基于时空双流卷积和长短期记忆网络的松耦合视觉惯性里程计

赵鸿儒[1)] 乔秀全[1)] 谭志杰[1)] 李研[2)] 孙恒[2)]

1) (北京邮电大学网络与交换技术国家重点实验室北京 100876)

2) (山西省交通规划勘察设计院有限公司 BIM 研发中心太原 030012)

摘要传统的松耦合视觉惯性里程计需要标定噪声和偏置等参数，而端到端学习的方法耦合性高、普适性低。因此，本文提出了一种由长短期记忆网络融合的端到端松耦合视觉惯性里程计 EE-LCVIO（End-to-End Loosely Coupled Visual-Inertial Odometry）。首先，在相机位姿和 IMU 融合部分，构建了一个时序缓存器和由一维卷积神经网络和长短期记忆网络相结合的融合网络；其次，为了解决现有单目深度视觉里程计难以利用长序列时域信息的问题，通过使用相邻图像对和帧间密集光流作为输入，设计了一种基于时空双流卷积的视觉里程计 TSVO（Visual Odometry With Spatial-Temporal Two-Stream Networks）。与 DeepVO 最多只能利用 5 帧图像信息相比，本文提出的视觉里程计可以利用连续 10 帧图像的时序信息。在 KITTI 和 EUROC 数据集上的定性和定量实验表明，TSVO 在平移和旋转方面超过了 DeepVO 的 44.6%和 43.3%，同时，在传感器数据没有紧密同步的情况下，本文的视觉惯性里程计 EE-LCVIO 优于传统单目 OKVIS（Open Keyframe-Based Visual-Inertial SLAM）的 78.7%和 31.3%，鲁棒性高。与现有单目深度视觉惯性里程计 VINet 相比，EE-LCVIO 获得了可接受的位姿精度，耦合性低，无需标定任何参数。

关键词　视觉惯性里程计；双流融合；长短期记忆网络；松耦合；时序缓存器

中图法分类号　TP391

Loosely Coupled Visual-Inertial Odometry based on Spatial-Temporal Two-Stream Convolution and Long Short-Term MemoryNetworks

ZHAO Hong-Ru[1)]　QIAOXiu-Quan[1)]　TAN Zhi-Jie[1)]　LI Yan[2)]　SUN Heng[2)]

[1)](State Key Laboratory of Networking and Switching Technology, Beijing University of Post and Telecommunications, Beijing 100876)

[2)](Shanxi Transportation Planning Survey and Design Institute BIM R&D Center, Taiyuan 030012)

AbstractVisual Odometry(VO) or Visual-Inertial Odometry(VIO) aims to predict six degrees of freedom (6-DOF) poses from motion sensors, which is a fundamental prerequisite for numerous applications in robotics, simultaneous localization and mapping (SLAM), automatic navigation, and augmented reality (AR). They have attracted a lot of attention over recent years due to the low cost and easy setup of cameras and inertial measurement unit (IMU) sensors. VIO is challenging due to the difficulties to model the complexity and diversity of real-world scenarios from a limited number of on-board sensors. Furthermore, since odometry is essentially a time-series prediction problem, how to properly handle time dependency and environment dynamics presents further challenges.Currently, types of VIO solutions are categorizedinto classical methods and learning-based methods.The classical loosely coupled visual-inertial odometry usually needs to calibrate parameters such as noise and bias, while the end-to-end learning-based method has tight coupling and low universality. Therefore, this

本课题得到国家重点研发计划课题(No.2018YFE0205503)、国家自然科学基金重点国际合作项目 (No. 61720106007)、高等学校学科创新引智基地(No. B18008)资助. **赵鸿儒**，博士研究生，主要研究领域为计算机视觉、同步定位与地图构建.E-mail: zhaohongru@bupt.edu.cn.**乔秀全（通信作者）**，博士，教授，中国计算机学会（CCF）会员（16846M），主要研究领域为未来网络架构、网络服务智能化、分布式神经网络和 Web AR/VR 研究.E-mail: qiaoxq@bupt.edu.cn.**谭志杰**，硕士研究生，主要研究领域为计算机视觉、Web AR/VR 研究、同步定位与地图构建.E-mail: besttangent@gmail.edu.cn.**李研**，正高级工程师，主要研究领域为结构分析与技术咨询、BIM 技术开发与应用.E-mail: 58093797@qq.com. **孙恒**，正高级工程师，主要研究领域为结构分析与技术咨询、BIM 技术开发与应用.E-mail: 332704733@qq.com.

paper presents an EE-LCVIO(End-to-End Loosely Coupled Visual-Inertial Odometry),which is integrated by long and short-term memory networks. Firstly, consideringthe fusion of camera pose and IMU, a sequential cache and a fusion network combined by one-dimensional convolutional neural networks and long short-term memory networks are constructed. Secondly, the existing learning-based monocular visual odometry is limited by remembering history knowledgefor long time. To address this dilemma, we propose a TSVO(Visual Odometry with Spatial-Temporal Two-Stream Networks) using the adjacent image pairs and inter-frame dense optical flow as inputs. Compared with DeepVO which can leverage no more than 5 frames, the proposed visual odometry can exploit the sequential information of 10 consecutive frames. Qualitative and quantitative experiments on the KITTI and EUROC datasets show that TSVO exceedsDeepVO by 44.6% and 43.3% in translation and rotation respectively. Meanwhile, In the case of without tight synchronized sensor data, EE-LCVIO in this paper surpassesthe traditional monocular OKVIS(Open Keyframe-Based Visual-Inertial SLAM)by 78.7% and 31.3% with high robustness. EE-LCVIO achieves an acceptable pose accuracy, fewer calibration parameters and lower coupling than VINet which is the state-of-the-art existing learning-based supervised monocular visual-inertial odometry.

Key wordsvisual-inertial odometry; two-stream fusion; long short-term memory network; loosely coupled; sequential cache

1 引言

六自由度运动估计是增强现实[1]、机器人导航和自动驾驶领域的一个关键挑战。由于相机和惯性传感器（Inertial Measurement Unit，IMU）成本低廉且易于安装，以此为基础的视觉里程计（Visual Odometry，VO）和视觉惯性里程计（Visual-Inertial Odometry，VIO）得到了广泛应用。传统的视觉里程计主要利用手工特征或光度一致性匹配从单目图片序列中计算相机位姿，例如 ORB-SLAM[2]（Oriented FAST and Rotated BRIEF）和 LSD-SLAM[3]（Large-Scale Direct SLAM）。然而当它们被部署在纹理缺失或光照过强环境中时，这些方法无法有效工作。为了克服传统视觉里程计的不足，融合视觉和惯性信息的视觉惯性里程计（VIO）吸引了许多学者的关注。在视觉惯性里程计中，IMU 不会受到低纹理、快速运动等条件影响而导致估计失败；此外，IMU 还可以提供高速率的惯性数据，在图像数据短时间缺失时也可以获得良好的位姿输出。依据是否把图像特征信息加入特征向量，当前的视觉惯性里程计可分为紧耦合和松耦合两种方式，例如 VINS-Mono[4]（Visual-Inertial Navigation System）和 Ethzasl_Msf[5]（Multi-Sensor Fusion）。尽管它们实现了非常高的估计精度，但是仍然严重依赖于传统的视觉里程计典型技术：特征检测、特征匹配和离群值剔除，容易出现特征丢失，跟踪失败等问题。同时传统的视觉惯性里程计需要精确的传感器数据同步工作，增加了算法的开发成本和调试周期。

近年来，鉴于神经网络具有强大的非线性拟合以及高层特征表达能力,已有研究人员通过深度学习来解决传统里程计面临的问题。DeepVO[6]使用深度循环卷积神经网络直接从图片序列中估计出相对位姿，而无需任何先验特征和参数信息。与此同时，文献[7]提出了第一个无需 IMU 和相机间手动同步和校准的端到端视觉惯性里程计网络 VINet。在 DeepVO 的基础上，使用一个小型 LSTM（Long Short-Term Memory[8]）网络处理两帧图像间的 IMU 数据，然后将经过处理得到的图像特征和 IMU 特征向量作为另一个较大 LSTM 网络的输入，最后通过全连接层将融合特征投影至 $SE(3)$ 空间。为了进一步增强基于深度神经网络的视觉惯性里程计的抗噪性能，文献[9]提出了基于软注意力和硬注意力的选择性融合方法，这种方法优于 VINet，同时在数据损坏情况下更加鲁棒。虽然现有基于深度学习的视觉惯性里程计在精度和鲁棒性上与同类传统方法相比具有显著优势，但它们仍然存在一些基本问题：首先，现有基于深度学习的视觉惯性里程计通常将视觉特征和惯性特征融合得到位姿，增加了数据的耦合性和计算量；而传统的松耦合方法需要标定噪声、偏置等参数，不适用于多种设备间的位姿估计问题。其次，里程计本质上是一个时间序列预测问题，由于图像数据的高维特性和 LSTM 的结

构缺陷，当输入图像序列长度超过 5 帧时，现有网络容易发生过拟合现象，限制了里程计的性能。除此之外，与图像相比，IMU 数据维度较低，同时含有噪声等因素，使用 LSTM 训练 IMU 很难收敛。

针对上述问题，本文提出了一种基于时空双流卷积和长短期记忆网络的松耦合视觉惯性里程计。与单纯依赖深度学习从连续图像和 IMU 序列中估计位姿不同，本文利用了行为识别任务中的双流卷积网络和传统的 IMU 积分算法，将深度学习应用于视觉位姿与惯性里程计位姿融合部分，在保留两种传感器数据异质性的同时，整个框架可以自动优化位置和姿态分量，在数据退化条件下更加鲁棒。而且与现有的经典松耦合方法相比，基于深度学习的融合方法可以隐式学习IMU位姿与VO估计间的标定参数，计算复杂度小。

本文在公开可用的 KITTI[10]和 EUROC[11]数据集上进行了实验，此外还评估了基于时空双流卷积的视觉里程计的性能。实验表明，基于时空双流卷积的视觉里程计优于 DeepVO，同时与传统的松耦合方法——MSF[5]相比，本文的视觉惯性里程计无需精确标定和校准参数。本文的主要贡献如下：

1、提出了一种基于时空双流卷积 VO 模块、IMU 积分和长短期记忆网络的端到端松耦合视觉惯性里程计 EE-LCVIO，在传感器数据没有紧密同步的情况下，本文的视觉惯性里程计 EE-LCVIO 优于传统的紧耦合单目 VIO 系统，鲁棒性高。与现有基于深度学习的视觉惯性里程计 VINet 相比，在获得可接受精度的同时，该方法耦合性低；同时相比于传统的松耦合方法，该方法可以隐式学习 IMU 位姿与 VO 估计间的联合标定参数，计算复杂度小。

2、受视频理解任务启发，本文设计了一种利用相邻图像帧和堆叠光流的时空双流视觉里程计 TSVO，在扩展现有视觉里程计输入序列长度的同时提升了现有基于监督学习的视觉里程计（DeepVO）的精度。

3、为了有效解决图片流和 IMU 数据流速率不同导致的 VO 估计和 IMU 积分位姿不同步问题，本文提出了一个时序缓存器以匹配VO和IMU单独计算得到的位姿。

本文的其余部分安排如下：

第二部分对现有视觉惯性里程计的相关工作进行了回顾，第三部分介绍了整体网络框架、基于时空双流卷积的视觉里程计 TSVO、IMU 积分算法、时序缓存器及融合网络、损失函数的设计，第四部分提供了在大型室外和室内公共数据集上实验结果和分析。最后在第五部分总结了本文的工作，并进行了未来的展望。

2 相关工作

本节回顾了视觉惯性里程计的相关工作，讨论了各种算法的优缺点。从所采用的技术和框架上来看，主要有三种类型的算法：基于几何的视觉惯性里程计、基于深度学习的视觉惯性里程计和基于几何—深度学习混合的视觉惯性里程计。

2.1 基于几何的视觉惯性里程计

依据是否将相机位姿信息与 IMU 融合，视觉惯性里程计可以被分为松耦合和紧耦合两种类型。松耦合是指将相机位姿信息与 IMU 的运动估计结果进行融合。如 Ethzasl_Msf[5]接受 VO 模块的位姿估计结果，并通过扩展卡尔曼滤波（Extended Kalman Filter, EKF）将其与 IMU 传播的状态进行融合和位姿更新。除了估计位姿、速度和 IMU 偏置外，它还保留了一个标量参数以估计单目 VO 的漂移比例。该方法虽然计算量较低，但仍需要手动初始化和测量参数以确保位姿尺度近似正确。紧耦合是指把将相机状态估计与 IMU 的运动估计进行联合优化。如 MSCKF[12]（Multi-State Constrained Kalman Filter）采用最小二乘优化方法对特征进行三角化，并在 EKF 中进行融合；OKVIS[13]（Open Keyframe-Based Visual-Inertial SLAM）和 VINS-Mono[4]则通过迭代非线性最小二乘优化来完成融合。紧耦合虽然在精度方面优于松耦合，但是整个融合过程状态向量的维度较高，需要传感器数据紧密同步，计算量大。

2.2 基于深度学习的视觉惯性里程计

与传统的方法相比，基于深度学习的方案在获得精确位姿的同时无需复杂的几何运算而受到大量关注。文献[6]使用递归卷积神经网络学习单目图片间的时间依赖关系，同时利用监督学习从大量图片中获取包含绝对尺度的相机轨迹。ESP-VO[14]通过将基于最大似然的损失函数纳入不确定性估计扩展了这项工作，进而证明了深度学习方法可以适应快速运动、运动模糊、曝光变化等挑战，弥补了传统方法的不足。随后，文献[15]提出了一个包含内存和细化模块的视觉里程计框架来解决累积误差引起的预测漂移。文献[8]第一个将VIO建模为序

列学习问题，它通过利用额外的LSTM网络来学习更好的特征表示，扩展了DeepVO框架以融合 IMU 数据，最终得到一个无需IMU和相机间手动同步和校准的端到端视觉惯性里程计网络VINet。文献[9]借助注意力机制研究了不同的传感器融合方法。除了监督学习外，不需要真实位姿数据参与训练的无监督学习(如文献[16])也有发展趋势，由于本文主要研究基于监督学习的视觉惯性里程计问题，因此不做讨论展开。尽管这些方法在准确率和鲁棒性方面具有竞争力，但与图像相比，IMU数据维度较低，不受外界环境影响，同时由于IMU数据含有噪声和偏置等特性，使用LSTM训练IMU很难收敛，应用深度网络来处理IMU数据是不必要的。

2.3 基于几何—深度学习混合的视觉惯性里程计

与仅仅依靠深度神经网络从数据中回归位姿不同，基于几何—深度学习混合的视觉惯性里程计结合了传统的几何理论与深度学习的优点，预测位姿也更加精确。最近许多研究者已经提出了将经典状态估计与深度神经网络混合的视觉惯性里程计。LS-Net[17]使用 LSTM 网络学习非线性最小二乘优化更新密集地图重建。Backprop KF[18]通过以位姿和速度作为状态向量，构建一个端到端可训练的EKF，并且使用由卷积和全连接层组成的深度网络获得的速度测量值进行状态更新，进而估计相机位姿。文献[19]提出了一个仅使用 IMU 的航迹推算系统与可学习的协方差用于伪测量 EKF 更新。文献[20]中提出了一种端到端可训练的直方图滤波器，并在简单的定位任务下证明了该滤波器的有效性。DPF[21]和 PF-Net[22]同时提出了端到端学习运动和测量模型的粒子滤波网络。Li[23]首次提出了一个通过 DeepVO 学习相对位姿和 EKF 融合 IMU 运动状态的端到端视觉惯性里程计，虽然整个系统的性能优于 DeepVO 和传统的 ORB[2]+MSF[5]方法，但这种结构仍需标定噪声和偏置等参数，计算复杂度高。不同于上述方法，本文提出了一种基于深度学习的松耦合方法，它通过由 CNN 和 LSTM 组成的深度神经网络对VO模块和IMU积分输出的位姿进行融合得到优化后的位姿。

3 EE-LCVIO 结构

3.1 整体网络框架

本文提出的视觉惯性里程计框架如图 1 所示，主要包含四个模块：基于时空双流卷积的视觉里程计 TSVO、IMU 积分算法、时序缓存器和长短期记忆网络。具体地，相邻图像对 I_{t-1} 与 I_t 和经高效光流提取网络 TV-Net[24]计算得到的帧间密集光流 $F_{t-L}\cdots F_{t-1}$ 分别被输入到视觉里程计的空间流网络分支和时间流网络分支中提取运动特征和时序特征，然后将两种特征相继输入全连接层和 SE(3)组合层得到视觉绝对位姿。同时对 IMU 加速度和角速度测量值进行积分得到每个时刻的 IMU 积分位姿，为了解决图片流和 IMU 数据流速率不同导致的 VO 估计和 IMU 积分位姿不同步问题，本文设计了一个时序缓存器以匹配位姿的位置向量和姿态向量，通过将固定大小和时间步长的滑动窗口组合的位姿数据依次输入到长短期记忆网络中得到每个时刻视觉惯性里程计的估计位姿。

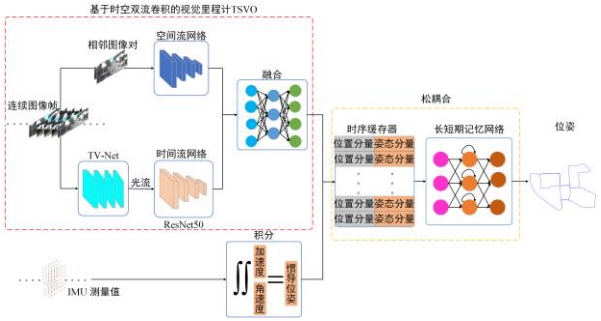

图 1 基于时空双流卷积和长短期记忆网络的松耦合视觉惯性里程计

3.2 TSVO

在行为识别任务中，文献[25]利用双流卷积网络结构（Two-Stream Convolutional Networks）分别提取静止帧和帧间运动的信息，而视觉里程计问题本质是从一系列连续图片序列（可以看作一段视频）中提取图像帧几何和时序信息回归出图像对间相对位姿。受行为识别任务的启发，本文的 TSVO 网络借鉴了双流网络将视频信息分解为空间流和时间流的设计思想，采用相邻图像帧和密集光流两路平行结构级联的方式，提出了一种将现有的仅依赖于单目图像序列输入的深度神经网络扩展为将相邻图像帧与堆叠的光流图作为网络输入的双流网络架构。如图 1 所示，空间流网络从相邻图像对中提取帧间几何特征，时间流网络使用 ResNet50 从

连续光流帧中建模运动特征的时序信息，然后将两路分支网络最后一个卷积层输出的两种特征进行融合得到由位移和欧拉角表示的六维帧间相对位姿，最终通过 $SE(3)$ 组合层得到每个时刻的绝对位姿。

3.2.1 空间流网络

相比于图像识别等领域，视觉里程计问题中特征提取需要体现出几何匹配特性，因此本文参考FlowNetSimple[26]的卷积层部分作为帧间特征提取器。空间流网络的输入为沿通道拼接起来的两张连续图像，网络参数如表1所示，该网络总共包含10层卷积和一层最大池化层，为了能够有效地进行梯度下降以及反向传播，避免梯度爆炸和梯度消失，本文在每层均采用非线性激活函数（Rectified Linear Unit，ReLU）。随着层数加深，卷积核的大小也从7x7逐渐减小至5x5，最后到3x3用以捕捉深层特征。

表 1　空间流网络参数列表

网络层	感受野	填充	步长	通道数
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
Cov6-1	3×3	1	1	1024
Max-Pool	2×2	0	2	—

3.2.2 时间流网络

时间流网络以连续堆叠的稠密光流帧作为输入提取光流图中相机运动的时间信息。在运动检测中，光流是连续帧I_{t-1}与I_t之间的一组像素位移矢量d_t，矢量的水平分量和垂直分量分别是速度向量的两个通道，两者共同描述每个像素点位置的运动向量。为了学习多帧图像间的时序关系，本文将L个连续帧的光流通道$d_t^{x,y}$堆叠起来形成$2L$个输入通道。假设w和h是光流图的水平分量和垂直分量，因此$F_\tau \in \mathbb{R}^{w\times h\times 2L}$表示$L$个连续光流帧堆叠后的网络输入。

为了将光流计算与时间流网络级联在一起，从而构成端到端的体系结构，同时不占用太多计算资源和存储成本，本文采用端到端TV-Net网络提取密集光流。它在获得精确光流的同时，无需任何真实光流的额外训练。在前期数据处理过程中，通过线性变换将光流数据离散到[0,255]的区间上以保证和图像数据分布同区间。由于时间流网络的输入为连续堆叠的稠密光流帧，为了增强提取时序信息的能力，同时避免网络加深造成的梯度消失问题，选择合适的网络框架至关重要。本文分别对比了GoogleNet[27]、ResNet50[28]和InceptionV3[29]三种常用的基本网络在输入图像分辨率为512*256条件下，平均平移和旋转准确率随堆叠光流帧数变化示意图及各自的模型大小、参数量和计算复杂度。由图2和表2可知，在这三个网络中，ResNet50性能最优，而且训练得到的模型参数量较少，计算复杂度较低，因此本文选择经过预训练的ResNet50提取时序特征。

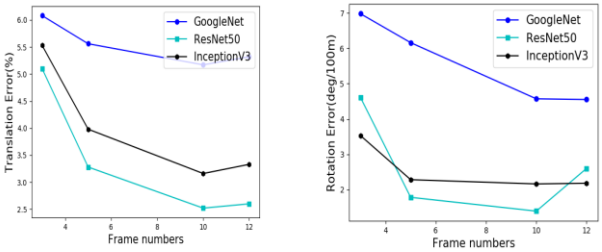

图 2：不同数量堆叠光流帧（L=3，5，10，12）下 GoogleNet、ResNet50 和 InceptionV3 的平均平移和旋转误差。

表 2　模型大小、参数量和浮点运算量对比

模型	大小(MB)	参数量(M)	浮点运算量(G)
FlowNetS-ResNet50	247	58.01	12.20
FlowNetS-GoogleNet	173	35.45	9.57
FlowNetS-InceptionV3	258	59.61	13.81

3.2.3 $SE(3)$ 组合层

将两路分支最后一层得到的空间特征和时间特征直接连接得到新的融合特征，然后通过两个全连接层进行特征压缩得到帧间相对位姿，全连接层隐藏单元数依次为 512，6。由于姿态的数值有正有负，因此仅在第一个全连接层后添加非线性激活函数 ReLU。为了恢复相机的运动轨迹，需要估计出每张图片的绝对位姿，即每张图片相对于初始位姿所定义的坐标系下的位姿。

绝对位姿通常表示为欧式变换矩阵群 $SE(3)$ 上的一个元素。$SE(3)$ 是定义在欧拉空间上的一个可微黎曼流形，它由旋转矩阵群 $SO(3)$ 的旋转矩阵 R 和一个平移分量 t 组成。如式（1）所示：

$$T=\left\{\begin{pmatrix} R & t \\ 0 & 1 \end{pmatrix} \middle| R\in SO(3), t\in \Re^3 \right\} \quad (1)$$

由于旋转矩阵需要满足正交约束，为了便于计算 $SE(3)$，本文首先计算 $SE(3)$ 的瞬时变换 $se(3)$，然后再通过指数映射将 $se(3)$ 转换为 $SE(3)$。$se(3)$ 如式（2）所示：

$$\frac{\xi}{dt}=\left\{\begin{pmatrix} [\omega]_\times & \nu \\ 0 & 1 \end{pmatrix} \middle| \omega\in so(3), \nu\in \Re^3 \right\} \quad (2)$$

3.3 IMU 积分算法

IMU 传感器使用三轴加速度计和三轴陀螺仪获得物体的加速度和角速度，通过对加速度和角速度分别积分可以得到载体的位置、速度和姿态。本文忽略地球自转的影响，所用 IMU 积分算法主要包含 IMU 运动模型和运动方程两部分。

3.3.1 IMU 运动模型

IMU 运动模型的微分方程[30]如式（3）所示：

$$\begin{aligned} \dot{R}_{WB} &= R_{WB}(w_B)^\wedge \\ \dot{v}_W &= a_W \\ \dot{p}_W &= v_W \end{aligned} \quad (3)$$

式中，下标B代表 IMU 坐标系，下标W代表世界坐标系。R_{WB} 为世界坐标系下 IMU 的旋转矩阵；$\dot{R}_{WB}$ 为世界坐标系下 IMU 旋转矩阵的一阶导数；$(w_B)^\wedge$ 为角速度的罗德格里斯公式；$\dot{p}_w$ 为世界坐标系下位置的一阶导数；$\dot{v}_w$ 为世界坐标系下速度的一阶导数。

3.3.2 IMU 运动方程

在 IMU 的时间戳与图像帧的时间戳对齐的情况下，由于 IMU 的采样频率远大于相机的采样频率，相邻两帧图像帧之间存在多组 IMU 测量数据，因此需要使用离散时刻下的运动积分。假设 IMU 时间间隔为 Δt，对载体在 t 到 $t+\Delta t$ 内积分可以得到 IMU 在世界坐标系下的位置 p、速度 v 和旋转 R，运动方程[30]为式（4）所示，其中 i 和 j 表示相邻时刻。

$$\begin{aligned} R_j &= R_i \prod_{k=i}^{j-1} Exp(w_k \Delta t) \\ v_j &= v_i + \sum_{k=i}^{j-1} a_k \Delta t \\ p_j &= p_i + \sum_{k=i}^{j-1} (v_k \Delta t + \frac{1}{2} a_k \Delta t^2) \end{aligned} \quad (4)$$

3.4 时序缓存器及融合网络

准确的相机和 IMU 空间位置关系是实现视觉里程计和 IMU 融合的基础，由于二者的相对位置固定不变，仅相差各自坐标系下的旋转矩阵，因此可以通过位姿变换将 IMU 积分得的位姿转化为相机坐标系下的位姿，如式（5）所示。其中，P_{IMU} 为 IMU 坐标系下的位姿，P^{IMU} 为变换到相机坐标系后的位姿，T_{IMU}^{cam} 为相机和 IMU 间的旋转矩阵。

$$p^{IMU} = T_{IMU}^{cam} p_{IMU} \quad (5)$$

由于 VO 估计和 IMU 积分速率不一致，进而导致两种数据不能紧密同步的问题，本文设计了一种时序缓存器以匹配两个模块产生的位姿，缓存器维度为 $T\times 6(N+1)$。如图 3 所示，j 时刻 VO 估计的位置分量 (V_j^x, V_j^y, V_j^z) 与 j-1 时刻到 j 时刻间的 N 个 IMU 积分得到的位置分量 $(I_{j1}^x, I_{j1}^y, I_{j1}^z, I_{j2}^x, I_{j2}^y, I_{j2}^z, \cdots I_{jN}^x, I_{jN}^y, I_{jN}^z)$ 被合并为 $1\times 3(N+1)$ 维位置向量，j 时刻 VO 估计的姿态分量 $(V_j^\phi, V_j^\psi, V_j^\chi)$ 与 j-1 到 j 时刻的 N 个 IMU 积分得到的姿态分量 $(I_{j1}^\phi, I_{j1}^\psi, I_{j1}^\chi, I_{j2}^\phi, I_{j2}^\psi, I_{j2}^\chi, \cdots I_{jN}^\phi, I_{jN}^\psi, I_{jN}^\chi)$ 被合并为 $1\times 3(N+1)$ 维姿态向量，然后沿时间维度将长度为 T 的所有位姿连接成 $T\times 6(N+1)$ 维向量后输入时序缓存器中。

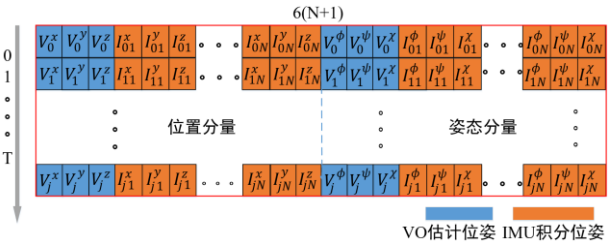

图 3 时序缓存器结构示意图

由于融合网络的输入为时序缓存器中依时间维度连接的位姿向量，一维卷积神经网络常用于时间序列数据的信息提取，而且位姿数据为时间连续分布序列。为了建模当前时刻位姿与之前若干时刻位姿间的依赖关系，本文的融合网络由一维卷积神经网络和LSTM结合而成。如图4所示，给定时序缓

存器中存储的$T\times6(N+1)$维位姿向量，以大小为$M\times6(N+1)$，步长为1的滑动窗口沿时间维度依次遍历时序缓存器中的所有位姿向量，每个时刻融合网络的输入为$M\times3(N+1)$维的位置分量和姿态分量。首先使用一维卷积分别对位置分量和姿态分量进行特征提取，卷积核的大小为$11\times3(N+1)$，通道数为64。在连续三个卷积层之后，使用大小为1×3的最大池化层聚合特征，然后将两路分支最大池化层的输出连接在一起并依次输入到LSTM网络中。为了进一步提高网络的表示能力及动态特性，实验中采用了两层LSTM，每层LSTM含有512个隐藏单元。由于来自CNN的数据分布为$[0,+\infty]$，而LSTM固有的激活函数Sigmoid会将输出限定在（-1,1），所以本文将激活函数改为ReLU。最后将每个时刻LSTM的输出经过两个全连接层（不添加ReLU）得到VO估计和IMU积分融合后的绝对位姿$(VI_x,VI_y,VI_z,VI_\phi,VI_\psi,VI_\chi)$，全连接层的隐藏单元数分别为128和6。为了避免过拟合，在每层LSTM后都添加一个dropout层。

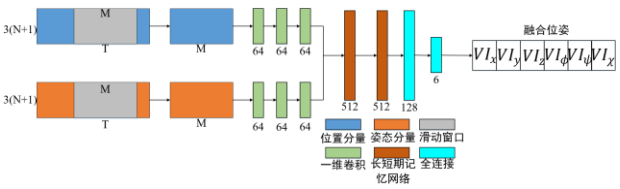

图 4 融合网络结构示意图

3.5 损失函数

位姿估计的关键挑战是设计一种能够同时学习位置和方向的损失函数。由于组成位姿的平移分量和旋转分量分别位于不同的欧式空间，因此通常将两种分量的损失函数加权求和为新的损失函数，如式（6）所示：

$$\begin{aligned} L_{pose} &= \alpha L_{trans} + \beta L_{rot} \\ L_{trans} &= \| \hat{p}_k - p_k \|_\gamma \\ L_{rot} &= \| \hat{\varphi}_k - \varphi_k \|_\gamma \end{aligned} \quad (6)$$

L_{trans} 和 L_{rot} 分别是平移误差和旋转误差，$\hat{p}_k,\hat{\varphi}_k$ 为网络估计的平移量和由欧拉角表示的旋转量，p_k,φ_k 为对应的真实平移量和旋转量，α 和 β 为平衡平移误差和旋转误差的尺度因子，γ 是范数。

在[6]中，作者通过实验发现当$\alpha:\beta=1:100$时可以得到最佳的网络模型，但是在不同的环境中仍然需要手动调整才能得到最优的超参数α和β。[31]提出了一种基于多任务同方差不确定性建模的损失函数，该损失函数鲁棒性高，无需手动调整尺度因子，可以不依赖于输入数据而对平移和旋转分量的不确定性进行测量，因此最终的损失函数如式（7）所示：

$$L_{pose} = L_{trans}\exp(-\hat{s}_t)+\hat{s}_t + L_{rot}\exp(-\hat{s}_r)+\hat{s}_r \quad (7)$$

在损失函数的$\hat{s}_t$，$\hat{s}_r$数值的选取问题上，本文采用和文献[31]相同的数值，因为一般情况下，相机的平移误差较大，具有较大的同方差，同时由于本文的损失函数对多任务同方差不确定性值的初始化选择鲁棒性高，结合两种误差的数量级关系，最终选取$\hat{s}_t=0.0$，$\hat{s}_r=-3.0$。

在视觉里程计实验过程中，本文发现使用 *L2* 范数可有效避免过拟合现象的产生，同时由 *L2* 范数构成的均方误差损失函数收敛更快，函数值震荡较小。因此本文中基于时空双流卷积的视觉里程计采用与[6]相同的均方误差函数作为平移分量和旋转分量各自的损失函数，最终 *N* 对图片真实相对位姿$y_k=(P_k,\varphi_k)$与估计相对位姿$\hat{y}_k=(\hat{p}_k,\hat{\varphi}_k)$的平移误差和旋转误差如式（8）所示：

$$\begin{aligned} L_{trans} &= \frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{3}\|\hat{p}_{ik}-p_{ik}\|_2^2 \\ L_{rot} &= \frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{3}\|\hat{\varphi}_{ik}-\varphi_{ik}\|_2^2 \end{aligned} \quad (8)$$

其中：P_k是平移分量，φ_k是由欧拉角表示的旋转分量，$\|\cdot\|_2$是 *L2* 范数，k 是每对图片相对位姿的状态量，由于平移和旋转分量各包含三个状态量，所以 k 的取值为 1 到 3。(p_{ik},φ_{ik})代表第i对相邻图像帧的平移和转角。

对于融合网络而言，本文通过实验发现使用 *L1* 范数时网络性能表现最好，因为它不会随幅度二次增加，也不会过度衰减较大残差。因此 *N*+1 张图片真实位姿$y_m=(P_m,\varphi_m)$与估计位姿$\hat{y}_m=(\hat{p}_m,\hat{\varphi}_m)$的平移误差和旋转误差如式（9）所示：

$$\begin{aligned} L_{trans} &= \frac{1}{N+1}\sum_{i=1}^{N+1}\sum_{m=1}^{3}\|\hat{p}_{im}-p_{im}\|_1 \\ L_{rot} &= \frac{1}{N+1}\sum_{i=1}^{N+1}\sum_{m=1}^{3}\|\hat{\varphi}_{im}-\varphi_{im}\|_1 \end{aligned} \quad (9)$$

其中P_m和φ_m分别是平移分量和旋转分量，$\|\cdot\|_1$是 *L2* 范数，m 是每张图片绝对位姿的状态量，m 的取值为 1 到 3，(p_{im},φ_{im})代表第i帧图像的绝对位置和姿态。

4 实验结果及分析

本文在公开可用的室外自动驾驶KITTI数据集和室内飞行器 EUROC 数据集上测试了所提出的方法。为了增加样本多样性，除了采用如文献[32]的两种数据增强技术：高斯模糊和椒盐噪声外，本文还使用了逆序图片对以训练基于时空双流卷积的视觉里程计网络。

实验使用 Pytorch 深度学习框架，硬件环境为四核 Intel(R) Xeon(R) Gold 5118 CPU @ 2.30GHz，内存 128GB，两张 NVIDIA Titan Xp，运行内存和显存均为 12GB，操作系统为 Ubuntu 18.04。

由于 IMU 速率为相机速率的 10 倍，因此选取 $M=60$，$N=10$，滑动窗口维度为 60×66。视觉里程计网络 batchsize 为 16，融合网络 batchsize 为 32，其他训练参数为：Adam 优化器，初始学习率为 $1e-4$（每 25 个周期下降 0.5），权重衰减为 0.005，dropout 系数为 0.5。所有的估计轨迹均采用文献[23]中提供的方法与真实轨迹对齐后进行评估。

4.1 KITTI数据集

KITTI VO/SLAM benchmark是评估VO和可视SLAM算法的最著名的公共室外汽车驾驶数据集之一。它提供了10Hz的相机数据、100Hz的IMU数据和从激光扫描仪和车载GPS装置中获得的10HZ的汽车真实轨迹。虽然里程计数据集中包含22个场景序列，但是只有00-02，04-10序列提供了汽车的真实位姿和原始的IMU测量值。本文仅将00、01、02、04、06、08、09左相机序列作为训练集，训练完的模型在05、07、09、10左相机序列上进行测试。通过数据增强，样本数量由原来的17987张变为53961张，输入图片维度被调整为512×256以适应显卡内存。为了测试基于时空双流卷积的视觉里程计性能，L 被分别设为3，5，10。

目前在 KITTI VO/SLAM 数据集上常用的评估指标是：不同长度（100m-800m）子序列的平均平移和旋转误差、不同速度子序列的平均平移和旋转误差。为了便于与现有的主流方法对比，本文最终了选择不同长度（100m-800m）子序列的平移误差与旋转误差作为评估指标。

本文将提出的模型与现有的主流 VO 和 VIO 方法进行了对比，包括：

基于学习的方法：DeepVO[6]、通过几何感知课程学习的单目视觉里程计 CL-VO[33]、基于 3D 卷积的单目视觉里程计 3DC-VO[34]、基于动态注意力机制的单目视觉里程计 DA-VO[35]和单目深度视觉惯性里程计 VINet[8]。

传统的方法：VISO-S[36]、VISO-M[36]、ORB-SLAM2[2]、VINS-Mono[4]、OKVIS[13]和EKF-VIO[23]。具体方法类型如表 3 所示。为了与本文模型进行客观比较，VINS-Mono、OKVIS 和 ORB-SLAM2 均不包含闭环检测。表 4 总结了测试集上的定量对比结果。由于利用同一时刻左右相机的图片和相邻时刻前后帧图片分别估计尺度和相机姿态，VISO-S 在各类算法中表现最优。但与基于深度学习的单目视觉里程计 DeepVO、CL-VO、3DC-VO、DA-VO 及 ORB-SLAM2（无闭环检测）、VISIO-M 相比，基于时空双流卷积的视觉里程计有更低的旋转误差和平移误差。在所有方法中，单独 IMU 积分的结果具有最低的旋转误差，当融合了 IMU 积分信息后，视觉-惯性网络的精度进一步提高。可以看出，EKF-VIO 始终比本文的视觉惯性里程计 EE-LCVIO 有更低的平移误差和旋转误差，与 VINet 相比，EE-LCVIO 有几乎一致的旋转误差。

表 3 现有 VO/VIO 方法类型总结

	方法	传感器	耦合性
VO	DeepVO	单目相机	—
	ORB-SLAM2	单目相机	
	VISO-M	单目相机	
	VISO-S	双目相机	
	CL-VO	单目相机	
	3DC-VO	单目相机	
	DA-VO	单目相机	
VIO	VINet	单目相机+IMU	紧耦合
	VINS-Mono	单目相机+IMU	紧耦合
	EKF-VIO	单目相机+IMU	松耦合
	OKVIS	单目相机+IMU	紧耦合

表 4 测试序列上的泛化性能比较

方法 \ 序列		05	07	09	10	Avg
DeepVO	$t_{rel}(\%)$	2.62	3.91	8.29	8.11	5.73
	$r_{rel}(°)$	3.61	4.60	6.88	8.83	5.98
CL-VO	$t_{rel}(\%)$	5.77	3.79	7.73	8.09	6.35
	$r_{rel}(°)$	2.00	3.00	7.29	7.94	5.06

3DC-VO	$t_{rel}(\%)$	2.70	3.01	6.54	6.75	4.75
	$r_{rel}(°)$	2.95	3.59	5.32	4.94	4.20
DAVO	$t_{rel}(\%)$	2.54	2.78	3.48	5.37	3.54
	$r_{rel}(°)$	1.09	1.98	2.06	1.64	1.69
VISO-M	$t_{rel}(\%)$	19.22	24.61	14.04	22.56	20.11
	$r_{rel}(°)$	17.58	19.11	13.32	12.99	15.75
VISO-S	$t_{rel}(\%)$	1.53	1.45	1.09	1.14	1.30
	$r_{rel}(°)$	1.60	1.91	1.39	1.30	1.55
ORB-SLAM2	$t_{rel}(\%)$	26.01	24.53	24.41	15.39	22.59
	$r_{rel}(°)$	10.62	10.83	2.08	3.20	6.68
TSVO(L=3)	$t_{rel}(\%)$	3.35	3.52	5.80	7.44	5.03
	$r_{rel}(°)$	4.41	3.60	4.07	6.65	4.68
TSVO(L=5)	$t_{rel}(\%)$	2.80	1.55	3.82	5.04	3.30
	$r_{rel}(°)$	1.08	2.76	1.56	1.99	1.85
TSVO(L=10)	$t_{rel}(\%)$	2.09	1.33	2.54	4.14	2.52
	$r_{rel}(°)$	0.76	1.04	1.61	2.15	1.39
IMU-Only	$t_{rel}(\%)$	35.26	31.13	29.96	24.03	30.10
	$r_{rel}(°)$	0.17	0.26	0.21	0.20	0.21
VINS-Mono	$t_{rel}(\%)$	31.90	15.39	17.35	20.35	21.25
	$r_{rel}(°)$	2.72	2.42	1.65	3.73	2.63
OKVIS	$t_{rel}(\%)$	13.77	9.65	5.69	10.82	9.98
	$r_{rel}(°)$	2.01	1.56	1.89	1.80	1.82
EKF-VIO	$t_{rel}(\%)$	2.10	0.93	1.24	1.17	1.36
	$r_{rel}(°)$	0.56	0.30	0.25	0.26	0.34
VINet	$t_{rel}(\%)$	1.24	1.10	2.16	2.03	1.64
	$r_{rel}(°)$	1.06	1.19	1.27	1.44	1.24
EE-LCVIO	$t_{rel}(\%)$	2.03	1.35	2.38	2.77	2.13
	$r_{rel}(°)$	1.16	0.98	1.33	1.54	1.25

$t_{rel}(\%)$：不同长度（100m-800m）的平均平移均方误差。

$r_{rel}(°/100m)$：不同长度（100m-800m）的平均旋转均方误差。

由于车载相机上下运动幅度变化不大，Y 方向位移较小，所以本文在 X-Z 平面上展示了 L=10 时各种算法在测试序列上的估计轨迹与真实轨迹，如图 5 所示，EKF-VIO 的轨迹曲线和真实轨迹最为贴近，基于时空双流卷积的视觉里程计 TSVO 估计的轨迹比 DeepVO 更加精确，并且优于传统的单目 VO 方法。同时，融合 IMU 后的轨迹和 VINet 相近，优于单独使用视觉数据估计的轨迹。产生上述结果的原因如下：

（1）本文的模型以相邻图像对和多帧密集光流作为输入，网络建模图像序列间依赖关系的能力得以提升，同时由于 ORB-SLAM2 存在累积误差，因此基于时空双流卷积的视觉里程计优于 ORB-SLAM2、 VISO-M 和 DeepVO。随着叠加光流帧数量的增加，视觉里程计的性能也逐渐提高，然而当堆叠光流帧的数量超过 10 帧时，网络性能不会显著改善，这是由于过多的光流帧包含了冗余信息，从而导致时间流网络学习能力的下降的原因。

（2）由于 KITTI 具有低噪声和偏置的陀螺仪，因此陀螺仪积分得到的角度估计误差较低，而噪声较大的加速度计积分得到的位置估计会迅速漂移。

（3）本文的 EE-LCVIO 在平移和旋转方面分别优于 OKVIS 的 78.7%和 31.3%，与 VINS-Mono 相比，平移和旋转精度也有所提升。造成这种结果

的原因是 OKVIS 和 VINS-Mono 都需要 IMU 测量值和图像之间的紧密同步，因此它们在 KITTI 数据集上表现不佳。此外， KITTI 数据集中的加速度测量值较小，进而导致了 OKVIS 和 VINS-Mono 的显著漂移。

（4）与 VINet 相比，EE-LCVIO 取得了可接受的精度。这是由于 VINet 是一种在中间特征层面上融合的视觉惯性里程计，网络对图像和 IMU 数据信息利用更加充分，因此在准确率方面稍优于本文的融合方法。但 EE-LCVIO 耦合性低，在图像或 IMU 数据缺失的情况下，能够单独利用视觉里程计或 IMU 积分估计位姿。

（5）传统的基于 EKF 的松耦合方法结合了加速度计和陀螺仪的偏置及位姿的协方差矩阵，通过状态方程对非线性系统进行优化，进而实时更新相机位姿，因此结果更加精确。但本文的 EE-LCVIO 无需联合标定相机和 IMU 参数，计算复杂度小。

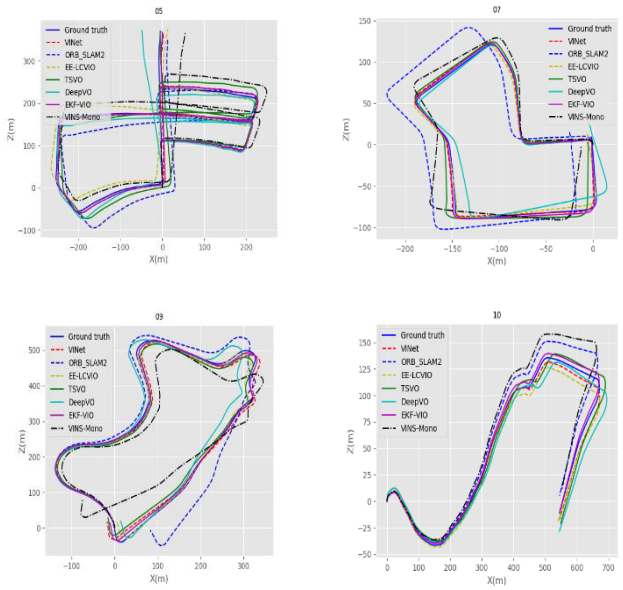

图 5：本文方法与现有方法在 05、07、09、10 序列上的测试轨迹，横轴代表 X 方向位移，纵轴代表 Z 方向位移。

4.2 EUROC数据集

EUROC 数据集提供了由微型飞行器在房间和工厂两种环境中采集到的 11 个不同序列，与 KITTI 数据集不同的是，由于强烈的光照变化，每种环境下采集的图像质量差异较大，例如， “MH_01”文件夹下的图像质量较高，而“MH_05”中的图像则更加模糊。为了与现有方法作对比，本文将相机的采样频率和 IMU 的采样频率分别下采样到 10HZ 和 100HZ，同时将 MH_01、MH_02、MH_03 及 V1、V2 作为训练序列，MH_04 和 MH_05 作为测试序列。输入图像大小被调整为 640×480，其余参数设置和评估指标均和 KITTI 数据集相同。

由于很少有基于学习的视觉里程计尝试处理 EUROC 数据集，同时不断变化的图像质量也对几何方法造成了挑战，本文对比了具有代表性的 DeepVO、VINet、VINS-Mono[4]、OKVIS[13]和 SVO[37]（Semi-direct Visual Odometry）+MSF 四种算法，其中 OKVIS 和 VINS-Mono 是利用关键帧优化位姿的紧耦合方法，SVO+MSF 是通过 MSF 框架实现 SVO 与 IMU 的松耦合方法。上述各类方法的定量对比结果如表 5 所示，由于 EUROC 序列相邻图片间变化非常小，因此估计位姿与真实位姿间的平移误差和旋转误差普遍较低。与在 KITTI 数据集上的表现相同，基于时空双流卷积的视觉里程计 $(L=10)$ 在平移和旋转方面超过了 DeepVO 的 44.6% 和 43.3%。融合 IMU 积分信息后的网络平移误差和 VINet 相当，但仍低于传统的松耦合方法，这是由于 EuRoC 数据集中的 IMU 数据包含较大的噪声和偏置，对视觉信息造成了干扰，当 IMU 数据质量较差且与视觉信息相互无法补充时，传统的方法会不断修正测量值以优化位姿。此外，由于 EUROC 数据集中相机和 IMU 数据紧密同步，基于关键帧优化的单目 OKVIS 和基于滑动窗口优化的 VINS-Mono 将视觉和 IMU 误差项一同加入损失函数进行优化，因此在精度方面优于本文方法。但 EE-LCVIO 可以处理时间松散同步的传感器数据，而无需明确估计它们的时间偏移量。图 6 在 X-Y 平面上展示了算法在测试序列上的估计轨迹，可以看出，OKVIS 的轨迹曲线和真实轨迹最接近，SVO+MSF 其次，视觉惯性里程计和 VINet 的轨迹相当，优于 DeepVO 和单独视觉产生的轨迹。

表 5 MH_04 和 MH_05 序列上的平均相对位姿误差比较

方法 \ 序列		MH_04	MH_05	Avg
DeepVO	t_{rel} (%)	4.86	4.24	4.55
	r_{rel} (°)	5.66	2.94	4.30
TSVO(L=10)	t_{rel} (%)	2.25	2.78	2.52
	r_{rel} (°)	3.23	1.65	2.44
VINet	t_{rel} (%)	1.56	1.33	1.45
	r_{rel} (°)	1.87	1.23	1.55
OKVIS	t_{rel} (%)	0.34	0.29	0.32
	r_{rel} (°)	0.76	0.81	0.79
VINS-Mono	t_{rel} (%)	0.23	0.25	0.24

	$r_{rel}(°)$	0.54	0.61	0.58
SVO+MSF	$t_{rel}(\%)$	1.38	1.01	1.20
	$r_{rel}(°)$	1.57	1.06	1.32
EE-LCVIO	$t_{rel}(\%)$	1.52	1.45	1.49
	$r_{rel}(°)$	2.10	1.66	1.88

$t_{rel}(\%)$：不同长度（100m-800m）的平均平移均方误差。

$r_{rel}(°/100m)$：不同长度（100m-800m）的平均旋转均方误差。

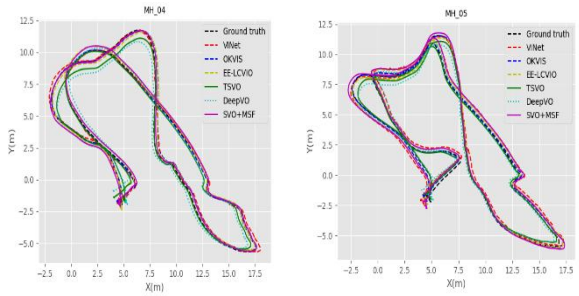

图 6：本文方法和 VINet、DeepVO、SVO+MSF、OKVIS 在 MH_04 和 MH_05 序列上的测试轨迹，横轴代表 X 方向位移，纵轴代表 Y 方向位移。

5 结论和展望

针对现有单目深度视觉里程计输入图像序列长度限制问题，本文将帧间密集光流和相邻图像对作为输入，利用双流网络融合不同层次特征，提出了基于时空双流卷积的视觉里程计。同时为了解决现有相机和 IMU 融合方法需要联合标定参数的问题，设计了一个由一维卷积神经网络和 LSTM 组成的融合网络。在 KITTI 和 EUROC 数据集上的实验表明，基于时空双流卷积的视觉里程计 TSVO 可以利用连续 10 帧图像的时序信息，在平移和旋转方面分别超过了 DeepVO 的 44.6%和 43.3%，运行时间为 14 帧/秒。同时，在传感器数据没有紧密同步的情况下，本文的视觉惯性里程计 EE-LCVIO 优于传统单目 OKVIS 的 78.7%和 31.3%，鲁棒性高。与现有单目深度视觉惯性里程计 VINet 相比，EE-LCVIO 获得了可接受的位姿精度，耦合性低。相比于传统的松耦合方法，本文的融合方法不需要标定相机和 IMU 的任何参数，计算复杂度低。由于未标定 IMU 的噪声和偏置，将 VO 估计与 IMU 积分得到的两种位姿直接融合，视觉惯性里程计的精度仍然有待提高，在未来的工作中，本文将专注于更稳健的融合策略以处理不完美的位姿数据。

参考文献

[1] XiuquanQiao, Pei Ren, SchahramDustdar, Ling Liu, Huadong Ma, Jun-Liang Chen.Web AR: A Promising Future for Mobile Augmented Reality - - State of the Art, Challenges, and Insights. Proceedings of the IEEE, 2019, 107(4):651–666.

[2] Mur-Artal R, Montiel J, et al. Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cameras. IEEE Transactions on Robotics, 2017, 33(5):1255–1262.

[3] Engel J, Schops T, and Cremers D. LSD-SLAM: Large-scale direct monocular slam// The European Conference on Computer Vision, Zurich, Switzerland,2014:834-849.

[4] Qin T, Li P, and Shen S. Vins-mono: A robust and versatile monocular visual-inertial state estimator. IEEE Transactions on Robotics, 2018, 34 (4): 1004-1020.

[5] Weiss S M. Vision based navigation for micro helicopters[Ph.D. thesis]. ETH Zurich,Switzerland, 2012.

[6] Wang S, Clark R, Wen H, et al. DeepVO: Towards end-to-end visual odometry with deep recurrent convolutional neural networks// IEEE International Conference on Robotics and Automation,Singapore,2017: 2043–2050.

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

[8] Donahue J, Hendricks L A, Guadarrama S, Rohrbach M, et al. Long-term recurrent convolutional networks for visual recognition and description . IEEE Transactions on Pattern Analysis and Machine Intelligence,1997,9(8):1735-17.

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

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

[11]Burri M, Nikolic J, Gohl P T, et al. The eurocmicro aerial vehicle datasets. The International Journal of Robotics Research, 2016,35(10): 1157-1163.

[12]Bloesch M, Omari S, Hutter M, et al. Robust visual inertial odometry using a direct EKF-based approach// Proceedings of IEEE International Conference on Intelligent Robots and Systems. Hamburg, Germany,2015: 298-304.

[13] Leutenegger S, Lynen S, Bosse M, et al. Keyframe-based visual-inertial odometry using nonlinear optimization . International Journal of Robotics Research, 2015, 34 (3): 314-334.

[14] Wang S, Clark R, Wen H, et al. End-to-end, sequence-to-sequence probabilistic visual odometry through deep neural networks. The International Journal of Robotics Research, 2018,37(5):513 – 542.

[15] Fang X, Wang X, Li S, et al. Beyond tracking: Selecting memory and refining poses for deep visual Odometry// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.Los Angeles , USA,2019:8575–8583.

[16] Li R, Wang S, Long Z, et al. Undeepvo: Monocular visual odometry through unsupervised deep learning// IEEE international conference of robotics and automation.Brisbane, Australia, 2018: 7286–7291.

[17] Clark R, Bloesch R, Czarnowski J, et al. Learning to solve nonlinear least squares for monocular stereo// The European Conference onComputer Vision.Munich, Germany,2018:284-299.

[18] Haarnoja T, Ajay A, Levine S, and Abbeel P. Backpropkf: Learning discriminative deterministic state estimators//Advances in Neural Information Processing Systems,Barcelona, Spain,2016:4376–4384.

[19] Brossard M, Barrau A, and Bonnabel S. AI-IMU deadreckoning. IEEE Transactions on Intelligent Vehicles, 2020, 5(4):585–595.

[20] Jonschkowski R and Brock O. End-to-end learnable histogram Filters// Conference and Workshop on Neural Information Processing Systems.Barcelona, Spain,2016:277-287.

[21] Karkus P, Hsu D, and Lee W S. Particle filter networks with application to visual localization. Proceedings of Machine Learning Research, 2018, 87(2):169–178.

[22] Jonschkowski R, Rastogi D, and Brock O. Differentiable Particle Filters: End-to-End Learning with Algorithmic Priors. Proceedings of Robotics: Science and Systems, 2018,6(2):1791-1799.

[23] Li C, Steven L, et al. Towards End-to-end Learning of Visual Inertial Odometry with an EKF // Conference on Computer and Robot Vision, Ottawa,Canada,2020:190-197.

[24] Fan L, Huang W, Gan C, et al. End-to-End Learning of Motion Representation for Video Understanding //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.Salty LakeCity,USA, 2018:8545–8556.

[25]Simonyan K, Zisserman A. Two-stream convolutional networks for action recognition in videos//Advances in neural information processing systems.Montreal, Canada,2014: 568-576.

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

[27] Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, et al. Going deeper with convolutions// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.Boston, USA,2015:1–9.

[28] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition// Proceedings of the IEEE conference on computer vision and pattern recognition, Las Vegas, USA, 2016: 770–778.

[29] Szegedy C, Ioffe S. Wojna Z, et al. Rethinking the inception architecture for computer vision// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.Las Vegas, USA, 2016:2818–2826.

[30] Forster C, Carlone L, Dellaert F, et al. IMU preintegration on manifold for efficient visual-inertial maximum-a-posteriori estimation[Ph.D. thesis]. Georgia Institute of Technology, Atlanta, USA,2015.

[31] Kendall A, Cipolla R, et al. Geometric loss functions for camera pose regression with deep learning// computer vision and pattern recognition.Honolulu,USA,2017:6555-6564.

[32] Huang Y K, Zhao H R, Qiao X Q, Tang J, Liu L. Towards Video Streaming Analysis and Sharing for Multi-Device Interaction with Lightweight DNNs // Proceedings of IEEE International Conference on Computer Communications.Toronto Canada,2021:1-10.

[33] Saputra M, Wang S, Markham A,et al. Learning monocular visual odometrythrough geometry-aware curriculum learning//IEEE International Conference on Robotics and Automation.Montreal, Canada,2019: 3549–3555.

[34] Alexender S, James A, Gaurav S, et al. Estimating metric scale visual odometry from videos using 3D convolutional networks// IEEE International Conference on Intelligent Robots and Systems.Macau, China, 2019:872-880.

[35] Kuo X Y, Liu C, Lin K C, et al, Dynamicattention-based visual odometry// Conference on Computer Vision and Pattern RecognitionWorkshops.Washington, USA, 2020.

[36] Geiger A, Ziegler J, and Stiller C. Stereoscan: Dense 3d reconstruction in real-time. IEEE Intelligent Vehicles Symposium, 2011,38(4):963–968.

[37] Forster C, Zhang Z, Gassner M, Werlberger M, and Scaramuzza D. SVO: Semidirect visual odometry for monocular and multicamera systems. IEEE Transactions on Robotics, 2017, 3(2):249–265.

ZHAO Hong-Ru, Ph.D. candidate.His current research interests include computer vision, simultaneous localization and mapping.

QIAO Xiu-Quan, Ph. D., professor. His research interests include Web AR/VR research，future network architecture, network service intelligence, and distributed neural network.

TAN Zhi-Jie, M.S. candidate.His research interests include computer vision, Web AR/VR research, simultaneous localization and mapping.

Li Yan,senior engineer. Her research interests include Structural analysis and technical consulting,AR-basedBIM technology development and application.

SunHeng,senior engineer. His research interests include Structural analysis and technical consulting,AR-basedBIM technology development and application.

Background

This work focuses on the visual-inertial odometry in the field of simultaneous localization and mapping by exploring the advantage of two-stream fusion, long short-term memory network, loosely coupled framework and provides an end-to-end loosely coupled visual-inertial odometry.

At the present, classical loosely coupled visual-inertial odometry usually needs to calibrate parameters such as noise and bias, while the end-to-end learning-based method has tight coupling and low universality. In this paper, we propose an EE-LCVIO(End-to-End Loosely Coupled Visual-Inertial Odometry) with deep neural networks, which achieves an acceptable result, fewer parameters and lower coupling than VINet which is the state-of-the-art existing learning-based monocular visual-inertial odometry. In the case of without tight synchronized sensor data, EE-LCVIO in this paper surpasses the traditional monocular OKVIS（Open Keyframe-Based Visual-Inertial SLAM）by 78.7% and 31.3% with high robustness. Compared with the classical loosely coupled method, it does not need to calibrate any parameters between camera and IMU. Meanwhile, in order to address the dilemma that DeepVO which is the learning-based monocular visual odometry can leverage no more than the temporal information of 5 frames, we propose a TSVO(Visual Odometry with Spatial-Temporal Two-Stream Networks) using the adjacent image pairs and inter-frame dense optical flow as inputs. The results show that TSVO exceeds DeepVO by 44.6% and 43.3% in translation and rotation respectively, and exploits the sequential information of 10 consecutive frames.

The main achievement of this paper is to solve a part of the theoretical problems of collaborative computing and collaborative service-oriented to human-computer integration in the National Key R&D Program of China(No. 2018YFE0205503), the Funds for International Cooperation and Exchange of NSFC (No.61720106007) and the 111 Project(No.B18008).