doi:10.19306/j.cnki.2095-8110.2020.04.006

基于注意力模型的视觉/惯性组合里程计算法研究

屈　豪,胡小平,陈昶昊,张礼廉

(国防科技大学智能科学学院,长沙 410073)

摘　要:由于外界环境的干扰和传感器精度的限制,视觉/惯性组合里程计的输入数据存在一定的噪声,这会增加里程计的解算误差,而且误差会随着时间积累。针对以上问题,设计了一种基于注意力模型的视觉/惯性组合里程计算法。该算法使用卷积神经网络和长短时记忆网络分别构建了视觉特征提取器与惯导信息特征提取器,同时引入了两种注意力模型:加权组合网络以及开关组合网络,对视觉特征信息和惯导特征信息的融合噪声进行降噪处理。通过在组合里程计算法中添加闭环校正环节,有效地抑制了里程计误差随时间的积累。对比实验结果表明,设计的组合里程计算法与其他算法相比,无论在性能上还是在精度上都有明显的提升。

关键词:深度学习;注意力模型;视觉里程计;自主导航

中图分类号:V249.32+8　　**文献标志码**:A　　**开放科学(资源服务)标识码(OSID)**:

文章编号:2095-8110(2020)04-0042-08

Research on Visual/Inertial Integrated Odometry Algorithm Based on Attention Models

QU Hao, HU Xiao-ping, CHEN Chang-hao, ZHANG Li-lian

(College of Intelligence Science and Technology, National University of Defense Technology, Changsha 410073, China)

Abstract: Due to the disturbance of the external environment and the limitation of the sensor accuracy, there is noise in the input data of the visual/inertial integrated odometry, which will increase odometry error, and the error will accumulate with time. To solve the above problems, this paper proposes an attention model-based visual/inertial integrated odometry algorithm. This algorithm uses convolutional neural network and long short-term memory network to build visual feature extractor and IMU information feature extractor. At the same time, two attention models, soft attention and hard attention network, are introduced to reduce the fusion noise of visual and inertial feature information. By adding loop closure optimization in the integrated odometry algorithm, the accumulation of odometry error with time is effectively restrained. The experimental results show that compared with other algorithms, the proposed algorithm has a significant improvement in performance and accuracy.

Key words: Deep learning; Attention model; Visual odometry; Autonomous navigation

收稿日期:2020-02-03;**修订日期**:2020-02-26

基金项目:国家自然科学基金(61573371)

作者简介:屈豪(1995-),男,博士生,主要从事仿生导航方面的研究。E-mail: quhao199541@163.com

通信作者:张礼廉(1985-)男,博士,副教授,主要从事仿生导航方面的研究。E-mail:lilianzhang@nudt.edu.cn

0　引言

单目相机因具有配置简单、成本低廉等特点，在民用精度级别的实时定位与建图(Simultaneous Localization and Mapping，SLAM)系统中应用较为广泛。针对单目相机开发的视觉里程计算法按照不同的基础理论可分为：基于几何的单目视觉里程计以及基于深度学习的单目视觉里程计。

基于几何的单目视觉里程计一般采用手工设计的算子识别图像中的特征点，并按照一定的准则辨识多帧图像中的匹配特征点；随后使用多视图几何模型描述时间上相邻图像匹配特征点之间的相对运动关系，同时采用异常值剔除算法排除不符合相机运动模型的特征点，进一步提升视觉里程计的性能。LIBVISO2[1]采用了典型的基于几何的视觉里程计，使用 sobel 算子构建特征描述子，并将相邻帧图像中描述子向量小于一定阈值的特征点视为匹配特征点，使用随机采样一致性(Random Sample Consensus，RANSAC)算法排除不符合相机运动模型的外点。为了能加快整体系统的运行效率以及精度，并行跟踪与映射算法(Parallel Tracking And Mapping，PTAM)[2]和 ORB-SLAM[3]采用轻量化算子在多种分辨率的图像中捕捉特征点。

然而在光线较暗或者存在大面积遮挡的情况下，单目视觉里程计的特征点捕获算法无法获得足够的图像特征，从而导致算法失败。现阶段一部分研究人员通过添加多传感器信息并开发组合里程计来增强 SLAM 系统的感知能力，最为常见的是视觉/惯性组合里程计，其中视觉惯性导航系统(Visual-Inertial Navigation System，VINS)[4]是典型的视觉/惯性组合里程计。VINS 采用预积分的方式处理惯性测量单元(Inertial Measurement Unit，IMU)的测量值，减小位姿更新的计算成本；同时采用非线性优化的方法耦合视觉与 IMU 信息求解相机位姿的过程，随后使用图优化算法结合闭环节点的位姿测量值来校正全局的位姿。VINS 适用于不同精度的相机/IMU 组合系统，对于多变的环境具有较好的鲁棒性，能集成于多种移动平台。

基于几何的视觉/惯性组合里程计根据成熟的多视图几何原理进行开发，在多种场景中都有较为稳定的性能；然而它需要较为理想的运行环境以及精确的参数设定，还需要标定和校准多传感器之间的位置关系与时间戳误差，这极大地提升了算法开发的成本和调试周期。同时视场中的遮挡以及 IMU 测量值的噪声会降低组合里程计的性能，基于几何的视觉/惯性组合里程计可通过滤波或者非线性优化的方式减少噪声对算法运行的干扰，或者预先对噪声进行建模，然后在算法运行的过程中排除噪声的影响。

鉴于深度学习神经网络具有强大的非线性拟合以及高层特征表达能力，已有研究人员尝试使用深度学习神经网络重构视觉里程计。文献[5]提出了 DeepVO 算法，使用光流提取网络(FlowNet[6])搭建视觉特征提取器，以此来取代传统视觉里程计中的特征点/光流提取算法。光流提取网络通过多层卷积层解析相邻帧图像中的高层特征，高层特征不易受到光照条件的干扰，因此具有一定的抗噪能力；同时使用长短时记忆网络(Long Short-Term Memory，LSTM[7])模拟传统视觉里程计的非线性优化模块，进一步提高了短时间内的位姿估计精度。使用全连接层构建位姿回归器，综合相邻帧图像的高层特征并投影为相对位姿估计值。通过长时间的训练，使得网络逐渐拟合图像与相机位姿之间的非线性关系。

文献[8]使用深度学习神经网络重构了视觉/惯性组合里程计，采用 FlowNet-Corr[6]作为视觉特征提取器，为前后两帧图像分别设计卷积层来提取高层特征；同时使用双层 LSTM 搭建了惯导信息特征提取器，将惯导信息特征提取器最后时刻的输出作为惯导信息的高层特征，特征的维度与 LSTM 的隐藏节点数一致。单独设计一个 SE(3)网络层将视觉/惯导信息组合特征投影至位姿标签空间，随后结合位姿标签值构建误差函数进行训练。

上述视觉/惯性组合里程计算法都无显式的抗噪模块，对噪声的适应能力较为有限。为了进一步增强基于深度学习的视觉/惯性组合里程计的抗噪性能，文献[9]提出了一种基于注意力机制的视觉/惯性组合里程计网络，在里程计网络中引入注意力模型，过滤融合特征中的噪声，从而进一步提高网络的性能。

综上所述，现阶段基于深度学习的视觉/惯性组合里程计算法都使用深度学习神经网络集成里程计的全部模块：特征提取、特征匹配、位姿求解以及短时间的位姿优化。但同时也缺少闭环优化的环节，随着时间的推移，位姿估计值误差会逐渐积累。

本文参考文献[9],设计了一种基于注意力模型的视觉/惯性组合里程计算法,针对视觉和惯导信息中可能存在的噪声,本文引入了两种注意力模型。针对里程计误差随时间积累的问题,本文参考传统 SLAM 系统的后端算法,引入闭环优化的环节,使用闭环节点的相对位姿测量值校正全局位姿。

1 基于视觉/惯导信息的特征提取器结构设计

使用深度学习重构视觉/惯性组合里程计可视为隐层特征投影的问题,针对不同种类的数据需设计不同结构的特征提取器来提取隐层特征,并通过前向传播将隐层特征投影至标签空间。本文将光流估计网络(FlowNetSimple[6])的卷积层部分作为视觉特征提取器,FlowNetSimple 网络能同时读取相邻两帧的图像,因此无需为每帧图像单独设计网络,这减轻了网络的计算成本。鉴于循环神经网络存在梯度爆炸的问题,本文使用 LSTM 搭建惯导信息特征提取器。

1.1 视觉特征提取器结构设计

FlowNetSimple 卷积层的结构如表 1 所示,在卷积核参数中,前 2 个维度表示卷积核在图像横纵两轴的尺寸,第 3 个参数表示输入通道数,最后 1 个参数表示输出通道数。图像在输入之前,将尺寸统一调整为(512,256)。

表 1 视觉特征提取器结构表

Tab. 1 Structure of visual feature extractor

网络层	卷积核参数	步长	输出尺寸
Conv1	(7,7,6,64)	2	(256,128,64)
Conv2	(5,5,64,128)	2	(128,64,128)
Conv3	(5,5,128,256)	2	(64,32,256)
Conv3_1	(3,3,256,256)	1	(64,32,256)
Conv4	(3,3,256,512)	2	(32,16,512)
Conv4_1	(3,3,512,512)	1	(32,16,512)
Conv5	(3,3,512,512)	2	(16,8,512)
Conv5_1	(3,3,512,512)	1	(16,8,512)
Conv6	(3,3,512,1024)	2	(8,4,1024)
Conv6_1	(3,3,1024,1024)	1	(8,4,1024)
Fc1	(8,4,1024,256)	—	(1,1,256)

本文在 FlowNetSimple 最后一层卷积层 Conv6_1 后加一个输出通道数为 256 的全连接层 Fc1,将相邻两帧图像的隐层特征压缩为(1,1,256)维度的张量。通过在每一层卷积层之后加上 LeakyReLU 函数来增强神经网络的非线性拟合能力。

1.2 惯导信息特征提取器结构设计

IMU 按照一定的采样频率在三轴方向输出线加速度和角速度,二者构成 6 个维度的惯导信息输入张量,因此需将 LSTM 的输入节点数设置为 6。同时为了进一步加强网络的高层特征表征能力,本文将 LSTM 的前后向隐藏节点输出的隐层特征在通道上进行联结。上述过程如图 1 所示,其中 LSTM 网络的隐含节点数为 128,总共设置两层 LSTM 网络节点,不同层之间的隐层特征相互联结,$[x_{t-1},x_t,x_{t+1}]$ 为输入序列,$[h^{\mathrm{L}}_{t-1},h^{\mathrm{L}}_t,h^{\mathrm{L}}_{t+1}]$ 为前向隐层特征序列,$[h^{\mathrm{R}}_{t-1},h^{\mathrm{R}}_t,h^{\mathrm{R}}_{t+1}]$ 为后向隐层特征序列,$[h_{t-1},h_t,h_{t+1}]$ 为输出的隐层特征序列。

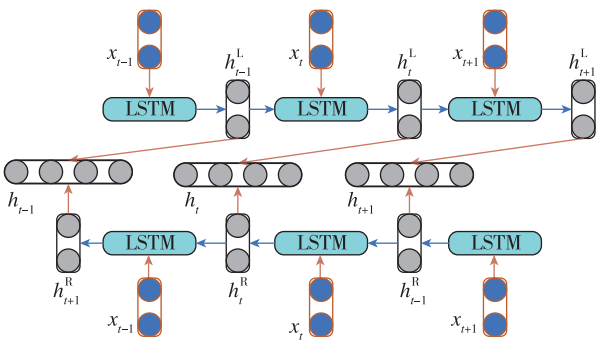


图 1 惯导信息特征提取器结构示意图

Fig. 1 Scheme of bidirectional LSTM

由于本文将前后向的隐层特征序列进行联结,因此 LSTM 网络最终输出 256 维度的隐层特征。为了进一步提高网络的泛化能力,减轻过拟合的现象,本文将 LSTM 网络的 dropout[10] 参数设置为 0.25。

2 基于注意力机制的掩膜网络结构设计

在传感器采集的原始数据中,可能存在一些由外界环境和器件本身所造成的噪声,例如图像中光照强度低或者纹理缺失的部分可能会减弱视觉特征提取器的性能,低精度 IMU 输出的惯导信息也存在一定量的白噪声和零偏。这些噪声的隐层特征若不经处理会降低整体网络的性能,因此本文设计了两种注意力网络:加权组合注意力网络以及开关组合注意力网络来过滤噪声的隐层特征。两种注意力网络的输出与视觉/惯导信息组合特征同维

度的权重掩膜，掩膜与组合特征按元素相乘，通过调节掩膜的数值从而改变组合特征中每一个元素占总体的比重。其中加权组合注意力网络的掩膜数值分布在(0,1)区间，开关组合注意力网络则输出只有 0 和 1 的二值掩膜。

2.1　加权组合注意力网络

在加权组合注意力网络输出的权重掩膜中，较大权重值对应的特征受噪声影响较小，而较小权重值对应的是受噪声影响较大的特征。通过长时间的训练使得权重掩膜的数据分布更符合组合特征中的噪声分布。

经过全连接层压缩的视觉隐层特征 α_{V} 和惯导信息隐层特征 α_{I} 的维度为(1,1,256)。本文首先将两种特征在通道上结合成(1,1,512)维度的组合特征 $\{\alpha_{\mathrm{I}}\,|\,\alpha_{\mathrm{V}}\}$；随后将组合特征代入输出通道数为 512 的全连接层 f_{all}；输出的特征再使用 sigmoid 激活函数 σ 进行非线性化处理，得到权重掩膜 M_{soft}；将掩膜与组合特征按元素相乘，得到经筛选的组合特征 α_{out}。上述过程如式(1)所示，网络结构示意图如图 2 所示。

$$\begin{aligned} M_{\mathrm{soft}} &= \sigma\{f_{\mathrm{all}}\{\alpha_{\mathrm{I}}\,|\,\alpha_{\mathrm{V}}\}\} \\ \alpha_{\mathrm{out}} &= M_{\mathrm{soft}} \odot \{\alpha_{\mathrm{I}}\,|\,\alpha_{\mathrm{V}}\} \end{aligned} \tag{1}$$

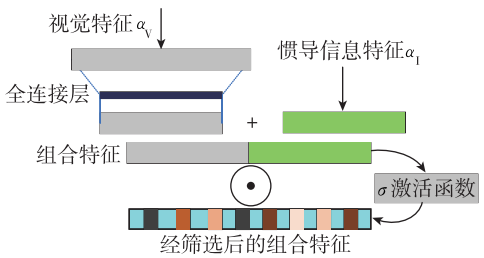


图 2　加权组合注意力模型示意图

Fig. 2　Scheme of soft attention model

2.2　开关组合注意力网络

开关组合注意力网络将组合特征的最大元素视为无噪声污染的特征，并将对应的掩膜元素设置为 1，其余设置为 0，本文使用重参数法来模拟求组合特征最大值的过程[11]。

首先，构建输出通道数与组合特征维度相同的全连接层 f_{all}，将组合特征输入到 f_{all} 得到张量 $\boldsymbol{v}$，$\boldsymbol{v}$ 中的每一个元素 v_i 代表对应位置的组合特征 $\{\alpha_{\mathrm{I}}\,|\,\alpha_{\mathrm{V}}\}_i$ 是无噪声污染的概率；随后将 v_i 与 Gumbel[11] 随机变量 κ_i 相加，并输入到 softmax[10] 函数中，得到的 M_{hard} 近似于二值掩膜。上述过程如式(2)所示

$$\begin{aligned} \boldsymbol{v} &= f_{\mathrm{all}}\{\alpha_{\mathrm{I}}\,|\,\alpha_{\mathrm{V}}\} \\ \kappa_i &= -\log(-\log(U_i)), U_i \sim \mathrm{Uniform}(0,1) \\ M_{\mathrm{hard}} &= \mathrm{softmax}(\kappa_i + v_i) \\ &= \frac{\exp((\kappa_i + v_i)/\mu)}{\sum_{j=1}^{n}\exp((\kappa_j + v_j)/\mu)}, i = 1,\cdots,n \\ \alpha_{\mathrm{out}} &= M_{\mathrm{hard}} \odot \{\alpha_{\mathrm{I}}\,|\,\alpha_{\mathrm{V}}\} \end{aligned} \tag{2}$$

其中，n 代表组合特征张量的长度，μ 是退火系数，调整 M_{hard} 与二值掩膜的相似程度，当 μ 的值较小时 M_{hard} 近似于二值掩膜，将 M_{hard} 与组合特征 $\{\alpha_{\mathrm{I}}\,|\,\alpha_{\mathrm{V}}\}$ 按元素点相乘得到经筛选的组合特征 α_{out}。网络结构示意图如图 3 所示。

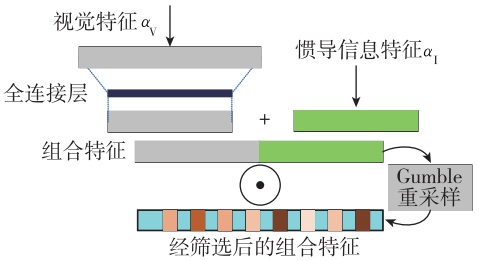


图 3　开关组合注意力模型示意图

Fig. 3　Scheme of hard attention model

3　窗口优化网络及闭环优化

3.1　窗口优化网络

为了对短时间内的位姿估计值进行优化，本文使用 LSTM 搭建窗口优化网络，LSTM 的层数为 2，隐藏节点为 512，dropout 参数为 0.25，窗口优化网络示意图如图 4 所示。将相邻多帧图像与 IMU 信息的组合特征输入到 LSTM 中，并输出经优化的组合特征，随后使用 2 个全连接层分别构成姿态回归器以及位置回归器，分别输出欧拉角形式的姿态增量 $\Delta\boldsymbol{\theta}_{i,i+1}$ 和平移矢量 $\Delta\boldsymbol{\rho}_{i,i+1}$。最后，将二者在通道上进行联结，构成 6 维度的相对位姿估计值张量。

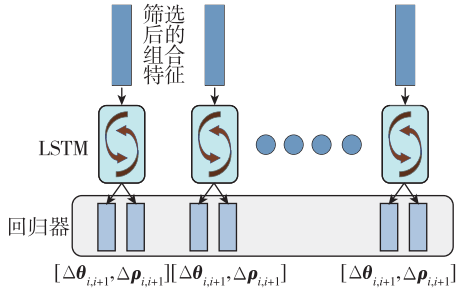


图 4　窗口优化网络示意图

Fig. 4　Scheme of windows optimization network

3.2 误差函数

为了能综合时间上下文信息，本文联合多帧的相对位姿估计值构建误差函数。如式(3)所示，将 n 对图像与 IMU 数据代入里程计网络，得到 n 个相对位姿估计值$\{\Delta\hat{\boldsymbol{\theta}}_j,\Delta\hat{\boldsymbol{\rho}}_j\}$，其中 $\Delta\hat{\boldsymbol{\theta}}_j$ 代表相对姿态估计值，$\Delta\hat{\boldsymbol{\rho}}_j$ 代表相对平移量估计值，结合 n 个相对位姿标签值$\{\Delta\boldsymbol{\theta}_j,\Delta\boldsymbol{\rho}_j\}$构成误差函数 L

$$L=\frac{1}{n}\sum_{j=1}^{n}\tau\|\Delta\boldsymbol{\theta}_j-\Delta\hat{\boldsymbol{\theta}}_j\|+\|\Delta\boldsymbol{\rho}_j-\Delta\hat{\boldsymbol{\rho}}_j\| \quad (3)$$

由于姿态与平移量的数值之间存在数量级的差异，因此需设置平衡系数 τ，本文将平衡系数设置为 100。

3.3 闭环优化

视觉/惯性组合里程计估计的相对位姿与标签值存在一定的误差，随着帧数的推进，恢复出的绝对位姿的误差会逐渐积累。参考传统 SLAM 框架，本文使用图优化算法校正位姿估计值的误差。首先使用 RTAB-Map[12]的闭环检测算法确定序列中的闭环节点，按照在时间轴上的距离，将闭环节点分为短闭环和长闭环；随后使用 monodepth2[13]的相对位姿估计网络 posenet 在 KITTI[14]数据集上进行训练，以此来测量闭环节点之间的相对位姿 $\hat{T}_{ij}$，并将其作为精度较高的闭环节点相对位姿估计值，结合里程计网络推算得到的闭环节点相对位姿值 $T_i^{-1}T_j$ 构建误差函数 L。上述过程如式(4)所示

$$\begin{aligned}L&=\sum_{i,j\in\varepsilon}e_{ij}^{T}e_{ij}\\ e_{ij}&=ln(\hat{T}_{ij}^{-1}T_i^{-1}T_j)^{\vee}\end{aligned} \quad (4)$$

其中，ε 代表相邻两帧以及闭环节点的索引号。

本文使用列文伯格-马尔夸特算法[15]对 L 进行优化，使用 g2o 工具包[16]实现算法。

4 实验结果与分析

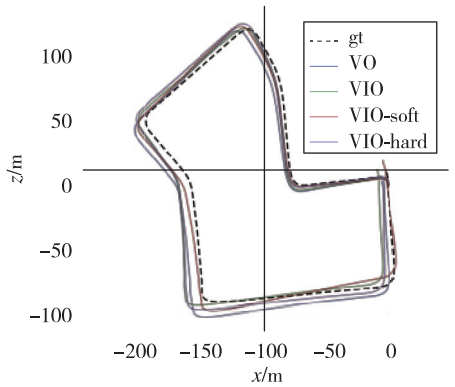
本节通过 3 组实验验证了本文设计的基于注意力模型的视觉/惯性组合里程计网络的有效性，第 1 组实验致力于比较不同注意力模型对里程计网络性能的影响，第 2 组实验致力于对比本文设计的里程计网络与主流视觉/惯性组合里程计算法 VINS 的性能，第 3 组实验验证了闭环优化算法对里程计网络性能的影响。

4.1 实验条件配置

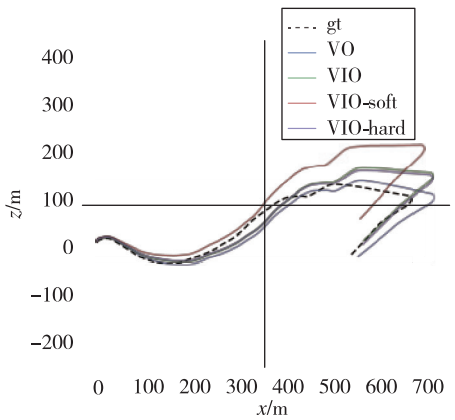
本文使用 KITTI 官网提供的原始数据进行实验，训练集为 00，01，02，06，08，09，验证集为 04 和 05，测试集为 07 和 10。按照文献[17]的提示制备实验数据的标签值，随后将网络在配备 4 块 Titan XP 的工作站上进行训练，总共耗费 13h。使用 pytorch 框架[18]开发本文设计的几种里程计网络。所有网络都使用动量为 0.9 的 Adam 梯度下降法[19]进行训练，学习率固定为 0.0001，批处理数为 48，训练轮数为 100epoch。

4.2 不同注意力网络性能对比实验

本节测试的网络分别为基于深度学习的视觉里程计 VO、基于深度学习的视觉/惯性组合里程计 VIO、基于加权组合注意力模型的视觉/惯性组合里程计 VIO-soft，以及基于开关组合注意力模型的视觉/惯性组合里程计 VIO-hard。其中 VIO 直接将视觉和惯导信息组合特征输入到窗口优化网络，VO 相较于 VIO 缺少惯导信息特征提取器，并且 VO 的视觉特征提取器的全连接层 Fc1 的输出通道数为 512，其余的网络结构与 VIO 相同。不同网络的轨迹对比图如图 5 所示，其中 gt 表示位姿标签。



(a)07 序列



(b)10 序列

图 5 不同注意力模型算法轨迹对比图

Fig. 5 Comparison of different attention models traces

本节使用 KITTI 官方提供的评价指标对不同里程计的性能进行评估，评估前不对轨迹进行对齐操作。在不同长度序列(100m，200m，…800m)上计算平移矢量和旋转量的均方误差，计算其均值并以此作为里程计定位与定姿的精度指标，误差的具体数值如表 2 所示。其中平移矢量误差 t_{rel} 的单位是(%)，旋转量误差 r_{rel} 的单位是((°)/100m)，加粗字段表示同组内的最小值。

表 2　不同注意力机制在 07，10 序列的误差汇总

Tab. 2　Error summary of different attention models in 07 and 10 sequences

序号	指标	07	10	均值
VO	t_{rel}/%	3.51	**7.94**	5.73
	r_{rel}/(°)	1.58	1.77	1.68
VIO	t_{rel}/%	3.69	8.01	5.85
	r_{rel}/(°)	1.71	**1.50**	1.61
VIO-hard	t_{rel}/%	3.24	8.14	5.69
	r_{rel}/(°)	**1.26**	1.63	**1.45**
VIO-soft	t_{rel}/%	**2.67**	8.04	**5.36**
	r_{rel}/(°)	1.78	2.25	2.02

从表 2 的均值项可以看出，添加 IMU 信息的 VIO 网络的定姿性能相对于 VO 有所提升，然而定位性能却有所下滑。值得注意的是，添加开关组合注意力机制的里程计网络 VIO-hard 能有效提高里程计网络的定位与定姿性能，而且它的定姿精度在四种网络中最高；同时添加加权组合注意力机制的 VIO-soft 也能在一定程度上提升定位精度，然而定姿精度却有所下滑。本文分析，原始 IMU 信息中可能存在一定量的噪声(时间戳误差、高斯白噪声与零偏等)，若不抑制噪声的隐层特征则会导致整体网络性能的下降。注意力机制的网络通过长时间的训练找到输入数据的噪声特征，并生成同维度的掩膜对其进行抑制，从而提升网络性能。

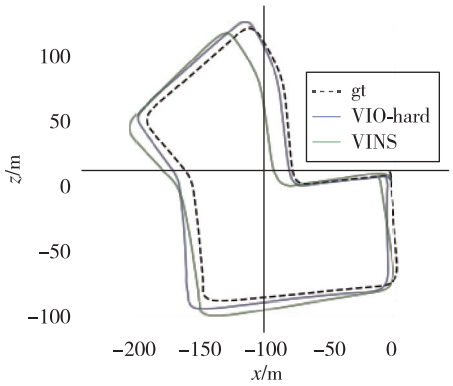
4.3　与 VINS 算法性能对比实验

从 4.2 节实验可以得出，在几种基于深度学习的里程计算法中，VIO-hard 能有效提高里程计网络的定姿和定位精度，因此将其与基于多视图几何的 VINS 算法进行对比，评价的准则与 4.2 节实验保持一致。鉴于 VIO-hard 网络无闭环优化的模块，因此使用无闭环的 VINS 进行实验，VINS 采用官方提供的程序[20]进行实现，算法的具体误差如表 3 所示，轨迹对比如图 6 所示。

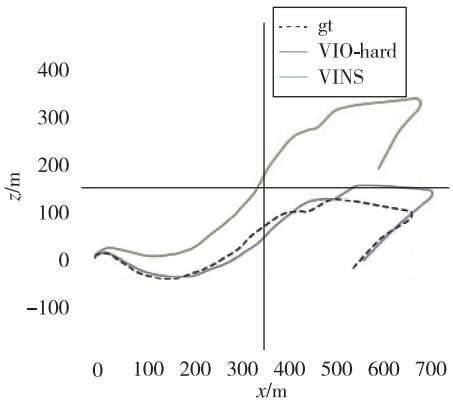
表 3　VIO-hard 与 VINS 的误差汇总

Tab. 3　Error summary of VIO-hard and VINS

序号	指标	07	10	均值
VIO-hard	t_{rel}/%	**3.24**	**8.14**	**5.69**
	r_{rel}/(°)	**1.26**	**1.63**	**1.45**
VINS	t_{rel}/%	3.30	17.76	10.53
	r_{rel}/(°)	1.56	3.52	2.54



(a)07 序列



(b)10 序列

图 6　VIO-hard 与 VINS 轨迹对比图

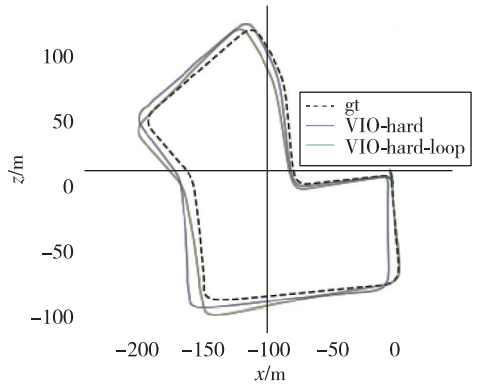
Fig. 6　Comparison of VIO-hard and VINS traces

从轨迹对比图中可以看出，VINS 的轨迹较为偏离轨迹标签，这可能是 VINS 在初始化阶段的性能不太稳定导致其在序列前段的位姿估计精度较低，从而加大了整体轨迹的误差。从误差汇总表和轨迹对比图可以看出，VIO-hard 在整个序列范围内性能较为稳定，具有更好的定姿定位精度。

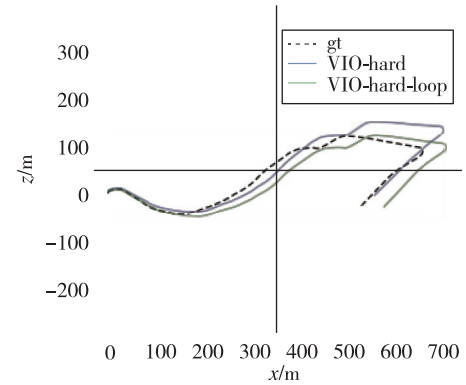
4.4　添加闭环优化算法的性能对比实验

本节致力于验证添加闭环优化算法对基于深度学习的视觉/惯性组合里程计性能的影响。以

VIO-hard 算法为例，添加闭环优化的 VIO-hard 算法在后文简称为 VIO-hard-loop，其中不同算法的轨迹对比如图 7 所示，误差汇总如表 4 所示。



(a)07 序列



(b)10 序列

图 7 添加闭环优化算法轨迹对比图

Fig. 7 Comparison of loop closure and none-loop closure trace

表 4 VIO-hard 和 VIO-hard-loop 的误差汇总

Tab. 4 Error of VIO-hard and VIO-hard-loop

序号	指标	07	10	均值
VIO-hard	t_{rel}/%	3.24	**8.14**	5.69
	r_{rel}/(°)	1.26	1.63	1.45
VIO-hard-loop	t_{rel}/%	**2.41**	8.55	**5.48**
	r_{rel}/(°)	**1.18**	**1.61**	**1.40**

从图 7(a)中可以看出，在 07 序列上 VIO-hard-loop 更贴近位姿标签，从误差对比表中也可以看出，VIO-hard-loop 性能较 VIO-hard 存在明显提升。分析原因，由于 07 序列存在一个明显的长闭环节点，因此全局位姿的校正效果较为明显。

从表 4 可以看出，添加闭环优化算法能够提高 VIO-hard 在 10 序列的定姿精度，然而定位精度却略有下滑。分析原因，短闭环节点的时间跨度较短，位姿估计值的误差积累量较小，与相对位姿估计网络 posenet 估计的节点相对位姿的差值也较小，导致全局位姿校正的效果并不明显；并且部分短闭环节点的相对位姿估计值可能存在一定的误差，从而降低了闭环优化的定位精度。

5 结论

针对已有基于深度学习的视觉/惯性组合里程计网络缺乏显式的降噪模块和闭环优化环节的现状，本文设计了一种基于注意力模型的视觉/惯性组合里程计网络。算法分析与实验验证表明：

1)直接添加原始的 IMU 数据，并且在网络中不对 IMU 数据的隐层特征进行处理，可能并不会提高深度学习视觉里程计的性能，这可能与 IMU 数据中存在的一定量的噪声有关。

2)开关组合注意力模型能有效提高基于深度学习的视觉/惯性组合里程计网络的定姿定位精度。加权组合注意力模型在一定程度上能够提高里程计网络的定位精度，但定姿精度却有所下滑。

3)添加开关组合注意力模型的 VIO-hard 网络相比于无闭环优化的 VINS 算法具有更稳定的性能，这表明基于深度学习的里程计算法可以达到与基于多视图几何的里程计算法相当的性能。

4)添加闭环优化算法能够显著提高里程计网络在长闭环序列中的性能，然而在短闭环序列中的效果并不理想，这表明闭环优化算法的效果受制于闭环节点相对位姿估计的精度。

参考文献

[1] Geiger A, Ziegler J, Stiller C, et al. StereoScan: Dense 3d reconstruction in real-time[C]// Proceedings of IEEE Intelligent Vehicles Symposium, 2011: 963-968.

[2] Klein G, Murray D W. Parallel tracking and mapping for small AR workspaces[C]// Proceedings of International Symposium on Mixed and Augmented Reality, 2007: 1-10.

[3] Murartal R, Montiel J M, Tardos J D, et al. ORB-SLAM: a versatile and accurate monocular SLAM system[J]. IEEE Transactions on Robotics, 2015, 31(5): 1147-1163.

[4] Qin T, Li P, Shen S, et al. VINS-Mono: a robust and versatile monocular visual-inertial state estimator [J]. IEEE Transactions on Robotics, 2018, 34(4):

1004-1020.

[5] Wang S, Clark R, Wen H, et al. DeepVO: towards end-to-end visual odometry with deep recurrent convolutional neural networks[C]// Proceedings of International Conference on Robotics and Automation, 2017: 2043-2050.

[6] Dosovitskiy A, Fischery P, Ilg E, et al. FlowNet: learning optical flow with convolutional networks[C]// Proceedings of International Conference on Computer Vision, 2015: 2758-2766.

[7] Hochreiter S, Schmidhuber J. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735-1780.

[8] Clark R, Wang S, Wen H, et al. VINet: visual inertial odometry as a sequence to sequence learning problem[C]// Proceedings of National Conference on Artificial Intelligence, 2017: 3995-4001.

[9] Chen C, Rosa S, Miao Y, et al. Selective sensor fusion for neural visual inertial odometry[C]// Proceedings of Computer Vision and Pattern Recognition, 2019: 10542-10551.

[10] Krizhevsky A, Sutskever I, Hinton G E, et al. ImageNet classification with deep convolutional neural networks[C]// Proceedings of Neural Information Processing Systems, 2012: 1097-1105.

[11] Maddison C J, Tarlow D, Minka T. A * sampling [C]// Proceedings of Advances in Neural Information Processing Systems, 2014: 1-10.

[12] Labbe M, Michaud F. RTAB-map as an open-source lidar and visual simultaneous localization and mapping library for large-scale and long-term online operation [J]. Journal of Field Robotics, 2019, 36(2): 416-446.

[13] Godard C, Aodha O M, Firman M, et al. Digging into self-supervised monocular depth estimation[C]// Proceedings of International Conference on Computer Vision, 2019: 3828-3838.

[14] Geiger A, Lenz P, Stiller C, et al. Vision meets robotics: the KITTI dataset[J]. The International Journal of Robotics Research, 2013, 32(11): 1231-1237.

[15] Moré J J. The Levenberg-Marquardt algorithm: implementation and theory[M]// Lecture Notes in Mathematics, 2006: 105-116.

[16] Kummerle R, Grisetti G, Strasdat H, et al. G2o: a general framework for graph optimization[C]// Proceedings of IEEE International Conference on Robotics and Automation. IEEE, 2011: 3607-3613.

[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 Linjian. Relative camera pose estimation using deep learning[D]. Hangzhou: Zhejiang University, 2018(in Chinese).

[18] Ketkar N. Introduction to PyTorch[M]// Deep Learning with Python, Springer, 2017: 195-208.

[19] Zou F, Shen L, Jie Z, et al. Weighted AdaGrad with unified momentum[J]. arXiv:1808.03408, 2018.

[20] Qin T, Li P, Shen S. VINS-Mono: a robust and versatile monocular visual-inertial state estimator[J]. IEEE Transactions on Robotics, 2018, 34(4): 1004-1020.