[14](Global Motion Aggregation)和本文方法的特征提取可视化结果. 从图4可以看出，与GMA相比，本文方法捕捉到了更为丰富且置信度更高的图像轮廓和纹理特征.

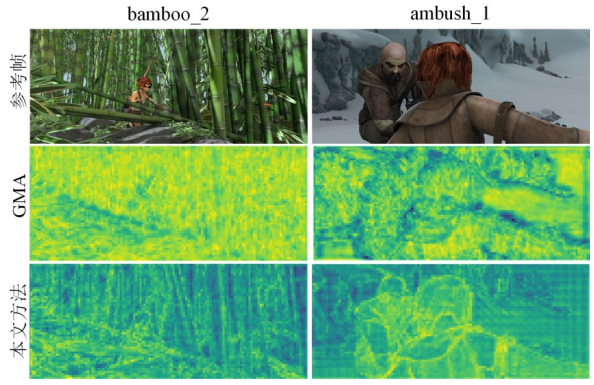


图4 MPI-Sintel数据集特征提取可视化结果对比

3.3 基于可学习的全局匹配优化模块

光流估计本质上是寻找连续图像序列帧间对应像素的匹配关系，并计算二者之间的位移场. 为了实现该

目标,深度学习光流估计模型一般通过逐像素构建匹配代价来计算像素间的特征相似性,找到最佳匹配像素.因此,该过程要求特征之间具有足够的差异性,确保准确匹配.当前主流的方法是采用RAFT[12](Recurrent All-pairs Field Transforms for ptical flow)提出的金字塔局部查找策略,从匹配代价体积中寻找到最佳匹配像素并回归光流.然而,当目标场景中包含大位移和弱纹理区域时,像素位移较大难以匹配以及缺乏必要的纹理信息,使该方案易出现局部匹配歧义而陷入局部最优,造成光流估计结果存在较严重的模糊现象.针对该问题,文献[25]将光流估计转化为一种显式问题,直接从全局匹配代价中解码得到全局光流场,但因遮挡区域匹配结果置信度较低,导致光流估计结果鲁棒性较差.

为此,本文提出了一种基于可学习的全局匹配优化模块.首先,计算全局匹配代价.令$\boldsymbol{I}_1,\boldsymbol{I}_2\in\boldsymbol{R}^{3\times H\times W}$表示输入的连续2帧图像序列,其中$H$表示图像高度,$W$表示图像宽度.经深度特征编码模块提取的特征表示为$\boldsymbol{F}_1,\boldsymbol{F}_2\in\boldsymbol{R}^{D\times H/8\times W/8}$,其中$D$表示特征图通道数.则全局匹配代价体积可计算如下:

$$\mathrm{Cost}(x,y,i,j)=\boldsymbol{F}_1(x,y)\odot\boldsymbol{F}_2(i,j) \tag{9}$$

式中,$(x,y)^{\mathrm{T}}$和$(i,j)^{\mathrm{T}}$分别表示特征图$\boldsymbol{F}_1$和$\boldsymbol{F}_2$中任意像素点的位置坐标;$\odot$表示逐元素相乘;$\mathrm{Cost}(x,y,i,j)\in\boldsymbol{R}^{H\times W\times H\times W}$为全局匹配代价体积,储存了2帧图像每个像素点之间的相关性.

使用对偶Softmax函数和坐标变换,将全局匹配代价体积转换为匹配置信度,过滤置信度较低的匹配像素,该计算过程可描述如下:

$$\begin{cases}\mathrm{Cost}_1(x,y,\cdot,\cdot)=\mathrm{Softmax}\big(\mathrm{Cost}(x,y,\cdot,\cdot)\big)\\ \mathrm{Cost}_2(\cdot,\cdot,i,j)=\mathrm{Softmax}\big(\mathrm{Cost}(\cdot,\cdot,i,j)\big)\\ \boldsymbol{P}_{\mathrm{C}}(x,y,i,j)=\mathrm{Cost}_1(x,y,\cdot,\cdot)\odot\mathrm{Cost}_2(\cdot,\cdot,i,j)\end{cases} \tag{10}$$

式中,$\boldsymbol{P}_{\mathrm{C}}(x,y,i,j)$表示全局匹配代价体积的匹配置信度.本文将匹配置信度最高的一对像素视为2帧中对应匹配点.

由于遮挡现象的存在,2帧间对应像素的丢失,导致该区域计算出的置信度并不可靠.针对该问题,本文利用匹配置信度$\boldsymbol{P}_{\mathrm{C}}$来确定可信匹配区域中非遮挡区域的方法,排除遮挡干扰.计算过程如下:

$$\begin{cases}\boldsymbol{M}_1(x,y)=\mathrm{Argmax}_{i,j}\big(\boldsymbol{P}_{\mathrm{C}}(x,y,i,j)\big)\\ \boldsymbol{M}_2(i,j)=\mathrm{Argmax}_{x,y}\big(\boldsymbol{P}_{\mathrm{C}}(x,y,i,j)\big)\\ \boldsymbol{M}_{\mathrm{C}}=\boldsymbol{M}_2\big(\boldsymbol{M}_1(x,y)\big)\end{cases} \tag{11}$$

式中,$\mathrm{Argmax}(\cdot)$表示返回输入指定维度的最大值;$\boldsymbol{M}_1(x,y)$表示在给定第2帧像素点坐标$(i,j)^{\mathrm{T}}$时,获取到的$\boldsymbol{I}_1$和$\boldsymbol{I}_2$中匹配概率最大的像素坐标;同理$\boldsymbol{M}_2(i,j)$表示在给定第1帧像素点坐标$(x,y)^{\mathrm{T}}$时,获取到的$\boldsymbol{I}_2$和$\boldsymbol{I}_1$中匹配概率最大的像素坐标.当$\boldsymbol{M}_1$和$\boldsymbol{M}_2$满足$\boldsymbol{M}_1(x,y)=\boldsymbol{M}_2(i,j)$时,即2帧图像中对应像素点坐标和匹配置信度都相同时,将该对像素点视为正确匹配点,其集合用$\boldsymbol{M}_{\mathrm{C}}$表示排除遮挡区域.

此外,由于遮挡区域无法获取置信度信息,其光流值是未知的.为了补充遮挡区域的光流值,本文通过屏蔽非遮挡区域$\boldsymbol{M}_{\mathrm{C}}$上的已知点获取遮挡区域掩码Mask.通过矩阵乘法,将前1对帧产生的光流与遮挡掩码进行计算,获取遮挡区域的初始光流值.最后将遮挡区域的光流和非遮挡区域的光流叠加,得到所有区域的光流初始值.由于光流可以通过计算对应像素坐标之间的差来获得,该过程如下:

$$\boldsymbol{f}_{\mathrm{LR}}^{\mathrm{Init}}(x,y)=\begin{cases}\boldsymbol{M}_{\boldsymbol{F}_1\to\boldsymbol{F}_2}(x,y)-(x,y), & (x,y)^{\mathrm{T}}\in\boldsymbol{M}_{\mathrm{C}}\\ \mathrm{Mask}(x,y)\times\boldsymbol{f}_{\mathrm{pre}}(x,y), & (x,y)^{\mathrm{T}}\in\mathrm{Others}\end{cases} \tag{12}$$

式中,$(x,y)^{\mathrm{T}}$表示第1帧中像素点坐标;$\boldsymbol{M}_{\boldsymbol{F}_1\to\boldsymbol{F}_2}(x,y)$表示第1帧像素点$(x,y)^{\mathrm{T}}$在第2帧中对应的匹配像素点坐标;$(x,y)^{\mathrm{T}}$表示满足集合$\boldsymbol{M}_{\mathrm{C}}$要求的匹配像素点位置坐标;$\boldsymbol{f}_{\mathrm{pre}}(x,y)$表示前1对帧的估计光流;$\mathrm{Mask}(x,y)\times\boldsymbol{f}_{\mathrm{pre}}(x,y)$表示用遮挡掩码和前1对帧估计光流预测的遮挡区域光流;$\boldsymbol{f}_{\mathrm{LR}}^{\mathrm{Init}}(x,y)$表示像素点$(x,y)^{\mathrm{T}}$处的低分辨率(Low Resolution,LR)初始光流(Initial optical flow,Init),即由全局匹配模块得到的初始光流.之后,再通过图1中的光流解码器进行迭代细化得到最终的输出光流.

3.4　联合全局与局部信息的光流损失函数

通常情况下,有监督光流计算模型大多采用L_1或L_2范数损失函数以约束模型优化.但这类损失函数过于关注对整幅图像像素的约束,忽略了局部区域差异.因此,本文提出了一种联合全局与局部信息的光流损失函数,通过联合局部光流损失、全局匹配损失和全局光流损失,更全面地约束模型优化,提升模型训练的可靠性.

具体来说,为了有效约束基于学习的全局匹配模块获取初始光流,本文使用光流真实值进行局部监督,构建了局部光流损失:

$$L_{\mathrm{Local}}=\left\|\boldsymbol{f}_{\mathrm{LR}}^{\mathrm{GT}}(x,y)-\boldsymbol{f}_{\mathrm{LR}}^{\mathrm{Init}}(x,y)\right\|_{L_1} \tag{13}$$

式中,$(x,y)^{\mathrm{T}}$表示匹配点的位置坐标;$\boldsymbol{f}_{\mathrm{LR}}^{\mathrm{Init}}(x,y)$表示计算出来的低分辨率下匹配区域光流值;$\boldsymbol{f}_{\mathrm{LR}}^{\mathrm{GT}}(x,y)$表示低分辨率下匹配区域光流真实值.在匹配区域,本文计算真实值与预测初始值的L_1范数来约束网络优化.

本文对匹配置信度施加惩罚,鼓励模型更加关注匹配置信度值最高的像素点对.为此,首先需要得到匹配真实值:

$$\boldsymbol{I}_1'(x,y)=\mathrm{Warp}\big(\boldsymbol{f}_{\mathrm{LR}}^{\mathrm{GT}}(x,y),\boldsymbol{I}_2(x,y)\big) \tag{14}$$

$$\boldsymbol{M}_{\mathrm{C}}^{\mathrm{GT}}(x,y)=\left|L_{\mathrm{ph}}\big(\boldsymbol{I}_1(x,y)-\boldsymbol{I}_1'(x,y)\big)\right|\leqslant 20 \tag{15}$$

式中，$(x,y)^{\mathrm{T}}$表示图像的任意像素点；Warp(·)表示变形操作，即利用光流真实值$\boldsymbol{f}_{\mathrm{LR}}^{\mathrm{GT}}(x,y)$将第2帧图像$\boldsymbol{I}_2(x,y)$向第1帧图像变形；$\boldsymbol{I}_1'(x,y)$表示变形得到的第1帧图像；$L_{\mathrm{ph}}(\cdot)$表示计算图像的亮度差异；$\boldsymbol{M}_{\mathrm{C}}^{\mathrm{GT}}$表示匹配区域真实值. 然后，使用负对数似然函数得到全局匹配损失：

$$L_{\mathrm{m}}=-\sum_{(x,y)\in \boldsymbol{M}_{\mathrm{C}}^{\mathrm{GT}}}\log \boldsymbol{P}_{\mathrm{C}}(x,y) \tag{16}$$

式中，$\boldsymbol{P}_{\mathrm{C}}(x,y)$表示匹配区域像素点置信度；$L_{\mathrm{m}}$表示匹配损失.

为了对全局的光流计算值进行有效约束，在每轮迭代优化过程中，利用光流真实值计算全局光流损失L_{g}：

$$\boldsymbol{f}_{\mathrm{g}}=\mathrm{VSGU}(\boldsymbol{f}_{\mathrm{LR}}) \tag{17}$$

$$L_{\mathrm{g}}=\sum_{t=1}^{N}\gamma^{N-t}\left\| \boldsymbol{f}_{\mathrm{g}}^{\mathrm{GT}}-\boldsymbol{f}_{\mathrm{g}}^{t}\right\|_1 \tag{18}$$

式中，t表示迭代优化次数；N表示迭代总次数；VSGU(·)[26](Visual Similarity Guided Upsampling)表示视觉相似引导的光流上采样网络，作用是将低分辨率光流$\boldsymbol{f}_{\mathrm{LR}}$上采样至原始分辨率$\boldsymbol{f}_{\mathrm{g}}$；$\boldsymbol{f}_{\mathrm{g}}^{\mathrm{GT}}$表示原始分辨率下光流真实值；$\gamma$表示每轮迭代优化过程中损失的权重系数，本文设置为0.85.

通过利用局部光流损失函数、全局匹配损失函数、全局光流损失函数构建总损失函数，在不同阶段约束整个网络模型的学习：

$$L_{\mathrm{total}}=\alpha L_{\mathrm{Local}}+\beta L_{\mathrm{m}}+L_{\mathrm{g}} \tag{19}$$

式中，α、β分别为局部光流损失与全局匹配损失的权重系数，本文分别设置为$\alpha=0.05$、$\beta=0.01$.

4 实验

4.1 数据集及评价指标

为了对提出的模型进行性能分析，本文选用光流估计领域通用的MPI-Sintel[27]和KITTI-2015[28](Karlsruhe Institute of Technology and Toyota technological Institute-2015)作为测试数据集对所提方法进行验证评估.

MPI-Sintel数据集分为Clean和Final这2个子数据集，它们都是由动画电影《Sintel》中包含的多种类型运动场景片段组成. 其中Clean子数据集包含了大位移、弱纹理、光照变化和遮挡等运动类型；Final子数据集是通过添加运动模糊、大气散射、焦散效果和景深变化等渲染效果，使场景更接近真实情况. 它们对算法性能的挑战性较大. 通常使用平均端点误差(Average End-Point-Error，AEPE)进行量化，该误差指标从整体性能上评估算法的准确性和鲁棒性. 此外，MPI-Sintel数据集官方还提供更加细化的$d_{0\text{-}10}$指标评估光流估计算法预测的光流与真实光流之间的误差在10像素或更少的范围内的像素百分比，用于反映算法在处理细节部分的精确度. $d_{10\text{-}60}$和$d_{60\text{-}140}$指标描述预测误差在10~60和60~140像素之间的像素百分比，用于评价算法在处理中等难度和复杂场景时的性能. $s_{0\text{-}10}$指标评估速度小于每帧10个像素的区域端点误差，$s_{10\text{-}40}$、s_{40+}指标评估速度在每帧10~40个像素和大于40个像素的区域端点误差，用于反映算法应对大位移运动的能力.

KITTI-2015数据集是一个专门用于训练和评估光流估计算法在真实场景下性能表现的数据集，包含动态对象(如移动的车辆、行人和自行车)以及静态场景元素(如建筑物、道路基础设施和植被). 这些场景具有真实的大位移运动、纹理变化、复杂的光照和自然遮挡等运动类型，为评估模型在实际应用中的性能提供了重要基准. 此外，该数据集更为关注光流估计的准确性而非精确的位移值，因此对于该数据集的量化评价通常使用数据集官方推荐的F_1-all(%)作为评价指标，它表示端点误差大于3个像素的光流预测值占比. 与MPI-Sintel数据集类似，KITTI官方也提供了更为细化的F_1-bg和F_1-fg用于评估方法在背景和前景区域的误差程度. 此外，需要说明的是KITTI官方并未使用AEPE(Average EndPoint Error)评价指标用于评估不同算法在KITTI-2015数据集上的性能，但为了与MPI-Sintel数据集指标统一，本文仍然采用AEPE指标用于辅助评价，原因在于AEPE在一定程度反映算法在KITTI-2015真实场景的性能.

4.2 模型训练和测试设置

遵循文献[13]中的基本参数设置，本文采用PyTorch框架和AdamW优化器在NVIDIA A100上进行混合精度训练. 模型权重采用了随机初始化策略，并使用了单周期线性学习率(One Cycle Learning Rate，One CycleLR). 为了防止梯度爆炸，提高训练的稳定性，本文在反向传播过程中执行梯度裁剪并将梯度的最大值设置为1.

其中，训练分为2个阶段：首先本文使用FlyingChairs[11]和FlyingThings[29]合成数据集对模型进行预训练，然后分别使用MPI-Sintel[27]、KITTI-2015[28]、HD1K[30](High-resolution Dataset with 1K image pairs)和VKITTI2[31]数据集对模型进行微调并在MPI-Sintel和KITTI-2015数据集上进行测试. 具体来说，本文在训练和测试阶段所使用的数据量如表1所示.

表1 模型训练与测试数据集数据量统计 单位：幅

阶段	Flying Chairs[11]	Flying Things[29]	MPI-Sintel[27]	KITTI-2015[28]	HD1K[30]	VKITTI2[31]
训练	22 232	96 336	2 082	200	1 047	42 420
验证	640	—	1 041	200	—	—
测试	—	—	1 104	200	—	—

此外,MPI-Sintel和KITTI-2015数据集中训练集和测试划分,都采用数据集提供方预设的划分方案,且该划分方法被广泛采用如PWC-Net+[12]、RAFT[13]、GMA[14]、GMFlow[25]等. 具体而言,MPI-Sintel的训练集包含23个序列片段(共2 082幅图像),测试集包含12个序列片段(1 104幅图像),每个序列都包含Clean和Final这2个版本,且划分是基于完整序列进行的,即同一序列不会同时出现在训练集和测试集中. KITTI-2015则包含200对训练用图像序列和200对测试用序列,仅训练集提供带有真实标签的光流数据. 同时,这2个数据集都是在线评测系统,测试集的真实标签均不公开,研究者需要将算法结果提交到各自的官方评测服务器进行性能评估. 这种机制可以确保评测的公平性,防止算法过拟合测试集.

4.3 MPI-Sintel数据集实验

首先,表2统计了本文方法和PWC-Net+[12](Pyramid Warping and Cost volume Network)、RAFT[13](Recurrent All-pairs Field Transforms for optical flow)、SeperableFlow[23]、GMA[14](Global Motion Aggregation)、GMFlow[25]、RAFT-OCTC[32](Occlusion Consistency and Transformation Consistency version of RAFT)、GMFlowNet[33](Global Matching Flow Network)、CRAFT[34](CRoss-Attentional Flow Transformer for robust optical flow)、AGFlow[35](Adaptive Graph reasoning for optical Flow)、Shu[26]等10种对比方法在MPI-Sintel训练集(train)和测试集(test)上的光流估计误差AEPE统计结果. 其中,PWC-Net+、RAFT和GMA为当前流行的深度学习光流估计基准,PWC-Net+采用特征金字塔与成本体积策略主要是为了解决大位移运动光流估计问题;RAFT采用循环递归策略在提升光流估计整体精度(如大位移、弱纹理区域等)的同时有效保护了边界;GMA方法将光流估计为特征匹配任务,通过构建全局匹配运动聚合模块有效改善了遮挡区域的光流估计准确性问题;其他对比方法如RAFT-OCTC、CRAFT以RAFT的循环递归策略为基准延伸而来,RAFT-OCTC是一种半监督光流估计方法,主要解决光流训练数据不够充分的问题,CRAFT通过引入语义平滑转换层和跨帧注意力机制有效缓解匹配歧义引起的大位移运动模糊问题. GMFlowNet、AGFlow和Shu则均是采用与GMA相似的匹配策略,通过利用局部场景内容和空间关系信息来缓解由像素特征匹配困难造成的误差. SeperableFlow采用与PWC-Net+相类似的特征金字塔和成本体积策略,通过将2D光流问题分解为2个独立的1D问题并引入非局部聚合层学习更精确的代价体,提高光流估计的计算效率和准确性.

表2中MPI-Sintel(train)是使用MPI-Sintel提供的2 082幅训练图像得到的AEPE误差结果,MPI-Sintel(test)是使用MPI-Sintel官网提供的1 104幅测试图像得到的光流误差AEPE结果. 从表2可以看出,本文方法在Clean和Final的训练集和测试集上平均端点误差均取得最优的估计精度. 证明本文方法光流估计具有较高的准确性和鲁棒性.

表2　MPI-Sintel数据集实验对比结果

对比方法	MPI-Sintel(train)		MPI-Sintel(test)	
	Clean	Final	Clean	Final
PWC-Net+[12]	1.71	2.34	3.45	4.60
RAFT[13]	0.76	1.22	1.61	2.86
SeperableFlow[23]	0.69	1.10	1.50	2.67
GMA[14]	0.62	1.06	<u>1.39</u>	2.47
GMFlow[25]	—	—	1.74	2.90
RAFT-OCTC[32]	0.73	1.23	1.82	3.09
GMFlowNet[33]	0.59	0.91	<u>1.39</u>	2.65
CRAFT[34]	0.60	1.06	1.45	2.42
AGFlow[35]	0.65	1.07	1.43	2.47
Shu[26]	<u>0.55</u>	<u>0.90</u>	1.50	<u>2.36</u>
本文方法	**0.52**	**0.76**	**1.28**	**2.25**

注:粗体数据表示最优结果,下划线数据表示次优结果.

进一步,表3分别统计了本文方法与对比方法在MPI-Sintel提供的1 104幅测试图像上的$d_{0\text{-}10}$、$d_{10\text{-}60}$、$d_{60\text{-}140}$和$s_{0\text{-}10}$、$s_{10\text{-}40}$、s_{40+}指标误差结果统计. 从表3可以看出,本文方法在$d_{10\text{-}60}$和$d_{60\text{-}140}$指标均取得了具有竞争力的光流估计精度,说明本文方法针对复杂场景光流估计具有较好效果. 尽管$d_{0\text{-}10}$指标本文方法未取得最优,但数值上仅略低于GMFlow. 在$s_{10\text{-}40}$和s_{40+}指标上本文方法同样也获得了最低的误差结果,说明本文方法相对其他对比方法能够准确地估计出大位移运动区域的光流结果.

最后,为了更直观地展示本文方法在不同区域下的光流估计效果,图5展示了MPI-Sintel测试集PERTURBED_market_3、temple_1和tiger的光流结果可视化效果. 其中图5(a)和图5(b)中的红色方框为弱纹理区域和大位移区域,图5(c)中红色方框区域既包含弱纹理,又存在大位移运动. 图5(d)~图5(f)为红色方框区域对应放大后的可视化结果.

从图5(a)可以看出,PWC-Net+方法没有估计出弱纹理区域的栅栏,RAFT虽估计出栅栏区域的部分形状,但存在较为严重的模糊现象,没有很好地预测出栅栏的轮廓和其中结构细节. GMFlow对于栅栏的细节纹理估计效果较差,GMA方法与RAFT估计效果相似. 本文所提方法较好地估计出了栅栏轮廓和内部空间结构,整体可视化效果更清晰. 针对大位移运动光流估计,从图5(b)中可以看出,对于飞龙所在的大位移区

表3 MPI-Sintel测试集细化类别指标误差统计结果

对比方法	MPI-Sintel Clean(test)						MPI-Sintel Final(test)					
	$d_{0\text{-}10}$	$d_{10\text{-}60}$	$d_{60\text{-}140}$	$s_{0\text{-}10}$	$s_{10\text{-}40}$	s_{40+}	$d_{0\text{-}10}$	$d_{10\text{-}60}$	$d_{60\text{-}140}$	$s_{0\text{-}10}$	$s_{10\text{-}40}$	s_{40+}
PWC-Net+[12]	3.91	1.25	0.49	0.75	2.23	19.85	4.78	2.05	1.23	0.95	2.98	26.62
RAFT[13]	1.62	0.52	0.30	0.34	1.04	9.29	3.11	1.13	0.77	0.63	1.82	16.37
SeperableFlow[23]	1.47	0.48	0.26	0.31	0.96	8.69	2.94	1.06	0.62	0.58	1.74	15.93
GMA[14]	1.54	0.46	0.28	0.33	0.96	7.66	2.86	1.06	0.65	0.57	1.82	13.49
GMFlow[25]	**1.23**	0.57	0.46	0.50	0.97	9.72	**2.49**	1.04	0.88	0.71	1.66	16.75
RAFT-OCTC[32]	1.46	0.44	**0.24**	**0.30**	0.94	8.12	0.88	1.05	0.67	0.58	1.70	14.59
GMFlowNet[33]	1.28	0.40	0.29	0.31	1.00	7.70	2.82	1.05	0.78	0.70	1.78	14.42
CRAFT[34]	1.57	0.55	0.25	0.31	0.99	8.13	2.84	1.01	0.55	**0.54**	1.62	13.66
AGFlow[35]	1.50	0.45	0.26	0.32	0.96	8.08	2.89	0.99	0.70	0.56	1.69	13.82
Shu[26]	—	—	—	—	—	—	2.75	0.97	0.55	0.55	1.57	13.49
本文方法	1.24	**0.39**	0.25	0.33	**0.84**	**7.05**	2.56	**0.84**	**0.48**	**0.54**	**1.54**	**12.44**

注:粗体数据表示最优结果,下划线数据表示次优结果.

(*a*) PERTURBED_market_3序列结果(弱纹理) (*b*) temple_1序列结果(大位移) (*c*) tiger序列结果(同时包含大位移和弱纹理) (*d*) 图(*a*)中红色矩形框区域放大图 (*e*) 图(*b*)中红色矩形框区域放大图 (*f*) 图(*c*)中红色矩形框区域放大图

图5 所提方法在弱纹理、大位移运动区域的光流估计可视化效果对比

域,所列举的对比方法估计准确度较差. 例如,RAFT和GMFlow并没有很好地估计出左侧翅膀的完整形状. 而本文方法光流估计结果更加贴近真实值. 最后,在具有挑战性的图5(*c*)区域,本文提出的方法对于人物腿部和狮子区域的光流估计精度较高. 结合上述结果表明,本文方法对于大位移和弱纹理区域的光流估计具有较

好的准确性和鲁棒性.

4.4 KITTI-2015数据集实验结果

为了验证模型在真实场景下的光流估计精度,本文在KITTI-2015测试集上对本文方法和对比方法进行详细比较. 由于表2和表3中部分方法并未提供在KITTI-2015测试集上的误差结果,因此,本文选择PWC-Net+[12]、RAFT[13]、GMA[14]、GMFlow[25]、CRAFT[34]、AGFlow[35]等6种方法作为对比方法,原因在于前3种算法为当前流行的基准模型,后3种方法均是在该3种基准模型上进一步拓展衍生,具有一定代表性. 误差统计结果如表4所示,其中,KITTI-2015(train)是使用KITTI-2015提供的200幅训练图像得到的AEPE和F_1-all误差结果. KITTI-2015(test)是使用KITTI-2015官网提供的200幅测试图像得到的光流误差结果.

根据表4中的结果,本文提出方法在KITTI-2015数据集的训练集和测试集F_1-all指标均领先所有对比方法,尽管在训练集本文方法优势较小,但通过AEPE指标辅助对比,可以看到本文方法仍然取得较好的光流估计结果. 表4还统计KITTI-2015测试集部分在F_1-bg和F_1-fg的指标统计结果,其中,F_1-bg和F_1-fg是仅针对背景和前景区域计算的异常值百分比. 通过这些指标,可以更加全面地评估光流估计在真实世界场景中的鲁棒性和准确性. 从表4可以看出,本文方法相较于其他对比方法实现了真实场景下的最优光流估计精度. 与同样使用全局匹配方法GMFlow[23]相比,本文的方法分别在前景、背景和所有区域的光流估计精度分别提升了55%、23%和51%. 这说明本文所提出的方法在真实复杂场景下对光流估计的有效性.

表4　KITTI-2015数据集实验结果

对比方法	KITTI-2015(train)		KITTI-2015(test)		
	AEPE	F_1-all/%	F_1-bg/%	F_1-fg/%	F_1-all/%
PWC-Net+[12]	1.50	5.3	7.69	7.88	7.72
RAFT[13]	0.63	1.5	4.74	6.87	5.10
GMA[14]	<u>0.57</u>	<u>1.2</u>	4.78	7.03	5.15
GMFlow[25]	—	—	9.67	7.57	9.32
CRAFT[34]	0.58	1.3	4.58	<u>5.85</u>	<u>4.79</u>
AGFlow[35]	0.58	<u>1.2</u>	<u>4.52</u>	6.75	4.89
本文方法	**0.53**	**1.1**	**4.35**	**5.84**	**4.60**

注:粗体数据表示最优结果,下划线数据表示次优结果.

为了进一步验证所提方法在大位移、弱纹理场景区域的光流估计效果,本文基于KITTI-2015训练集(KITTI-2015测试集由KITTI官网提供整个数据集的平均结果,无法更为细化地选取具体运动场景,故此处使用训练集作为实验数据支撑),选取了具有代表性的大位移运动场景序列(000008、000039、000050)、弱纹理场景运动序列(000000、000006、000007)和同时包含大位移与弱纹理运动场景(000128、000089、000067),以验证所提方法的有效性. 实验结果统计如表5所示(由于部分算法没有提供源代码和权重,因此表5选择RAFT、GMA、GMFlow和CRAFT作为对比方法). 从表5可以看出,本文所提方法在大位移、弱纹理和同时包含大位移与弱纹理的运动场景AEPE和$F_{1\text{-all}}$误差指标最低,这进一步验证了所提方法针对大位移、弱纹理区域的光流估计具有较好效果. 尽管在000050序列本文方法并未取得最优,但是数据结果仅略低于对比方法GMA.

表5　KITTI-2015不同运动场景光流估计误差统计

运动类型	序列	RAFT[13]		GMA[14]		GMFlowNet[33]		CRAFT[34]		本文方法	
		AEPE	F_1-all/%	AEPE	F_1-all/%	AEPE	F_1-all/%	AEPE	F_1-all/%	AEPE	F_1-all/%
大位移	000008	0.29	0.84	0.27	0.83	0.28	0.87	0.27	0.77	**0.25**	**0.70**
	000039	0.74	1.66	0.72	1.50	0.76	1.90	0.69	1.33	**0.68**	**1.25**
	000050	0.05	0.19	**0.04**	**0.14**	0.05	0.18	0.05	0.19	0.05	0.15
	平均值	0.36	0.90	0.34	0.82	0.36	0.98	0.34	0.76	**0.32**	**0.70**
弱纹理	000000	0.63	1.08	0.59	1.19	0.64	1.08	0.61	1.03	**0.56**	**0.84**
	000006	0.50	0.63	0.52	0.48	0.62	0.75	0.50	0.55	**0.45**	**0.48**
	000007	0.35	0.86	0.33	0.80	0.34	0.76	0.33	0.76	**0.30**	**0.61**
	平均值	0.49	0.86	0.48	0.82	0.53	0.86	0.48	0.78	**0.44**	**0.64**
大位移+弱纹理	000128	0.50	0.50	0.63	1.01	0.51	0.58	0.53	0.64	**0.45**	**0.29**
	000089	0.10	0.23	**0.07**	0.15	0.09	0.58	0.07	0.13	**0.07**	**0.12**
	000067	2.59	10.79	2.13	8.98	2.59	11.43	2.19	9.49	**1.91**	**7.66**
	平均值	1.06	3.84	0.94	3.38	1.06	4.20	0.93	3.42	**0.81**	**2.69**

注:粗体数据表示最优结果.

为了定性对比各方法的光流估计效果,图6以000050、000000、000067序列为例展示了各对比方法在上述运动区域的光流估计可视化图. 从图6可以看出,本文所提方法取得了最佳估计效果,例如大位移运动

区域，本文方法估计的汽车轮廓更为贴近真实值，其他对比方法存在明显的边缘扩张．在弱纹理区域，本文方法较好地预测出标签区域的形状，而对比方法如RAFT、GMA和CRAFT存在明显的误差．

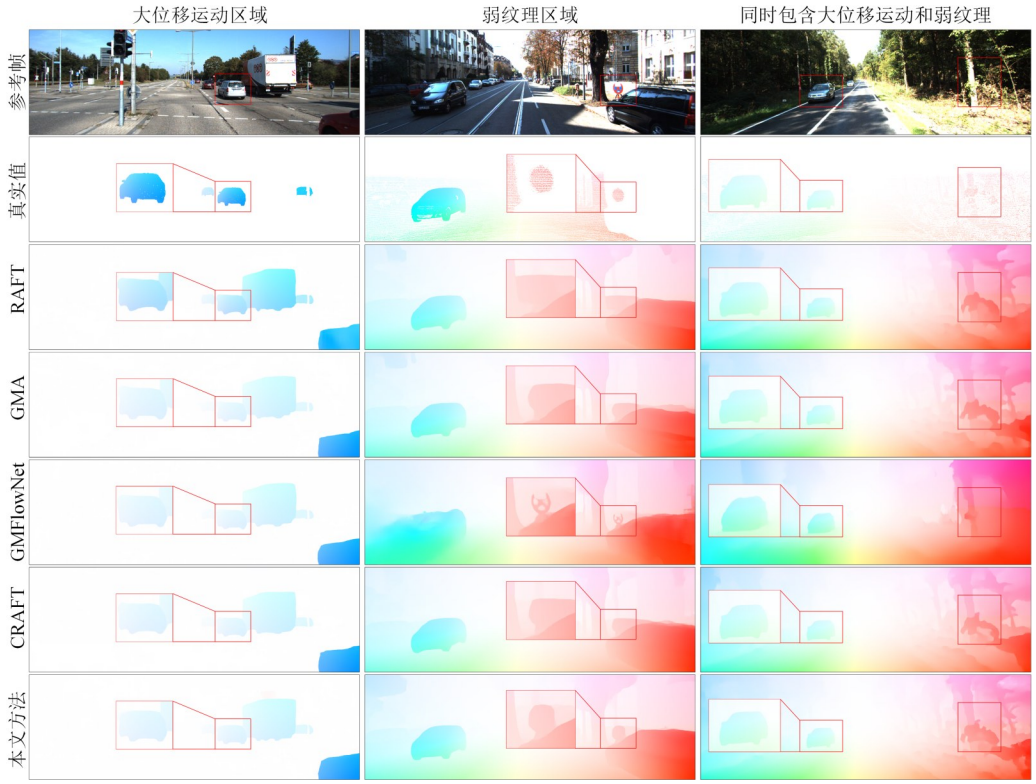


图6 KITTI-2015不同运动场景光流估计可视化效果对比

4.5 消融实验

为了分析本文所提各模块对所提方法的作用，本文在MPI-Sintel数据集上进行了消融实验．为了合理比较模块之间的作用，所有模型都将使用相同的训练步骤和策略．即先在FlyingChairs和FlyingThings3D数据集上进行预训练，然后使用MPI-Sintel数据集进行对比和分析．

本文的基线模型为GMA算法，其使用全局损失函数(L_g)约束网络训练，用base表示．表6统计了本文所提各个模块之间的消融实验统计结果．其中，JDM(Joint Depth-separable residual and Multi-scale dual-channel attention)表示联合深度可分离残差与多尺度双通道注意力的深度特征编码模块，GM表示基于可学习的全局匹配优化模块，L_{Local}表示局部光流损失，L_{m}表示全局匹配损失．

从表6可以看出，每个网络组件的加入都能提升光流计算的整体精度．例如，加入联合深度可分离残差与多尺度双通道注意力的深度特征编码模块JDM后，模型②由于获得更准确的图像序列特征，在Clean和Final序列上均取得了5%的提升．然后，在引入基于可学习的全局匹配优化模块GM(Global Matching optimization module)后，得益于准确的初始运动光流信息，模型③相比于基线模型获得了5%的性能提升．损失函数的L_{Local}和L_{m}的分别引入，在一定程度提升了模型的光流估计性能(见模型④和⑤)．在全模型⑥下光流计算的精度最高，相比于基线模型，在MPI-Sintel训练集的Clean和Final序列上分别取得了19.2%和14.2%的提升．

表6 消融实验对比结果

序号	模型	MPI-Sintel(train)	
		Clean	Final
①	Base	1.30	2.74
②	Base+JDM	1.24	2.61
③	Base+JDM+GM	1.22	2.54
④	Base+JDM+GM +L_{Local}	1.16	2.42
⑤	Base+JDM+GM +L_{m}	1.10	2.45
⑥	Base+JDM+GM +L_{m}+L_{Local}	**1.05**	**2.35**

注：粗体数据表示最优结果．

4.6 复杂性分析

为了分析所提算法模型的复杂度，本文参考文献[36, 37]中的分析方案，引入FLOPs(FLoating point OPerations, FLOPs)指标评估网络复杂性并用参数量和推理时间辅助分析，更加综合、客观地对所提方法复杂性进行评价．其中FLOPs表示浮点运算次数是衡量计算复杂度

的通用指标,数值越大表明模型的计算复杂度越高.

首先,本文使用NVIDIA RTX3090 GPU测试MPI-Sintel训练集来分析所提方法不同模块对模型算法复杂性的影响,实验结果如表7所示. 从表7可以看出,在FLOPs指标方面,本文提出的JDM模块有效降低了基线模型的计算复杂度,原因在于深度可分离卷积可以将标准卷积分解为2个更小的深度卷积和逐点卷积操作,减少模型的计算复杂度. 所提的GM模块因为需要进行稠密的匹配任务,增加了基线模型的计算复杂度. 全模型主要得益于JDM模块在降低复杂度方面的优势,使相对于基线模型在计算复杂度方面仍具有明显下降. 在参数量和推理时间方面,由于本文方法是在基线模型基础上的进一步深化和改进,因此,JDM和GM模块的额外引入一定程度增加模型的参数量和推理时间,但增加幅度相对较低是可接受的.

本文又统计了所提模型与表2中7种对比方法在FLOPs、参数量和推理时间的复杂度对比(由于RAFT-OCTC、AGFlow和SeperableFlow原论文未提供FLOPs、参数量以及推理时间3个指标中的至少2个指标数据,故未统计),结果如表8所示. 从表8可以看出,在FLOPs指标PWC-Net+取得了最低的计算复杂度结果,但光流性能相对较差. 与其他方法相比,本文方法FLOPs指标取得了最优,说明本文方法的计算复杂度相对较低. 此外,GMFlowNet在参数量和推理时间指标方面数值最大,本文方法虽然通过使用深度可分离卷积减少了FLOPs,但参数量相对较大导致推理时间增加. 然而,与算法在测试数据集的指标提升相比,本文方法相对其他方法在模型计算复杂度和提高光流估计性能之间取得了良好的权衡.

表7　不同模块对算法模型复杂度的影响

评价指标	模型			
	基线模型	+JDM	+GM	全模型
FLOPs/G	798	740	806	775
参数量/M	5.9	6.3	6.5	7.0
推理时间/ms	150	171	160	182

表8　本文方法与对比方法的复杂度指标统计结果

评价指标	PWC-Net+[12]	RAFT[13]	GMA[14]	GMFlowNet[33]	GMFlow[25]	CRAFT[34]	Shu[26]	本文方法
FLOPs/G	180	805	798	780	1 109	814	822	775
参数量/M	8.75	5.3	5.9	9.3	4.7	6.3	6.76	7.0
推理时间/ms	30	110	150	200	120	177	—	182

4.7　模型通用性分析

本文遵循文献[38]中所提的通用性分析方案,额外引入了Middlebury光流基准数据集[39]对所提方法和RAFT、GMA、GMFlowNet、CRAFT等4种对比方法进行模型通用性对比分析. 其中,Middlebury数据集主要包含的是室内运动场景,样本数量较少(共12组测试图像序列),主要用于测试深度学习算法在小样本环境下的泛化性能. 实验结果如表9所示,需要注意的是,本文并没有在Middlebury数据集上微调所提出光流估计模型,而是直接使用Middlebury数据集进行测试. 从表9可以看出,本文所提方法取得了最佳的光流误差估计精度,该结果有力地验证了所提方法具有较强的通用性和泛化能力.

表9　Middlebury数据集光流估计误差统计

评价指标	RAFT[13]	GMA[14]	GMFlowNet[33]	CRAFT[34]	本文方法
AEPE	0.695	0.635	0.644	0.621	**0.611**

注:粗体数据表示最优结果.

5　结论

本文提出一种联合深度可分离残差与多尺度双通道注意力的全局匹配优化光流估计方法. 首先,利用深度可分离残差块与基于多尺度双通道注意力的Transformer模块,构建深度特征提取模块,在平衡参数量与运算速度的同时获取连续帧间更为准确且全面的深度特征信息. 然后,针对局部匹配歧义问题,构建一种基于可学习的全局匹配优化光流估计策略,通过排除遮挡并高效利用全局匹配信息,有效缓解因匹配歧义引起的光流估计运动模糊. 最后,为了提高模型的训练稳定性与泛化能力,本文提出一种联合全局与局部的光流损失函数,约束模型训练. 通过丰富的实验与对比分析,证明了本文所提出方法的有效性,特别是在大位移运动和弱纹理区域. 在未来的研究工作中,将引入教师-学生网络对所提模型进行知识蒸馏,以提升模型对遮挡和剧烈光照变化场景的光流估计能力,同时通过剪枝等方式进一步提高模型的推理速度.

参考文献

[1] 柯逍, 缪欣, 郭文忠. 基于时空交叉感知的实时动作检测方法[J]. 电子学报, 2024, 52(2): 574-588.
KE X, MIAO X, GUO W Z. Real-time action detection based on spatio-temporal interaction perception[J]. Acta Electronica Sinica, 2024, 52(2): 574-588. (in Chinese)

[2] 王正文, 宋慧慧, 樊佳庆, 等. 基于语义引导特征聚合的

显著性目标检测网络[J]. 自动化学报, 2023, 49(11): 2386-2395.
WANG Z W, SONG H H, FAN J Q, et al. Semantic guided feature aggregation network for salient object detection[J]. Acta Automatica Sinica, 2023, 49(11): 2386-2395. (in Chinese)

[3] 杨鑫, 杨春玲. 基于MAP的多信息流梯度更新与聚合视频压缩感知重构算法[J]. 电子学报, 2023, 51(11): 3320-3330.
YANG X, YANG C L. MAP-based multi-information flow gradient update and aggregation for video compressed sensing reconstruction[J]. Acta Electronica Sinica, 2023, 51(11): 3320-3330. (in Chinese)

[4] 李公平, 陆耀, 王子建, 等. 基于模糊核估计的图像盲超分辨率神经网络[J]. 自动化学报, 2023, 49(10): 2109-2121.
LI G P, LU Y, WANG Z J, et al. Blurred image blind super-resolution network *via* kernel estimation[J]. Acta Automatica Sinica, 2023, 49(10): 2109-2121. (in Chinese)

[5] ZHENG Z H, NIE N, LING Z, et al. DIP: Deep inverse patchmatch for high-resolution optical flow[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2022: 8915-8924.

[6] 张聪炫, 史世栋, 葛利跃, 等. 基于遮挡优化的金字塔块匹配光流估计方法[J]. 电子学报, 2023, 51(9): 2539-2548.
ZHANG C X, SHI S D, GE L Y, et al. Pyramid patch-matching optical flow estimation method based on occlusion optimization[J]. Acta Electronica Sinica, 2023, 51(9): 2539-2548. (in Chinese)

[7] ZHAI M L, XIANG X Z, LV N, et al. Optical flow and scene flow estimation: A survey[J]. Pattern Recognition, 2021, 114: 107861.

[8] 江峤, 陈震, 危水根, 等. 基于结构张量的变分光流计算方法[J]. 南昌航空大学学报(自然科学版), 2011, 25(2): 48-53.
JIANG D, CHEN Z, WEI S G, et al. A variational calculation for optical flow based on structure tensor[J]. Journal of Nanchang Hangkong University (Natural Sciences), 2011, 25(2): 48-53. (in Chinese)

[9] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. ImageNet classification with deep convolutional neural networks[J]. Communications of the ACM, 2017, 60(6): 84-90.

[10] 范兵兵, 何庭建, 张聪炫, 等. 联合遮挡约束与残差补偿的特征金字塔光流计算方法[J]. 电子学报, 2023, 51(3): 648-657.
FAN B B, HE T J, ZHANG C X, et al. Feature pyramid optical flow estimation method jointing occlusion constraint and residual compensation[J]. Acta Electronica Sinica, 2023, 51(3): 648-657. (in Chinese)

[11] DOSOVITSKIY A, FISCHER P, ILG E, et al. FlowNet: Learning optical flow with convolutional networks[C]// 2015 IEEE International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2015: 2758-2766.

[12] SUN D Q, YANG X D, LIU M Y, et al. Models matter, so does training: An empirical study of CNNs for optical flow estimation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 42(6): 1408-1423.

[13] TEED Z, DENG J. RAFT: Recurrent all-pairs field transforms for optical flow[M]//Computer Vision-ECCV 2020. Cham: Springer International Publishing, 2020: 402-419.

[14] JIANG S H, CAMPBELL D, LU Y, et al. Learning to estimate hidden motions with global motion aggregation[C]// 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2021: 9752-9761.

[15] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2016: 770-778.

[16] SZEGEDY C, LIU W, JIA Y Q, et al. Going deeper with convolutions[C]//2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2015: 1-9.

[17] ILG E, MAYER N, SAIKIA T, et al. FlowNet 2.0: Evolution of optical flow estimation with deep networks[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2017: 1647-1655.

[18] RANJAN A, BLACK M J. Optical flow estimation using a spatial pyramid network[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2017: 2720-2729.

[19] HUI T W, TANG X O, LOY C C. LiteFlowNet: A lightweight convolutional neural network for optical flow estimation[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 8981-8989.

[20] HUR J, ROTH S. Iterative residual refinement for joint optical flow and occlusion estimation[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recog-

nition (CVPR). Piscataway: IEEE, 2019: 5747-5756.

[21] YANG G S, DEVA R. Volumetric correspondence networks for optical flow[J]. Neural Information Processing Systems, 2019, 1: 545-554.

[22] ZHAO S Y, SHENG Y L, DONG Y, et al. MaskFlownet: Asymmetric feature matching with learnable occlusion mask[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2020: 6277-6286.

[23] ZHANG F H, WOODFORD O J, PRISACARIU V, et al. Separable flow: Learning motion cost volumes for optical flow estimation[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2021: 10787-10797.

[24] SUN S K, CHEN Y Q, ZHU Y, et al. Skflow: Learning optical flow with super kernels[C]//Advances in Neural Information Processing Systems. Red Hook: Curran Associates, 2022: 11313-11326.

[25] XU H F, ZHANG J, CAI J F, et al. GMFlow: Learning optical flow via global matching[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2022: 8111-8120.

[26] 舒铭奕, 张聪炫, 陈震, 等. 基于局部-全局建模与视觉相似引导的光流估计方法[J]. 中国科学: 信息科学, 2023, 53(10): 1945-1964.
SHU M Y, ZHANG C X, CHEN Z, et al. Optical flow estimation based on local-global modeling and visual similarity guidance[J]. Scientia Sinica (Informationis), 2023, 53(10): 1945-1964. (in Chinese)

[27] BUTLER D J, WULFF J, STANLEY G B, et al. A naturalistic open source movie for optical flow evaluation[M]// Computer Vision-ECCV 2012. Berlin: Springer Berlin Heidelberg, 2012: 611-625.

[28] MENZE M, GEIGER A. Object scene flow for autonomous vehicles[C]//2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2015: 3061-3070.

[29] MAYER N, ILG E, HÄUSSER P, et al. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2016: 4040-4048.

[30] KONDERMANN D, NAIR R, HONAUER K, et al. The HCI benchmark suite: Stereo and flow ground truth with uncertainties for urban autonomous driving[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Piscataway: IEEE, 2016: 19-28.

[31] TANG X, YANG M, SUN P H, et al. PaReNeRF: Toward fast large-scale dynamic NeRF with patch-based reference[C]//2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2024: 5428-5438.

[32] JEONG J, LIN J M, PORIKLI F, et al. Imposing consistency for optical flow estimation[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2022: 3171-3181.

[33] ZHAO S Y, ZHAO L, ZHANG Z X, et al. Global matching with overlapping attention for optical flow estimation[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2022: 17571-17580.

[34] SUI X C, LI S H, GENG X, et al. CRAFT: Cross-attentional flow transformer for robust optical flow[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2022: 17581-17590.

[35] LUO A, YANG F, LUO K, et al. Learning optical flow with adaptive graph reasoning[C]//Proceedings of the AAAI Conference on Artificial Intelligence(AAAI). Menlo Park: AAAI, 2022, 36(2): 1890-1898.

[36] CHENG R, HE R A, JIANG X H, et al. Context-aware iteration policy network for efficient optical flow estimation[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2024, 38(2): 1299-1307.

[37] FENG M J, JIA H, YAN Z Q, et al. APCAFlow: All-pairs cost volume aggregation for optical flow estimation[J]. IEEE Transactions on Multimedia, 2024, 26: 9060-9069.

[38] WANG H, FAN R, LIU M. CoT-AMFlow: Adaptive modulation network with co-teaching strategy for unsupervised optical flow estimation[C]//Conference on Robot Learning (CoRL). New York: PMLR, 2021: 143-155.

[39] BAKER S, SCHARSTEIN D, LEWIS J P, et al. A database and evaluation methodology for optical flow[J]. International Journal of Computer Vision, 2011, 92(1): 1-31.

作者简介

王子旭 男,1999 年 8 月生,河南洛阳人. 西北工业大学博士研究生. 主要研究方向为图像检测与智能识别.
E-mail: wangzixu0827@163.com

张聪炫 男,1984 年 7 月生,河南焦作人. 南昌航空大学教授. 主要研究方向为图像处理与计算机视觉.
E-mail: zcxdsg@163.com

陈弘烨 男,2000 年 3 月生,浙江杭州人. 南昌航空大学硕士研究生. 主要研究方向为光流估计.
E-mail: 2308080400005@stu.nchu.edu.cn

陈 震 男,1969 年 11 月生,江西九江人. 南昌航空大学教授. 主要研究方向为图像理解与测量.
E-mail: dr_chenzhen@163.com

葛利跃 男,1992 年 10 月生,安徽蚌埠人. 实验师,北京航空航天大学博士研究生. 主要研究方向为机器视觉与智能感知.
E-mail: lygeah@163.com

王梓歌 女,1998 年 4 月生,河北晋州人. 南昌航空大学助理实验师. 主要研究方向为计算机视觉.
E-mail: Wangzggg@163.com