控制与决策
Control and Decision

**结合改进密集模块深度估计网络和多视几何的视觉里程计**

彭道刚, 欧阳海林, 戚尔江, 王丹豪

引用本文:
彭道刚,欧阳海林,戚尔江,王丹豪. 结合改进密集模块深度估计网络和多视几何的视觉里程计[J]. 控制与决策, 2023, 38(4): 980−988.

在线阅读 View online: https://doi.org/10.13195/j.kzyjc.2021.1264

---

# 结合改进密集模块深度估计网络和多视几何的视觉里程计

彭道刚[1,2†], 欧阳海林[1], 戚尔江[1,2], 王丹豪[1]

(1. 上海电力大学 自动化工程学院，上海 200090；2. 上海发电过程智能管控工程技术研究中心, 上海 200090)

**摘 要:** 以多视图几何原理为基础,有效结合卷积神经网络进行图像深度估计和匹配筛选,构造无监督单目视觉里程计方法. 针对主流深度估计网络易丢失图像浅层特征的问题,构造一种基于改进密集模块的深度估计网络,有效地聚合浅层特征,提升图像深度估计精度. 里程计利用深度估计网络精确预测单目图像深度,利用光流网络获得双向光流,通过前后光流一致性原则筛选高质量匹配. 利用多视图几何原理和优化方式求解获得初始位姿和计算深度,并通过特定的尺度对齐原则得到全局尺度一致的6自由度位姿. 同时,为了提高网络对场景细节和弱纹理区域的学习能力,将基于特征图合成的特征度量损失结合到网络损失函数中. 在KITTI Odometry数据集上进行实验验证,不同阈值下的深度估计取得了85.9 %、95.8 %、97.2 % 的准确率. 在09和10序列上进行里程计评估,绝对轨迹误差在0.007 m. 实验结果验证了所提出方法的有效性和准确性,表明其在深度估计和视觉里程计任务上的性能优于现有方法.

**关键词:** 无监督深度学习；视觉里程计；深度估计；光流估计；多视图几何；密集模块

**中图分类号:** TP242.6 **文献标志码:** A

**DOI:** 10.13195/j.kzyjc.2021.1264

**引用格式:** 彭道刚,欧阳海林,戚尔江,等. 结合改进密集模块深度估计网络和多视几何的视觉里程计[J]. 控制与决策, 2023, 38(4): 980-988.

# Visual odometry combined with depth estimation network of improved dense block and multi-view geometry

*PENG Dao-gang*[1,2†], *OUYANG Hai-lin*[1], *QI Er-jiang*[1,2], *WANG Dan-hao*[1]

(1. College of Automation Engineering，Shanghai University of Electric Power，Shanghai 200090，China；2. Shanghai Engineering Research Center of Intelligent Management and Control for Power Process，Shanghai 200090，China)

**Abstract:** An unsupervised monocular visual odometry based on the principle of multi-view geometry and effective combination of the convolutional neural network for image depth estimation and correspondences selection is proposed. Aiming at the problem that mainstream depth estimation networks tend to lose the shallow features of images, a depth estimation network based on improved dense blocks is constructed to effectively aggregate shallow features and improve the accuracy of image depth estimation. The odometry uses the depth estimation network to accurately predict the depth of the monocular image, uses the optical flow network to obtain forward-backward optical flow, and select high-quality correspondences based on the principle of forward and backward optical flow consistency. The initial pose and calculated depth are obtained by using multi-view geometric principles and optimization methods, and a 6-degree-of-freedom pose with the fixed global scale is obtained through a specific scale alignment principle. At the same time, in order to improve the network's ability to learn scene details and the information of weak texture regions, the feature metric loss based on feature map synthesis is combined into the network loss function. On the KITTI Odometry dataset, the depth estimation under different thresholds has achieved accuracy rates of 85.9 %, 95.8 %, and 97.2 %, and the absolute trajectory error of the odometry evaluation on the 09 and 10 sequences is 0.007m. Experimental results show the effectiveness and accuracy of the proposed method, and prove that it is superior to the existing methods on the task of visual odometry.

**Keywords:** unsupervised deep learning；visual odometry；depth estimation；optical flow estimation；multi-view geometry；dense block

**收稿日期:** 2021-07-20；**录用日期:** 2022-01-28.

**基金项目:** 上海市“科技创新行动计划”高新技术领域项目(21511101800).

**责任编委:** 谢晖.

†通讯作者. E-mail: pengdaogang@126.com.

## 0 引　言

同时定位与建图(simultan eous localization and mapping, SLAM)是使机器人具备自主定位与导航能力的关键技术,视觉里程计作为经典视觉SLAM框架中的前端部分,通常利用相机获取的图像信息进行深度估计和位姿估计. 随着深度学习在图像处理等诸多计算机视觉任务中取得优异的性能,一系列研究表明,深度学习在单目视觉里程计任务中也能取得不俗的表现. 基于深度学习的单目图像深度估计和相机运动估计[1]通常分为有监督和无监督两类,前者需要图像真值进行网络训练,后者通过构造特定的损失函数作为监督信号训练网络. Konda等[2]提出了将视觉里程计看作归类问题,使用CNN (convolutional neural network)处理输入图像,实现视觉里程计; DeepVO[3]将循环神经网络与CNN结合,增加图像序列的时序建模,端到端地实现位姿估计. 上述监督学习方式的真值数据集获取成本较高,且现有的监督训练数据集数量有限,而无监督学习方式可自动构造监督信号,无需真值数据集,泛化性更好. 因此,基于无监督学习方式的单目视觉里程计在近些年受到了更为广泛的关注和研究. Zhou等[4]提出了利用视图合成构造监督信号,同时学习图像深度及相机位姿的无监督方法. 因此,主流无监督深度估计和相机运动估计方法[5]通过视图合成理论,以图像合成的光度损失构造监督信号.

以往无监督方法通常采用CNN设计网络提取图像特征,并基于图像光度损失构造监督信号. 基于CNN的框架在反向传播过程中易丢失浅层特征,无法关联全局信息. 光度损失在弱纹理、过曝或者低光照情况下易陷入局部极小值,无法获取充分的监督信号训练网络. 基于上述原因,以往方法不能获得准确的深度结果及贴合真实轨迹的位姿. Huang等[6]提出的DenseNet通过密集模块中层与层之间的特征直接拼接,保留了图像的语义信息和结构细节,提升了浅层特征的传递; Guo等[7]利用密集神经网络结合惯性传感器的方法,提出了一种轻量级的视觉里程计系统; Huang等[8]提出了一种利用预训练的DenseNet-121模型作为编码器的稀疏深度估计网络. 密集网络提取的特征优于CNN特征,包含更全面的场景结构,基于这一优势, Yu等[9]利用密集网络提取图像深度特征解决了自运动机器人SLAM中的闭环检测问题. 因此,本文采取轻量化思路[10],设计基于改进密集模块的深度估计网络和视觉里程计方法,并通过实验验证方法的有效性. 改进的密集模块重用特征图,将浅层的特征通过跳跃连接级联至更高层,有效地避免了浅层特征丢失导致的深度图模糊. 为了更全面地训练网络,本文添加一种基于特征图形式的视图合成损失,即特征度量损失[11],增加网络对弱纹理场景的学习能力. 除此之外,针对单目视觉里程计固有的尺度漂移问题,端到端的位姿估计网络完全忽视了图像间的几何信息,导致位姿轨迹精度存在不足. 本文利用光流网络获得双向光流,进行匹配筛选,通过多视图几何方法求解深度和初始位姿,并结合预测深度和计算深度构造尺度对齐原则,得到全局尺度一致的位姿估计. 本文的主要贡献如下:

1) 提出一种改进的密集模块并将其用于构造深度估计网络,复用浅层特征,恢复更丰富的图像结构信息,提升深度估计精度.

2) 将深度估计网络预测结果结合多视图几何原理,提出一种尺度对齐原则,得到尺度一致位姿估计,提高单目视觉里程计鲁棒性.

3) 添加新颖的特征度量损失训练整体系统,有效地提升网络对图像序列间特征信息的使用效率以及对场景结构的学习能力,提高位姿估计精度.

## 1 算法框架

本文构建的单目视觉里程计主要由深度估计网络、光流网络和特征图网络组成. 准确的深度估计是保证视觉里程计尺度一致的前提,而以往的CNN网络模型易丢失浅层特征,无法关联全局信息,造成深度估计精度不足. 本文利用改进的密集模块构造深度估计网络,提升网络训练过程中的信息传递,鼓励浅层特征复用,保留更精细的结构细节,获得准确的深度估计结果. 同时,利用光流网络生成相邻帧间双向光流,通过前后一致性原则筛选得到高质量2D-2D匹配. 结合预测深度,构造PnP问题并求解得到初始位姿. 考虑到单目视觉里程计的尺度漂移问题,利用三角测量得到的计算深度与深度估计网络输出的预测深度进行尺度对齐,固定全局尺度. 除此以外,本文基于视图合成理论,通过预训练的VGG网络生成单目图像特征图,添加一种新颖的基于特征图形式的特征度量损失用于训练整体网络,保证系统输出准确的深度和尺度一致位姿. 系统输入为图像序列中相邻两帧组成的图像对,设前一帧为源帧$\boldsymbol{I}_s$,后一帧为目标帧$\boldsymbol{I}_t$. 系统整体算法框架如图1所示.

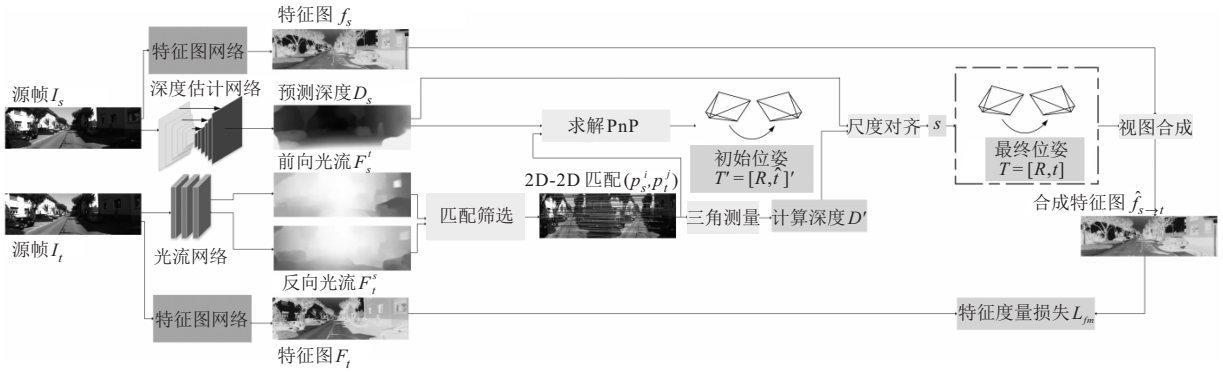



**图 1 系统框架**

### 1.1 基于改进密集模块的深度估计网络

#### 1.1.1 改进密集模块

图像浅层特征包含丰富的结构信息,因此,DenseNet将浅层特征与更高层的特征互补,复用浅层特征图. 在DenseNet中,当前层的特征都将通过跳跃连接级联至此后的每一层. 虽然这样尽可能地避免了特征丢失和反向传播过程中的梯度消失问题,但是由于过多的连接和特征图堆叠,导致特征冗余. 因此,本文结合密集网络和跳跃连接的优点,采取轻量化的思路设计4层改进密集模块,如图2所示.

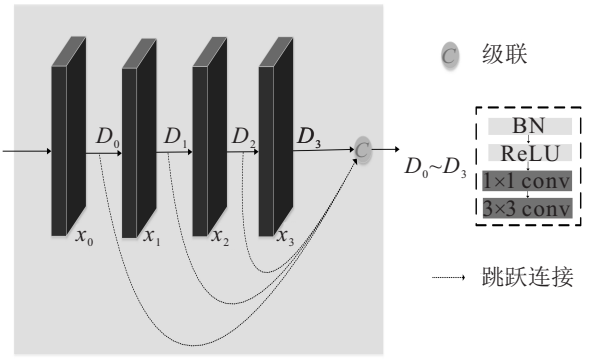



**图 2 改进密集模块**

具体的轻量化改进的密集模块中,无需将当前层的输出级联至往后的每一层,只需将其通过跳跃连接级联至最后一层输出即可. 因此,只有最后一层采取了与DenseNet相同的操作. 具体的:$\boldsymbol{x}_0 \sim \boldsymbol{x}_3$表示每一层的特征,根据密集模块的增长率决定每一层级联至最后一层的特征图的数量;$D_0 \sim D_3$为非线性变换函数. 定义如下:

$$\boldsymbol{x}_l = D_{l-1}(\boldsymbol{x}_0, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_{l-1}),\ l = 1,2,3,4, \quad (1)$$

其中$l$为卷积层的序号. 在每一层中,具体执行批量化归一操作(BN)、ReLU激活函数操作、1×1卷积、3×3卷积. 在改进密集模块中,每一层的非线性优化函数$D_{l-1}$都产生$k$个特征图,那么一个模块的输出就有$k_0+(l-1)\times k$个特征图作为下一模块的输入,其中$k_0$是模块输入层$l_0$的特征通道数. 轻量化的密集模块已足够保留场景的特征信息,在视觉里程计场景下,能够降低特征冗余,减小模型参数,相较于以往的CNN模型,保证了网络的灵活性和特征的持续性.

#### 1.1.2 深度估计网络结构设计

本文的深度估计网络采用编码器-解码器架构,如图3所示.

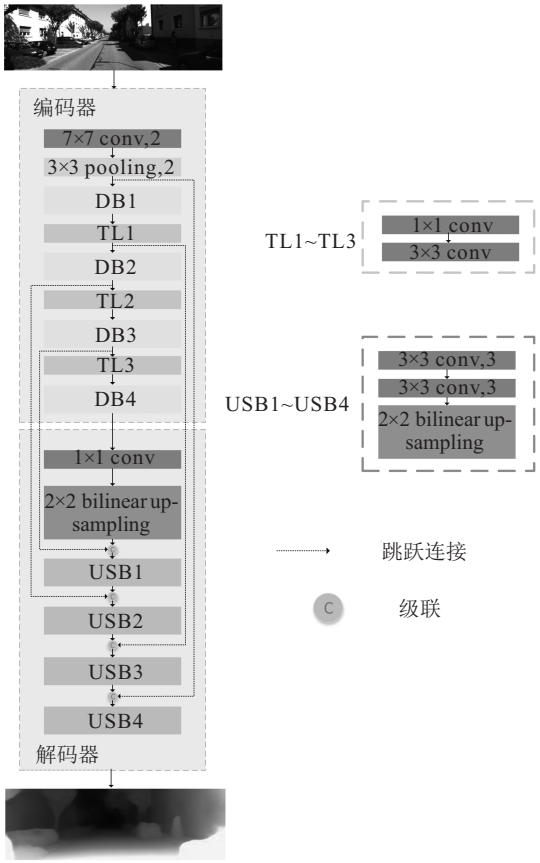



**图 3 深度估计网络**

编码器添加改进密集模块提升网络对场景细节信息的编码,复用浅层特征,将单目图像浅层特征和更高层特征以有效的方式连接,使得深度估计网络获取图像更精细的特征表达. 编码器依次由一个步长

为2的7×7卷积、一个步长为2的3×3池化,以及4个密集模块和3个过渡层(tansition layer, TL)组成;解码器依次由一个1×1的卷积、2×2的双线性上采样和4个上采样模块(up-sampling block, USB)组成. 其中:过渡层由1×1卷积与3×3卷积组成,上采样模块由两个3×3卷积与一个2×2双线性上采样组成. 在编码器与解码器之间添加了跳跃连接,保证特征传递的流通. 通过深度估计网络对单目图像进行深度估计,输出深度图,如图3所示.

### 1.2　基于光流的匹配筛选和位姿估计

#### 1.2.1　匹配提取及筛选

光流包含了丰富的场景信息,依据光流原理,对于输入光流网络的相邻图像对$(\boldsymbol{I}_s, \boldsymbol{I}_t)$,光流描述了$\boldsymbol{I}_s$中的像素运动,这给出了$\boldsymbol{I}_s$中$\boldsymbol{I}_t$的所有像素的对应关系,可以从光流中提取帧间2D-2D匹配. 因此,本文依靠光流网络生成的帧间双向光流提取特征匹配,通过光流中丰富的场景信息,获得鲁棒的匹配结果. 基于深度学习的光流网络已经受到广泛的研究和应用,本文采用轻量级光流网络LiteFlowNet[12]提取输入图像相邻帧的前向和后向光流. 考虑到并非所有像素都能找到一致的高精度匹配,并且对每一帧中所有像素计算稠密光流十分耗费资源,本文采用前后光流一致性准则[13],筛选优质的光流匹配. 基于一致性越高、匹配精度越高的前提,考虑相邻图像对$(\boldsymbol{I}_s, \boldsymbol{I}_t)$,设前向光流为$\boldsymbol{F}_s^t$,后向光流为$\boldsymbol{F}_t^s$,则光流一致性计算为

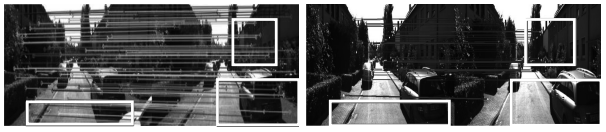
$$\boldsymbol{C} = -\boldsymbol{F}_s^t - w[\boldsymbol{F}_t^s, \boldsymbol{p}_f \boldsymbol{F}_s^t]. \tag{2}$$

其中:$\boldsymbol{p}_f$为前后向光流中像素的对应关系;$w$表示对光流中像素的变换操作,即

$$w[\boldsymbol{F}_t^s(\boldsymbol{p}), \boldsymbol{p}_f \boldsymbol{F}_s^t(\boldsymbol{p})] = \boldsymbol{F}_t^s[\boldsymbol{p} + \boldsymbol{F}_s^t(\boldsymbol{p})], \tag{3}$$

这里$\boldsymbol{p}$为图像中像素点. 计算光流一致性后,通过Best-$N$选取策略,筛选出不一致性较小、表现最好的$N$对2D-2D匹配$(\boldsymbol{p}_s^i, \boldsymbol{p}_t^j)$,$i$和$j$表示像素上标索引. 依据实践经验,本文$N$选取2 500.

本文方法的匹配结果如图4(a)所示,并与传统人工设计的ORB特征提取匹配结果进行对比.



(a) 双向光流匹配筛选　　(b) ORB提取匹配筛选

**图4　匹配对比**

从图4中白色框选对比区域可以看出,传统ORB特征提取匹配结果偏向图像中光度较亮和无遮挡区域,且匹配分布不均匀. 本文筛选后所得的结果更为清晰、均匀且包含更多场景信息,提供了丰富和鲁棒的匹配.

#### 1.2.2　位姿计算和尺度对齐

2D-2D匹配求解位姿的对极几何方法需要至少8个点对,且存在初始化和纯旋转问题. 因此,本文通过上述构造的深度估计网络获得源帧$\boldsymbol{I}_s$中匹配点处的预测深度$\boldsymbol{D}_s$,结合$\boldsymbol{I}_t$中对应的2D匹配点生成3D-2D匹配,构造PnP (perspective-n-point)问题. 然后,通过非线性优化的方式求解PnP,计算相机初始位姿在李群SE(3)上的表示$\boldsymbol{T}'$. 考虑某个经光流前后一致性筛选出的空间点,它的齐次坐标为$\boldsymbol{p}_i = (X_i, Y_i, Z_i)^{\mathrm{T}}$,它在$\boldsymbol{I}_s$中的归一化平面齐次投影点坐标为$\boldsymbol{u}_i = (u_i, v_i)^{\mathrm{T}}$. 由单目相机的理论模型可知,二者的关系如下:

$$s_i \boldsymbol{u}_i = \boldsymbol{K}\boldsymbol{T}'\boldsymbol{p}_i. \tag{4}$$

其中:$s_i$为该3D点的深度,$\boldsymbol{K}$为单目相机内参矩阵. 由于相机观测与实际计算存在一定误差,将该3D点由初始位姿变换$\boldsymbol{T}'$重投影至$\boldsymbol{I}_t$,重投影处的像素位置与$\boldsymbol{I}_t$中的观测位置存在误差,即重投影误差. 基于此构建非线性优化问题如下:

$$\boldsymbol{T}^* = \arg\min_{\boldsymbol{T}} \frac{1}{2} \sum_{i=1}^{n} \left\| \boldsymbol{u}_i - \frac{1}{s_i} \boldsymbol{K}\boldsymbol{T}'\boldsymbol{p}_i \right\|_2^2. \tag{5}$$

使用李代数构建无约束优化,通过BA (bundle adjustment)进行求解,最终获得初始位姿$\boldsymbol{T}' = [\boldsymbol{R}, \hat{\boldsymbol{t}}]$. 由于单目视觉相机在建模过程中丢失了图像平移尺度,在位姿轨迹中易出现尺度漂移现象,初始位姿中平移的尺度需要得到对齐.

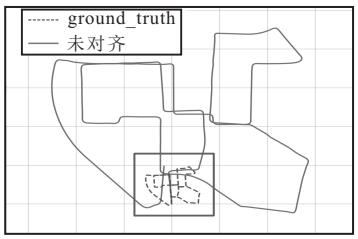
传统单目视觉里程计中,通常将当前帧与上一帧中通过三角测量得到的3D路标进行尺度对齐,而这一操作将随着相机运动累积误差,导致全局尺度漂移. 如图5(a)所示,估计位姿的轨迹形状与真值保持一致,但由于未能将平移尺度固定,导致轨迹的全局尺度陷入较大误差中. 本文通过深度估计网络预测输入图像深度$\boldsymbol{D}$,作为参考值. 利用三角化计算匹配点对深度值$\boldsymbol{D}'$,与参考值$\boldsymbol{D}$进行对齐. 本文设计的对齐原则如下:第0帧图像在第$i$处的预测深度为$\boldsymbol{D}_0^i$,通过三角测量获得的对应的计算深度为$\boldsymbol{D}_0^{i'}$,则令

$$s_0 = \frac{1}{N} \sum_{i=1}^{N} \frac{|\boldsymbol{D}_0^i - \boldsymbol{D}_0^{i'}|}{\boldsymbol{D}_0^i} \tag{6}$$

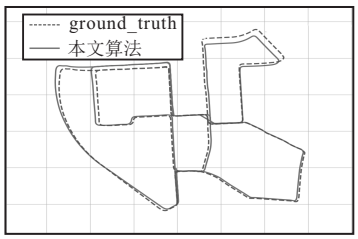
作为初始尺度对齐因子. 而后,对每一帧执行相同的操作,即

$$s_t = \frac{1}{N}\sum_{i=1}^{N}\frac{|\boldsymbol{D}_t^i - \boldsymbol{D}_t^{i'}|}{\boldsymbol{D}_t^i},\ 0<t<M, \tag{7}$$

其中$M$为图像帧的数量. 若相邻图像序列$\boldsymbol{I}_s$与$\boldsymbol{I}_t$间的尺度对齐因子保持一致,即认为$s_s \approx s_t$,则随着帧间匹配尺度对齐的约束传递,位姿的全局尺度得到固定,如图5(b)所示.




(a) 尺度未对齐




(b) 尺度对齐

**图5 尺度对齐**

为了进一步加强位姿尺度的固定,将尺度对齐因子加入深度一致性损失中,作为深度估计网络训练过程的监督,进行尺度矫正,从而固定全局尺度. 里程计最终输出尺度一致的位姿$\boldsymbol{T}=[\boldsymbol{R},\boldsymbol{t}]$.

算法整体伪代码如表1所示.

**表1 算法伪代码列表**

输入: 图像序列$[\boldsymbol{I}_0,\boldsymbol{I}_1,\ldots,\boldsymbol{I}_n]$;
输出: 相机位姿$[\boldsymbol{T}_0,\boldsymbol{T}_1,\ldots,\boldsymbol{T}_n]$.
step 1: 初始化$\boldsymbol{T}_0=\boldsymbol{I}, i=0$,
while $i<n$;
step 2: $\boldsymbol{I}_s=\boldsymbol{I}_i, \boldsymbol{I}_t=\boldsymbol{I}_{i+1}$;
step 3: 获取预测深度值$\boldsymbol{D}_s$,前向光流$\boldsymbol{F}_s^t$,后向光流$\boldsymbol{F}_t^s$,特征图$\boldsymbol{f}_s$、$\boldsymbol{f}_t$;
step 4: 计算光流一致性$(\boldsymbol{F}_s^t,\boldsymbol{F}_t^s)$;
step 5: 筛选$N$对2D-2D匹配$(\boldsymbol{p}_s^i,\boldsymbol{p}_t^j)$;
step 6: 通过BA求解PnP问题,得到初始位姿$\boldsymbol{T}_s^{t'}=[\boldsymbol{R},\hat{t}]$;
step 7: 三角测量计算的计算深度$\boldsymbol{D}_s^{'}$;
step 8: 计算深度与预测深度对齐固定尺度因子$s$;
step 9: 得到尺度固定位姿$\boldsymbol{T}_i=[\boldsymbol{R},s\hat{t}]$;
end

## 2 损失函数设计

弱纹理区域的像素具有十分相近的光度值,因此光度误差容易陷入局部极小值导致错误的深度估计和位姿估计结果,本文添加一种新颖的特征度量损失来训练系统网络. 具体的,本文利用预训练VGG架构的特征图提取网络,对源帧$\boldsymbol{I}_s$和目标帧$\boldsymbol{I}_t$提取相对应的特征图$\boldsymbol{f}_s$和$\boldsymbol{f}_t$,在图像的特征图表示形式下基于视图合成理论构造新颖的特征度量损失,提升网络对场景细节的学习能力. 除此之外,为了进一步保证深度估计的准确性,使用深度一致性损失进行系统网络训练. 网络整体损失函数将由特征度量损失、光度损失和深度一致性损失组成.

### 2.1 特征度量损失

在弱纹理区域,像素间较小的光度差异无法保证网络深度和位姿估计的精度. 可以在网络损失中添加一阶平滑损失或二阶平滑损失增加深度的传播,但这通常会造成边界区域的过度平滑. 本文通过预训练模型输出深度特征图,对每个像素使用特征形式表示,根据深度特征图重建定义特征度量损失,即使在无纹理区域也明确限制其具有判别性,有效地弥补了光度损失或平滑损失的不足. 根据视图合成理论,设$\boldsymbol{p}_s$为源帧中某一像素的齐次坐标,$\hat{\boldsymbol{p}}_t$为合成帧中对应像素的齐次坐标,则根据相机位姿和深度值可得

$$\hat{\boldsymbol{p}}_t(\boldsymbol{K},\boldsymbol{D}_s,\boldsymbol{T}_s^t)=\boldsymbol{K}\boldsymbol{T}_s^t\boldsymbol{K}^{-1}\boldsymbol{p}_s\boldsymbol{D}_s(\boldsymbol{p}_s). \tag{8}$$

其中:$\boldsymbol{K}$为相机的内参矩阵,$\boldsymbol{D}_s$为源帧中网络输出的像素深度,$\boldsymbol{T}_s^t$为源帧至目标帧的位姿矩阵. 由式(8)通过源帧的特征图可得合成特征图$\hat{\boldsymbol{f}}_s$,则特征图合成损失如下:

$$L_f=|\phi_s(\hat{\boldsymbol{p}})-\phi_t(\boldsymbol{p})|_1, \tag{9}$$

其中$\phi(\boldsymbol{p})$为图像的特征表示. 为了保证网络在场景的弱纹理区域依然能学习到梯度变化较大的特征,使用判别式损失函数

$$L_d=-\sum_{\boldsymbol{p}}\mathrm{e}^{-|\nabla^1\boldsymbol{I}(\boldsymbol{p})|_1}|\nabla^1\boldsymbol{\phi}(\boldsymbol{p})|_1, \tag{10}$$

其中$\nabla^1$为一阶微分算子. 为了平滑特征梯度,引入对二阶梯度的惩罚项,即收敛损失

$$L_c=\sum_{\boldsymbol{p}}|\nabla^2\boldsymbol{\phi}(\boldsymbol{p})|_1, \tag{11}$$

其中$\nabla^2$为二阶微分算子. 综上,特征度量损失$L_{fm}$如下:

$$L_{fm}=\lambda_f L_f+\lambda_d L_d+\lambda_c L_c. \tag{12}$$

其中:$\lambda_f$为特征图合成损失权重,按经验取值,设置为1;$\lambda_d$为深度损失权重,参考文献[11],设置为1e-3;$\lambda_c$为收敛损失权重,参考文献[11],设置为1e-3.

### 2.2 光度损失

基于视图合成理论,光度损失以合成帧$\hat{\boldsymbol{I}}_s$与目标帧$\boldsymbol{I}_t$之间的光度误差作为监督信号训练网络,采用如下形式:

$$L_{ph}=0.85(1-\mathrm{SSIM}[\boldsymbol{I}_s(\hat{\boldsymbol{p}})-\boldsymbol{I}_t(\boldsymbol{p})])/2+$$

$$0.15\sum_{\boldsymbol{p}}|\phi_s(\hat{\boldsymbol{p}})-\phi_t(\boldsymbol{p})|_1. \tag{13}$$

其中: $\boldsymbol{p}$为图像像素; SSIM[14]为鲁棒的图像相似性评估函数,具体形式为

$$\mathrm{SSIM}(x,y)=\frac{(2u_xu_y+c_1)(2\sigma_{xy}+c_2)}{(u_x^2+u_y^2+c_1)(\sigma_x^2+\sigma_y^2+c_2)}. \tag{14}$$

这里: $x$和$y$分别为原始图和对比图, $u_x$和$\sigma_x^2$为$x$的平均值和方差, $u_y$和$\sigma_y^2$为$y$的平均值和方差, $\sigma_{xy}$为二者的协方差; $c_1$和$c_2$为稳定常数,取$0.01^2$和$0.03^2$.

### 2.3　深度一致性损失

为了进一步巩固里程计尺度一致性,令$\boldsymbol{I}_s$的预测深度为$\boldsymbol{D}_s$, $\boldsymbol{I}_t$的预测深度为$\boldsymbol{D}_t$,由式(8)可得$\hat{\boldsymbol{I}}_s$的深度为$\hat{\boldsymbol{D}}_s$. 结合预测深度与计算深度对齐所得尺度对齐因子,采用逆深度的形式构造深度一致性损失,形式为

$$L_{dc}(\boldsymbol{D}_t,\hat{\boldsymbol{D}}_s)=\left|\frac{1}{s\boldsymbol{D}_t}-\frac{1}{\hat{\boldsymbol{D}}_s}\right|. \tag{15}$$

综上,系统整体损失由特征度量损失、光度损失和深度一致性损失组成,即

$$L=\lambda_{fm}L_{fm}+\lambda_{ph}L_{ph}+\lambda_{dc}L_{dc}. \tag{16}$$

其中: $\lambda_{ph}$为光度损失权重,按经验取值,设置为5; $\lambda_{dc}$为深度一致性损失权重,按经验取值,设置为1; $\lambda_{fm}$为本文引入的特征度量损失权重. 随着训练迭代次数增加,系统总体损失逐渐减小. 因此,针对不同$\lambda_{fm}$取值下系统收敛时的迭代次数进行实验对比,取最优表现的权重作为$\lambda_{fm}$的值. 从表2中结果可以看出, $\lambda_{fm}$设置为1.

**表2　迭代次数比较**

| $\lambda_{fm}$ | 迭代次数 |
|---|---|
| 0.5 | $1.80\times10^5$ |
| 1 | $1.57\times10^5$ |
| 1.5 | $1.65\times10^5$ |

## 3　实验评估

### 3.1　实验平台参数

本文采用PyTorch框架实现网络架构,为了有效地进行网络训练和优化,获得准确的网络参数,采用高性能NVIDIA Quadro P4000 GPU训练系统网络. 同时,工作站配有Intel Xeon W-2123 3.6 GHz CPU,安装Ubuntu 16.04操作系统. 本文使用Adam算法进行网络优化. 在网络训练参数中,学习率影响网络训练收敛的速度和精度,较大的学习率易导致目标函数波动过大无法收敛甚至发散,较小的学习率则导致网络收敛速度过慢. 在Adam算法中,学习率控制权重的更新比率,根据经验值,本文将学习率设置为$\alpha=$ 1e-4. Adam算法计算梯度的指数移动均值,一阶矩估计衰减率和二阶矩估计衰减率控制移动均值的衰减率,通常设置为接近1的值,使得矩估计的偏差接近0. 因此,1阶矩估计指数衰减率设置为$\beta_1=0.9$, 2阶矩估计指数衰减率设置为$\beta_2=0.999$. 网络训练的批规模是每次迭代训练中使用的样本数量,影响训练的速度和模型收敛,按经验值设置为4. 密集模块增长率表示密集模块中每一层生成的特征图数量,考虑视觉里程计任务的实用性,一个较小的增长率即能获得不错的效果,将密集模块的增长率$k$设置为12.

采用KITTI Odometry数据集作为实验数据来源. KITTI Odometry数据集是目前最为广泛使用的里程计测试评估基准,由安装在行驶的汽车上的相机以每秒10帧/s的速率采集到的图像序列组成,其中包含了众多具有挑战性的场景,例如十字路口和街道. 该数据集包含了22组图像序列,其中00～10序列提供了轨迹真值. 本文使用KITTI Odometry数据集中的00～08序列进行网络训练,09和10序列进行测试. 数据集中原始图像大小均调整为640×480.

与文献[10]提出的RDenseCNN-12-147进行模型复杂度比较,如表3所示. 通过计算量和访存量两个指标进行模型的时间复杂度和空间复杂度分析. 计算量即FLOPs,为浮点数运算次数; 访存量即Parameters,为模型的参数数量. 在客观上,进一步分析了本文方法在降低计算参数上的优势,体现了改进密集模块的轻量化.

**表3　模型复杂度比较**

| 方法 | FLOPs | Parameters |
|---|---|---|
| RDenseCNN-12-147[10] | 110.8 M | 1.10 M |
| 本文算法 | 78.2 M | 0.57 M |

### 3.2　单目图像深度估计精度评估

本文在同一实验机器上采用文献[15]方法对KITTI划分后的子集(共697幅图像)进行实验,将本文方法与现有的监督方法、无监督立体图像训练方法、无监督单目图像训练方法进行评价对比. 评价指标包括绝对相对误差(absolute relative difference, Abs Rel)、平方相对误差(squared relative difference, Sq Rel)、均方根误差(root mean squared error, RMSE)和对数均方根误差(RMSE lg),以及在阈值分别为$\delta<1.25$、$\delta<1.25^2$和$\delta<1.25^3$时的准确率,以下为每个指标的计算公式:

$$\mathrm{Abs\,Rel}=\frac{1}{N}\sum_{i=1}^{N}\frac{|d_i-d_i^*|}{d_i}, \tag{17}$$

$$\mathrm{Sq\,Rel}=\frac{1}{N}\sum_{i=1}^{N}\frac{|d_i-d_i^*|^2}{d_i}, \tag{18}$$

$$\mathrm{RMSE}=\sqrt{\frac{1}{N}\sum_{i=1}^{N}|d_i-d_i^*|^2},\tag{19}$$

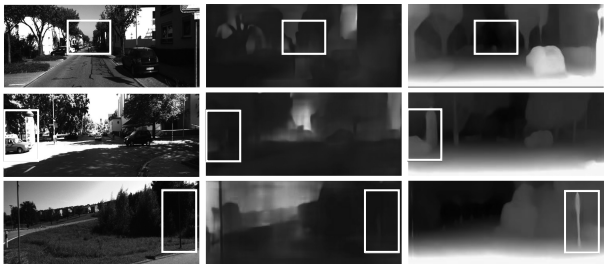
$$\mathrm{RMSE\,lg}=\sqrt{\frac{1}{N}\sum_{i=1}^{N}\frac{|\lg d_i-\lg d_i^*|}{d_i}}.\tag{20}$$

其中: $d_i$ 为真实的深度值, $d_i^*$ 为预测深度值. 在比较真实深度值与预测深度值时, 统计使 $\frac{d_i}{d_i^*}$ 和 $\frac{d_i^*}{d_i}$ 中二者最大值小于阈值的像素点占总体像素点的百分比, 越接近1表明估计效果越好. 通常, 阈值取1.25、$1.25^2$ 和 $1.25^3$. 因此, 阈值越大情况下, 所得结果越好, 即阈值为 $1.25^3$ 时的结果要好于阈值为1.25时的结果. KITTI Odometry数据集上深度估计结果比较如表4所示. 从表4结果可看出, 在定量评估中, 本文方法均优于主流深度估计算法.

**表4 KITTI Odometry数据集上深度估计结果比较**

| 方法 | 学习方式 | 误差(越低越好) | | | | 准确率(越高越好) | | |
|---|---|---|---|---|---|---|---|---|
| | | Abs Rel | Sq Rel | RMSE | RMSE lg | $\delta<1.25$ | $\delta<1.25^2$ | $\delta<1.25^3$ |
| 文献[15](粗) | 监督 | 0.214 | 1.605 | 6.563 | 0.292 | 0.673 | 0.884 | 0.957 |
| 文献[15](精) | 无监督 | 0.203 | 1.548 | 6.307 | 0.282 | 0.702 | 0.890 | 0.958 |
| 文献[16] | 无监督 | 0.151 | 1.226 | 5.849 | 0.246 | 0.784 | 0.921 | 0.967 |
| 文献[17] | 无监督 | 0.148 | 1.344 | 5.927 | 0.247 | 0.803 | 0.922 | 0.964 |
| 文献[18] | 无监督 | 0.144 | 1.391 | 5.869 | 0.241 | 0.803 | 0.928 | 0.969 |
| 文献[4] | 无监督 | 0.208 | 1.768 | 6.856 | 0.283 | 0.678 | 0.885 | 0.957 |
| 文献[19] | 无监督 | 0.182 | 1.481 | 6.501 | 0.267 | 0.725 | 0.906 | 0.963 |
| 本文算法 | 无监督 | **0.139** | **1.191** | **5.549** | **0.240** | **0.859** | **0.958** | **0.972** |

为评估密集模块的有效性, 在相同的数据集和实验环境下, 将本文方法与全卷积残差网络(FCRN)[20]进行比较, 并将结果可视化, 对比结果如图6所示.



(a) 输入图 (b) FCRN (c) 本文方法

**图6 深度估计结果对比**

FCRN在单目深度估计中做出了一定的贡献. 该网络也采用了编码器-解码器架构, 通过添加预训练的ResNet50结构, 设计了残差(tes-nlock), 达到更深的网络深度, 取得了较好的估计精度. 从对比结果可以看出, 基于本文改进密集模块的深度估计网络取得了更好的表现. 在图6(a)的第1辐图中, 离视角较远范围的物体在本文方法结果中也能显示出较为清晰的轮廓, 优于FCRN输出的结果. 从图6(a)第2幅和第3幅的框选区域可以看出, 在曝光过度或者背光阴影处, 由于图像光度相似性较强, FCRN无法正确估计图像深度, 结果模糊, 而本文方法仍然能清晰恢复该区域的物体深度轮廓.

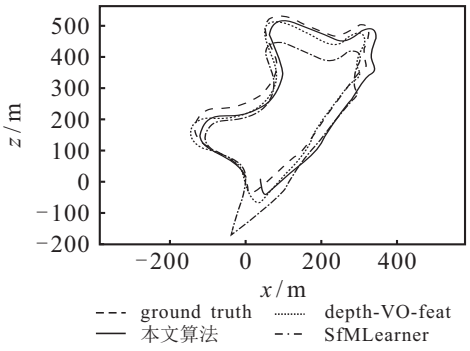
除此之外, 针对本文提出的特征度量损失, 为了验证其对深度估计网络训练的有效性, 在相同的网络结构设计下, 对不同损失组合训练下的网络进行深度估计结果对比. $t_{\mathrm{err}}$ 表示平均平移均方根误差, 以"%"度量, $r_{\mathrm{err}}$ 表示平均旋转均方根误差, 以"/(100 m)"度量, 如表5所示. 表5中数据显示, 相比于仅使用光度损失, 深度一致性损失巩固了里程计系统对尺度的估计, 特征度量损失加大了对错误估计的惩罚, 进一步提升了深度估计性能.

**表5 不同损失函数组合深度估计结果比较**

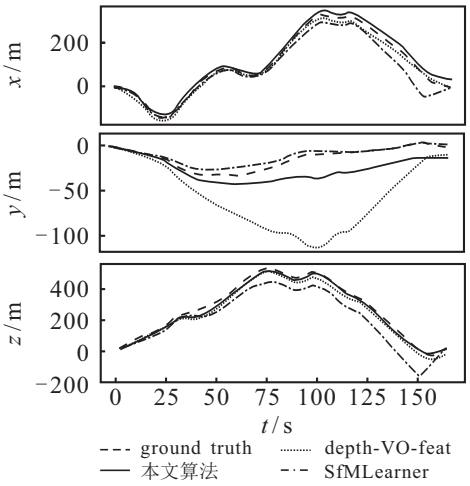
| 损失 | 误差 | | | | 准确率 | | | 序列09 | | 序列10 | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Abs Rel | Sq Rel | RMSE | RMSE lg | $\delta<1.25$ | $\delta<1.25^2$ | $\delta<1.25^3$ | $t_{\mathrm{err}}$ | $r_{\mathrm{err}}$ | $t_{\mathrm{err}}$ | $r_{\mathrm{err}}$ |
| $L_{ph}$ | 0.174 | 1.243 | 5.682 | 0.317 | 0.595 | 0.817 | 0.901 | 3.02 | 0.91 | 2.05 | 1.11 |
| $L_{ph}+L_{dc}$ | 0.159 | 1.230 | 5.594 | 0.270 | 0.610 | 0.856 | 0.947 | 2.74 | 0.87 | 2.12 | 1.01 |
| $L_{ph}+L_{fm}$ | 0.154 | 1.198 | 5.551 | 0.258 | 0.638 | 0.923 | 0.960 | 2.56 | 0.84 | 1.99 | 0.95 |
| $L_{fm}+L_{ph}+L_{dc}$ | **0.139** | **1.191** | **5.549** | **0.240** | **0.859** | **0.958** | **0.972** | **2.03** | **0.78** | **1.98** | **0.81** |

### 3.3 视觉里程计精度评估

为了评估本文提出的视觉里程计方法性能, 采用Wang等[3]提出的评估方法对本文方法与主流方法进行评估. 采用提供真实轨迹的KITTI Odometry数据集中的09和10序列进行测试, 二者的长度分别为1 591帧和1 201帧, 测试结果如图7和图8所示.

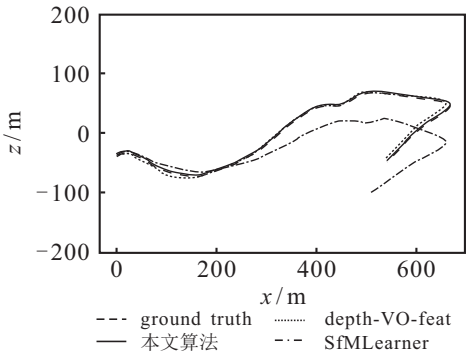



(a) 轨迹对比

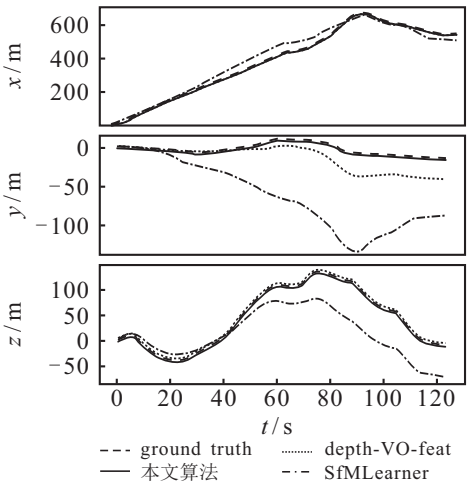



(b) 误差对比

**图 7　序列09上的轨迹估计**




(a) 轨迹对比




(b) 误差对比

**图 8　序列10上的轨迹估计**

由图7和图8可知,本文方法得到的位姿估计轨迹更贴合真实轨迹,且在$x$、$y$、$z$方向上的结果都与真值更为贴近.

除此之外,本文采用绝对轨迹误差(absolute trajectory error, ATE)作为位姿估计精度的衡量指标,数值越小精度越高,位姿估计结果越准确. 如表6所示,数据对比结果进一步验证了本文方法优于其他方法.

**表 6　KITTI Odometry数据集上绝对轨迹误差**

| 方法 | 序列09/m | 序列10/m |
|---|---|---|
| ORB-SLAM[21] | 0.014±0.008 | 0.012±0.011 |
| SfMLearner[4] | 0.021±0.017 | 0.020±0.015 |
| SGANVO[22] | 0.015±0.006 | 0.014±0.009 |
| LSTMVO[23] | 0.014±0.007 | 0.012±0.008 |
| depth-VO-feat[18] | 0.013±0.009 | 0.013±0.008 |
| GeoNet[13] | 0.012±0.007 | 0.012±0.009 |
| Monodepth2[24] | 0.017±0.008 | 0.015±0.010 |
| 本文算法 | **0.007±0.004** | **0.007±0.002** |

## 4　结　论

本文提出了一种基于改进密集模块的深度估计和无监督单目视觉里程计方法,并通过实验验证了其有效性. 本文将简化密集模块融入深度估计网络中,提升了网络对图像浅层特征的复用. 进一步,在系统网络训练中添加了特征度量损失,提升了网络对场景细节的学习能力和在弱纹理区域下的鲁棒性. 针对单目视觉里程计存在的尺度漂移问题,通过光流网络筛选匹配点对,结合预测深度构造PnP问题,利用非线性优化方式求解位姿,并通过本文提出的尺度对齐原则固定尺度,最终得到全局尺度一致位姿估计. 将本文方法在KITTI数据集上进行了验证,不同阈值下的深度估计准确率分别为85.9%、95.8%、97.2%,在09和10序列上里程计绝对轨迹误差均在0.007 m左右. 实验表明,本文方法在深度估计和位姿估计性能上均优于主流方法. 未来,计划对网络资源消耗优化以及场景存在大量快速移动物体的问题进行研究处理,进一步提高里程计性能.

**参考文献**(References)

[1] Zhou H Z, Ummenhofer B, Brox T. DeepTAM: Deep tracking and mapping with convolutional neural networks[J]. International Journal of Computer Vision, 2020, 128(3): 756-769.

[2] Konda K, Memisevic R. Learning visual odometry with a convolutional network[C]. Proceedings of the 10th International Conference on Computer Vision Theory and Applications. Berlin, 2015: 486-490.

[3] Wang S, Clark R, Wen H K, et al. DeepVO: Towards end-to-end visual odometry with deep

recurrent convolutional neural networks[C]. 2017 IEEE International Conference on Robotics and Automation. Singapore, 2017: 2043-2050.

[4] Zhou T H, Brown M, Snavely N, et al. Unsupervised learning of depth and ego-motion from video[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR). Piscataway: IEEE, 2017: 6612-6619.

[5] Bian J W, Li Z C, Wang N Y, et al. Unsupervised scale-consistent depth and ego-motion learning from monocular video[J/OL]. 2019, arXiv: 1908.10553.

[6] Huang G, Liu Z, van der Maaten L, et al. Densely connected convolutional networks[C]. 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, 2017: 2261-2269.

[7] Guo Z B, Yang M K, Chen N H, et al. LightVO: Lightweight inertial-assisted monocular visual odometry with dense neural networks[C]. 2019 IEEE Global Communications Conference. Waikoloa, 2019: 1-6.

[8] Huang K, Qu X T, Chen S Q, et al. Superb monocular depth estimation based on transfer learning and surface normal guidance[J]. Sensors, 2020, 20(17): 4856-4878.

[9] Yu C, Liu Z X, Liu J X, et al. Dense-loop: A loop closure detection method for visual SLAM using DenseNet features[C]. The 1st International Workshop on the Semantic Descritor, Semantic Modelingard Mapping for Humanlike Perceptionand Navigation of Mobile Robots towards Large Scale Long-Term Autonomy (SDMM19). Piscataway: IEEE, 2019: 27-37.

[10] Fooladgar F, Kasaei S. Lightweight residual densely connected convolutional neural network[J]. Multimedia Tools and Applications, 2020, 79(35/36): 25571-25588.

[11] Shu C, Yu K, Duan Z X, et al. Feature-metric loss for self-supervised learning of depth and egomotion[M]. Computer Vision-ECCV 2020. Cham: Springer International Publishing, 2020: 572-588.

[12] 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. Salt Lake City, 2018: 8981-8989.

[13] Yin Z C, Shi J P. GeoNet: Unsupervised learning of dense depth, optical flow and camera pose[C]. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, 2018: 1983-1992.

[14] 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.

[15] Eigen D, Puhrsch C, Fergus R. Depth map prediction from a single image using a multi-scale deep network[J/OL]. 2014, arXiv: 1406.2283.

[16] Garg R, Vijay Kumar B G, Carneiro G, et al. Unsupervised CNN for single view depth estimation: Geometry to the rescue[C]. The 14 th European Conference on Computer Vision. Berlin, 2016: 740-756.

[17] Godard C, Aodha O M, Brostow G J. Unsupervised monocular depth estimation with left-right consistency[C]. 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, 2017: 6602-6611.

[18] Zhan H Y, Garg R, Weerasekera C S, et al. Unsupervised learning of monocular depth estimation and visual odometry with deep feature reconstruction[C]. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, 2018: 340-349.

[19] Yang Z H, Wang P, Xu W, et al. Unsupervised learning of geometry with edge-aware depth-normal consistency[J/OL]. 2017, arXiv: 1711.03665.

[20] Laina I, Rupprecht C, Belagiannis V, et al. Deeper depth prediction with fully convolutional residual networks[C]. 2016 Fourth International Conference on 3D Vision (3DV). Stanford, 2016: 239-248.

[21] Mur-Artal R, Montiel J M M, Tardós J D. ORB-SLAM: A versatile and accurate monocular SLAM system[J]. IEEE Transactions on Robotics, 2015, 31(5): 1147-1163.

[22] 叶星余, 何元烈, 汝少楠. 基于生成式对抗网络及自注意力机制的无监督单目深度估计和视觉里程计[J]. 机器人, 2021, 43(2): 203-213.
(Ye X Y, He Y L, Ru S N. Unsupervised monocular depth estimation and visual odometry based on generative adversarial network and self-attention mechanism[J]. Robot, 2021, 43(2): 203-213.)

[23] 陈宗海, 洪洋, 王纪凯, 等. 基于循环卷积神经网络的单目视觉里程计[J]. 机器人, 2019, 41(2): 147-155.
(Chen Z H, Hong Y, Wang J K, et al. Monocular visual odometry based on recurrent convolutional neural networks[J]. Robot, 2019, 41(2): 147-155.)

[24] Godard C, Aodha O M, Firman M, et al. Digging into self-supervised monocular depth estimation[C]. 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, 2019: 3827-3837.

## 作者简介

彭道刚 (1977−), 男, 教授, 博士, 从事低碳智能发电、电力巡检智能机器人等研究, E-mail: pengdaogang@126.com;

欧阳海林 (1994−), 男, 硕士生, 从事同时定位与建图、电力巡检智能机器人定位导航的研究, E-mail: yanghailin1221@163.com;

戚尔江 (1991−), 男, 硕士, 从事电力巡检智能机器人、嵌入式软件开发的研究, E-mail: xinbdzh@163.com;

王丹豪 (1992−), 男, 博士生, 从事综合能源系统、火电运行优化、能源互联网的研究, E-mail: damhao.wang@qq.com.

(责任编辑：闫 妍)