外观融合运动感知的运动目标分割算法

徐邦武，吴秦，周浩杰

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.221200153

外观融合运动感知的运动目标分割算法

徐邦武[1]　吴　秦[1,2]　周浩杰[1]
1 江南大学人工智能与计算机学院　江苏 无锡 214122
2 江苏省模式识别与计算智能工程实验室　江苏 无锡 214122
(1658576022@qq.com)

摘　要　现实场景中的运动目标分割旨在分割当前场景下的运动物体,对于许多计算机视觉应用有着至关重要的作用。现有的运动目标分割算法大多通过2D光流图中的运动信息来分割运动物体,然而,这些方法还存在一些问题。当运动物体在极面内运动或者其3D运动方向和背景一致时,很难通过光流图分割得到;另外,错误的光流预测也会影响分割的结果。为了解决以上问题,提出了不同的运动代价,以提升运动目标分割的正确率。针对和背景共线或共面运动的物体,设计均衡重投影代价和多角度光流对比代价,通过运动物体的2D光流与背景2D光流的差异来检测运动物体。针对自我运动退化,设计差异单应性代价。最后,提出了一种基于外观融合的运动感知结构,以分割各种场景下的运动物体。采用多模态共同注意力门控,更有效地捕获运动特征和外观特征的关系,以促进外观特征和运动特征更好地交互。此外,为了突出运动的物体,提出了多层运动注意力模块,以减少冗余的外观特征对结果的影响。实验结果表明,所提方法在KITTI,JNU-UISEE,KittiMoSeg和Davis-2016数据集上均能获得较优的运动目标分割结果。

关键词:运动目标分割;均衡重投影代价;多角度光流对比代价;多模态共同注意力门控;多层运动注意力模块
中图分类号　TP391.413

Appearance Fusion Based Motion-aware Architecture for Moving Object Segmentation

XU Bangwu[1],WU Qin[1,2] and ZHOU Haojie[1]
1 School of Artificial Intelligence and Computer Science,Jiangnan University,Wuxi,Jiangsu 214122,China
2 Jiangsu Provincial Engineering Laboratory of Pattern Recognition and Computational Intelligence,Wuxi,Jiangsu 214122,China

Abstract　Moving object segmentation aims to segment all moving objects in the current scene,and it is of critical significance for many computer vision applications. At present,many moving object segmentation methods use the motion information from 2D optical flow maps to segment moving objects,which have many defects. For moving objects moving in the epipolar plane or moving objects whose 3D motion direction are consistent with the background,it is difficult to identify these objects by the 2D optical flow maps. Besides,incorrect 2D optical flow also effects the result of moving object segmentation. To solve the above problems,this paper proposes different motion costs to improve the performance of moving object segmentation. In order to detect moving objects with coplanar and collinear motion,this paper proposes a balanced reprojection cost and a multi-angle optical flow contrast cost,which measures the difference between the 2D optical flow of moving objects and that of the background. For ego-motion degeneracy,this paper designs a differential homography cost. To segment moving objects in complex scenes,this paper proposes an appearance fusion based motion-aware architecture. In this architecture,in order to effectively fuse appearance features and motion features of objects,the multi-modality co-attention gate is adapted to achieve better interaction between appearance and motion cues. Besides,to emphasize moving objects,this paper introduces a multi-level motion based attention module to suppress redundant and misleading information. Extensive experiments are conducted on the KITTI dataset,the JNU-UISEE dataset,the KittiMoSeg dataset and the Davis-2016 dataset,and the proposed method achieves excellent performance.

Keywords　Moving object segmentation,Balanced reprojection cost,Multi-angle optical flow contrast cost,Multi-modality co-attention gate,Multi-level motion based attention module

到稿日期:2022-12-27　返修日期:2023-06-05
基金项目:国家自然科学基金(61972180)
This work was supported by the National Natural Science Foundation of China(61972180).
通信作者:周浩杰(zhouhaojie@jiangnan.edu.cn)

1 引言

运动目标分割旨在分割给定视频中的运动物体，是计算机视觉任务中的一项关键的研究课题。在静态环境下，很容易检测出运动的物体，但在相机运动的复杂场景下，例如在自动驾驶的场景下，为了保证汽车安全地行驶，自动驾驶系统需要检测当前环境下的所有运动物体，运动目标的检测则变得十分复杂。目前，主流的自动驾驶系统主要依靠实例分割或目标检测的方法来保障汽车行驶的安全。然而，这些方法仅仅依靠物体的外观特征，不能判断物体是否运动，也不能预测动态场景下的运动物体的运动趋势。

传统的运动目标分割方法主要依靠几何方法或粒子匹配方法来分割运动目标[1-3]。这些方法在一些简单的场景下已经取得了显著的效果，但在复杂场景下的运动目标分割效果不佳，例如在共面运动退化或共线运动退化的场景下，传统的方法难以分割沿极线运动的物体。近年来，随着深度学习的快速发展，很多基于卷积神经网络(Convolutional Neural Networks，CNNs)的运动目标分割算法被提出。Dave 等[4]融合物体的外观信息和运动信息来分割运动物体。Tokmakov 等[5]提出了一个双流网络，分别从 RGB 原图和光流图中提取物体的外观特征和运动特征，并采用一个 ConvGRU 模块以获得更好的分割结果。然而，这些方法都是从 2D 光流图中提取物体的运动特征，难以分割与背景 2D 运动相似的运动物体。Yang 等[6]提出了一种能在共线运动场景下分割运动物体的方法。但是，该方法需要使用单目深度估计算法来估计图像的深度，而单目深度估计器预测的深度的好坏将直接影响该方法的运动目标分割效果。

为了能够分割各种场景下的运动物体，本文提出了不同的运动代价。为了检测共面运动和共线运动的运动物体，提出了均衡重投影代价和多角度光流对比代价，通过几何方法计算物体的 2D 运动和背景的 2D 运动的差异。针对自我运动退化的场景，设计了差异单应性代价。此外，本文还提出了一种基于外观融合的运动感知结构，以有效地分割运动物体。为了更有效地融合物体的外观特征和运动特征，采用多模态共同注意力门控(Multi-Modality Co-Attention Gate，MCG)[7]来建立物体外观特征和运动特征之间的联系，并通过自适应的方法调整外观特征和运动特征的权重，以提高网络对有用特征的关注。同时，考虑到由于错误的运动估计，浅层次的运动特征存在较多的运动噪声，本文提出了一种多层运动融合模块(Multi-level Motion Fusion Module，MMFM)，以促进浅层次运动特征和深层次运动特征的交互，并丰富浅层次运动特征的语义信息。最后，为了突出运动物体，并抑制显著的静态物体，设计了一个基于运动的注意力模块(Motion based Attention Module，MBAM)，利用物体的运动信息来减少干扰的外观信息对分割结果的影响。多层运动融合模块和基于运动的注意力模块组成了本文提出的多层运动注意力模块(Multi-level Motion based Attention Module，MMBAM)。本文的主要贡献如下：

(1)针对不同场景下运动目标分割的难点，提出了不同的运动代价，如均衡重投影代价、多角度光流对比代价和差异单应性代价等，以更有效地分割各种场景下的运动物体。

(2)提出了基于外观融合的运动感知结构，结合多模态共同注意力门控，更有效地提取并融合运动物体的外观特征和运动特征，提高网络对有用信息的关注。

(3)提出了多层运动注意力模块，以丰富浅层次运动特征的语义信息，并结合物体的运动信息，以减少无用的外观特征对分割结果的影响。

(4)在 KITTI，JNU-UISEE，KittiMoSeg 和 Davis-2016 数据集上的大量实验都表明了本文方法的有效性。

2 相关工作

2.1 运动目标分割

传统的运动目标分割算法主要依靠几何方法分割视频中的运动目标[1,3,8]。Zhu 等[9]提出了一种时间连续性约束的低秩分解背景更新模型来检测视频中的运动目标。这些方法在简单场景下效果显著。但是，在复杂场景下难以获得较好的效果，例如在共面运动或共线运动的场景下，传统的方法难以分割在极面内运动的物体。Zhang 等[10]采用无线传感器和质心定位算法来确定运动目标的位置。近年来，随着深度学习在计算机视觉领域内取得成功，一些基于深度学习的运动目标分割算法被提出[4-6,11-13]。Dave 等[4]提出了一个双分支的网络，分别从 RGB 原图和光流图中提取外观特征和运动特征，以融合外观特征和运动特征来分割运动物体。Tokmakov 等[5]结合一个双流网络和一个 ConvGRU 模块，以获得更好的运动目标分割结果。但是，这些方法都是从光流图中提取运动特征，对光流图中错误的光流估计敏感，也难以分割沿极线运动的物体。Yang 等[6]提出深度对比代价用于检测共线运动的物体。然而，深度对比代价并非从运动的角度来解决问题，且依赖单目深度估计器估计的深度，错误的深度估计将会影响分割的结果。本文从物体运动的特性出发，设计能够在各种场景下检测运动物体的方法，并巧妙地融合运动物体的外观特征与运动特征，以获得更好的结果。

2.2 无监督的视频目标分割

视频目标分割的主要目标是分割视频帧中的显著物体，很多无监督的视频目标分割方法结合物体的外观特征和运动特征来分割物体[14-16]。FusionSeg[14]建立一个双流网络来分别处理外观特征和运动特征。然而，这些方法仅仅依靠物体的 2D 光流，没有考虑物体运动的几何特性，也忽略了光流图中预测错误的光流。本文则从物体的运动特性出发，利用几何方法设计不同的运动代价以处理不同运动退化的场景，如均衡重投影代价和多角度光流对比代价等。此外，本文还提出多层运动融合模块来融合不同层次的运动特征，以丰富浅层次运动特征的语义信息。

2.3 运动外观特征交互

目前很多运动目标分割和视频目标分割方法都采用运动特征与外观特征融合的方式来取得更优的效果。早期的方法直接通过按通道级联或逐元素相加等方式来融合物体的外观

特征和运动特征[17-18]，但光流图中存在很多无用和误导的信息，所以早期的融合方式会限制分割的准确性。也有一些方法提出单向交互策略[19-21]，即建立一个基于运动的注意力策略来增强外观特征，或者通过外观特征来增强运动特征。但是，这些方法也存在一些问题，例如通过运动来增强外观特征，会使增强后的特征太过依赖通过光流图提取的运动特征。在这种情况下，错误的光流预测将会影响最终结果。近年来，一些将运动特征与外观特征自适应融合的方法被提出[7,22]。FS-Net[22]采用一种全双工策略来进行双向信息传输。Yang等提出的AMC-Net[7]中添加了多模态共同注意力门控，以自适应地促进外观特征和运动特征的有效融合。本文在编码阶段采用多模态控制门控来融合物体的外观特征和运动特征。与AMC-Net不同的是，本文并非简单地从2D光流图中提取运动特征，以与外观特征进行信息交互，而是利用均衡重投影代价和多角度光流对比代价等运动代价来获得更准确的运动信息，以更好地分割各种场景下的运动物体。

3 外观融合运动感知模型

本文提出的基于外观融合的运动感知结构如图1所示，主要包含4个部分：代价图和运动特征生成模块(Cost Map and Motion Feature Generation Module，CMMFGM)、编码器、多层运动注意力模块和解码器。其中，代价图和运动特征生成模块用于计算不同的运动代价(均衡重投影代价、多角度光流对比代价、差异单应性代价和桑普森代价)，以分割不同场景下的运动物体。编码器用于提取不同层次的运动特征和外观特征，其结构如图2所示。编码阶段采用多模态共同注意力门控，以更有效地促进运动特征与外观特征的信息交互。多层运动注意力模块由多层运动融合模块和基于运动的注意力模块组成。多层运动融合模块用于融合不同层次的运动特征，以丰富浅层次特征的语义信息。基于运动的注意力模块用于减少冗余的外观信息对分割结果的影响，以进一步强调运动的物体。解码器则采用DLASeg[23]的解码结构，以获得运动物体的掩模。

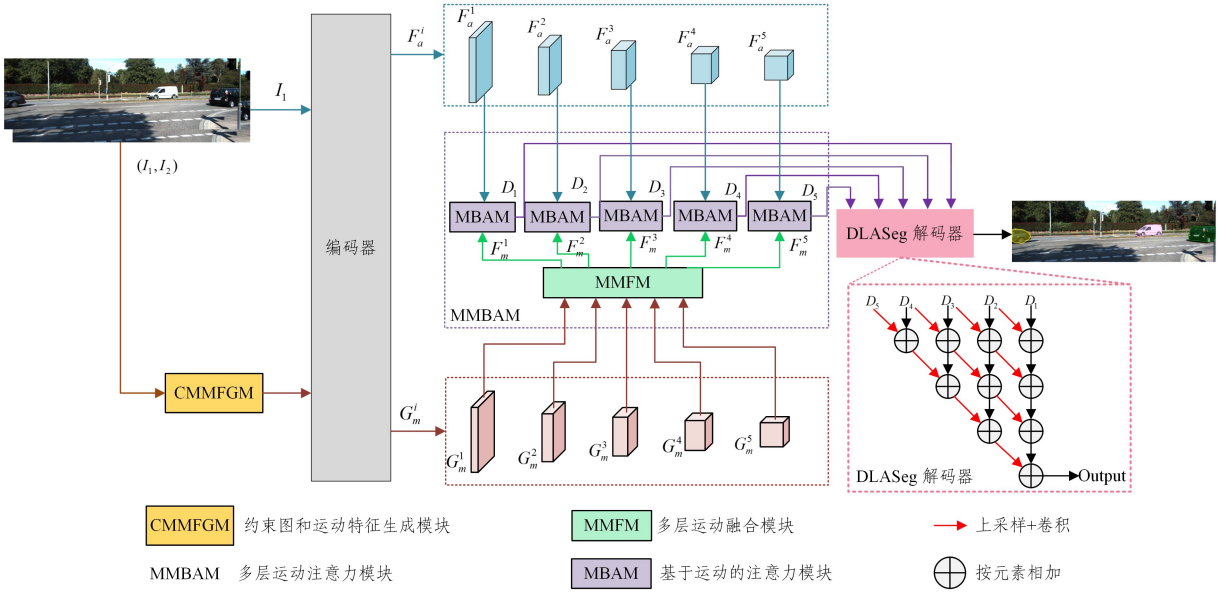


图1 基于外观融合的运动感知结构

Fig. 1 Appearance fusion based motion-aware architecture

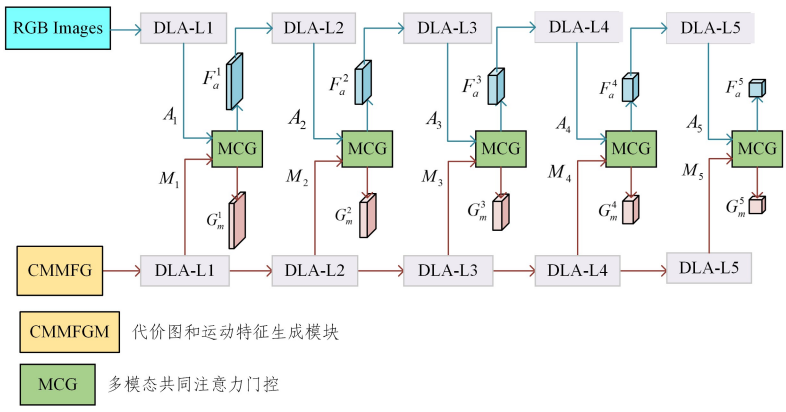


图2 编码器结构

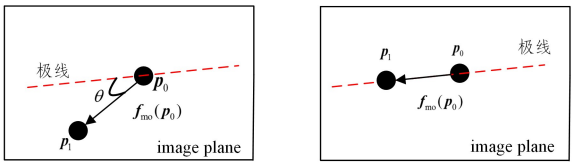
Fig. 2 Architecture of encoder

基于外观融合的运动感知结构的输入为连续的两帧RGB图片($\boldsymbol{I}_1$，$\boldsymbol{I}_2$)。$\boldsymbol{I}_1$和$\boldsymbol{I}_2$首先通过代价图和运动特征生成

模块，以生成不同的运动代价图和运动特征图。在编码阶段，使用两个 DLA102[24] 主干网络来分别提取物体的外观特征与运动特征。如图 2 所示，DLA-Li（$i=1,2,3,4,5$）为 DLA102 主干网络的第 i 层。首先，每一层的运动特征与外观特征经过一个多模态共同注意力门控[7]，以获得增强的外观特征 $\boldsymbol{F}_a^i$ 和运动特征 $\boldsymbol{G}_m^i$。然后，运动特征 $\boldsymbol{G}_m^i$ 通过一个多层运动融合模块，以生成增强的运动特征 $\boldsymbol{F}_m^i$。接着，$\boldsymbol{F}_a^i$ 和 $\boldsymbol{F}_m^i$ 经过一个基于运动的注意力模块，以获得有效的时空融合特征 $\boldsymbol{D}_i$。最后，$\boldsymbol{D}_i$ 被输入一个 DLASeg[23] 的解码器以获得运动物体的掩模。

3.1 均衡重投影代价和多角度光流对比代价

为了检测图像中的运动物体，传统的方法主要计算运动物体的 2D 光流与极线之间的角度偏差。如图 3(a)所示，对于运动的像素点 p_0，$\boldsymbol{f}_{\mathrm{mo}}(p_0)$ 为 p_0 的 2D 光流，p_1 为 p_0 在下一帧中的对应像素点，p_0 的 2D 光流与极线之间的角度偏差记为 θ。在共面运动退化或共线运动退化的场景下，p_0 沿着极线运动[6]，如图 3(b)所示。这种情况下，p_0 的 2D 运动方向与极线的方向一致，因此，计算光流与极线之间角度偏差的方法将会失效。



(a)正常运动 (b)运动退化

图 3 正常运动与运动退化

Fig. 3 Normal motion and motion degeneracy

为了解决共面运动退化和共线运动退化引发的运动目标分割问题，本文测量运动物体的 2D 运动与背景的 2D 运动的差异。对于任意运动的像素点 $p_0=(x_0,y_0)$，假设在下一帧中的对应像素点为 $p_1=(x_1,y_1)$，如果它的 2D 光流和背景光流不一致，则能够断定 p_0 是运动的。此时，p_0 满足式(1)：

$$\boldsymbol{f}_{\mathrm{mo}}(p_0)\neq\boldsymbol{f}_{\mathrm{bg}}(p_0) \tag{1}$$

其中，$\boldsymbol{f}_{\mathrm{mo}}(p_0)$ 为 p_0 的 2D 光流，$\boldsymbol{f}_{\mathrm{bg}}(p_0)$ 代表在 p_0 处的背景光流。参考文献[6]的做法，$\boldsymbol{f}_{\mathrm{bg}}(p_0)$ 能够通过 NG-RANSAC[25] 计算，且其尺度能够通过光学膨胀[26]和单目深度估计。根据以上的运动分析，对于运动的像素点 p_0，可以考虑通过它的 2D 运动与背景的 2D 运动之间的绝对差异检测，如式(2)所示，并定义式(2)为光流差异代价：

$$C_{\mathrm{gap}}=\|\boldsymbol{f}_{\mathrm{mo}}(p_0)-\boldsymbol{f}_{\mathrm{bg}}(p_0)\| \tag{2}$$

光流差异代价能够有效地检测检近处的运动物体。但是，对于具有相同 3D 运动的两个物体，远处的运动物体由于 2D 光流更小，因此比近处的运动物体更难通过光流差异代价检测。为此，考虑到近处的背景光流模长较长，远处的背景光流模长较短，针对远处且低速运动的物体，本文提出均衡重投影代价，利用背景光流的模长来调整远近物体的代价范围，均衡重投影代价的定义如式(3)所示：

$$C_{\mathrm{bre}}=\frac{\|\boldsymbol{f}_{\mathrm{mo}}(p_0)-\boldsymbol{f}_{\mathrm{bg}}(p_0)\|}{\|\boldsymbol{f}_{\mathrm{bg}}(p_0)\|+\varepsilon} \tag{3}$$

其中，ε 为一个大于 0 的常量，避免发生分母为 0 的情况；$\boldsymbol{f}_{\mathrm{mo}}(p_0)$ 由光流估计算法[27]估计。

同时，为了获得更丰富的运动特征，提高运动目标分割的准确率，可以考虑通过运动物体的 2D 运动与背景的 2D 运动的相对差异，构建光流相对差异代价用于提高运动物体的检测效果，如式(4)所示：

$$C_{r\mathrm{gap}}=\frac{\|\boldsymbol{f}_{\mathrm{mo}}(p_0)\|}{\|\boldsymbol{f}_{\mathrm{bg}}(p_0)\|+\varepsilon} \tag{4}$$

借助光流相对差异代价，能够检测不同距离的运动物体。但是，对于 2D 光流模长和背景一样的运动物体，即使其 2D 运动方向与背景的不一致，光流相对差异代价也会失效。为此，针对 2D 光流模长和背景一致但方向不同的运动物体，本文提出多角度光流对比代价，其定义如式(5)所示，该代价结合了运动物体的 2D 运动速度和方向信息。

$$C_{\mathrm{mofc}}=\left|\frac{\|\boldsymbol{f}_{\mathrm{mo}}(p_0)\|*(\cos\theta+1)}{\|\boldsymbol{f}_{\mathrm{bg}}(p_0)\|+\varepsilon}-2\right| \tag{5}$$

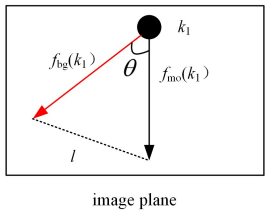


图 4 均衡重投影代价和多角度光流对比代价的对比

Fig. 4 Comparison between balanced reprojection cost and multi-angle optical flow contrast cost

对于任意像素点 p_0，如果它的 2D 光流模长与背景的不相等或 2D 运动方向和背景的不一致，则 p_0 能够通过多角度光流对比代价检测为运动的像素点。

为了分析均衡重投影代价和多角度光流对比代价的效果，本文引入图 4 中的案例。如图 4 所示，k_1 为运动的特征点，其 2D 光流为 $\boldsymbol{f}_{\mathrm{mo}}(k_1)$，$k_1$ 处的背景光流为 $\boldsymbol{f}_{\mathrm{bg}}(k_1)$，$\theta$ 为 $\boldsymbol{f}_{\mathrm{mo}}(k_1)$ 和 $\boldsymbol{f}_{\mathrm{bg}}(k_1)$ 之间的夹角，l 为 $\boldsymbol{f}_{\mathrm{mo}}(k_1)$ 和 $\boldsymbol{f}_{\mathrm{bg}}(k_1)$ 之间的欧氏距离。因此，k_1 的均衡重投影代价由式(6)表示：

$$C_{\mathrm{bre}}(\boldsymbol{k}_1)=\frac{\|\boldsymbol{f}_{\mathrm{mo}}(k_1)-\boldsymbol{f}_{\mathrm{bg}}(k_1)\|}{\|\boldsymbol{f}_{\mathrm{bg}}(k_1)\|}=\frac{l}{\|\boldsymbol{f}_{\mathrm{bg}}(k_1)\|} \tag{6}$$

同理，k_1 的多角度光流对比代价由式(7)表示：

$$C_{\mathrm{mofc}}(k_1)=\left|\frac{\|\boldsymbol{f}_{\mathrm{mo}}(k_1)\|(\cos\theta+1)}{\|\boldsymbol{f}_{\mathrm{bg}}(k_1)\|}-2\right| \tag{7}$$

当 k_1 的取值使得式(8)成立时，即 $C_{\mathrm{bre}}(k_1)<C_{\mathrm{mofc}}(k_1)$，比起均衡重投影代价，多角度光流对比代价的表现更好；反之，当 k_1 的取值使得式(9)成立时，即 $C_{\mathrm{bre}}(k_1)>C_{\mathrm{mofc}}(k_1)$，均衡重投影代价的表现更好。

$$l<\left|\left[\|\boldsymbol{f}_{\mathrm{mo}}(k_1)\|(\cos\theta+1)-2\|\boldsymbol{f}_{\mathrm{bg}}(k_1)\|\right]\right| \tag{8}$$

$$l>\left|\left[\|\boldsymbol{f}_{\mathrm{mo}}(k_1)\|(\cos\theta+1)-2\|\boldsymbol{f}_{\mathrm{bg}}(k_1)\|\right]\right| \tag{9}$$

此外，为了处理自我运动退化的场景，本文采用了 Yang 等[6]提出的单应性代价。单应性代价的定义如式(10)所示：

$$C_{\mathrm{hom}}=d(\tilde{\boldsymbol{p}}_0,\boldsymbol{H}_R\tilde{\boldsymbol{p}}_1)^2+d(\tilde{\boldsymbol{p}}_1,\boldsymbol{H}_R^{-1}\tilde{\boldsymbol{p}}_0)^2 \tag{10}$$

其中，$\tilde{\boldsymbol{p}}_0$ 和 $\tilde{\boldsymbol{p}}_1$ 为像素点 p_0 和 p_1 的齐次坐标，$d(\cdot,\cdot)$ 为两个像素点的欧几里得距离，$\boldsymbol{H}_R$ 为旋转单应矩阵。在相机快速移动的场景下，单应性代价容易将近处且静止的物体误判为运动物体。为了解决这一问题，本文提出差异单应性代价，旨在运用光流差异代价来削弱静止物体的显著特征。差异单应

性代价的定义如式(11)所示：

$$C_{\text{rhom}}=C_{\text{gap}}*C_{\text{hom}} \tag{11}$$

对于正常运动的像素点，本文计算每个像素点的桑普森误差[28]作为桑普森代价。桑普森代价的定义如式(12)所示：

$$C_{\text{sam}}=\frac{(\tilde{\boldsymbol{p}}_1^{\mathrm{T}}\boldsymbol{F}\tilde{\boldsymbol{p}}_0)^2}{(\boldsymbol{F}\tilde{\boldsymbol{p}}_0)_1^2+(\boldsymbol{F}\tilde{\boldsymbol{p}}_0)_2^2+(\boldsymbol{F}^{\mathrm{T}}\tilde{\boldsymbol{p}}_1)_1^2+(\boldsymbol{F}^{\mathrm{T}}\tilde{\boldsymbol{p}}_1)_2^2+\varepsilon} \tag{12}$$

其中，$\boldsymbol{F}$ 为相机的基础矩阵。各种运动代价图由代价图和运动特征生成模块进行计算，如图 5 所示。

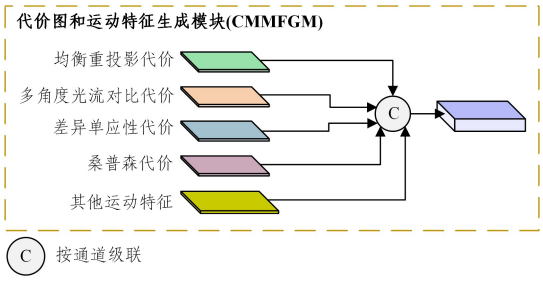


图 5 代价图和运动特征生成模块

Fig. 5 Cost map and motion feature generation module

最后，参照 Rigid-mask[6]，本文将第一帧图片的 3D 点、3D 光流、光流和光学膨胀的不确定性估计[25,27]合并为一个 8 通道的运动特征图，用于分割 2D 运动特征较弱的运动物体。如图 5 所示，运动特征图由代价图和运动特征生成模块负责生成。

3.2 多模态共同注意门控

有效的运动特征与外观特征的交互能够帮助网络同时关注到物体的外观信息和运动信息，最常用的方法是采用按通道级联或逐元素相加等方式实现运动特征与外观特征的融合。但是，这些方法往往忽略了错误的运动估计，同时也对图片中静止物体的显著外观敏感。为此，本文采用多模态共同注意门控来融合物体的运动特征与外观特征，以促进运动特征与外观特征之间的信息交互并获得有用的信息，其结构如图 6 所示。

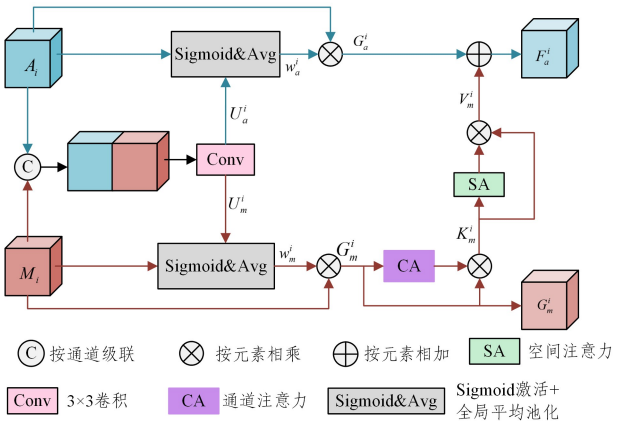


图 6 多模态共同注意力门控结构

Fig. 6 Architecture of multi-modality co-attention gate

多模态共同注意门控的输入为编码阶段第 $i(i=1,2,3,4,5)$ 层的外观特征 $\boldsymbol{A}_i$ 和运动特征 $\boldsymbol{M}_i$。其中 $\boldsymbol{A}_i$ 和 $\boldsymbol{M}_i$ 分别为外观特征提取分支和运动特征提取分支所得。为了提取 $\boldsymbol{A}_i$ 和 $\boldsymbol{M}_i$ 之间的相对关系，$\boldsymbol{A}_i$ 和 $\boldsymbol{M}_i$ 首先通过一个跨通道级联和一个卷积操作，以获得初步融合特征 $\boldsymbol{U}_i$。为了分别计算外观特征和运动特征的重要度，$\boldsymbol{U}_i$ 接着按通道均匀分成两部分，分别为 $\boldsymbol{U}_a^i$ 和 $\boldsymbol{U}_m^i$。同时，$\boldsymbol{U}_a^i$ 和 $\boldsymbol{U}_m^i$ 被依次输入到一个 Sigmoid 函数中进行标准化，并采用全局平均池化操作以获得运动特征权重 $\boldsymbol{w}_m^i$ 和外观特征权重 $\boldsymbol{w}_a^i$。整个过程如式(13)和式(14)所示：

$$(\boldsymbol{U}_a^i,\boldsymbol{U}_m^i)=split(Conv(Cat(\boldsymbol{A}_i,\boldsymbol{M}_i))) \tag{13}$$

$$(\boldsymbol{w}_a^i,\boldsymbol{w}_m^i)=Avg(\sigma(\boldsymbol{U}_a^i,\boldsymbol{U}_m^i)) \tag{14}$$

其中，Avg 为全局平均池化，$Conv$ 为卷积操作，Cat 为按通道级联操作，$split$ 为按通道均匀分成两部分，σ 为 Sigmoid 激活函数。最终，通过 $\boldsymbol{w}_a^i$ 和 $\boldsymbol{w}_m^i$ 获得加权后的外观特征 $\boldsymbol{G}_a^i$ 和运动特征 $\boldsymbol{G}_m^i$，如式(15)所示：

$$(\boldsymbol{G}_a^i,\boldsymbol{G}_m^i)=(\boldsymbol{w}_a^i*\boldsymbol{A}_i,\boldsymbol{w}_m^i*\boldsymbol{M}_i) \tag{15}$$

该过程旨在抑制干扰信息对结果的影响，并增强有用的信息。通过给运动特征和外观特征分配不同的权重，网络能够自适应地关注运动特征或外观特征，以减少误导信息对最终结果的影响。考虑到运动估计的不确定性，运动特征 $\boldsymbol{G}_m^i$ 接着经过通道注意力模块获得对每一个通道自适应的关注，以提高运动特征的表示能力，然后通过空间注意力模块，以提高对运动区域的关注，其具体过程如式(16)和式(17)所示：

$$\boldsymbol{K}_m^i=CA(\boldsymbol{G}_m^i)*\boldsymbol{G}_m^i \tag{16}$$

$$\boldsymbol{V}_m^i=SA(\boldsymbol{K}_m^i)*\boldsymbol{K}_m^i \tag{17}$$

其中，SA 为空间注意力，CA 为通道注意力。为了获得增强的外观特征，将 $\boldsymbol{G}_a^i$ 与 $\boldsymbol{V}_m^i$ 按逐元素相加的方式结合，以互补地聚合运动特征与外观特征，其具体过程如式(18)所示：

$$\boldsymbol{F}_a^i=\boldsymbol{G}_a^i+\boldsymbol{V}_m^i \tag{18}$$

最终，多模态共同注意力门控将输出增强后的外观特征 $\boldsymbol{F}_a^i$ 和加权后的运动特征 $\boldsymbol{G}_m^i$。其中，$\boldsymbol{F}_a^i$ 被输入到外观编码分支的下一个编码模块，以获得更高级的语义信息，$\boldsymbol{G}_m^i$ 被输入到多层运动融合模块，以与不同层次的运动特征进一步融合。

3.3 多层运动注意力模块

在编码阶段合理地利用多模态共同注意力门控，能够获得有效的时空特征。但是，多模态共同注意力门控融合了物体的外观信息和运动信息，一些静态物体由于具有显著的外观特征会被误判为运动物体。为此，本文提出了多层运动注意力模块。多层运动注意力模块由多层运动融合模块和基于运动的注意力模块组成，其中多层运动融合模块用于促进不同层次运动特征之间的信息交互，以丰富浅层次的运动特征的语义信息；基于运动的注意力模块则利用物体的运动信息来抑制显著的静态物体。

多层运动融合模块如图 7 所示，其输入为多模态共同注意力门控输出的运动特征 $\boldsymbol{G}_m^i$，对于前四层的运动特征 $\boldsymbol{G}_m^i$ $(i=1,2,3,4)$，该层以下的每一层运动特征经过上采样和卷积操作，以获得与 $\boldsymbol{G}_m^i$ 相同的尺寸。随后，将这些特征与 $\boldsymbol{G}_m^i$ 以按通道级联的方式融合，以充分融合不同层次的运动特征，并接上一个卷积操作得到增强的运动特征 $\boldsymbol{F}_m^i$。对于最深层次的运动特征 $\boldsymbol{G}_m^5$，通过一个卷积操作得到 $\boldsymbol{F}_m^5$，并保持其原有尺度不变。经过多层运动融合，$\boldsymbol{F}_m^i$ 比 $\boldsymbol{G}_m^i$ 有着更少的运动噪声以及更丰富的语义信息，这有利于使用物体的运动信息突显运动的物体。为了使用增强的运动特征 $\boldsymbol{F}_m^i$ 来强调运动的物体，本文提出了基于运动的注意力模块，其结构如图 8 所示。

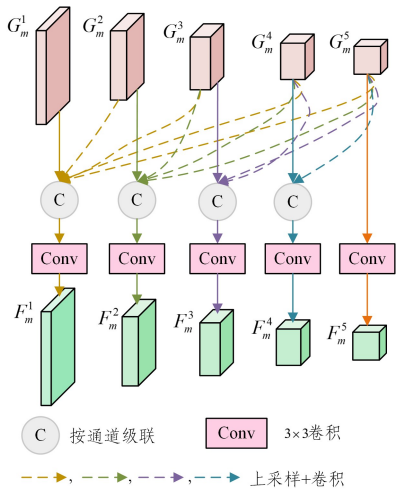


图 7 多层运动融合模块结构

Fig. 7 Architecture of multi-level motion fusion module

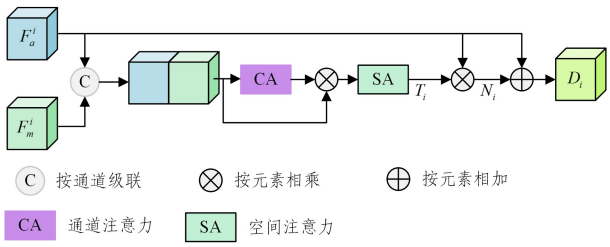


图 8 基于运动的注意力模块结构

Fig. 8 Architecture of motion based attention module

基于运动的注意力模块的输入为增强的外观特征 $\boldsymbol{F}_a^i$ 和增强的运动特征 $\boldsymbol{F}_m^i$。首先将 $\boldsymbol{F}_a^i$ 和 $\boldsymbol{F}_m^i$ 以按通道级联的方法结合，以促进增强运动特征和增强外观特征的交互，并获得时空融合特征 $\boldsymbol{F}_i$。接着采用通道注意力以加强网络对 $\boldsymbol{F}_i$ 的重要通道的关注，同时，使用空间注意力以捕捉运动的区域并加强关注，其具体过程如式(19)和式(20)所示：

$$\boldsymbol{F}_i = Cat(\boldsymbol{F}_a^i, \boldsymbol{F}_m^i) \tag{19}$$

$$\boldsymbol{T}_i = SA(CA(\boldsymbol{F}_i) * \boldsymbol{F}_i) \tag{20}$$

$\boldsymbol{T}_i$ 即通过基于运动的注意力模块得出的运动注意力图，用于有效地过滤背景噪声，以促进运动物体的特征学习。随后，使用逐元素相乘的方式将运动注意力图 $\boldsymbol{T}_i$ 与增强的外观特征 $\boldsymbol{F}_a^i$ 结合得到中间特征图 $\boldsymbol{N}_i$，以抑制 $\boldsymbol{F}_a^i$ 中显著的静态物体的信息。再将 $\boldsymbol{N}_i$ 与 $\boldsymbol{F}_a^i$ 按逐元素相加的方式融合，以防止有效信息丢失，其具体过程如式(21)所示：

$$\boldsymbol{D}_i = \boldsymbol{F}_a^i + \boldsymbol{T}_i * \boldsymbol{F}_a^i \tag{21}$$

多层运动融合模块和基于运动的注意力模块组成了本文的多层运动注意力模块。最终，多层运动注意力模块输出有效的时空融合特征 $\boldsymbol{D}_i$，$\boldsymbol{D}_i$ 将被输入到一个 DLA-Seg 的解码模块以获得运动物体的掩模。

3.4 损失函数

在训练过程中，运动注意力图 $\boldsymbol{T}_i$ 用于重点突出运动的区域。为获得准确的运动注意力图，使用二值化后的标签图 $\boldsymbol{B}$ 来对运动注意力图进行指导。对于 $\boldsymbol{T}_i$ 和 $\boldsymbol{B}$ 使用二值交叉熵损失(Binary CrossEntropy Loss)，如式(22)所示：

$$L_b^i = \frac{-1}{H * W} \sum_{x=0}^{W-1} \sum_{y=0}^{H-1} [\boldsymbol{B}(x,y) \log \boldsymbol{T}_i(x,y) + (1-\boldsymbol{B}(x,y)) \log(1-\boldsymbol{T}_i(x,y))] \tag{22}$$

其中，W 和 H 分别为图像的宽度和高度，$B(x,y)$ 和 $T_i(x,y)$ 分别为在像素点 (x,y) 处的二值化标签图 B 和运动注意力图 T_i 的值。

此外，对于任意运动物体，使用 focal loss[23] 和 polar loss[29]，其中 focal loss 由式(23)计算：

$$L_f = \frac{-1}{H * W} \sum_{x=0}^{W-1} \sum_{y=0}^{H-1} \begin{cases} (1-\boldsymbol{P}(x,y))^{\mu} \log \boldsymbol{P}(x,y), & \text{if } \boldsymbol{G}(x,y)=1 \\ (1-\boldsymbol{G}(x,y))^{\eta} \boldsymbol{P}(x,y)^{\mu} \log(1-\boldsymbol{P}(x,y)), & \text{otherwise} \end{cases} \tag{23}$$

其中，μ 和 η 均为超参数，$\boldsymbol{G}(x,y)$ 和 $\boldsymbol{P}(x,y)$ 分别为像素点 (x,y) 处的标签目标中心点热力图和预测的目标中心点热力图的值。遵循文献[23]的做法，μ 和 η 分别设置为 2 和 4。为计算 polar loss，将运动物体的标签轮廓转成极坐标，该极坐标由从物体中心点均匀出发的 K 条射线与物体的轮廓共同决定。polar loss 由式(24)计算：

$$L_p = \frac{1}{K * M} \sum_{m=1}^{M} \sum_{k=1}^{K} |d_{m,k} - d_{m,k}^*| \tag{24}$$

其中，M 代表运动物体的数量，对于第 m 个运动物体，$d_{m,k}^*$ 表示其第 k 条射线的标签长度，$d_{m,k}$ 表示其第 k 条射线的预测长度。最终，整体的训练损失如式(25)所示：

$$L = \beta_1 \sum_{i=1}^{5} L_b^i + \beta_2 L_f + \beta_3 L_p \tag{25}$$

与 Rigid-mask[6] 一样，超参数 β_1，β_2 和 β_3 分别设置为 1×10^{-4}，1×10^{-3} 和 1×10^{-7}。

4 实验与分析

本章首先介绍使用的数据集和实验细节，然后可视化一些运动代价图的样例，并分析不同场景下不同运动代价的效果。最后，在 KITTI[30]，KittiMoSeg[31]，JNU-UISEE 和 Davis-2016[32] 数据集上评估本文提出的方法。此外，本文还通过消融实验来验证每一个模块的有效性。

4.1 数据集

本文在 5 个数据集上进行实验，分别为 Scene Flow Datasets(SFD)[33]，KITTI[30]，KittiMoSeg[31]，JNU-UISEE 和 Davis-2016[32]。SFD 是一个合成的数据集，主要当作训练集来使用。SFD 由 FlyingThings3D，Driving 和 Monkaa 这 3 部分组成，一共包含 39 049 帧，提供相机的参数、每一帧的深度、2D 光流、3D 光流和目标的掩模。KITTI 是一个街道驾驶的数据集，由 800 张图片和 200 个场景组成，每个场景都存在多个运动的车辆，并且也提供相机参数、2D 光流、视差图和运动目标的掩模。JNU-UISEE 是现实场景下的自动驾驶数据集，由江南大学和驭势科技共同收集和标注。JNU-UISEE 包含 366 帧自动驾驶场景下的图片，每一帧都有运动物体，同时也提供运动物体的掩模和相机的参数。KittiMoSeg 由 Siam 等[31]提供，包含 1 649 张自动驾驶场景下的图片，其中 1 300 张作为训练集，349 张作为测试集。KittiMoSeg 中包含多种运动物体，如汽车、行人等，它提供运动目标的掩模、2D 光流和运动物体的边界框。Davis-2016 由 50 个视频序列组成，共

包含 3455 帧，其中 2079 帧作为训练集，1376 帧作为测试集。其运动目标的种类繁多，包括汽车、狗、行人、船只等。

4.2 实验细节

本文使用 Pytorch 框架实现所提出的方法。为了公平比较，遵循文献[6]的做法，首先在 SFD 数据集上进行训练，并在 KITTI，JNU-UISEE 和 KittiMoSeg 上进行测试。在数据增强阶段，加入水平翻转和随机裁剪。模型采用 Adam 优化器，初始学习率设置为 1×10^{-3}，Batch size 设置为 12。对于 Davis-2016 数据集，由于该数据集没有提供相机的参数，因此设置图片的中心点为相机的主点，相机的焦距设置为两倍的图片宽度。参照文献[7]，在 Davis-2016 的训练集上进行训练，使用 Davis-2016 的测试集来评估本文方法。

4.3 运动代价图的定性分析

为了进一步分析均衡重投影代价的作用，图 9 展示了均衡重投影代价和光流差异代价在两个不同场景下的可视化对比。

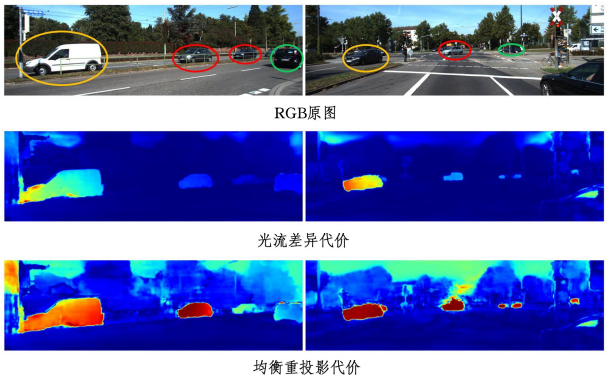


图 9 光流差异代价与均衡重投影代价的对比(电子版为彩图)

Fig. 9 Comparison between optical flow difference cost and balanced reprojection cost

如图 9 所示，正常运动的物体用绿色圈标注，近处共面运动的物体用黄色圈标注，远处共面运动的物体用红色圈标注。对于近处的运动物体，即位于黄色圈内的运动物体，光流差异代价和均衡重投影代价均取得了较优的性能。但是，对于远处的运动物体，如位于红色圈内的运动物体，由于其 2D 光流模长较小，光流差异代价也相对较小。因此，远处的运动物体比近处的运动物体更难通过光流差异代价检测。为此，均衡重投影代价利用远处较小的背景的 2D 光流，以提高远处运动物体的辨识度。相比光流差异代价，均衡重投影代价能更有效地检测远距离的运动物体。

为进一步分析多角度光流对比代价的效果，图 10 展示了光流相对差异代价与多角度光流对比代价在两个不同场景下的可视化对比。如图 10 所示，对于红色圈标注的运动物体，其 2D 光流模长与背景的 2D 光流模长相等，但 2D 运动方向和背景的 2D 运动不一致。由于光流相对差异代价只考虑运动物体的 2D 光流模长信息，因此，光流相对差异代价难以检测红色圈内的运动物体。多角度光流对比代价结合运动物体的 2D 运动速度和方向信息，能够有效地检测红色圈内的运动物体，如图 10 右侧所示。

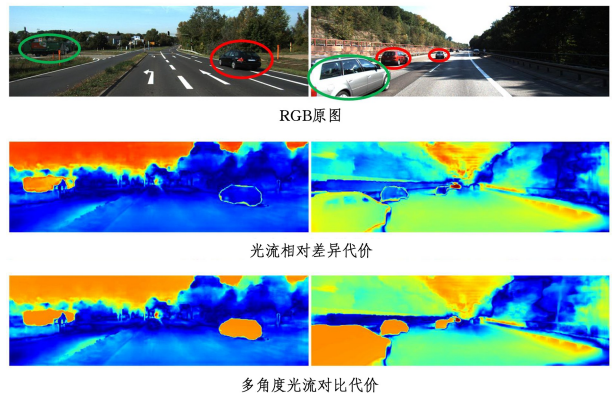


图 10 光流相对差异代价与多角度光流对比代价的对比(电子版为彩图)

Fig. 10 Comparison between optical flow relative difference cost and multi-angular optical flow contrast cost

4.4 实验结果对比分析

4.4.1 在 KITTI 数据集上的实验结果

参照文献[6]，本文在 KITTI 数据集上使用 F-measure (F)[34]和交并比(IoU)作为评价指标。对于运动物体(obj)，计算 F-measure；对于背景区域(bg)，计算 IoU。将本文提出的方法同 5 种比较先进的方法进行对比，对比的方法包括 MAT-Net[16]，COSNet[15]，CC-Motion[35]，AMC-Net[7]和 Rigid-mask[6]。表 1 列出了本文方法与 5 种对比方法在 KITTI 数据集上的对比结果。可以看出，本文方法在各项指标上都获得了最好的结果。相比现有最好的方法 Rigid-mask，本文方法在 F-measure 和 IoU 指标上分别提升了 2.73% 和 1.19%。

表 1 在 KITTI 数据集上的运动分割结果

Table 1 Motion segmentation results on KITTI dataset

Method	*F*(obj)	*IoU*(bg)
MAT-Net[16]	68.40	93.08
COSNet[15]	66.67	93.03
CC-Motion[35]	42.99	74.06
AMC-Net[7]	77.00	96.32
Rigid-mask[6]	90.71	97.05
ours	**93.44**	**98.24**

4.4.2 在 JNU-UISEE 数据集上的实验结果

在 JNU-UISEE 数据集上，本文也采用运动物体的 F-measure 和背景的 IoU 来评估本文方法。如表 2 所列，在 JNU-UISEE 数据集上，本文方法也取得了最好的结果。相比 Rigid-mask，本文方法在 F-measure 指标上提高了 3.53%，在 IoU 指标上提高了 0.19%。

表 2 在 JNU-UISEE 数据集上的运动分割结果

Table 2 Motion segmentation on JNU-UISEE dataset

Method	*F*(obj)	*IoU*(bg)
AMC-Net[7]	70.47	99.18
Rigid-mask[6]	85.79	99.05
ours	**89.32**	**99.24**

4.4.3 在 KittiMoSeg 数据集上的实验结果

在 KittiMoSeg 数据集上，本文方法的各项指标都获得了最优的结果。如表 3 所列，相比 Rigid-mask，本文方法的

F-measure 提高了 8.49%，IoU 指标提高了 0.49%。

表 3 在 KittiMoSeg 数据集上的运动分割结果

Table 3 Motion segmentation results on KittiMoSeg dataset

Method	*F*(obj)	*IoU*(bg)
MODNet-Separate[31]	54.25	—
MODNet-Joint[31]	62.46	—
U2-ONet[36]	64.23	—
AMC-Net[7]	50.07	95.21
Rigid-mask[6]	67.14	97.71
ours	**75.63**	**98.20**

注："—"表示该方法没有采用背景的 IoU 来评估。

4.4.4 在 Davis-2016 数据集上的实验结果

在 Davis-2016 数据集上，参照以往的方法 AMC-Net[7]，采用 Mean J，Mean F 和 J&F 这 3 个指标来评价本文的方法。

如表 4 所列，本文方法的综合评价指标最优，相比 AMC-Net，Mean F 和 J&F 分别提高了 1.4%和 0.6%。

表 4 在 Davis-2016 数据集上的运动分割结果

Table 4 Motion segmentation results on Davis-2016 dataset

Method	Mean *J*	Mean *F*	*J*&*F*
Rigid-mask[6]	67.4	66.0	66.7
FSNet[22]	83.4	83.1	83.3
F2Net[37]	83.1	84.4	83.7
AMC-Net[7]	**84.5**	84.6	84.6
COSNet[38]	81.1	79.7	80.4
FAMINet[39]	82.4	83.4	82.9
BMVOS[40]	82.9	81.4	82.2
ours	84.4	**86.0**	**85.2**

为了直观地展示本文提出的方法的效果，图 11 展示了本文方法与其他先进方法的可视化对比，从左到右依次为标签图、本文方法的预测结果、Rigid-mask 方法的结果、AMC-Net 方法的结果。

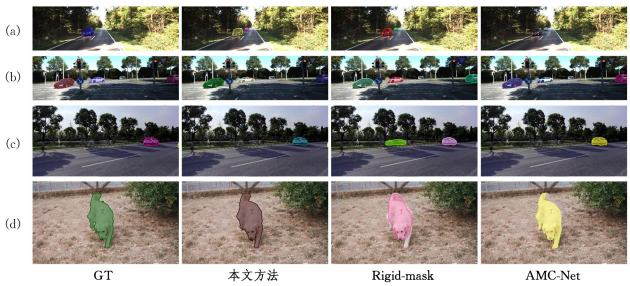


图 11 本文方法与其他先进方法的可视化对比

Fig. 11 Visual comparison between the proposed method and other advanced methods

图 11(a)和图 11(b)展示的是共线运动退化和共面运动退化场景下的运动目标分割结果，AMC-Net 由于没有考虑到共面运动退化和共线运动退化的情况，因此难以分割一些共线运动或共面运动的物体。得益于本文提出的均衡重投影代价和多角度光流对比代价，本文方法能够有效地分割共面运动和共线运动的物体。由图 11(c)可以看出，由于 Rigid-mask 没有对运动物体的错误运动估计做细化处理，因此，受到错误运动估计的影响，Rigid-mask 将一些背景的区域误判为运动区域。得益于多层运动融合模块，本文方法能够很好地区分运动区域和背景区域。由图 11(d)可以看出，由于 Rigid-mask 没有融合物体的外观信息和运动信息，对一些非刚性运动物体的分割效果不佳。得益于多模态共同注意力门控，本文提出的方法结合了物体的外观信息和运动信息，能够更好地分割非刚性运动物体。

4.5 消融实验结果与分析

为了验证每个模块的有效性，本文在数据集 Davis-2016 上进行了一系列的消融实验。首先，本文只采用按通道级联的方法来融合运动物体的外观特征和运动特征，并将融合的特征直接输入解码模块进行结果的预测。接着使用多模态共同注意力门控 MCG 来促进物体的外观特征与运动特征更好地交互，以获得更有效的时空融合特征。如表 5 第二行所列，引入多模块共同注意门控 MCG 之后，得益于更有效的时空特征，运动物体的分割结果也有显著的提升，J&F 指标由 79.2%提升到 80.7%。随后，加入基于运动的注意力模块 MBAM，以抑制静止物体的显著外观特征。如表 5 第三行所列，实验结果得到进一步的提升，J&F 指标提高了 1.8%。最后，加入多层运动融合模块 MMFM，以降低浅层次运动特征的运动噪声。如表 5 第四行所列，J&F 提升了 2.7%。

表 5 不同模块的消融实验

Table 5 Ablation study of different modules

CMMFGM	MCG	MBAM	MMFM	*J*&*F*
√				79.2
√	√			80.7
√	√	√		82.5
√	√	√	√	**85.2**

同时，为了验证本文的代价图和运动特征生成模块(CMMFGM)的有效性，对运动特征提取分支提供不同的输入进行一系列的消融实验。首先，运动特征提取分支仅从光流图中(Flow map)提取运动特征，并采用按通道级联的方式来融合外观特征与运动特征。随后，加入代价图和运动特征生成模块 CMMFGM，为运动特征提取分支提供更准确的运动信息，如表 6 的第二行所列，加入代价图和运动特征生成模块之后，J&F 由 78.8%提升到 79.2%。接着，采用多模态共同注意力门控 MCG，以更有效地融合运动特征和外观特征。如表 6 的第三行和第四行所列，相比光流图提供的运动信息，代价图和运动特征生成模块提供的运动信息能够帮助网络获得更好的分割结果，J&F 也提高了 1.4%。

表 6 运动特征提取分支不同输入的消融实验

Table 6 Ablation study for different inputs of motion feature extraction branch

Flow map	CMMFGM	MCG	MBAM	MMFM	*J*&*F*
√					78.8
	√				79.2
√		√	√	√	83.8
	√	√	√	√	**85.2**

此外，为了验证本文提出的均衡重投影代价和多角度光流对比代价的有效性，本文将去掉了均衡重投影代价和多角度光流对比代价的代价图和运动特征生成模块(CMMFGM)命名为其他运动特征生成模块(Other Motion Feature Generation Module，OMFGM)，并在此模块的基础上依次加入均衡重投影代价和多角度光流对比代价进行实验。如表 7 的第二行和第三行所列，在已有的运动特征基础上分别加入均衡重

投影代价和多角度光流对比代价，运动目标分割的效果均得到了不同层次的提升。加入均衡重投影代价后，J&F 指标提升了 0.8%；加入多角度光流对比代价后，J&F 指标提升了 0.6%。在已有的运动特征基础上同时加入均衡重投影代价和多角度光流对比代价，结果得到了进一步提升，如表 7 第四行所列，J&F 指标提升了 1.2%。同时加入均衡重投影代价和多角度光流对比代价，网络能够获取到更丰富的运动信息，有利于提高运动目标分割的效果。

表 7 均衡重投影代价和多角度光流对比代价的消融实验

Table 7 Ablation study for the balanced reprojection cost and the multi-angle optical flow contrast cost

OMFGM	C_{bre}	C_{mofc}	$J\&F$
√			84.0
√	√		84.8
√		√	84.6
√	√	√	**85.2**

结束语 针对不同场景下的运动目标分割的难点，本文提出了不同的解决办法，如均衡重投影代价、多角度光流对比代价和差异单应性代价等。为了更好地在各种复杂的场景下分割运动物体，本文提出了一种基于外观融合的运动感知结构。其中多模态共同注意力门控用于融合物体的运动特征和外观特征，以获得有效的时空特征。为了抑制冗余的外观特征对最终结果的影响，本文提出了多层运动注意力模块，以融合不同层次的运动特征并利用物体的运动信息来强调运动的物体。在 KITTI，JNU-UISEE，KittiMoSeg 和 Davis-2016 数据集上的一系列实验都表明了本文方法的有效性。未来的工作当中，将考虑设计一种全新的注意力策略，来捕获不同运动代价之间的联系，以便动态更新不同运动代价的权重。

参 考 文 献

[1] PHILIP H S T. Geometric motion segmentation and model selection[J]. Philosophical Transactions-Royal Society Mathematical，Physical and Engineering Sciences，1998，356(1740)：1321-1340.

[2] RAHMATI H，DRAGON R，AAMO O M，et al. Weakly supervised motion segmentation with particle matching[J]. Computer Vision and Image Understanding，2015，140：30-42.

[3] THAKOOR N，GAO J，DEVARAJAN V. Multibody Structure-and-Motion Segmentation by Branch-and-Bound Model Selection [J]. IEEE Transactions on Image Processing，2010，19(6)：1393-1402.

[4] DAVE A，TOKMAKOV P，RAMANAN D. Towards Segmenting Anything That Moves[C]// Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW). Piscataway，NJ：IEEE，2020：1493-1502.

[5] TOKMAKOV P，SCHMID C，ALAHARI K. Learning to Segment Moving Objects[J]. International Journal of Computer Vision，2019，127(3)：282-301.

[6] YANG G，RAMANAN D. Learning to Segment Rigid Motions from Two Frames[C]// Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway，NJ：IEEE，2021：1266-1275.

[7] YANG S，ZHANG L，QI J，et al. Learning Motion-Appearance Co-Attention for Zero-Shot Video Object Segmentation[C]// Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway，NJ：IEEE，2022：1544-1553.

[8] TORR P H S，FITZGIBBON A W，ZISSERMAN A. The Problem of Degeneracy in Structure and Motion Recovery from Uncalibrated Image Sequences[J]. International Journal of Computer Vision，1999，32(1)：27-44.

[9] ZHU X，WANG L，ZHANG C，et al. Moving Object Detection Based on Continuous Constraint Background Model Deduction [J]. Computer Science，2019，46(6)：311-315.

[10] ZHANG N，SHI J H，YI J，et al. Real-time tracking method of underground moving target based on weighted centroid positioning[J]. Journal of Jilin University(Engineering and Technology Edition)，2023，53(5)：1458-1464.

[11] XU Y K，CHEN T Y，CHEN S Y，et al. Multi-object Tracking and Segmentation Algorithm by Fusing Motion Feature Embedding[J]. Journal of Chinese Computer Systems. 2023，44(6)：1304-1310.

[12] TOKMAKOV P，ALAHARI K，SCHMID C. Learning Motion Patterns in Videos[C]// Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway，NJ：IEEE，2017：531-539.

[13] RASHED H，RAMZY M，VAQUERO V，et al. FuseMODNet：Real-Time Camera and LiDAR Based Moving Object Detection for Robust Low-Light Autonomous Driving[C]// Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW). Piscataway，NJ：IEEE，2020：2393-2402.

[14] JAIN S D，XIONG B，GRAUMAN K. FusionSeg：Learning to Combine Motion and Appearance for Fully Automatic Segmentation of Generic Objects in Videos[C]// Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition(CVPR). Pisca-taway，NJ：IEEE，2017：2117-2126.

[15] LU X，WANG W，MA C，et al. See More，Know More：Unsupervised Video Object Segmentation With Co-Attention Siamese Networks[C]// Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway，NJ：IEEE，2020：3618-3627.

[16] ZHOU T，WANG S，ZHOU Y，et al. Motion-Attentive Transition for Zero-Shot Video Object Segmentation[C]// Proceedings of the Proceedings of the AAAI Conference on Artificial Intelligence(AAAI). Menlo Park，CA：AAAI，2020，34(7)：13066-13073.

[17] CHENG J，TSAI Y H，WANG S，et al. SegFlow：Joint Learning for Video Object Segmentation and Optical Flow[C]// Proceedings of the 2017 IEEE International Conference on Computer Vision(ICCV). Piscataway，NJ：IEEE，2017：686-695.

[18] LI S，SEYBOLD B，VOROBYOV A，et al. Unsupervised Video Object Segmentation with Motion-Based Bilateral Networks [C]// Proceedings of the Computer Vision(ECCV 2018). New York，NY：Springer International Publishing，2018：215-231.

[19] HU P，WANG G，KONG X，et al. Motion-Guided Cascaded Re-

finement Network for Video Object Segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 42(8): 1957-1967.

[20] PENG Q, CHEUNG Y M. Automatic Video Object Segmentation Based on Visual and Motion Saliency[J]. IEEE Transactions on Multimedia, 2019, 21(12): 3083-3094.

[21] LI H, CHEN G, LI G, et al. Motion Guided Attention for Video Salient Object Detection[C]// Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Pisca-taway, NJ: IEEE, 2020: 7273-7282.

[22] JI G P, FU K, WU Z, et al. Full-Duplex Strategy for Video Object Segmentation[C]// Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway, NJ: IEEE, 2022: 4902-4913.

[23] ZHOU X, WANG D, KRäHENBüHL P. Objects as points[J]. arXiv: 1904, 07850, 2019.

[24] YU F, WANG D, SHELHAMER E, et al. Deep Layer Aggregation[C]// Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). Piscataway, NJ: IEEE, 2018: 2403-2412.

[25] BRACHMANN E, ROTHER C. Neural-Guided RANSAC: Learning Where to Sample Model Hypotheses[C]// Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision(ICCV). Piscataway, NJ: IEEE, 2020: 4321-4330.

[26] YANG G, RAMANAN D. Upgrading Optical Flow to 3D Scene Flow Through Optical Expansion[C]// Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). Piscataway, NJ: IEEE, 2020: 1331-1340.

[27] YANG G, RAMANAN D. Volumetric Correspondence Networks for Optical Flow[C]// Proceedings of the Proceedings of the 33rd International Conference on Neural Information Processing Systems (NIPS). New York, NY: Curran Associates Inc., 2019: 794-805.

[28] CHUM O, PAJDLA T, STURM P. On Geometric Error for Homographies[J]. Computer Vision and Image Understanding, 2005, 97: 86-102.

[29] XIE E, SUN P, SONG X, et al. PolarMask: Single Shot Instance Segmentation With Polar Representation[C]// Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). Piscataway, NJ: IEEE, 2020: 12190-12199.

[30] GEIGER A, LENZ P, URTASUN R. Are we ready for autonomous driving? The KITTI vision benchmark suite[C]// Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition(CVPR). Piscataway, NJ: IEEE, 2012: 3354-3361.

[31] SIAM M, MAHGOUB H, ZAHRAN M, et al. MODNet: Motion and Appearance based Moving Object Detection Network for Autonomous Driving[C]// Proceedings of the 2018 21st International Conference on Intelligent Transportation Systems (ITSC). Piscataway, NJ: IEEE, 2018: 2859-2864.

[32] PERAZZI F, PONT-TUSET J, MCWILLIAMS B, et al. A Benchmark Dataset and Evaluation Methodology for Video Object Segmentation[C]// Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition(CVPR). Piscataway, NJ: IEEE, 2016: 724-732.

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

[34] OCHS P, MALIK J, BROX T. Segmentation of Moving Objects by Long Term Video Analysis[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014, 36(6): 1187-1200.

[35] RANJAN A, JAMPANI V, BALLES L, et al. Competitive Collaboration: Joint Unsupervised Learning of Depth, Camera Motion, Optical Flow and Motion Segmentation[C]// Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ: IEEE, 2020: 12232-12241.

[36] WANG C, LI C, LIU J, et al. U2-ONet: A Two-Level Nested Octave U-Structure Network with a Multi-Scale Attention Mechanism for Moving Object Segmentation [J]. arXiv: 2007.13092, 2020.

[37] LIU D, YU D, WANG C, et al. F2Net: Learning to Focus on the Foreground for Unsupervised Video Object Segmentation[C]// Proceedings of the AAAI Conference on Artificial Intelligence (AAAI). Menlo Park, CA: AAAI, 2021, : 2109-2117.

[38] LU X, WANG W, SHEN J, et al. Zero-Shot Video Object Segmentation With Co-Attention Siamese Networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(4): 2228-2242.

[39] LIU Z, LIU J, CHEN W, et al. FAMINet: Learning Real-Time Semisupervised Video Object Segmentation With Steepest Optimized Optical Flow[J]. IEEE Transactions on Instrumentation and Measurement, 2022, 71: 1-16.

[40] CHO S, LEE H, KIM M, et al. Pixel-Level Bijective Matching for Video Object Segmentation[C]// Proceedings of the 2022 IEEE/CVF Winter Conference on Applications of Computer Vision(WACV). Piscataway, NJ: IEEE, 2022: 1453-1462.

XU Bangwu, born in 1998, postgraduate, is a member of CCF (No. N9250G). His main research interests include computer vision and deep learning.

ZHOU Haojie, born in 1981, Ph.D, associate professor, is a member of CCF (No. 19225S). His main research interests include system architecture, intelligent system and distributed computing.

(责任编辑:何杨)