第 58 卷 第 4 期/2021 年 2 月/激光与光电子学进展 研究论文

激光与光电子学进展 先进成像

一种基于深度学习的视觉里程计算法

张再腾，张荣芬，刘宇红*

贵州大学大数据与信息工程学院，贵州 贵阳 550025

摘要 近年来，视觉里程计广泛应用于机器人和自动驾驶等领域，传统方法求解视觉里程计需基于特征提取、特征匹配和相机校准等复杂过程，同时各个模块之间要耦合在一起才能达到较好的效果，且算法的复杂度较高。环境噪声的干扰以及传感器的精度会影响传统算法的特征提取精度，进而影响视觉里程计的估算精度。鉴于此，提出一种基于深度学习并融合注意力机制的视觉里程计算法，该算法可以舍弃传统算法复杂的操作过程。实验结果表明，所提算法可以实时地估计相机里程计，并具有较高的精度和稳定性以及较低的网络复杂度。

关键词 机器视觉；深度学习；视觉里程计；注意力机制；多任务学习

中图分类号 TP391.4；TP181 **文献标志码** A **doi**：10.3788/LOP202158.0415001

Visual Odometry Algorithm Based on Deep Learning

Zhang Zaiteng, Zhang Rongfen, Liu Yuhong*

College of Big Data and Information Engineering, Guizhou University, Guiyang, Guizhou 550025, *China*

Abstract Recently, visual odometry has been widely used in robotics and autonomous driving. Traditional methods for addressing visual odometry are based on complex processes such as feature extraction, feature matching, and camera calibration. Moreover, each module must be integrated to achieve improved results, and the algorithm is high complexity. The interference of environmental noise and the accuracy of the sensor affect the feature extraction accuracy of the traditional algorithm, thereby affecting the estimation accuracy of the visual odometer. In this context, a visual mileage calculation method based on deep learning and fusion attention mechanism is proposed. The proposed method can eliminate the complicated operation process of traditional algorithms. Experimental results show that the proposed algorithm can estimate the camera odometer in real time achieves improved accuracy and stability and reduced network complexity.

Key words machine vision; deep learning; visual odometry; attention mechanism; multi-task learning

OCIS codes 150.1135; 110.4153; 110.4155

1 引 言

目前，深度学习技术广泛应用在目标检测[1]和目标追踪等诸多计算机视觉领域，以及同步定位与建图(SLAM)领域。视觉里程计(VO)算法[2]是通过视觉技术来获取相机位姿，而 SLAM 的思想是构建整个环境的地图，但 VO 算法对于位姿的估计精度直接影响最终地图的构建效果。近年来，通过一系列有着时间先后顺序的图像来推算相机位姿的方法受到了越来越多研究者的关注[3]，而且该方法广泛应用在各类机器人导航定位以及自动驾驶的领域。

早期，VO 算法是针对火星探索计划进行研究的。2004 年，Nister 等[2]提出了 VO 算法并搭建了最早的 VO 系统，为后续研究提供了优秀的范例和参考。传统 VO 系统的工作过程主要包括：首先校准相机，然后对输入图像进行特征提取及检测并对相邻序列的图像进行特征匹配，接着剔除图像中的

收稿日期：2020-06-10；**修回日期**：2020-07-06；**录用日期**：2020-08-06

基金项目：贵州省科技计划项目(黔科合平台人才[2016]5707)

* **E-mail**：liuyuhongyx@sina.com

异常值，将匹配好的特征进行运动估计和比例估计，最后对其进行局部优化并输出位姿。由此可见，传统 VO 系统的处理过程相对复杂，涉及模块较多。一些基于传统算法的变体在准确性和鲁棒性方面虽然表现较为出色，但通常需要很大的工程量，并且每个模块都需要精准细调才能确保其在特定环境下正常工作。对于单目 VO 系统来说，由于其缺少绝对的信息尺度，所以必须使用一些额外的信息（如相机的高度）才能更好地抑制漂移，从而进行姿态估计。

随着深度学习技术的飞速发展以及卷积神经网络（CNN）在图像识别和分割领域的巨大成功，使得研究者使用 CNN 来处理 VO 问题成为了可能。CNN 可以自动对图像进行不同尺度的特征提取，省去了传统机器学习中繁琐的特征提取过程。但是由于 VO 算法考虑了连续图像序列的相关信息，需要处理和发现图像之间更多的低层几何变换信息，因此在处理 VO 问题仅仅使用 CNN 是不够的。

基于以上问题，本文提出一种基于深度学习的 VO 算法。首先将 RGB（Red，Green，Blue）图像序列输入网络中，对图像数据使用多层 CNN 进行由局部底层几何变换信息到全局高层几何变换信息的提取，同时使用融合注意力机制进一步提取图像中的几何变换信息。然后将提取的特征通过 CNN 降维后连接两个单独的全连接网络再进行多任务学习。最后网络输出模块分别回归位置信息以及角度信息。所提算法构建的网络可进行端到端的训练，从而舍弃传统方法中的特征提取及匹配等复杂过程。网络在降低复杂度的同时可以提高 VO 算法的定位精度，说明所提算法具有较高的稳定性。

2 相关原理

单目 VO 是机器人和自动驾驶领域的重要研究问题之一，实现单目 VO 的方法主要分为基于几何特征的方法和基于学习的方法两类。

2.1 基于几何特征的方法

基于几何特征的方法可进一步划分为稀疏特征法和直接法。稀疏特征法的代表性方法为 LIBVISO2（Library for Visual Odometry 2）法[4]，采用该方法对连续图像帧之间的特征点进行提取和匹配，得到显著的特征点后进行运动估计，稀疏特征法的基本框架如图 1 所示。随着时间的推移，图像帧之间的误差就会产生累积，从而造成一定的漂移，导致 VO 算法的准确度下降。为了进一步解决这个问题，科研学者提出了视觉 SLAM 优化连续特征图的方法。早期，单目视觉 SLAM 是借助于卡尔曼滤波器（EKF）来实现位姿估计[5]，但 EKF 的计算复杂度和线性变化具有不确定性，所以实时性难以得到保证。文献[6]通过滤除动态物体的方法来提高 VO 精度，但计算复杂度很高，同时具有特征点法的弊端。总的来说，基于特征点的方法对于特征点的提取和匹配十分耗时而且计算复杂度很高，同时提取到的特征点不具有全局性，也会丢失一部分信息。当图像不具备明显的纹理信息时，基于特征点的方法很难提取到特征点。虽然直接法的提取速度更快，但其依赖于像素的变化，而且在光照强度发生变化等情形下，VO 精度很容易受到影响。

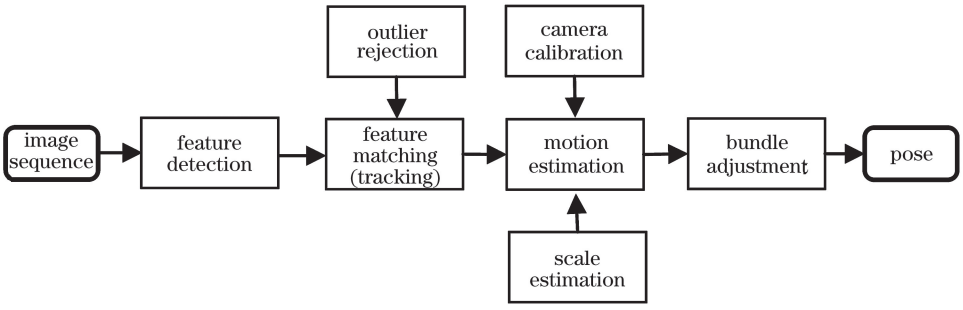

图 1 稀疏特征法的提取框架

Fig. 1 Extraction framework of sparse feature method

2.2 基于学习的方法

基于几何特征的方法主要是从图像中提取特征信息以进行运动估计，而基于学习的方法是从数据中推断运动情况。Roberts 等[7]首先基于学习的方法将每帧图像划分为网格单元并计算每个单元的光流，然后采用 k 最近邻（KNN）算法来估计当前位姿变化情况，虽然该方法不如几何方法准确，但其具有可行性。随后，一些研究人员使用 CNN 从光流图像帧序列中估计运动位姿，采用的方法有 P-CNN VO 方法[8]、Flowdometry 方法[9] 和 LS-VO 方法[10]等。由于计算 RGB 图像的光流十分耗时，而且对系统整体的实时性有一定影响。DeepVO 和

GCN(Geometric Correspondence Network)[11]等是将 CNN 与循环神经网络(RNN)组合,用于从图像序列中提取运动信息,但训练 RNN 需要很大的计算成本,同时网络的复杂度也会很高。

综合以上分析,传统的基于几何特征的方法在图片特征不明显以及光照突变的情况下会导致准确度下降,同时会出现误差累积的现象。基于学习的方法是最近几年发展起来的,由于传统机器学习算法处理数据的能力有限,所以使用基本 CNN 从图像序列中提取特征成为了趋势,然而仅仅使用 CNN 提取表层特征是不够的,通过 RNN 进一步提取特征尽管在效果上有所提高,但这会增加很大的计算压力,复杂度较高。鉴于此,提出一种新型轻量级端到端的网络架构来处理 VO 问题,同时使用融合注意力机制来提取图像序列中的几何变换信息,可以弥补仅使用 CNN 存在的不足。与文献[12]相似,所提网络增加多任务输出位姿的层,分别回归位置信息和角度信息。所提网络的框架如图 2 所示,其中 FC 为全连接。

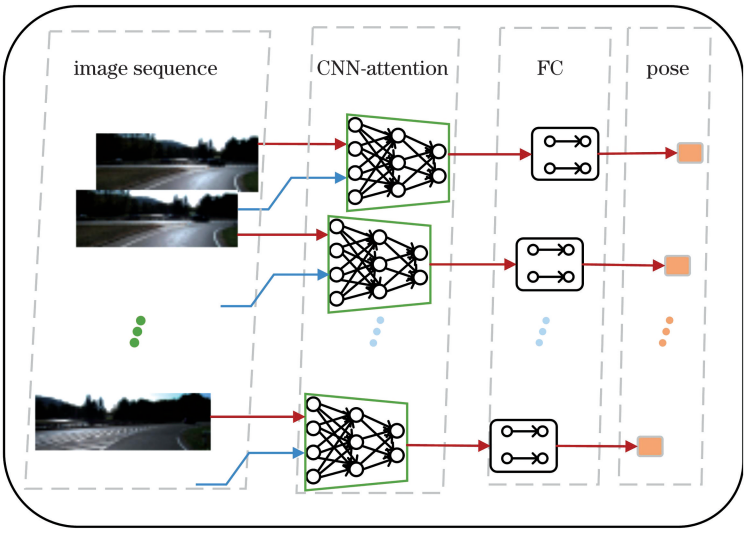

图 2　所提网络的框架

Fig. 2　Framework of proposed network

3　所提算法

所提的网络架构可以进行端到端的训练并且可以输出相机位姿,主要过程包括 CNN 特征的提取、基于注意力机制对层级的图像几何关系特征进行提取,以及使用连接两个分离的全连接网络对多任务特征进行降维和浓缩处理。

3.1　网络框架

所提的神经网络结构如图 3 所示。虽然已有很多有用的 CNN 架构用于处理计算机视觉任务,如 VGGNet、GoogLeNet、ResNet 和 DenseNet 等,但均主要用于分类或者目标检测等。通过图像来计算相机的位姿,但位姿依赖于图像的几何特征信息,这是深度学习中的回归问题之一,因此使用诸如 VGGNet 结构解决 VO 问题的效果较差。综上所述,提出一种能够更好地学习图像几何特征表达的网络架构,该架构可以解决 VO 问题以及其他的图像几何问题。

在网络中输入相邻两帧的图像,两张图像均为数据集中的原始 RGB 图像。为了适应所提的网络架构,将图像尺寸修改为 1280 pixel×384 pixel,并对两张图像进行第三维度上的串联,组成第三维度为 6 的数据并输入网络中,通过 CNN 以及注意力模块对其进行特征提取,再经过两个分离的 FC 层进行单独的降维及特征浓缩处理,最后网络输出这两张图像之间的相对位姿。

由于 VO 问题更依赖于从图像中提取的几何特征信息,因此为了更综合地学习两张图像之间的几何变换关系,实验参考文献[13]的网络架构来设计 CNN 以提取图像特征,并在 CNN 的最后连接两个单独的 FC 层进行多任务的位姿回归。

所设计的 CNN 共有 9 层卷积层,网络内部参数如表 1 所示,对 CNN 中每一层的输出图像使用 ReLU 激活函数进行非线性处理。卷积核的由大变小且通道数的由少变多,可以使网络更能关注图像的局部特征和多种融合特征。经过所有卷积层后再通过最大池化操作可以获得 10×3×1024 的特征图,其中通道数为 1024,特征图的尺寸为 10 pixel×3 pixel,并将其转换为一维的特征向量后先输入到一层共享的 FC 层,再输入各自的 FC 层后进行位姿

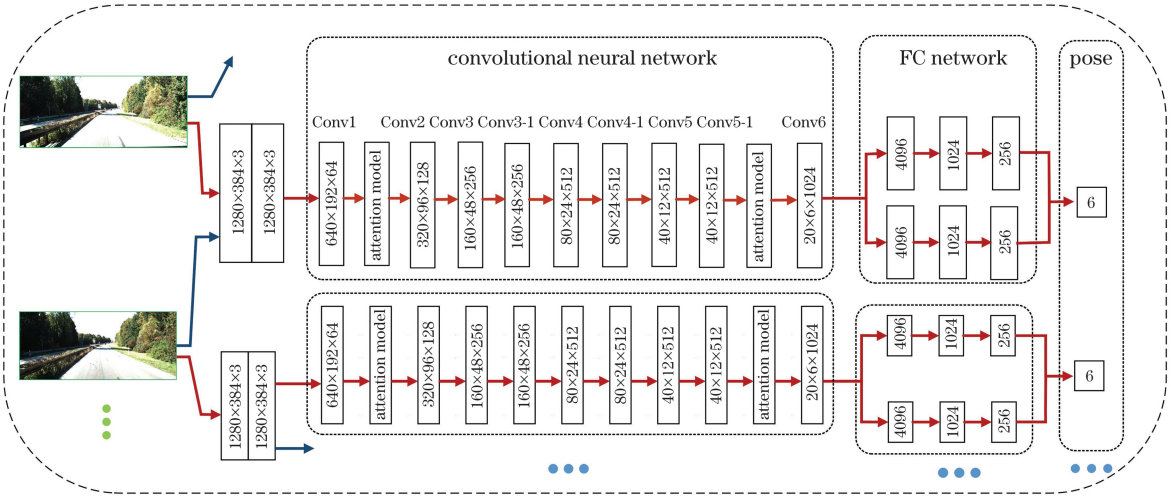

图 3　所提的神经网络结构

Fig. 3　Proposed neural network structure

表 1　CNN 参数

Table 1　Parameters of CNN

Layer	Receptive field size	Padding	Stride	Number of channels
Conv1	7×7	3	2	64
Conv2	5×5	2	2	128
Conv3	5×5	2	2	256
Conv3_1	3×3	1	1	256
Conv4	3×3	1	2	512
Conv4_1	3×3	1	1	512
Conv5	3×3	1	2	512
Conv5_1	3×3	1	1	512
Conv6	3×3	1	2	1024

回归，最终网络输出得到 6 自由度的位姿数据。

3.2　注意力模块

注意力模块可嵌入 CNN 中并进行一种简单而又有效的注意力机制部署，其主要包含通道注意力模块和空间注意力模块。使用注意力机制来增强网络架构的表达能力，可以进一步表征图像之间的几何变换关系，从而使网络可以更智能化地学习更重要的特征并同时关注那些特征区域，从而减少学习一些不重要的特征，这也是注意力机制的本质。由于卷积运算是通过融合各个通道以及空间平面特征图的信息来分离特征，故卷积注意力模块也围绕着通道维度以及特征图的空间维度分别设置注意力机制，这可以使神经网络能够学习更重要的特征信息。受到文献[14]的启发，将基于卷积的注意力模块集成到所提的 CNN 架构中并进行端到端的训练，其基本结构如图 4 所示，其中 $\boldsymbol{F}\in\mathbf{R}^{C\times H\times W}$ 为输入的特征图，$\otimes$ 为逐个元素相乘的符号，$\boldsymbol{F}'$ 为注意力模块内部优化的特征图，$\boldsymbol{F}''$ 为注意力模块优化后输出的特征图，$\boldsymbol{M}_{\mathrm{C}}\in\mathbf{R}^{C\times1\times1}$ 为一维的通道注意力图，$\boldsymbol{M}_{\mathrm{S}}\in\mathbf{R}^{1\times H\times W}$ 为二维的空间注意力图，MLP 为多层感知器，C、H 和 W 分别为特征图的通道数、高和宽，$\oplus$ 为逐个元素相加的符号，$\ominus$ 为 Sigmoid 函数。

输入一张图像，假设这张图像为神经网络内部的某个特征图 $\boldsymbol{F}\in\mathbf{R}^{C\times H\times W}$，使用注意力机制先后生成一个 $\boldsymbol{M}_{\mathrm{C}}\in\mathbf{R}^{C\times1\times1}$ 以及一个 $\boldsymbol{M}_{\mathrm{S}}\in\mathbf{R}^{1\times H\times W}$。总的注意力机制处理过程可以表示为

$$\boldsymbol{F}'=\boldsymbol{M}_{\mathrm{C}}(\boldsymbol{F})\otimes\boldsymbol{F},\tag{1}$$

$$\boldsymbol{F}''=\boldsymbol{M}_{\mathrm{S}}(\boldsymbol{F}')\otimes\boldsymbol{F}'。\tag{2}$$

使用通道注意力模块对输入的特征图分别进行空间维度的全局最大池化操作和全局平均池化操作，得到两个维度为 $C\times1\times1$ 的特征图并将其送入 MLP 网络中，得到两个优化后的特征图并对其进行元素级别的加和操作，再通过 Sigmoid 激活函数对其进行激活从而得到 $\boldsymbol{M}_{\mathrm{C}}$。

使用空间注意力模块对输入的特征图分别进行通道维度的全局最大池化操作和全局平均池化操作，得到两个维度为 $1\times H\times W$ 的特征图并对其进行通道维度的拼接处理，得到维度为 $2\times H\times W$ 的特征图并通过一层卷积网络降维为 1 个通道，再通过 Sigmoid 函数得到维度为 $1\times H\times W$ 的 $\boldsymbol{M}_{\mathrm{S}}$。

将注意力机制与所提的 CNN 结合，并在卷积运算的基础上融合注意力机制，使网络能够自主地关注具有信息量的特征以及特征区域，并且可以学习不同通道以及空间特征图之间的相互关联性，从而在不增加太多网络复杂度的同时可以使神经网络输出的相机位姿更准确。

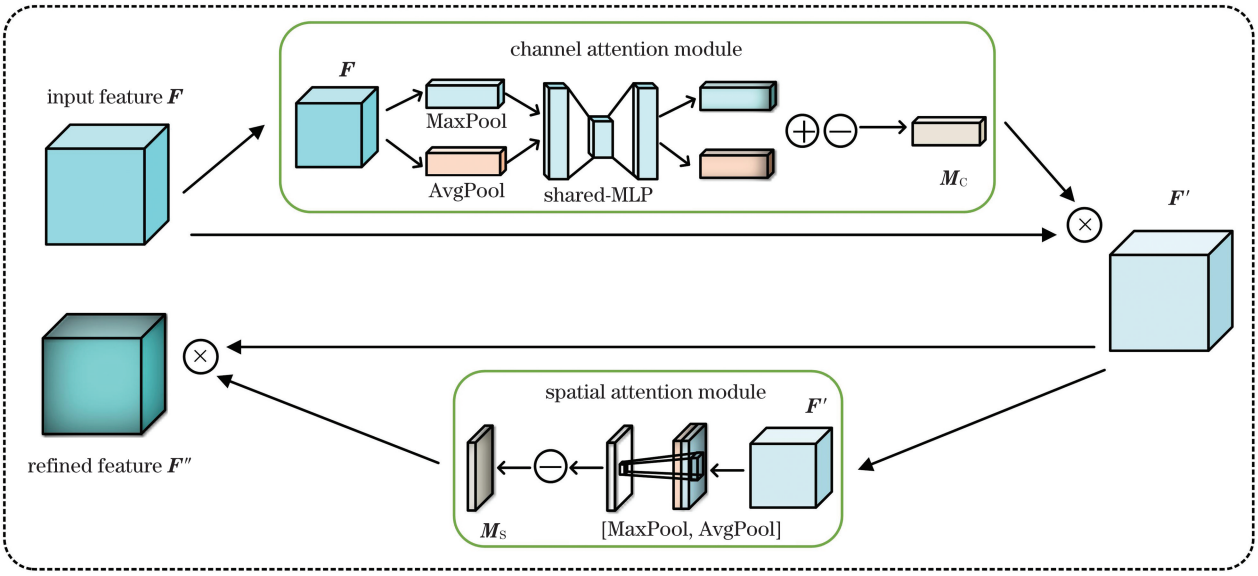

图 4　注意力模块的结构

Fig. 4　Structure of attention module

3.3　多任务学习机制

使用 CNN 以及注意力机制对图像的几何特征进行浓缩处理，后连接两个单独的 FC 层对特征进行进一步的降维，并分别执行回归相机的位移以及转角两个不同的监督任务，从而使模型具有更好的泛化能力。两个单独的 FC 层分别输出相机的位移 (x,y,z) 以及转角 (α,β,γ)，其中 α 为俯仰角，β 为航向角，γ 为滚动角。通过多任务学习机制可以使神经网络更专注于各自的回归任务，从而获得更丰富的特征信息。两个 FC 层的网络结构均采用相同的参数，两个网络输出的自由度均为 3，然后对其进行拼接以组成 6 自由度的位姿数据，从而进行后续误差的计算以及反向传播优化神经网络参数。

3.4　损失函数及优化

所提的网络架构是通过计算条件概率来实现对相机位姿的准确预测。假设 $\boldsymbol{Y}_t=(y_1,y_2,\cdots,y_t)$ 为网络预测输出的位姿，$\boldsymbol{X}_t=(x_1,x_2,\cdots,x_t)$ 为单目图像序列，t 为时刻，t 时刻的条件概率 $P(\boldsymbol{Y}_t\mid\boldsymbol{X}_t)=P(y_1,y_2,\cdots,y_t\mid x_1,x_2,\cdots,x_t)$，通过最大化概率 $P(\boldsymbol{Y}_t\mid\boldsymbol{X}_t)$ 来得到最优的参数，表达式为

$$\boldsymbol{\theta}^*=\underset{\boldsymbol{\theta}}{\operatorname{argmax}}\,P(\boldsymbol{Y}_t\mid\boldsymbol{X}_t;\boldsymbol{\theta}),\tag{3}$$

式中：$\boldsymbol{\theta}^*$ 为网络的最优参数；$\boldsymbol{\theta}$ 为网络参数。为了得到最优的参数，假设网络输出的相机位移为 $\hat{\boldsymbol{p}}_{ki}$，相机转角为 $\hat{\boldsymbol{\varphi}}_{ki}$，地面真值位姿为 $(\boldsymbol{p}_{ki},\boldsymbol{\varphi}_{ki})$，并计算 N 对样本图像的均方误差(MSE)，其中 $\boldsymbol{p}_{ki}$ 为地面真值位姿的相机位移，$\boldsymbol{\varphi}_{ki}$ 为地面真值位姿的相机转角。总的损失函数可以表示为

$$L=\underset{\boldsymbol{\theta}}{\operatorname{argmin}}\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{t}\|\hat{\boldsymbol{p}}_{ki}-\boldsymbol{p}_{ki}\|_2^2+\kappa\|\hat{\boldsymbol{\varphi}}_{ki}-\boldsymbol{\varphi}_{ki}\|_2^2,\tag{4}$$

式中：$\|\cdot\|_2$ 为二范数；κ 为一个尺度因子，用来平衡位移以及转角之间的差异。通过对神经网络进行梯度下降处理，可以获得网络模型的最优参数。相机的转角可以通过欧拉角来描述，这可以使神经网络求解起来简单方便。

4　实验验证与分析

4.1　数据集

KITTI Visual Odometry 是由 Geiger 等[4]构建的汽车驾驶数据集，广泛用于评估各种 VO 算法或 V-SLAM(Visual SLAM)算法的性能。KITTI VO benchmark 共包含 22 个场景图像，每个场景都包含由双目摄像机拍摄的一系列图像，实验只使用双目数据集中的单目图像数据。其中前 11 个场景不仅包含双目图像数据，还包含汽车行驶轨迹的真值，其他 11 个场景仅包含原始的传感器数据。KITTI 数据集中的图像采集频率约为 10 Hz，部分场景中包含动态移动物体以及明暗显著变换的图像。当汽车起步时，相邻两帧图像之间的变化较小，当汽车行驶速度较快或者转弯时，相邻两帧图像之间的变化较大，部分场景中汽车的行驶速度高达 90 km/h。因此上述情况对 VO 的估计精度有相对较大的影响，这对相机的位姿估计更具有挑战性。

4.2　实验环境及参数配置

实验采用的显卡为 Nvidia Geforce Titan Xp

Pascal，用来训练和测试网络，CPU 为 Intel 志强 E5-2673-V3，在深度学习框架 PyTorch 中进行相关算法的设计。使用 Adam（Adaptive Moment Estimation）优化器进行 100 个周期的训练，并将学习率设为 10^{-4}，网络权重初始化方式采用 Xavier 的方式，同时引入防止过拟合和提前停止的方法以防止模型过拟合。神经网络输入的预处理图像尺寸为 1280 pixel×384 pixel，训练过程中同时采用两块 GPU 进行训练，训练一个周期的时间约为 0.15 h。

4.3 实验结果与分析

训练后的 VO 模型采用 KITTI VO/SLAM 官方的评价方式来评测。测试的路径长度范围为 100～800 m，在汽车速度不同（不同场景中汽车的行驶速度不同）的所有子场景中计算平移和旋转误差的方均根误差（RMSE）。

网络模型在场景 00、01、02、08 和 09 中进行训练，在场景 03、04、05、06、07 和 10 中进行测试。VISO2_S 与 VISO2_M 均采用传统方式求解 VO 精度，两者不同之处在于 VISO2_M 为单目算法，VISO2_S 为双目算法。为了对比不同方法的实验效果，在 KITTI 数据集的场景中对 VISO2_S 与 VISO2_M 进行测试，所提算法在不同路径长度和速度下平移和旋转角度的 RMSE 如图 5 所示。从图 5 可以看到，所提算法的平均平移误差比 VISO2_M 大，比 VISO2_S 小；在高速的场景下，所提算法的平均平移误差比 VISO2_M 大，原因在于 00、02、08 和 09 场景中的最大速度都低于 50 km/h，没有太多的训练样本可供训练，从而导致高速场景下的曲线有较大漂移，在训练集中扩充足够多的样本可以使神经网络学习相应的特征，从而提高模型的泛化能力。

图 6 为不同场景下的 VO 轨迹。从图 6 可以看到，所提算法相比于地面真值具有相对准确的精度；与未添加注意力模块的 VO 轨迹相比，所提算法具有集成注意力模块的优势，可以使网络学习位姿变换的深层表示，进一步提高 VO 精度。

不同算法的参数计算量如表 2 所示，其中 FLOPs 为浮点运算次数，t_{rel} 为平移 RMSE，r_{rel} 为

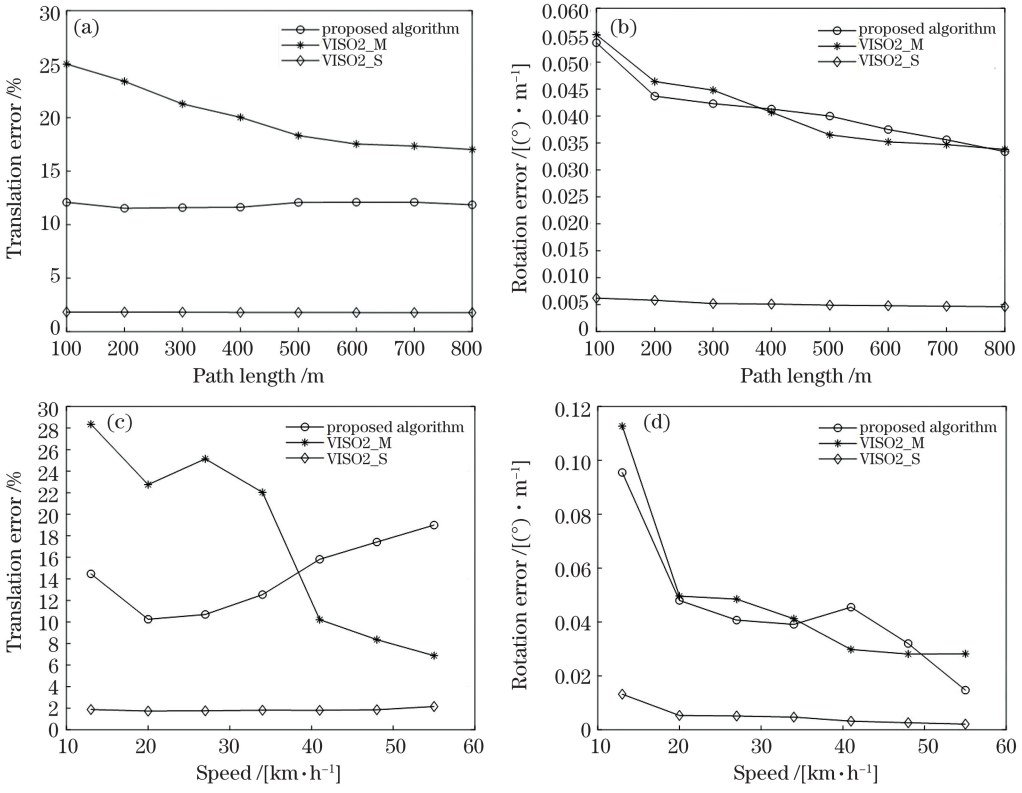

图 5　不同情况下的平均平移误差和平均旋转误差。(a)不同路径长度下的平均平移误差；(b)不同路径长度下的平均旋转误差；(c)不同速度下的平均平移误差；(d)不同速度下的平均旋转误差

Fig. 5　Mean translation errors and mean rotation errors under different conditions. (a) Mean translation errors under different path lengths; (b) mean rotation errors under different path lengths; (c) mean translation errors under different speeds; (d) mean rotation errors under different speeds

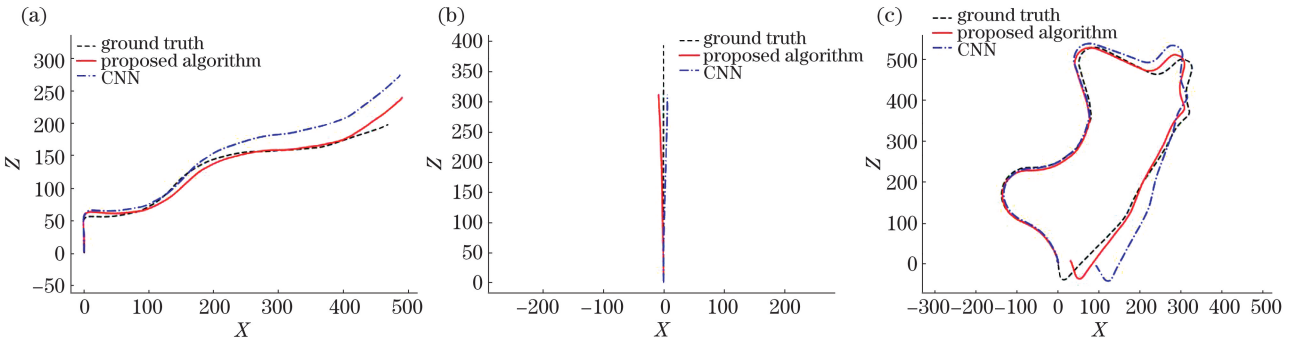

图 6　不同场景下的 VO 轨迹。(a)场景 03；(b)场景 04；(c)场景 09

Fig. 6　VO trajectories in different scenarios. (a) Scenario 03; (b) scenario 04; (c) scenario 09

旋转 RMSE。从表 2 可以看到，DeepVO 算法在不包括两层 LSTM(Long Short-Term Memory)乘加运算的情况下，所需的计算量与所提算法相当，但所提算法所需的参数量仅为其一半；在精度下降较小的情况下，所提算法的计算量和参数量显著下降，这有利于进一步使用硬件来实现；所提算法的平移误差比 VISO2_M 算法减少 49.37%，这可以提升传统算法的效果，不过在旋转误差上有待进一步减小。

为了验证所提算法的有效性，在实际的室外场景下进行实验，首先使用相机录制一段视频，为了尽可能地构造与训练场景符合的原始数据，对视频进行跳帧处理并将得到的连续图像送入神经网络模型中。使用模型预测相机轨迹，得到的实际场景如图 7 所示，实际的地面真值数据需要专业设备来获取，所以将实际的地图作为参考。

从图 7 可以看到，与地图数据相比，所提算法的结果较为准确，通过对比分析可以证明所提算法的有效性。由于手持相机的波动性较大，跳帧后图像反映的实际运动速度比训练集低很多，因为相机在运动过程中难免会受到一些因素的影响，如车辆的快速流动，而且其未行走于路中等。这在一定程度上会影响网络预测的结果，造成结果存在误差。

表 2　不同算法的参数计算量

Table 2　Parameter calculation of different algorithms

Algorithm	Constitute	FLOPs /10^{10}	Parameter /10^{8}	t_{rel}/%	r_{rel}/(°)
VISO2_M	Conventional (monocular)	\	\	17.48	16.52
VISO2_S	Conventional (stereo)	\	\	1.89	1.96
DeepVO	CNN+LSTM	5.143(not include LSTM)	5.1016	5.96	6.12
Proposed algorithm	CNN+attention	5.186	2.6031	8.85	15.68

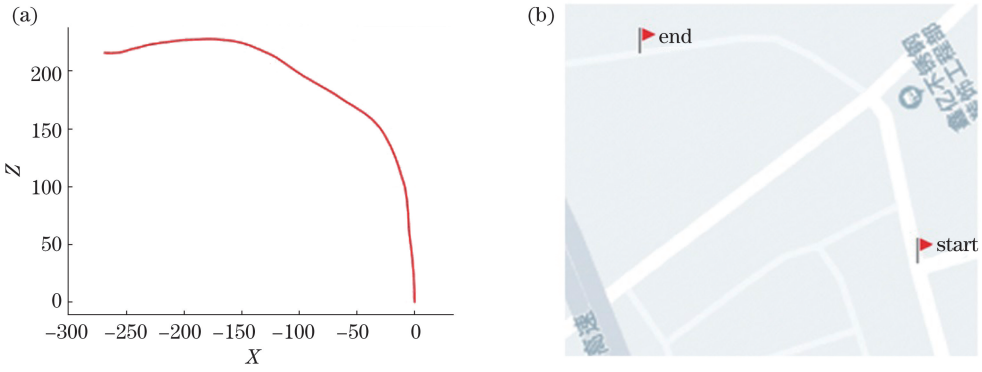

图 7　实际实验。(a)实际场景下的 VO 轨迹；(b)实际地图

Fig. 7　Practical experiment. (a) VO trajectory in actual scene; (b) practical map

对于单目相机获取的 RGB 图像，采用深度学习的算法并融入注意力模块可以直接端到端地输出相机位姿，而且复杂度较低，无需 VISO2 中的特征提取和匹配以及相机标定等复杂步骤。在场景中含有

动态物体的情况下，特征匹配会造成较大的误差，而所提算法能够提取更深的特征信息，使其对动态对象不敏感。如果能够扩充训练集的样本数据，并且数据集中含有不同情况下的数据，将会使网络模型具有更强的泛化能力，并获得更好的效果。

5 结 论

提出一种新型端到端的基于深度学习并融入注意力机制的多任务 VO 算法。通过融合 CNN 以及注意力机制，可以使网络模型能够学习图像之间的深层几何变换信息。相比于传统算法，所提算法可以得到更准确的结果，同时可以舍弃相机标定和特征提取等复杂过程。在不同的场景下，所提算法更易于实现，具有较高的稳定性，并且网络的复杂度低，将来可以使用硬件来实现，具有更广泛的商业价值。

参 考 文 献

[1] He Q Q, Zhang R F, Liu Y H. Human detection algorithm optimization in machine vision[J]. Laser & Optoelectronics Progress, 2020, 57(10): 101006.
何倩倩, 张荣芬, 刘宇红. 机器视觉中的人体检测算法优化[J]. 激光与光电子学进展, 2020, 57(10): 101006.

[2] Nister D, Naroditsky O, Bergen J. Visual odometry [C]//Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition 2004, CVPR 2004, June 27-July 2, 2004, Washington, DC, USA. New York: IEEE, 2004: 652-659.

[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 (ICRA), May 29-June 3, 2017, Singapore, Singapore. New York: IEEE, 2017: 17058210.

[4] Geiger A, Ziegler J, Stiller C. StereoScan: dense 3D reconstruction in real-time[C]//2011 IEEE Intelligent Vehicles Symposium (IV), June 5-9, 2011, Baden-Baden, Germany. New York: IEEE, 2011: 963-968.

[5] Davison A J, Murray D W. Simultaneous localization and map-building using active vision [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2002, 24(7): 865-880.

[6] Lin F C, Liu Y H, Zhou J F, et al. Optimization of visual odometry algorithm based on ORB feature[J]. Laser & Optoelectronics Progress, 2019, 56(21): 211507.
林付春, 刘宇红, 周进凡, 等. 基于 ORB 特征的视觉里程计算法优化[J]. 激光与光电子学进展, 2019, 56(21): 211507.

[7] Roberts R, Nguyen H, Krishnamurthi N, et al. Memory-based learning for visual odometry[C]//2008 IEEE International Conference on Robotics and Automation, May 19-23, 2008, Pasadena, CA, USA. New York: IEEE, 2008: 47-52.

[8] 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(1): 18-25.

[9] Muller P, Savakis A. Flowdometry: an optical flow and deep learning based approach to visual odometry [C]//2017 IEEE Winter Conference on Applications of Computer Vision (WACV), March 24-31, 2017, Santa Rosa, CA, USA. New York: IEEE, 2017: 624-631.

[10] Costante G, Ciarfuglia T A. LS-VO: learning dense optical subspace for robust visual odometry estimation [J]. IEEE Robotics and Automation Letters, 2018, 3(3): 1735-1742.

[11] Tang J X, Folkesson J, Jensfelt P. Geometric correspondence network for camera motion estimation [J]. IEEE Robotics and Automation Letters, 2018, 3(2): 1010-1017.

[12] Zhang M H, Zhang B, Gao C C. Object classification based on multitask convolutional neural network[J]. Laser & Optoelectronics Progress, 2019, 56(23): 231502.
张苗辉, 张博, 高诚诚. 一种多任务的卷积神经网络目标分类算法[J]. 激光与光电子学进展, 2019, 56(23): 231502.

[13] Dosovitskiy A, Fischer P, Ilg E, et al. FlowNet: learning optical flow with convolutional networks[C]//2015 IEEE International Conference on Computer Vision (ICCV), December 7-13, 2015, Santiago, Chile. New York: IEEE, 2015: 2758-2766.

[14] Woo S, Park J, Lee J Y, et al. CBAM: convolutional block attention module[M]//Ferrari V, Hebert M, Sminchisescu C, et al. Computer vision-ECCV 2018. Lecture notes in computer science. Cham: Springer, 2018, 11211: 3-19.