融合动态区域检测的自监督视觉里程计方法

马 伟, 贾兆款, 米 庆
(北京工业大学信息学部, 北京 100124)

摘 要: 为解决室外场景中动态区域对视觉里程计计算过程的干扰,获得准确的相机位姿和场景深度,提出一种自监督深度学习框架下融合动态区域检测的视觉里程计算法. 给定相邻2帧图像,首先,采用深度估计网络计算2幅图像对应深度图,采用位姿估计网络获得二者初始相对位姿. 然后,借助视点变换,计算两视角深度图像之间的差异,确定动态区域. 在此基础上,对输入图像中动静态区域进行分离. 之后,匹配两视角图像静态区域特征,计算最终相机位姿. 从光度、平滑度以及几何一致性三方面构造损失函数,并在损失函数中融入动态区域信息,对所构造网络模型进行端到端自监督训练. 在KITTI数据集上验证了所提算法,并将其与最近2年提出的相关算法进行比较. 实验结果表明,该算法能够更好地应对动态场景,实现更高精度的相机姿态估计和细小物体深度估计.

关键词: 动态区域检测; 视觉里程计; 深度信息; 深度学习; 自监督; 动静态信息分离
中图分类号: TP 391 **文献标志码**: A **文章编号**: 0254 - 0037(2021)05 - 0444 - 11
doi: 10.11936/bjutxb2020120036

Self-supervised Visual Odometry Method Based on Dynamic Region Detection

MA Wei, JIA Zhaokuan, MI Qing
(Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China)

Abstract: A robust visual odometry in the framework of deep self-supervision, which can overcome the interference of dynamic regions in camera pose estimation and scene depth computation, was proposed. Two consecutive frames of images were selected. Depth estimation network was first used to calculate the depth maps of the two images, and pose estimation network was used to calculate the relative poses. Then, dynamic regions were obtained by comparing a depth map with the one synthesized by warping the other depth map into the current view. Based on the detected dynamic regions, the dynamic and static parts in input images were seperated with this algorithm. Then, the features in the static regions of the two images were matched to obtain the final camera poses. The loss function was composed of photometric error, smoothness error, and geometric consistency error, and dynamic region information was integrated into the loss function. Finally, this loss function was used to train the proposed network in a self-supervised end-to-end manner. Experiments were carried out on the widely used KITTI dataset. Specifically, the proposed model was compared with the state-of-the-art ones proposed in recent two years. Experimental results show that this algorithm can deal with the dynamic scenes more robustly and achieve higher accuracy in camera pose estimation and scene depth computation.

收稿日期: 2020-12-31
基金项目: 国家自然科学基金资助项目(61771026)
作者简介: 马 伟(1980—), 女, 副教授, 主要从事计算机视觉、图像处理、文化遗产数字化展示与修复方面的研究, E-mail: mawei@bjut.edu.cn

Key words：dynamic region detection；visual odometry；depth message；deep learning；self-supervision；separation of dynamic and static information

视觉里程计是机器人和计算机视觉领域一项关键技术. 视觉里程计通过分析相邻 2 张图像间静态特征的对应关系，获得相机位置和姿态（简称位姿）以及其他相关数据，例如深度图. 精确稳定的静态特征提取和匹配对于视觉里程计尤为重要. 然而，在实际应用场景中存在诸多非静态物体和区域，本文将这些非静态物体和区域统称为动态区域，并将动态区域归纳为 2 类：一类是连续运动的物体，另一类是由于遮挡或视角范围改变造成的突然出现或消失的区域. 由动态物体内部特征对应关系确定的是相机和动态物体之间的相对运动，而突然出现或消失的区域不存在帧间对应关系，因此，上述 2 类区域的存在都将干扰静态特征匹配. 针对上述问题，本文提出一种融合动态区域检测的视觉里程计算法，以有效去除动态区域干扰，提升视觉里程计计算精度和细小物体的深度估计.

现有视觉里程计算法可以分为 2 类：传统视觉里程计[1-2]和基于深度学习的视觉里程计[3-4]. 静态场景下的传统视觉里程计[5-10]通过提取相邻 2 张图像中经过手工设计的特征点，建立相应特征点的对应关系，并使用随机抽样一致（random sample consensus，RANSAC）算法滤除误匹配，进而实现相机位姿计算. 这种算法在静态场景中能够得到精确的特征匹配关系，进而保证里程计计算精度. 但在动态场景或遮挡物较多的场景中，经常存在匹配不一致情况，而且 RANSAC 算法不能有效地将这些匹配不一致情况滤除，导致传统方法在动态场景中不够鲁棒. 因此，国内外研究人员对此问题提出了各种解决方案[11-14]. 其中 Sheng 等[11]提出的算法在传统方法基础上融合了动态物体检测算法，手动去除动态物体提供的特征匹配关系，根据静态特征匹配关系计算相机位姿，进而提升传统方法在动态场景中的鲁棒性. 传统方法直接基于像素值或手工设计特征进行匹配，利用的图像信息不够充分. 深度学习技术基于强大的特征提取和表达能力，为研究人员提供了一种新的充分利用图像信息的思路，因此，大量研究人员开始探究深度学习技术在视觉里程计中的应用.

基于深度学习的视觉里程计算法[15-19]以先进的网络模型（例如 GoogLeNet[20]和 AlexNet[21]）为基础，实现特征提取及匹配工作，最终利用一个全卷积网络建立回归模型，实现视觉里程计计算. 实验结果表明，使用神经网络提取的特征对模糊、亮度和对比度异常问题更加鲁棒，并且这些算法能够在一定程度上应对场景中动态信息所带来的干扰，但是效果有限. 此后，研究人员提出一系列基于自监督的视觉里程计算法[22-29]. 这些算法通过构建深度估计网络[30-31]和位姿估计网络实现自监督学习，降低了获取监督信息的高昂代价，同时，可以利用深度和相机位姿之间的关系实现场景中动态信息捕捉. 然而，这些算法仅将捕捉到的动态信息融入监督信号中以提升模型训练效果. 例如，Bian 等[29]提出一种面向动态场景的自监督视觉里程计算法（本文将其作为基准算法），该算法首先利用深度信息实现动态物体检测；然后，将动态物体信息融入损失函数中，借此优化训练过程以提升模型的训练效果. 该算法限于通过优化损失函数和训练过程来提升模型能力，这也是现存算法的通病. 除此之外，现存算法中，位姿估计网络模块的输入是包含动静态混合信息的原始图像，图像中存在的动态信息对视觉里程计的计算造成不良影响.

相对于传统视觉里程计，基于深度学习的视觉里程计具有更强的特征提取能力，能够更鲁棒地应对模糊、亮度和对比度异常问题. 不同于其他算法使用动态信息对损失函数进行优化，本文所提算法使用动态信息优化输入图像和特征，从而优化视觉里程计计算过程，使其计算直接利用静态信息，不受动态信息干扰. 综合上述 2 种优势，本文提出一种自监督深度学习框架下融合动态区域检测的视觉里程计算法. 此算法使用卷积神经网络构建视觉里程计模型，并且在此基础上融入动态区域检测，使用检测得到的动态信息优化视觉里程计的输入信息，实现输入信息的动静态信息分离，最终使得视觉里程计的计算仅依赖可靠的静态信息，消除动态信息的干扰. 上述计算可实现对视觉里程计计算过程的优化. 同时，本算法使用动态信息优化损失函数，以此优化训练过程，进而提升模型能力. 最终，本文将视觉里程计、动态区域检测和动静态信息分离构建成一个能够实现端到端训练的深度学习网络模型.

1　算法框架

本文提出的融合动态区域检测的视觉里程计

算法框架如图 1 所示. 图中 D_t 和 D_s 分别表示图像 I_t 和图像 I_s 对应的深度图. 本算法分为 4 个子模块:深度估计模块、位姿估计模块、动态区域检测模块和动静态信息分离模块. 输入为 2 张相邻图像,深度估计网络分别计算 2 张图像的深度信息;位姿估计网络 1 根据原始图像计算相机位姿变换;动态区域检测模块利用深度信息和位姿信息计算动态信息;动静态信息分离模块利用获得的动态信息实现对原始图像动静态信息的分离;位姿估计网络 2 根据去除动态信息的图像计算最终的相机位姿. 本文从光度误差、平滑度误差、几何一致性误差 3 个角度对损失函数进行约束,实现自监督网络模型的训练.

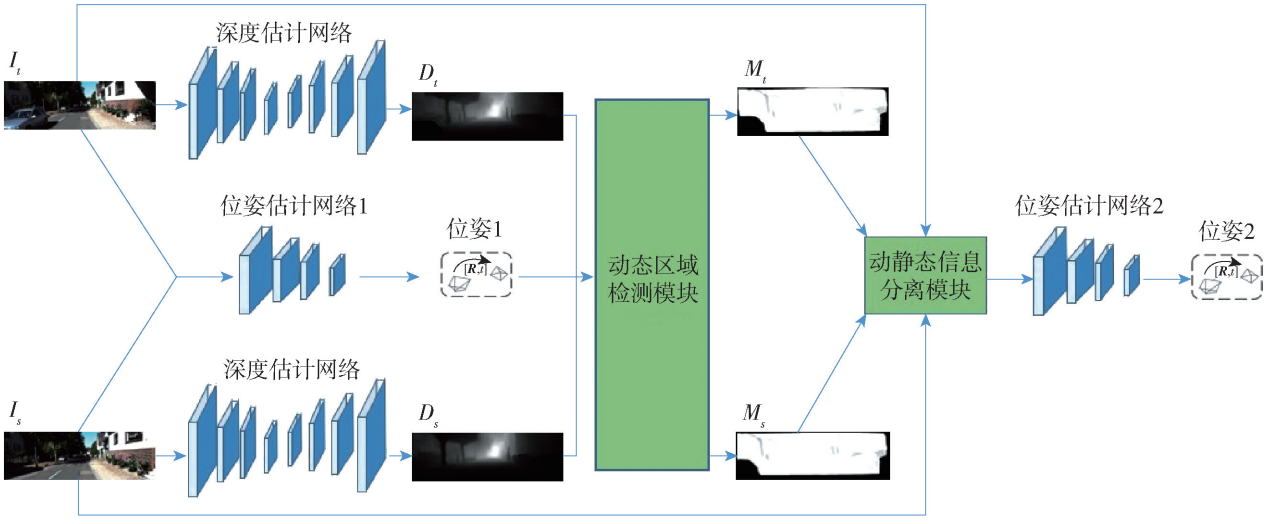


图 1 本文算法框架

Fig. 1 Framework of the proposed algorithm

2 动态区域检测和动静态信息分离

场景中存在的动态区域影响视觉里程计计算,因此,如何实现场景中动静态信息分离,去除动态信息干扰,利用静态信息实现相机位姿计算成为关键.

现有实现动态区域检测的方法有 2 种:一种是基于深度信息的动态区域检测,另一种是基于光流信息的动态区域检测. 对于光流信息,根据相机成像原理,距离相机较远物体所表现出来的光流和距离相机较近物体所表现出来的光流不一致,因此,造成光流信息整体不均匀但局部均匀的现象. 虽然动态物体的出现造成光流信息局部不均匀现象,基于此可实现动态信息检测,但是在进行动静态信息分离时,光流的整体不均匀性导致动静态信息难以分离. 考虑到光流信息的局限性,本文利用深度信息实现动态区域检测,进而去除场景中的动态信息,将静态信息输入相机位姿估计网络,实现更为精确的视觉里程计计算.

利用深度信息实现动态信息检测的方法如图 2 所示. D_s^t 表示利用相机位姿变换,将 D_t 投影到图像 I_s 对应视角的深度图像. 此时,D_s 和 D_s^t 均表示在图像 I_s 视角下对应的深度图. 由于图像中的汽车是运动的,造成 D_s^t 中汽车的深度和 D_s 中的不一致. 因此,对 D_s^t 和 D_s 做差,计算得到图 I_t 中存在的动态信息. 综上所述,实现动态区域检测分为 2 步:首先,利用公式 $D_s^t = D_t T_{t\to s}$,计算 D_t 在图像 I_s 视角下对应的深度信息 D_s^t;其次,利用动态区域计算公式

图 2 动态区域检测

Fig. 2 Dynamic object detection

$$M_{\mathrm{d}}(p)=\frac{|D_s^t(p)-D_s(p)|}{D_s^t(p)+D_s(p)} \tag{1}$$

计算图像 I_t 中存在的动态信息 $M_{\mathrm{d}}(p)$. 式中：$M_{\mathrm{d}}(p)$为计算得到的动态信息；$D_s^t(p)$为将图像 I_t 在坐标 p 处对应深度值投影到图像 I_s 视角下的深度值；$D_s(p)$为图像 I_s 在坐标 p 处对应的深度值；p 为图像中的像素坐标值.

视觉里程计根据图像信息计算相机位姿. 动态场景中获得的图像蕴含动静态 2 种信息，相较于动态信息，静态信息是可靠信息. 因此，本文所提算法利用动静态信息分离算法去除图像中动态信息，保留静态信息. 根据前人思想[26,28]，本文将动静态信息分离定义为

$$I_{\mathrm{static}}=I\times(1-M_{\mathrm{d}}) \tag{2}$$

式中：I_{static}为分离得到的静态信息；I 为原始图像. 如图 3 所示，图像 I_t 和 I_s 中存在连续运动的汽车，根据式(1)计算得到此汽车对应的动态区域信息 M_t 和 M_s. 根据式(2)，首先通过$(1-M_{\mathrm{d}})$对包含动态区域的遮罩图像 M_{d} 取反，由此计算图像 I_t 和 I_s 对应的静态区域遮罩图像. 然后利用静态区域遮罩图像计算原始图像中对应的静态信息 I_{tm}和 I_{sm}，由此实现动静态信息分离，去除原始图像中汽车所对应动态区域信息，提取得到原始图像中的静态信息. 更多实验结果见图 4，其中第 1 列是原始图像，第 2 列是动态区域检测结果，第 3 列是经过动静态信息分离后提取得到的静态信息. 从图中可以看到，经过动静态分离后，图像中汽车等动态信息被去除，建筑物等静态信息被保留.

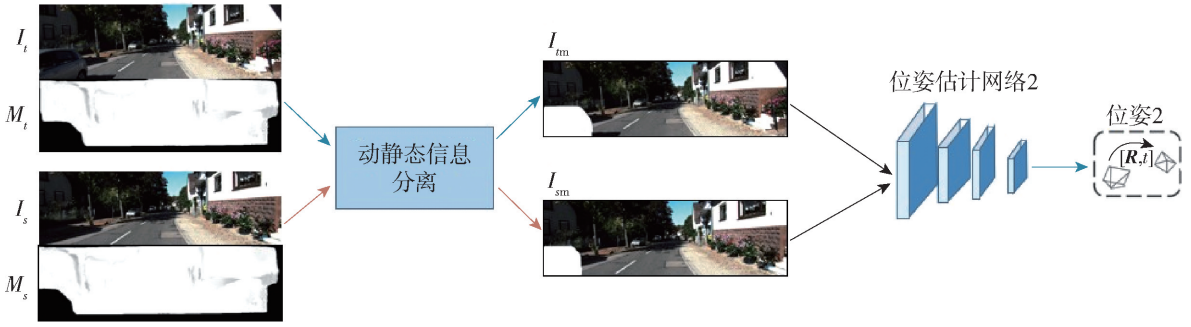


图 3　动静态信息分离

Fig. 3　Separation of dynamic and static information

图 4　动静态信息分离结果

Fig. 4　Results of dynamic and static information separation

最终，如图 3 所示，将提取得到的相邻 2 张图像对应静态信息 I_{tm}和 I_{sm}输入相机位姿估计网络 2，实现最终相机位姿估计.

3　损失函数设计

对卷积神经网络的训练需要设置全面合理的损失函数. 有监督卷积神经网络利用自身计算结果和真值间的差异构建损失函数. 自监督卷积神经网络通过算法本身构建监督信息，进而实现对网络的训练. 自监督卷积神经网络不需要真值，因此，不要求数据集提供真值信息，最终使得数据集获取难度大大降低. 在损失函数设计上，损失函数携带越多有效信息，越有利于网络模型的学习. 因此，根据前人提出的各种算法[27-29]，本文从光度误差 L_{b}、平滑度

误差 L_s 和几何一致性误差 L_g 三个角度出发,将损失函数定义为

$$L_{all} = \lambda_b L_b + \lambda_s L_s + \lambda_g L_g \tag{3}$$

式中 λ 为不同类型误差的权重.

3.1 光度误差损失函数

光度误差利用合成图像的思想进行构建,是实现自监督训练的核心. 光度误差损失函数构建流程如下.

给定相邻 2 张图像 I_t、I_s,利用图像 I_t 计算图像 I_s 的合成图像 $\hat{I}_s$,利用图像 I_s 和 $\hat{I}_s$ 的相似性构建损失函数. 光度误差损失函数定义为

$$L_b = \frac{1}{N}\sum_{p\in I} \| I_s(p) - \hat{I}_s(p) \|_1 \tag{4}$$

式中 N 表示图像中像素的个数.

对于 $\hat{I}_s$ 的计算,需要用到图像 I_t、I_s 对应的深度图像 D_s 及图像 I_s 到图像 I_t 的相机位姿变换 $T_{s\to t}$. 具体的计算公式为

$$\hat{p}_t \sim \boldsymbol{K} T_{s\to t} D_s(p_s) \boldsymbol{K}^{-1} p_s \tag{5}$$

式中:p_s 为在图像 I_s 上某一像素点的坐标;$\boldsymbol{K}$ 为相机的内参矩阵;$D_s(p_s)$ 为像素坐标 p_s 处的深度信息. 式(5)从右向左看,首先利用内参矩阵 $\boldsymbol{K}^{-1}$ 将像素坐标 p_s 投影到相机的归一化平面上,其次利用深度信息 $D_s(p_s)$ 计算得到像素坐标 p_s 对应的三维深度坐标 R_s,然后利用相机位姿变换 $T_{s\to t}$ 将三维深度坐标 R_s 变换到图像 I_t 对应视角下的三维坐标 R_t,最后利用相机内参 $\boldsymbol{K}$ 将 R_t 投影到图像 I_t 上,从而计算得到像素坐标 p_s 在图像 I_t 中对应的坐标 $\hat{p}_t$. 坐标 $\hat{p}_t$ 不会落于图像 I_t 所在坐标系的整数坐标处,因此,需要利用可微的双线性差值[32]计算坐标 $\hat{p}_t$ 对应的像素值,即:在图像 I_t 上取 $\hat{p}_t$ 的 4 个相邻像素计算 $\hat{p}_t$ 处的像素值. 通过上述算法,利用图像 I_t 计算得到图像 I_s 的合成图像 $\hat{I}_s$.

使用式(4)训练网络能得到可接受的结果,但实际场景中存在非朗伯体反射等现象(例如街道两旁的窗户和车面造成的反射),其引发的光强变化将严重影响损失函数的计算和网络训练. 因此,本文加入结构相似性(structural similarity, SSIM)[33]计算. 结构相似性指数能够将结构信息独立于亮度和对比度,反映场景中物体的结构属性,进而更好地应对光强变化,更好地衡量原始图像和合成图像之间的相似程度. 据此,将式(4)更新为

$$L_b = \frac{1}{N}\sum_{p\in I}\left(\alpha \| I_s(p) - \hat{I}_s(p) \|_1 + (1-\alpha)\frac{1-\mathrm{SSIM}_{s\hat{s}}(p)}{2}\right) \tag{6}$$

根据前人工作[24-26],设置 $\alpha = 0.15$.

3.2 平滑度误差损失函数

使用光度误差构造的损失函数存在梯度局部性问题. 梯度局部性问题是指在图像低纹理区域和均匀区域,光度误差给出的信息非常少,在这些区域上计算得到的梯度信息几乎没有任何实际意义. 使用光度误差损失函数训练深度网络,并使用此网络计算图像的深度信息,会造成在低纹理区域或均匀区域深度信息不连续,因此,要减少或抑制低纹理区域或均匀区域提供的梯度信息,或迫使这部分区域携带更多有意义的信息. 有 2 种方式可以在一定程度上缓解上述问题:一是使用编解码网络实现深度信息计算. 编解码网络的瓶颈结构能够扩大感受野,并且该网络的编解码过程能够实现信息的逐步扩散,进而使得低纹理区域携带更多有意义的信息. 二是融合多尺度信息并引入平滑度损失函数. 这种方式同样能够扩大感受野,使低纹理区域携带更多有意义的信息,进而增强深度信息的平滑度. 现有绝大部分算法都采用第 2 种方式来解决梯度局部性问题,本文同样利用平滑度损失函数约束网络模型的训练,进而解决梯度局部性问题. 平滑度损失函数定义为

$$L_s = \sum_p \left(e^{-\nabla I_a(p)} \nabla D_a(p)\right)^2 \tag{7}$$

式中:p 为像素坐标;∇ 为求导操作. 平滑度损失函数对深度信息进行约束使得此函数携带更多有效梯度信息,从而促进网络的训练.

3.3 几何一致性损失函数

在单目视觉里程计中,由于单视点的原因,视觉里程计无法计算得到具有真实尺度信息的相机位姿,而且会造成在一个序列上计算得到的各个相机位姿尺度不一致. 单目深度估计同样存在上述尺度问题. 为解决这一问题,本文引入几何一致性损失函数

$$L_g = \frac{1}{N}\sum_{p\in I} M_d(p) \tag{8}$$

此函数通过对比相邻图像之间深度信息,能够确保计算得到的深度信息尺度是一致的,进而能够保证相机位姿变换的尺度也是一致的. 基于此,本文在训练的过程中,先对 I_{s-1} 和 I_t 进行一次计算,再对 I_t 和 I_{s+1} 进行一次计算,接着依次对 I_{s+1} 和 I_{s+2}、I_{s+2} 和 I_{s+3}……进行计算,直到对整个序列完成相应计算. 这种逐步扩散方式,能够保证图像 I_{s-1}、I_t、I_{s+1}、I_{s+2}、

I_{s+3}……对应的深度信息尺度是一致的，进而能够确保在整个序列上计算的深度和相机位姿尺度一致.

3.4　整体损失函数

本文利用深度信息实现动态区域检测，并且实现动静态信息分离，进而降低动态信息的干扰，使得视觉里程计的计算只利用静态信息提升视觉里程计算法在动态场景中的鲁棒性. 为进一步提升训练效果，本文对损失函数进行设计时，在损失函数中融入动态信息，降低动态信息对训练造成的影响. 据此，本文将式(3)更新为

$$L_{\text{all}} = \lambda_{\text{b}} M_{\text{s}} L_{\text{b}} + \lambda_{\text{s}} L_{\text{s}} + \lambda_{\text{g}} M_{\text{s}} L_{\text{g}} \tag{9}$$

式中 $M_{\text{s}} = (1 - M_{\text{d}})$.

现有算法[27-29]根据深度估计网络和位姿估计网络 1 实现上述损失函数的计算，完成对网络模型的训练. 对比现有算法，本文增加了位姿估计网络 2，因此，本文根据深度估计网络和位姿估计网络 2 实现上述损失函数的计算. 根据训练过程中梯度回传的原理，仅利用深度估计网络和位姿估计网络 2 计算损失函数，不能很好地完成位姿估计网络 1 的训练. 因此，本文在损失函数中的主要贡献是，首先利用深度估计网络和位姿估计网络 1 计算一次损失函数，然后利用深度估计网络和位姿估计网络 2 计算一次损失函数，最终对上述 2 次损失函数进行整合，以此实现对网络模型的训练，获得最好的训练效果.

4　算法步骤

至此，本文已对算法框架、动态区域检测、动静态信息分离和损失函数分别进行了详细的说明. 为更完整说明所提算法，本文对其进行总结，算法如下.

输入：相邻图像 I_t 和图像 I_s.

输出：相机位姿 $T_{t\to s}$、深度信息 D_t 和 D_s.

1) 输入图像 I_t 和 I_s 到深度估计网络，计算对应深度信息 D_t 和 D_s.

2) 输入图像 I_t 和 I_s 到位姿估计网络 1，粗略计算相机位姿 $T'_{t\to s}$ 和 $T'_{s\to t}$.

3) 利用上述信息及式(1)，计算动态区域 M_t 和 M_s.

4) 根据式(2)实现动静态信息分离，计算图像 I_t 和 I_s 中对应的静态信息 I^t_{static} 和 I^s_{static}.

5) 将 I^t_{static} 和 I^s_{static} 输入位姿估计网络 2，计算最终的相机位姿 $T_{t\to s}$.

5　实验结果与分析

5.1　实验数据与评价标准

本文选用 KITTI Odometry[34-35]数据集进行实验，此数据集是 KITTI[34-35]数据集的一个子集. KITTI 数据集由德国卡尔斯鲁厄理工学院和丰田美国技术研究院联合创办，是目前国际上最大的自动驾驶场景下的计算机视觉算法评测数据集. KITTI 数据集包含市区、乡村和高速公路等场景采集的真实图像数据，每张图像中最多达 15 辆车和 30 个行人，还有各种程度的遮挡与截断. KITTI Odometry 数据集是由立体图像采集得到的. 针对 22 个室外场景，提供了对应的 22 个视频帧序列. 每个序列中包含一个左视图视频帧序列和一个右视图视频帧序列以及相机内参. 其中前 11 个序列提供了相机轨迹的真值. 同其研究人员的设置一样，本文选用这 11 个序列中的前 9 个序列作为训练集，后 2 个序列作为测试集.

对 KITTI Odometry 数据集提供的原始图像，做了如下增强处理. KITTI Odometry 数据集原始图像宽高为 1 241 ×376 或 1 242 ×375 或 1 226 ×370. 首先将图像缩放至 832 ×256 大小，其次使用随机水平翻转、尺度缩放和裁剪 3 种技术增强数据集.

本文将从量化指标、相机运动轨迹、动态区域检测效果 3 个角度进行实验分析. 为便于实验对比分析，在量化指标方面，本文选用绝对轨迹误差(absolute trajectory error，ATE)对实验结果进行量化测评. 绝对轨迹误差定义为

$$E_i = Q_i^{-1} S T_i \tag{10}$$

其计算的是真实位姿和估计位姿之间的差值，能够反映出算法的精度和轨迹的全局一致性. 式中：E_i 为第 i 时刻的绝对轨迹误差(包括平移误差和旋转误差)；Q_i 为相机位姿的真值；S 为尺度因子；T_i 为相机位姿的估计值. 通常使用均方根误差(root mean square error，RMSE)对绝对轨迹误差进行统计，并以此作为最终的结果，其公式定义为

$$\text{RMSE}(E_{1:n}) = \sqrt{\frac{1}{n}\sum_{i=1}^{n} \| \text{trans}(E_i) \|^2} \tag{11}$$

式中 $\text{trans}(E_i)$ 为平移误差. $\text{RMSE}(E_{1:n})$ 的单位为 hm.

5.2　参数设置及模型选择

本文提出的网络模型涉及 3 个子模型：深度估计网络、位姿估计网络 1、位姿估计网络 2. 对于深度估计网络模型，本文选用 DispResNet[25]网络，此

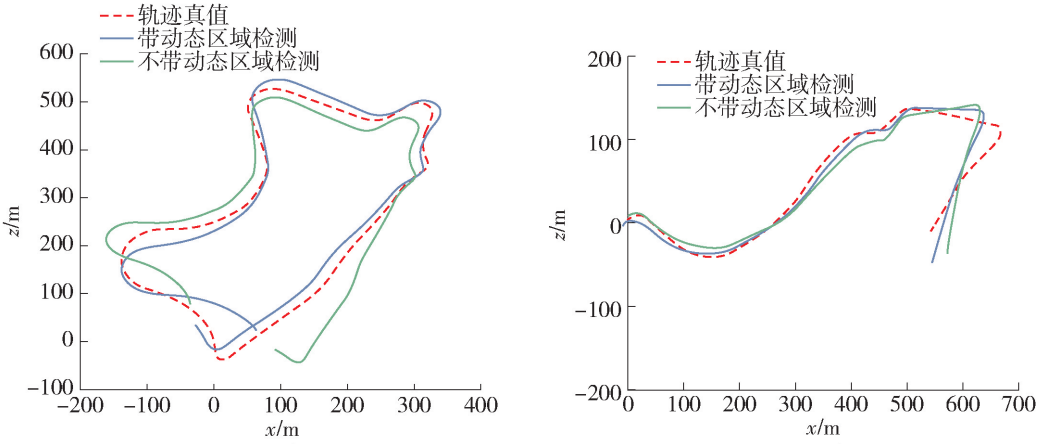
网络是一种编解码网络,编码器用于对原始图像进行特征的逐步提取和凝练,解码器用于对提取得到的特征进行解析,使其符合预期结果. 在编解码器之间会使用跳跃连接对相同尺度的特征进行连接和融合以提升网络性能. 此网络的输入为一张 RGB 图像,输出为对应的逐像素的深度信息. 对于位姿估计网络 1 和位姿估计网络 2,本文选用 PoseNet[22] 网络. 此网络是一个简单的编码器,输入是按颜色通道连接的相邻 2 张 RGB 图像,输出是一个六自由度的相机位姿.

本文所提算法使用 PyTorch 框架实现. 使用 Adam 优化器实现网络训练,其中 $\beta_1 = 0.900$,$\beta_2 = 0.999$;学习率设置为 10^{-4}. 对于光度、平滑度和几何一致性 3 种误差,光度误差最为重要,是构建自监督学习的依据,同时也是完成整个网络模型训练的核心所在. 因此,对于光度误差权重需要设置为最大. 本文利用平滑度误差解决由光度误差造成的梯度局部性问题,目的是优化网络模型训练,使计算得到的深度信息更为平滑. 平滑度误差主要是为优化网络训练而设计,因此,其权重可以设置小一些. 本文利用几何一致性误差实现尺度一致性. 尺度一致性对视觉里程计实际应用非常重要,对整个网络模型的训练是一个强约束项,因此,其权重占比较大. 综上所述,光度误差和几何一致性误差是功能项,决定网络模型实现哪些功能. 平滑度误差是优化项,可优化网络模型训练. 由此,经过实验验证,本文将上述 3 种损失函数对应权重设置为 $\lambda_b = 1.0$,$\lambda_s = 0.1$,$\lambda_g = 0.5$. 最后,对于整个网络,一共训练 200 代,每代 1 000 个批次,每个批次的大小为 4.

5.3 消融实验

本文核心贡献之一是实现了场景的动静态信息分离,由此提升视觉里程计精度. 为进一步验证其有效性,本文进行消融实验,对比融合动态区域检测方法前后视觉里程计轨迹差异,以此说明融合动态区域检测后能够提升视觉里程计计算精度.

图 5 所示为相机在 KITTI Odometry 数据集的 09 和 10 两个场景中的运动轨迹,展示了融合动态区域检测方法前后视觉里程计轨迹对比结果. 不带动态区域检测的视觉里程计算法对应位姿估计网络 1. 由图中所示对比结果可以看出,动态区域检测模块能够有效提升相机运动轨迹结果,即融合动态区域检测的视觉里程计算法能够更好地应对动态场景,实现更为精确的视觉里程计计算. 为进一步验证上述说明,本文对动态区域检测结果进行可视化.



(a) 09序列 (b) 10序列

图 5 融合动态区域检测前后的视觉里程计轨迹对比

Fig. 5 Comparison of visual odometry trajectory with and without dynamic region detection

图 6 所示是利用深度信息实现的动态区域检测,彩色图像是原始图像,灰度图像是动态区域检测结果. 图 6 对 2 类动态区域检测结果进行展示,前 2 行展示的是场景中连续运动的物体. 对于这种动态物体区域,物体边缘与周围静态环境信息距离较近,边缘区域会发生较严重的特征误匹配,对视觉里程计的计算造成较大干扰. 针对此问题分析相邻 2 帧图像中连续运动物体的深度信息,在物体边缘上深度信息变化较大. 基于上述特性,本文算法能够通过分析深度信息的差异检测到此动态区域,进而减少其对视觉里程计的干扰. 图 6 中后 2 行展示的是在距离相机较近的区域,由于遮挡或视角变换造成的突然出现或消失的区域. 此区域同样会造成较为严重的特征误匹配情况,但是本文算法同样能够对

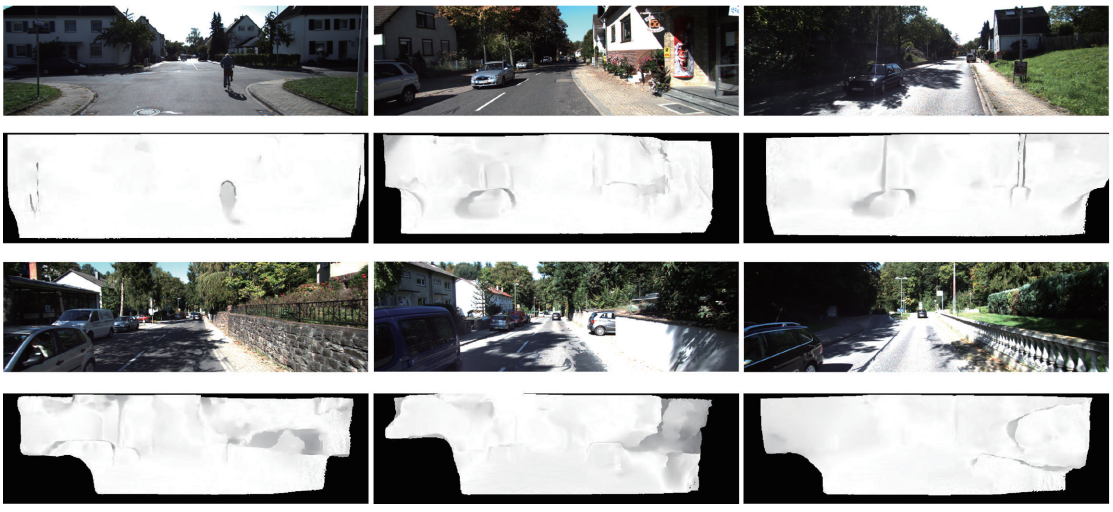


图 6　基于深度信息的动态区域检测

Fig. 6　Dynamic object detection based on depth information

此区域进行有效检测. 分析相邻 2 帧图像中突然出现或消失区域的深度信息，此区域深度信息变化剧烈且较为完整，因此，能够通过深度信息的差异检测得到动态区域，进而可以减少误匹配信息，提升视觉里程计计算精度. 综上所述，本文算法能够较为准确地检测到场景中的动态区域，并且能够因此提升视觉里程计计算精度.

5.4　视觉里程计性能

随着深度学习的出现，涌现出了一大批基于深度学习的视觉里程计算法. 本文选取 4 个具有代表性的算法进行对比实验. Yin 等[24]利用光流实现了动态信息估计，Luo 等[26]利用场景流的方式实现动态信息估计，Casser 等[28]直接利用 Mask R-CNN[36]实现潜在动态物体估计，Bian 等[29]利用深度信息实现动态物体检测. 这 4 个算法都实现了动态物体检测，但是这些算法仅是将动态信息融入损失函数中，通过优化训练过程来提升训练效果. 本文算法利用估计得到的动态信息，首先对场景中的动静态信息进行分离，然后仅利用静态信息进行视觉里程计的计算，最终同样将动态信息融入损失函数中，优化模型的训练.

本文从 ATE 和视觉里程计轨迹 2 个方面对视觉里程计性能进行评估. 为便于实验分析，本文同文献[24,26,28-29]一样在 KITTI Odometry 数据集的 09 和 10 两个序列上进行测试，实验对比结果如表 1 所示. 本文提出的算法在 09 序列上得到了最好的实验结果，相较于 Bian 等[29]的算法提升了 13%. 在 10 序列上，本文算法的性能与其他算法相当. 从实验结果来看，本文所提算法能够提升视觉里程计计算精度.

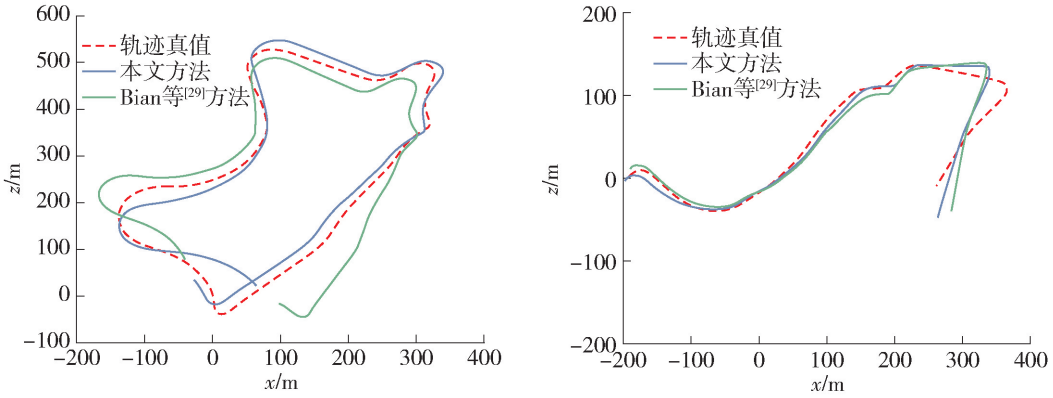
表 1　绝对轨迹误差对比结果

Table 1　Absolute trajectory error comparison results

算法	ATE	
	09	10
Yin 等[24]	0.012 0	0.012 0
Luo 等[26]	0.013 0	0.012 0
Casser 等[28]	0.011 0	0.011 0
Bian 等[29]	0.007 1	0.015 0
本文	0.006 2	0.011 9

从量化指标上来看，本文提出的算法取得了更好的性能. 为了获得更直观的结果，本文对视觉里程计轨迹进行了可视化. 本文以 Bian 等[29]所提算法为基准，因此，对 Bian 等[29]的算法、本文算法和真值的相机运动轨迹进行了可视化，结果如图 7 所示. 从图中可以看出，本文所提算法优于 Bian 等[29]算法.

对比图 7 和图 5，对于图 7 所示 Bian 等[29]所提视觉里程计和图 5 所示不带动态区域检测的视觉里程计，二者计算的视觉里程计轨迹相似. 这是因为本文以 Bian 等[29]所提算法为基准，对其算法进行了改进，融入了动态物体检测、动静态信息分离和位姿估计网络 2 模块. 因此，上述 2 个视觉里程计模型相同，而且输入信息均为原始未经过动静态信息分离的图像. 由此，计算得到的视觉里程计轨迹相似.



(a) 09序列 (b) 10序列

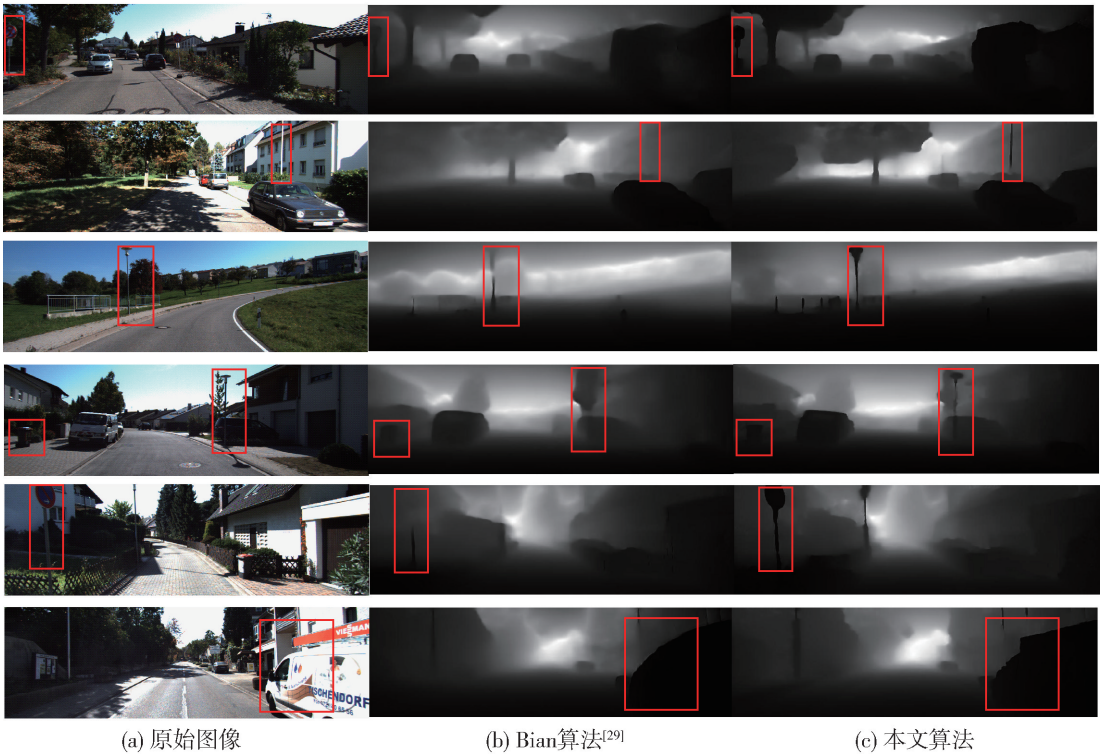
图 7 本文算法与 Bian 等[29]所提算法视觉里程计轨迹对比

Fig. 7 Comparison of visual odometry trajectory between the method in this paper and the method proposed by Bian et al. [29]

5.5 深度信息估计

本文所提算法是基于自监督的视觉里程计算法,包含了深度信息估计模型和位姿信息估计模型,并完成了对二者的联合训练. 其中,任何一个模块性能的提升都会影响另外一个模块的训练,使得另一个模块性能同样得到提升. 两者相互促进,不断提升 2 个模块的性能. 上述视觉里程计相关实验证明,本文算法能够提升视觉里程计计算精度. 因此,深度信息估计模块的性能同样应该得到提升.

如图 8 所示,为验证上述预测,本文给出了图像对应深度信息可视化结果,并和 Bian 等[29]算法进行对比. 从图像中可以看出,对于细小物体,例如图中所示路标和电杆,本文算法能够得到更为准确的估计. 图 8 所示结果证明,视觉里程计精度的提升促进深度估计更为准确. 深度信息精度的提升会使得动态区域检测更为完整. 综上所述,对本文所提算法的训练,将实现一个正向循环,促进整个网络模型的正向学习,提升各模块的性能.



(a) 原始图像 (b) Bian算法[29] (c) 本文算法

图 8 深度图像

Fig. 8 Depth map

6　结论

1）在自监督深度学习框架下，提出了一种融合动态区域检测的视觉里程计模型. 在 KITTI 数据集上和现有先进算法进行对比实验，验证本文所提算法的有效性. 实验结果表明，本文所提算法能够更好地应对动态场景，实现动静态信息的分离，利用静态信息计算相机运动，进而使得视觉里程计计算更为准确，同时提升场景深度估计的准确性.

2）本文所提算法使用相邻图像间深度信息的差异性实现动态区域检测，并进一步实现动静态信息分离. 此算法计算速度快、效率高且能够实现较为准确和全面的动态区域检测. 对动静态信息的分离，消除了动态信息对视觉里程计计算过程的干扰，有效地促进了视觉里程计计算精度的提升.

3）本算法使用光度误差、平滑度误差和几何一致性误差构建损失函数，并且使用分离得到的静态信息进一步约束光度误差和几何一致性误差，优化损失函数，提升训练效果.

4）通过动态区域可视化结果可知，当前对动态区域的检测依然不够完整. 对连续运动物体，更多地是检测到其边缘部分，但是，其内部信息同样会对视觉里程计计算造成一定的干扰. 因此，未来可以从种子区域生长、融合光流信息、引入注意力机制等方面实现对动态区域的进一步检测，以实现更为准确和完整的动态区域估计.

参考文献：

[1] SMITH R C, CHEESEMAN P. On the representation and estimation of spatial uncertainty[J]. International Journal of Robotics Research, 1986, 5(4): 56-68.

[2] DAVISON A, REID I, MOLTON N, et al. Monoslam: real-time single camera SLAM[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2007, 29(6): 1052-1067.

[3] KENDALL A, GRIMES M, CIPOLLA R. PoseNet: a convolutional network for real-time 6-DOF camera relocalization[C]// IEEE International Conference on Computer Vision. Piscataway: IEEE, 2015: 2938-2946.

[4] KENDALL A, CIPOLLA R. Modelling uncertainty in deep learning for camera relocalization[C]// IEEE International Conference on Robotics and Automation. Piscataway: IEEE, 2016: 4762-4769.

[5] KLEIN G, MURRAY D. Parallel tracking and mapping for small AR workspaces[C]// 2007 6th IEEE and ACM International Symposium on Mixed and Augmented Reality. Piscataway: IEEE, 2007: 225-234.

[6] MUR-ARTAL R, MONTIEL J, TARDOS J D. ORB-SLAM: a versatile and accurate monocular SLAM system[J]. IEEE Transactions on Robotics, 2015, 31(5): 1147-1163.

[7] ENGEL J, SCHOPS T, CREMERS D. LSD-SLAM: large-scale direct monocular SLAM[C]// European Conference on Computer Vision. Berlin: Springer, 2014: 834-849.

[8] FORSTER C, PIZZOLI M, SCARAMUZZA D. SVO: fast semi-direct monocular visual odometry[C]// IEEE International Conference on Robotics and Automation. Piscataway: IEEE, 2014: 15-22.

[9] LABBE M, MICHAUD F. Online global loop closure detection for large-scale multi-session graph-based SLAM[C]// IEEE International Conference on Intelligent Robots and Systems. Piscataway: IEEE, 2014: 2661-2666.

[10] KERL C, STURM J, CREMERS D. Dense visual SLAM for RGB-D cameras[C]// IEEE International Conference on Intelligent Robots and Systems. Piscataway: IEEE, 2013: 2100-2106.

[11] SHENG C, PAN S G, ZENG P, et al. Monocular SLAM system in dynamic scenes based on semantic segmentation[C]// Image and Graphics. Berlin: Springer, 2019: 593-603.

[12] 张合新，徐慧，姚二亮，等. 动态场景下一种鲁棒的立体视觉里程计算法[J]. 仪器仪表学报，2018，39(9)：246-254.
ZHANG H X, XU H, YAO E L, et al. Robust stereo visual odometry algorithm in dynamic scenes[J]. Chinese Journal of Scientific Instrument, 2018, 39(9): 246-254. (in Chinese)

[13] 牛文雨，李文锋. 基于动态物体特征点去除的视觉里程计算法[J]. 计算机测量与控制，2019，27(10)：218-222.
NIU W Y, LI W F. Visual odometry algorithm based on dynamic object feature point removal[J]. Computer Measurement & Control, 2019, 27(10): 218-222. (in Chinese)

[14] 张华琰. 动态环境下仿人机器人视觉定位与运动规划方法研究[D]. 北京：北京建筑大学，2018.
ZHANG H Y. Research on visual localization and motion planning for humanoid robot in dynamic environment[D]. Beijing: Beijing University of Civil Engineering and Architecture, 2018. (in Chinese)

[15] COSTANTE G, MANCINI M, VALIGI P, et al. Exploring representation learning with CNNs for frame-to-frame ego-motion estimation[J]. IEEE Robotics and Automation Letters, 2016(1): 18-25.

[16] MOHANTY V, AGRAWAL S, DATTA S, et al. DeepVO: a deep learning approach for monocular visual odometry[EB/OL]. [2016-11-18]. https://arxiv.

org/abs/1611. 06069.

[17] 张林箭. 基于深度学习的相机相对姿态估计[D]. 杭州: 浙江大学, 2018.
ZHANG L J. Relative camera pose estimation using deep networks[D]. Hangzhou: Zhejiang University, 2018. (in Chinese)

[18] 苏健鹏, 黄影平, 赵柏淦, 等. 基于深度卷积神经网络的视觉里程计研究[J]. 光学仪器, 2020, 42(4): 33-40.
SU J P, HUANG Y P, ZHAO B G, et al. Research on visual odometry using deep convolution neural network [J]. Optical Instruments, 2020, 42(4): 33-40. (in Chinese)

[19] 张再腾, 张荣芬, 刘宇红. 一种基于深度学习的视觉里程计算法[J]. 激光与光电子学进展, 2021, 58(4): 0415001.
ZHANG Z T, ZHANG R F, LIU Y H. A visual odometry algorithm based on deep learning [J]. Laser & Optoelectronics Progress, 2021, 58(4): 0415001. (in Chinese)

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

[21] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. ImageNet classification with deep convolutional neural networks[C] // The 25th International Conference on Neural Information Processing Systems. Piscataway: IEEE, 2012: 1097-1105.

[22] ZHOU T H, BROWN M, SNAVELY N, et al. Unsupervised learning of depth and ego-motion from video [C]// IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2017: 1851-1860.

[23] ZHAN H Y, GARG R, WEERASEKERA C S, et al. Unsupervised learning of monocular depth estimation and visual odometry with deep feature reconstruction[C] // IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 340-349.

[24] YIN Z C, SHI J P. GeoNet: unsupervised learning of dense depth, optical flow and camera pose[C] // IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 1983-1992.

[25] RANJAN A, JAMPANI V, BALLES L, et al. Competitive collaboration: joint unsupervised learning of depth, camera motion, optical flow and motion segmentation[C]// IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2019: 12240-12249.

[26] LUO C X, YANG Z H, WANG P, et al. Every pixel counts + +: joint learning of geometry and motion with 3D holistic understanding[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 42(10): 2624-2641.

[27] ZHU A Z H, LIU W X, WANG Z Y, et al. Robustness meets deep learning: an end-to-end hybrid pipeline for unsupervised learning of egomotion[EB/OL]. [2018-09-20]. https: //arxiv. org/abs/1812. 08351.

[28] CASSER V, PIRK S, MAHJOURIAN R, et al. Depth prediction without the sensors: leveraging structure for unsupervised learning from monocular videos [C] // Thirty-Third AAAI Conference on Artificial Intelligence. Palo Alto, California: AAAI, 2019: 8001-8008.

[29] BIAN J W, LI Z C, WANG N Y, et al. Unsupervised scale-consistent depth and ego-motion learning from monocular video[C] // Advances in Neural Information Processing Systems. Piscataway: IEEE, 2019: 35-45.

[30] 马伟, 李曈, 龚超凡, 等. 结合 CNN 与分割约束的立体匹配算法[J]. 北京工业大学学报, 2019, 45(5): 413-420.
MA W, LI T, GONG C F, et al. Stereo matching with CNN and constraints from segmentation[J]. Journal of Beijing University of Technology, 2019, 45(5): 413-420. (in Chinese)

[31] 李曈, 马伟, 徐士彪, 等. 适应立体匹配任务的端到端深度网络[J]. 计算机研究与发展, 2020, 57(7): 1531-1538.
LI T, MA W, XU S B, et al. Task-adaptive end-to-end networks for stereo matching[J]. Journal of Computer Research and Development, 2020, 57(7): 1531-1538. (in Chinese)

[32] JADERBERG M, SIMONYAN K, ZISSERMAN A, et al. Spatial transformer networks[J]. Advances in Neural Information Processing Systems, 2015, 18(1): 2017-2025.

[33] WANG Z, BOVIK A C, SHEIKH H R, et al. Image quality assessment: from error visibility to structural similarity[J]. IEEE Transactions on Image Processing, 2004, 13(4): 600-612.

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

[35] GEIGER A, LENZ P, STILLER C, et al. Vision meets robotics: the KITTI dataset[J]. International Journal of Robotics Research, 2013, 32(1): 1231-1237.

[36] HE K, GKIOXARI G, DOLLAR P, et al. Mask R-CNN [C] // IEEE International Conference on Computer Vision. Piscataway: IEEE, 2017: 2961-2969.

(责任编辑 梁 洁)