计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
2025,34(8):125−138 [doi: 10.15888/j.cnki.csa.009941] [CSTR: 32024.14.csa.009941]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于注意力机制与局部交互的视觉惯性里程计①

王顺兰, 沈　艳

(成都信息工程大学 计算机学院, 成都 610225)
通信作者: 沈　艳, E-mail: sheny@cuit.edu.cn

摘　要: 视觉惯性里程计 (visual-inertial odometry, VIO) 通过融合视觉和惯性数据来实现位姿估计. 在复杂环境中, 惯性数据受噪声干扰, 长时间运动会导致累积误差, 同时大多数 VIO 忽略了模态间局部信息交互, 未充分利用不同模态的互补性, 从而影响位姿估计精度. 针对上述问题, 本文提出了一种基于注意力机制与局部交互的视觉惯性里程计 (attention and local interaction-based visual-inertial odometry, ALVIO) 模型. 首先, 该模型分别提取到视觉特征和惯性特征. 其次, 保留惯性特征的历史时序信息, 并通过基于离散余弦变换 (discrete cosine transform, DCT) 的通道注意力机制增强低频有效特征, 抑制高频噪声. 接着, 设计了多模态局部交互与全局融合模块, 利用改进的分散注意力机制与 MLP-Mixer 逐步实现模态间的局部交互与全局融合, 根据不同模态的贡献调节局部特征权重, 实现模态间互补, 再在全局维度上整合特征, 得到统一表征. 最后, 将融合的特征进行时间建模和位姿回归得到相对位姿. 为了验证模型在复杂环境下的有效性, 对公开数据集 KITTI 和 EuRoC 进行了低质量处理并实验, 实验表明, ALVIO 相较于直接特征拼接模型、多头注意力融合模型、软掩码融合模型, 平移误差分别减少了 49.92%、32.82%、37.74%, 旋转误差分别减少了 51.34%、25.96%、29.54%, 且具有更高的效率和鲁棒性.

关键词: 视觉惯性里程计; 位姿估计; 通道注意力; 分散注意力; MLP-Mixer

引用格式: 王顺兰,沈艳.基于注意力机制与局部交互的视觉惯性里程计.计算机系统应用,2025,34(8):125–138. http://www.c-s-a.org.cn/1003-3254/9941.html

Visual-inertial Odometry Based on Attention Mechanism and Local Interaction

WANG Shun-Lan, SHEN Yan

(School of Computer Science, Chengdu University of Information Technology, Chengdu 610225, China)

Abstract: Visual-inertial odometry (VIO) achieves pose estimation by fusing visual and inertial data. In complex environments, inertial data are prone to noise interference, and long-term motion leads to cumulative errors. Additionally, most VIO models overlook local information interaction between modalities and fail to fully utilize their complementary nature, thereby compromising pose estimation accuracy. To address these issues, this study proposes an attention and local interaction-based visual-inertial odometry (ALVIO) model. First, the model extracts visual features and inertial features. Then, the historical time-series information of the inertial features is preserved, and a channel attention mechanism based on discrete cosine transform (DCT) is applied to enhance low-frequency effective features and suppress high-frequency noise. Next, a multi-modal local interaction and global fusion module is designed, which gradually achieves local interaction and global fusion between modalities through improved split-attention mechanism and MLP-Mixer. This module adjusts the local feature weights based on the contributions of different modalities to realize inter-modal complementarity and then integrates the features globally to obtain a unified representation. Finally, the fused features are used for temporal modeling and pose regression to obtain relative poses. To verify the effectiveness of the

① 基金项目: 国家自然科学基金 (62172061); 四川省揭榜挂帅项目 (2023YFG0374)
收稿时间: 2024-12-10; 修改时间: 2025-02-12; 采用时间: 2025-03-06; csa 在线出版时间: 2025-06-20
CNKI 网络首发时间: 2025-06-23

model in complex environments, this paper conducts experiments on processed low-quality versions of the public KITTI and EuRoC datasets. The results show that, compared to the direct feature concatenation model, the multi-head attention fusion model, and the soft mask fusion model, ALVIO reduces the translation error by 49.92%, 32.82%, and 37.74%, respectively, and the rotation error by 51.34%, 25.96%, and 29.54%, respectively, while also demonstrating higher efficiency and robustness.

Key words: visual-inertial odometry (VIO); pose estimation; channel attention; split-attention; MLP-Mixer

1 相关工作

随着机器人导航和自动驾驶技术的快速发展, 里程计已经广泛应用于自主定位与导航任务[1]. 早期的研究主要基于单一传感器模式, 如视觉里程计 (visual odometry, VO)、惯性里程计. 然而, 单独使用相机或惯性传感器都有其局限性. 相机依赖于环境光照和丰富的纹理信息, 而惯性传感器则容易受到噪声干扰导致累积误差. 因此, 越来越多研究人员寻求多传感器融合的方法. VIO 通过融合相机和惯性传感器两种模态的数据, 已经成为实现位姿估计的关键技术之一. 现有的 VIO 主要分为传统的方法和基于深度学习的方法. 传统的方法通常遵循标准的处理流程, 包括特征检测与跟踪、传感器融合优化等. 这些方法主要依赖于手工设计的特征, 并通过滤波[2,3]或非线性优化[4−7]来融合视觉和惯性传感器的数据. 基于滤波的方法, 比如改进的 MSCKF 算法[2]使用扩展卡尔曼滤波器来进行状态的传播与更新, 虽然可以有效处理传感器数据, 但在特征提取和匹配错误、传感器噪声以及漂移等问题上表现欠佳. 基于非线性优化的方法, 如 DM-VIO[4]、VINS-Mono[6]和 ORB-SLAM3[7], 采用局部视觉惯性调整进行位姿估计, 通过光度或重投影误差构建运动约束, 但手动定义的数据关联和惯性测量单元 (inertial measurement unit, IMU) 建模偏差在复杂场景中往往不够准确. 尽管传统的方法在一定程度上解决了 VIO 问题, 但它们仍然面临特征提取和匹配错误、传感器噪声和漂移以及环境变化对视觉惯性特征的影响, 导致位姿估计结果的准确性受限.

深度学习为 VIO 研究提供了新思路, 展现出更强的鲁棒性. 与传统方法相比, 深度学习模型通过神经网络自适应训练实现高精度导航, 避免了数学模型与实际应用不符的问题, 此外, 深度学习的方式擅长提取代表自我运动的高级特征, 能够提升复杂环境中的导航性能[8]. 基于深度学习的 VIO 通常采用端到端的系统设计, 包括前端特征提取、后端特征融合、时间建模与位姿回归等模块. 前端特征提取方面, 视觉特征通常使用卷积神经网络从图像中提取. 例如 DeepVO[9]通过 FlowNet 编码器处理图像对, 并通过长短期记忆 (long short-term memory, LSTM) 网络层和全连接 (fully connected, FC) 网络层进行位姿回归. VFE-VO[10]采用卷积神经网络提取视觉特征, 通过中心差分卷积, 先激活后池化增强了特征表达能力, 实现了视觉特征增强的光流法视觉里程计. 惯性特征则通常采用 LSTM 处理 IMU 数据, 捕捉时间序列中的动态变化, 如 RNIN-VIO[11]通过 ResNet18 和 LSTM 网络建模惯性特征, 但仅使用最后一个时刻的惯性特征, 忽略了之前的历史信息, 以及惯性数据中噪声干扰, 长时间运动会导致累积误差, 导致位姿估计的准确性较低. VINet[12]通过两层双向长短期记忆网络 (bidirectional long short-term memory, BiLSTM) 建模惯性时序特征, 两层 BiLSTM 结构可以得到复杂的时间依赖关系, 但同样没有考虑到历史信息以及存在累积误差, 从而影响位姿估计的准确性.

在后端特征融合的研究中, 最简单的方法是直接将视觉和惯性特征拼接在一起. 例如, VINet 作为首个基于深度学习的端到端视觉惯性里程计, 在后端融合中直接拼接视觉和惯性特征, 而这种直接融合的方法无法充分利用不同模态数据的互补信息, 导致位姿估计准确度不高. 为了提高特征融合的可解释性和系统鲁棒性, Chen 等[13]提出在视觉惯性特征向量拼接后使用软融合和强融合两种模式进行特征选择, 但这种方式仍缺少显式联系, 在复杂情况下的位姿估计性能有所欠缺. SelfOdom[14]采用软注意力机制融合视觉和惯性特征, 当数据质量较低时会导致融合结果较差. Shinde 等[15]基于多头注意力机制建模后端融合模型, 能够实现特征显示融合, 但全局融合方式忽略了模态间局部信息, 并且多头注意力机制所需参数量较大, 计

算代价高. EMA-VIO[16]采用了一种外部注意力机制全局融合视觉和惯性特征, 同样忽略了模态间局部信息, 未能有效利用不同模态的互补性, 同时外部记忆存储会增加额外的内存需求.

综上所述, 尽管基于深度学习的 VIO 在复杂环境中鲁棒性较好, 但仍存在不足: (1) 在复杂环境中, 惯性数据受噪声干扰, 长时间运动会导致累积误差; (2) 忽略了模态间局部信息交互, 未充分利用不同模态的互补性, 从而影响位姿估计精度. 为此, 本文提出了一种基于注意力机制与局部交互的视觉惯性里程计模型 (attention and local interaction-based visual-inertial odometry model, ALVIO). 本文的主要贡献如下.

(1) 充分利用了惯性特征的时序信息. 采用两层 BiLSTM 提取惯性特征, 考虑到只使用最后一个时刻的惯性特征会导致缺乏完整的时序信息, 本文将每个时刻的惯性特征进行拼接, 保留了最后一个时刻之前的历史信息, 有助于更全面地描述物体的运动状态.

(2) 提出了基于离散余弦变换 (discrete cosine transform, DCT) 的通道注意力机制用于优化惯性特征. 首先, 通过 DCT 将惯性特征的时域信息转换成频域信息, 并通过频域注意力权重强调低频特征, 从而抑制高频噪声, 增强对长时间稳定运动的表示, 为位姿估计提供更可靠的特征.

(3) 设计了一种多模态局部交互与全局融合模块, 实现模态间的有效互补性与一致性表达. 首先, 将传统分散注意力扩展为支持多模态的形式, 通过对不同模态特征进行局部特征块划分并形成联合表达, 根据不同模态的贡献程度赋予特征块不同的权值, 增强模态间互补性. 最后, 通过 MLP-Mixer 在全局维度上对局部交互后的特征进行进一步整合, 得到不同模态的统一表征.

2 模型设计

本文提出的 ALVIO 模型其整体结构如图 1 所示, 该模型由输入层、特征提取层、特征融合层、输出层组成.

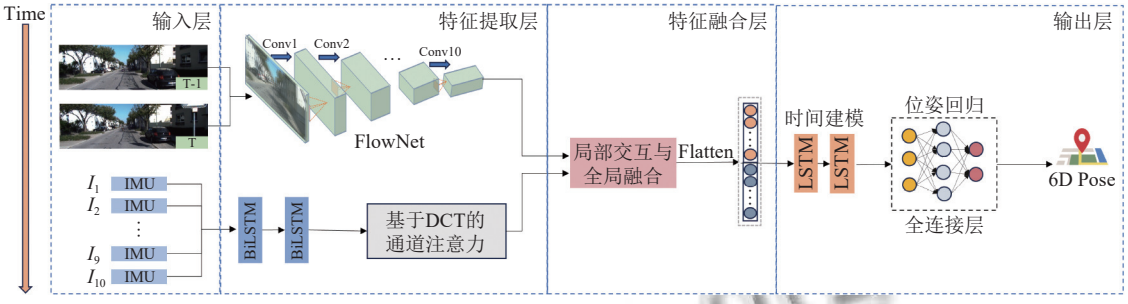


图 1 ALVIO 模型

2.1 输入层

输入数据主要由视觉传感器和惯性传感器所采集的多模态数据组成, 其中, 每次输入的视觉数据为连续的两帧单目图像, 即 T–1 和 T 时刻的图像. 由于提供的图像采集频率为 10 Hz, IMU 采集频率为 100 Hz, 因此, 输入惯性数据为两帧间所对应的 10 个的加速度和角速度信息, 即给定 t 时刻下的惯性数据为 I_t.

$$I_t = [a_x^t, a_y^t, a_z^t, \omega_x^t, \omega_y^t, \omega_z^t] \tag{1}$$

其中, $t \in \{1,2,3,\cdots,10\}$, a_x、a_y、a_z 表示在 xyz 方向的加速度分量, ω_x、ω_y、ω_z 表示在 xyz 方向的角速度分量.

2.2 特征提取层

2.2.1 视觉特征提取

由于 FlowNet 网络擅长处理光流预测[17], 能够从输入的连续图像中提取具有几何意义的特征, 因此, 视觉特征提取部分采用了 FlowNet 网络, 其网络结构如图 2 所示. 该网络由 10 个卷积层组成, 其中卷积核设计方面采用逐步缩小的感受野, 第 1 层的卷积核感受野为 7×7, 之后两层的感受野减少为 5×5, 再减少到 3×3, 从第 3 层开始, 每层后面都有一个 ReLU 非线性激活函数, 以增强特征的表达能力. 随着卷积核感受野的减少, 特征的通道数量增加, 可以使网络提取到更丰富更细节的视觉特征.

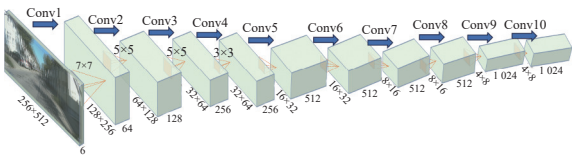


图 2 FlowNet 网络

2.2.2 惯性特征提取

惯性特征提取部分采用两层 BiLSTM 处理 IMU 数据, 每层 BiLSTM 由前向 LSTM 和后向 LSTM 组成, 其结构如图 3 所示.

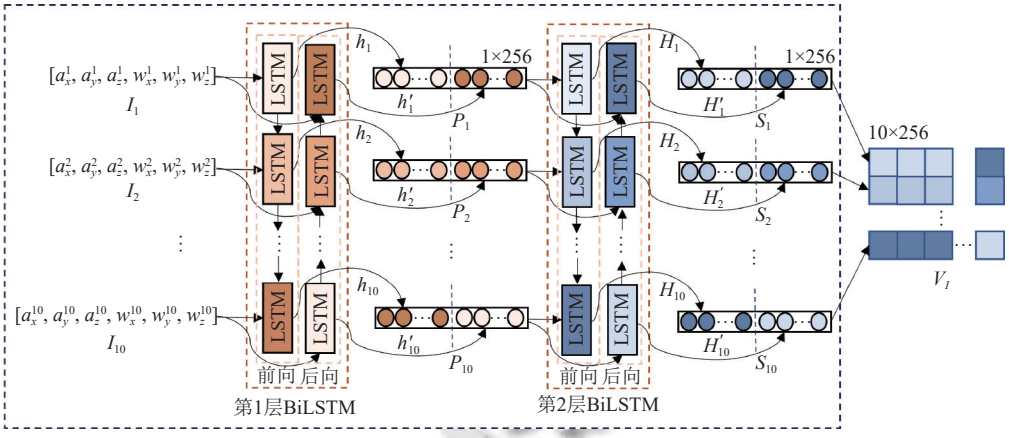


图 3 两层 BiLSTM

惯性数据 I_t 输入到第 1 层 BiLSTM. 在 BiLSTM 中的 LSTM 内部结构如图 4 所示.

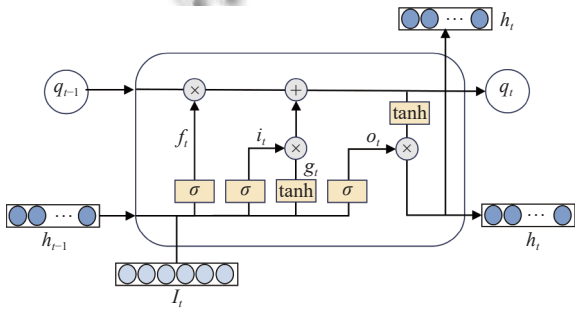


图 4 LSTM 内部结构

前向 LSTM 的过程表示为:

$$\begin{cases} f_t = \sigma(W^{fI}I_t + W^{fh}h_{t-1} + b_f) \\ i_t = \sigma(W^{iI}I_t + W^{ih}h_{t-1} + b_i) \\ g_t = \tanh(W^{gI}I_t + W^{gh}h_{t-1} + b_g) \\ q_t = g_t \otimes i_t \oplus q_{t-1} \otimes f_t \\ o_t = \sigma(W^{oI}I_t + W^{oh}h_{t-1} + b_o) \\ h_t = o_t \otimes \tanh(q_t) \end{cases} \tag{2}$$

其中, h 表示前向的隐藏状态, f 表示遗忘门, i 表示输入门, q 表示前向的记忆变量, o 表示输出门, W^{fI}、W^{iI}、W^{gI}、W^{oI} 为输入 I_t 对应的权重, W^{fh}、W^{ih}、W^{gh}、W^{oh} 为隐藏状态变量 h_{t-1} 对应的权重, b_f、b_i、b_g、b_o 为偏置向量, σ 为 Sigmoid 激活函数, tanh() 为 tanh 激活函数, $\oplus$和$\otimes$表示为矩阵加法运算和乘法运算.

后向 LSTM 的过程和前向类似, 区别在于输入不再是前一刻的隐藏状态而是下一刻的隐藏状态. 后向 LSTM 的过程表示为:

$$\begin{cases} f'_t = \sigma(W^{fI'}I_t + W^{fh'}h'_{t+1} + b'_f) \\ i'_t = \sigma(W^{iI'}I_t + W^{ih'}h'_{t+1} + b'_i) \\ g'_t = \tanh(W^{gI'}I_t + W^{gh'}h'_{t+1} + b'_g) \\ q'_t = g'_t \otimes i'_t \oplus q'_{t+1} \otimes f'_t \\ o'_t = \sigma(W^{oI'}I_t + W^{oh'}h'_{t+1} + b'_o) \\ h'_t = o'_t \otimes \tanh(q'_t) \end{cases} \tag{3}$$

同样, h' 表示后向隐藏状态, f' 表示遗忘门, i' 表示输入门, q' 表示前向的记忆变量, o' 表示输出门, $W^{fI'}$、$W^{iI'}$、$W^{gI'}$、$W^{oI'}$ 为输入 I_t 对应的权重, $W^{fh'}$、$W^{ih'}$、$W^{gh'}$、$W^{oh'}$ 为隐藏状态变量 h_{t+1} 对应的权重, b'_f、b'_i、b'_g、b'_o 为偏置向量.

每层 LSTM 有 128 个隐藏层, 因此得到的前后隐藏状态是通道数为 128 的向量, 将前后向的隐藏状态拼接可以得到 I_t 经过第 1 个 BiLSTM 的输出向量 P_t, $P_t \in R^{1\times256}$:

$$P_t = [h_t, h'_t] \tag{4}$$

其中, 前向隐藏状态 h_t 反映了开始时刻到 t 时刻的时间序列信息, 后向隐藏状态 h'_t 反映了最后时刻到 t 时刻的时间序列信息.

受 IONet[18]的启发, 对于复杂的惯性数据时, 单层 BiLSTM 只能提取到较基础的时序特征, 即加速度和角速度随时间的变换. 而通过堆叠两层 BiLSTM, 可以建模数据间存在的隐藏物理变量, 如初速度、重力, 有助于模型更好地理解惯性数据之间的复杂运动模式和

转换关系. 因此, 本文采用两层 BiLSTM 的结构, 将第一层的输出 P_t 作为下一层的输入, 得到第 2 层的前向隐藏状态 H 和后向隐藏状态 H', 同样进行拼接可以得到第 2 层的输出向量 S_t, $S_t \in R^{1\times256}$:

$$S_t = [H_t, H'_t] \tag{5}$$

在得到的惯性特征 S_t 中即包含开始时刻到 t 时刻的时间序列信息, 也包含最后时刻到 t 时刻的时间序列信息. 考虑到 LSTM 会遗忘一部分历史信息, 而完整的时序信息有助于更全面地描述物体的运动状态, 因此对所有时刻的输出特征进行拼接, 得到惯性特征矩阵 V_I, $V_I \in R^{10\times256}$, 其中 256 表示通道数:

$$V_I = concat[S_1, S_2, \cdots, S_{10}] \tag{6}$$

然后, 将得到的惯性特征矩阵 V_I 通过基于 DCT 的通道注意力层进行优化. 频率是分析时间序列的天然辅助手段[19]. 该层中, 通过 DCT 将特征中的时域信息转换成频域信息, 最后通过注意力强调低频有效信息, 抑制高频噪声. 基于 DCT 的通道注意力结构如图 5 所示.

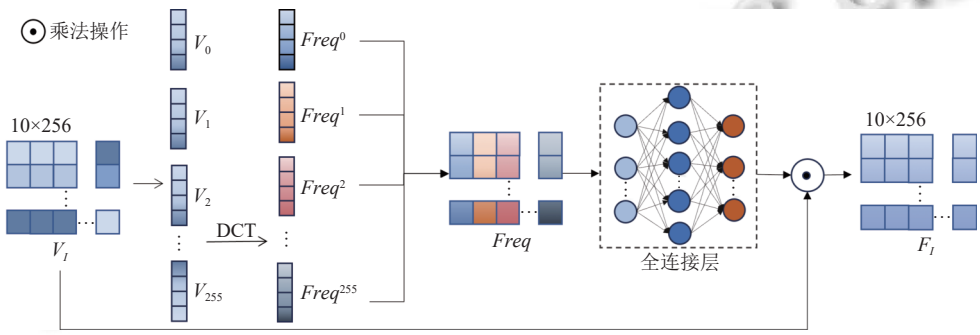


图 5 基于 DCT 的通道注意力

首先, 将特征矩阵 V_I 沿通道划分为 256 个子组, 其中 $V_i \in R^{1\times10}$, $i \in \{0,1,2,\cdots,255\}$.

其次, 将第 i 个子组在第 j 个时刻的特征 V_j^i 通过 DCT 将特征矩阵的时域信息转换成频域信息 $Freq_l^i$, 公式表达如下:

$$Freq_l^i = \sum_{j=0}^{L-1} V_j^i \cos\left(\frac{\pi l}{L}\left(j+\frac{1}{2}\right)\right) \tag{7}$$

其中, $j \in \{0,1,2,\cdots,9\}$, l 为频率索引, $l \in \{0,1,2,\cdots,9\}$, 在 $Freq_l^i$ 中, 低频部分表示较长期稳定的运动变换特征, 而高频部分则反映细节的运动变化或噪声信息, 将低频和高频特征进行拼接得到频域特征矩阵 $Freq$:

$$Freq = [Freq^0, Freq^1, \cdots, Freq^{255}] \tag{8}$$

然后将频域特征矩阵 $Freq$ 输入到两个全连接层中. 其中, 在第 1 个全连接层之后通过 ReLU 激活函数引入非线性, 在第 2 个全连接层之后引入 Sigmoid 激活函数, Sigmoid 激活函数将输出限制在 [0, 1] 范围内. 最后生成通道注意力权重 att:

$$att = \sigma(W_2\delta(W_1 Freq + b_1) + b_2) \tag{9}$$

其中, σ 是 Sigmoid 激活函数, δ 是 ReLU 激活函数, W_1 和 W_2 是全连接层的权重矩阵, b_1 和 b_2 是全连接层的偏置向量.

最后, 通过得到的注意力权重 att 来调整原始时序特征 V_I, 得到优化的惯性特征 F_I.

$$F_I = V_I \odot att \tag{10}$$

注意力权重能够可以动态地调整每个通道的重要性, 强调低频通道中与稳定运动变换相关的重要特征, 增强惯性特征的时序信息.

2.3 特征融合层

在特征融合中, 最简单的方法是直接融合或重新加权它们, 但这类方法无法捕捉模态之间的复杂交互, 导致特征融合不够充分, 不能有效利用模态间的关联信息. 在后续研究中, 逐渐倾向于采用深度神经网络进行特征融合, 如果训练得当, 基于深度神经网络的融合方式能够优先选择最有可能的假设, 有助于改善模型的归纳偏差, 能够更好地挖掘多模态特征间的关联性, 并增强模型的表达能力, 然而, 现有的特征融合通常通过全局进行融合, 没有考虑到局部信息之间的交互, 未充分发挥不同特征之间的互补性.

因此, 本文设计了一种多模态局部交互与全局融合模块, 实现视觉和惯性特征的有效互补性与一致性表达. 其整体结构如图 6 所示, 该模块主要包含改进的

分散注意力层和全局 MLP-Mixer 层.

通过特征提取层得到特征映射 $F_m \in R^{N_m \times K_m}$, $m \in \{V, I\}$, 其中 V、I 代表视觉和惯性两个模态, N_m 是模态 m 的通道数, K_m 对于视觉模态指的是特征图的宽×高, 对于惯性模态指的是时序长度, 即 K_V=32, K_I=10. 在该模块, 先将得到的特征映射 F_m 输入到局部分散注意力层进行局部交互, 再通过全连接层分别得到视觉与惯性的全局特征, 再通过全局 MLP 层实现统一表征.

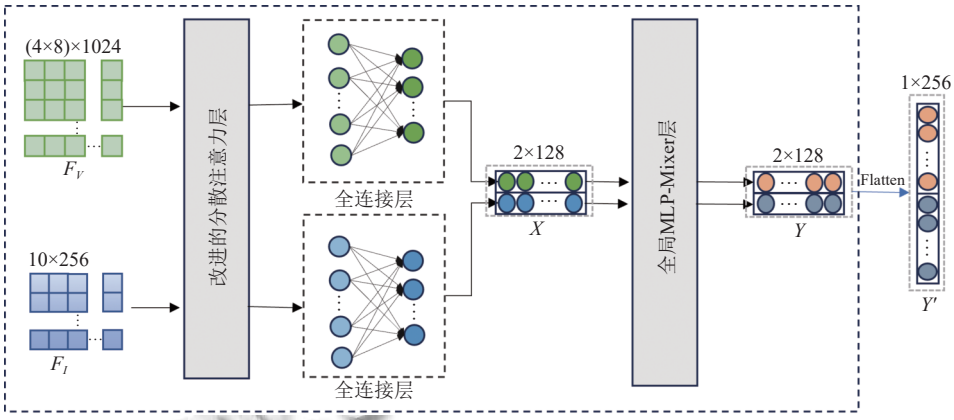


图 6 局部交互与全局融合模块

2.3.1 改进的分散注意力层

在改进的分散注意力层中, 受到 ResNeSt 网络[20]中 Split-Attention 的启发, 本文采用改进的 Split-Attention 来实现视觉和惯性特征之间的局部交互. 原始的 Split-Attention 能够让同一模态的不同特征进行交互, 本文改进的 Split-Attention 可以接受不同模态进行局部交互. 该层主要分为 3 个步骤: 分离、联合、强调, 结构如图 7 所示.

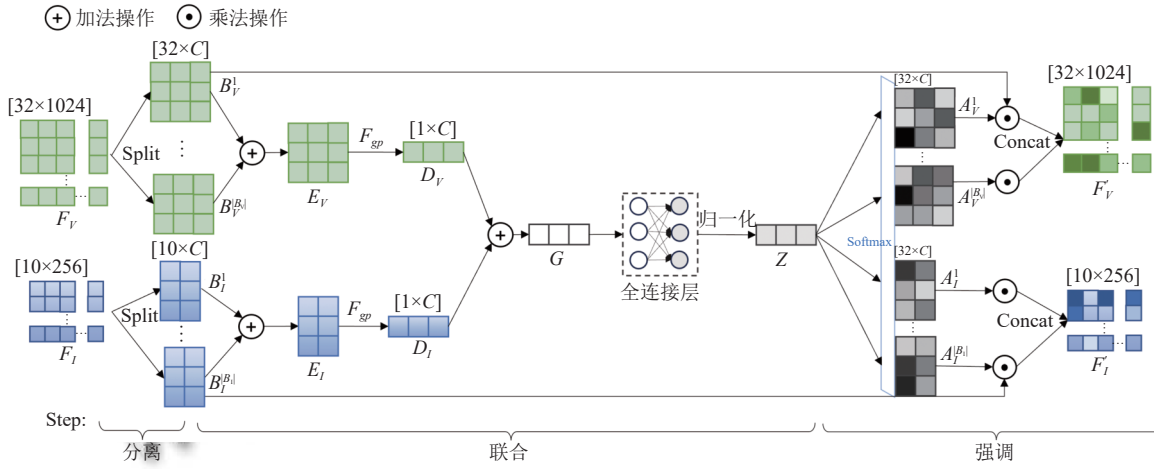


图 7 改进的分散注意力层

(1) 分离: 将特征映射 F_m 按通道分割成等通道的特征块, C_m 表示模态 m 的总通道数, C_V=1024, C_I=256. B_m 表示模态 m 的特征块集合, $|B_m|$表示模态 m 的特征块数量, 每个特征块包含固定数量的通道 C, 表示公式如下:

$$|B_m| = \lceil C_m / C \rceil \tag{11}$$

其中, 当总的通道数 C_m 不是 C 的倍数时, 最后一个特征块在缺失的通道中用 0 填充.

分离操作确保了不同模态的数据在相同的尺度上进行交互, 有助于后续操作, 而且特征块划分方式有效减少了模型计算量.

(2) 联合: 先将分割后的特征块集合 B_m 进行求和, 得到模态内的联合表达 E_m, 并在 k_m 维度上进行全局平均池化 F_{gp}, 得到不同模态的通道描述符 D_m, 表达公式

如下:

$$E_m = \sum_{i=1}^{|B_m|} B_m^i \tag{12}$$

$$D_m = F_{gb}(E_m) \tag{13}$$

$$= \frac{1}{K_m} \sum_{j=1}^{K_m} \sum_{i=1}^{|B_m|} B_m^i(j) \tag{14}$$

其中, B_m^i 是 B_m 中第 i 个特征块, $i \in \{1, 2, \cdots, |B_m|\}$, j 是 k_m 维度上的位置索引.

其次, 将视觉和惯性模态描述符的元素和形成多模态通道描述符 G. 再将得到的多模态通道描述符 G 通过全连接层, 并进行批量归一化操作, 之后使用 $ReLU$ 激活函数. $ReLU$ 函数能够将多模态通道描述符 G 中负值置为 0, 正值保持不变. 最后, 得到模态间联合表示 Z, 包含了不同模态的局部关联性, 表达公式如下:

$$Z = ReLU(W_Z G + b_Z) \tag{15}$$

$$= ReLU(W_Z(D_V + D_I) + b_Z) \tag{16}$$

其中, W_Z 全连接层的权重矩阵, b_Z 是全连接层的偏置向量.

联合步骤通过提取并整合各模态的通道描述符, 得到跨模态的共同特征, 形成视觉与惯性特征的联合表达, 以实现模态间的局部交互.

(3) 强调: 首先, 对于第 i 个特征块 B_m^i, 通过对模态间联合表示 Z 应用线性变换生成相应的 Logits U_m^i, Logits U_m^i 可以表示为 m 模态中第 i 个特征块的重要性分数. 再将得到 Logits U_m^i 之后使用 Softmax 激活函数获得块级注意力权重 A_m^i, 表达公式如下:

$$U_m^i = W_m^i Z + b_m^i \tag{17}$$

$$A_m^i = \frac{\exp(U_m^i)}{\sum_{k}^{M} \sum_{j}^{|B_k|} \exp(U_k^j)} \tag{18}$$

其中, W_m^i 表示 Logits U_m^i 所对应的权重, b_m^i 表示偏置, 在式 (18) 中分子表示 U_m^i 的指数函数, 分母表示对其中所有指数值做归一化操作, 从而将特征块得分转换为注意力权重.

其次, 通过得到的注意力权重 A_m^i 来调整原始时序特征块 B_m^i, 得到优化的特征块 $B_m^{i\prime}$, 表达公式如下:

$$B_m^{i\prime} = A_m^i \odot B_m^i \tag{19}$$

最后, 通过拼接分别将属于视觉和惯性优化的特征块合并, 生成优化的特征映射 F_m':

$$F_m' = \left[B_m^{i\prime}, B_m^{i\prime}, \cdots, B_m^{|B_m|\prime}\right] \tag{20}$$

强调步骤根据各个模态的贡献程度自适应地为视觉和惯性特征块分配不同的注意力权重, 特别是在特征缺失或者质量较差时, 选择更适合当前位姿估计的模态和重要特征信息, 实现模态间有效互补.

2.3.2 全局特征提取

由式 (20) 得到优化的视觉惯性特征分别通过全连接操作提取全局特征, 同时统一模态间的维度, 得到视觉和惯性的特征向量, 并进行拼接得到视觉惯性特征矩阵 X, $X \in R^{2\times 128}$, 2 表示视觉和惯性两个模态, N 表示通道数量:

$$\tilde{F}_m = W_m F_m' + b_m \tag{21}$$

$$X = concat(\tilde{F}_V, \tilde{F}_I) \tag{22}$$

其中, W_m 是 m 模态的权重矩阵, b_m 是 m 模态的偏置.

2.3.3 全局 MLP 层

局部分散注意力中特征块划分方式有效减少了模型计算量, 但不同特征也局限于特征块之间, 只能进行局部交互, 缺乏全局关联, 因此在该层之后引入全局 MLP-Mixer 层. 考虑到 MLP-Mixer[21–23]模型可实现全局多模态特征融合, 且具有较低的计算复杂度和较高的灵活性. 因此, 本文采用 MLP-Mixer 进行全局融合.

将式 (22) 得到的特征矩阵 X 输入全局 MLP-Mixer 层, 结构如图 8 所示. 该层由多个相同的 MLP 层组成, 每个 MLP 层由两个全连接层和一个 GELU 激活函数组成.

特征矩阵 X 先进行通道混合, 实现模态内通道之间相互融合. 选择特征 X 的视觉模态/惯性模态的所有通道的数据记为 $X_{i,*}$, $i \in \{V, I\}$. 先将 X 通过 Layer Norm 归一化处理, 再将 $X_{i,*}$ 通过 MLP1 层. 在 MLP1 层中, $X_{i,*}$ 通过一层全连接层线性变换之后, 引入 GELU 激活函数增加模型非线性, 再次通过全连接层, 得到通道融合后的特征 $X_{i,*}'$. 最后采用 Skip-connections 跳跃连接结合原始通道信息 $X_{i,*}$ 和融合之后的通道信息 $X_{i,*}'$, 得到通道混合后的特征矩阵 $Z_{i,*}$. 表达公式如下:

$$X_{i,*}' = W_2 \Phi(W_1 Norm(X)_{i,*}) \tag{23}$$

$$Z_{i,*} = X_{i,*} + X'_{i,*} \tag{24}$$

其中, W_1 和 W_2 是 MLP1 中全连接层的权重矩阵, Φ 表示 GELU 激活函数, $Norm()$ 表示归一化层.

得到矩阵 Z 之后, 同样的先进行归一化处理, 再将 $Z_{*,j}$ 通过 MLP2 进行模态混合, $Z_{*,j}$ 表示选择第 j 个通道的不同模态的数据, j 范围为 1–N, 在图 8 中用转置表示所选择特征的维度不一样. MLP2 的结构与 MLP1 一样, 通过 MLP2 之后得到模态融合后的特征 $Z'_{*,j}$. $Z_{*,j}$ 最后同样采用 Skip-connections, 得到通道混合后的特征矩阵 $Y_{*,j}$, 最终实现特征的充分融合:

$$Z'_{*,j} = W_4\Phi(W_3 Norm(Z)_{*,j}) \tag{25}$$

$$Y_{*,j} = Z_{*,j} + Z'_{*,j} \tag{26}$$

其中, W_3 和 W_4 是 MLP2 中全连接层的权重矩阵, Φ 表示 GELU 激活函数, $Norm()$ 表示归一化层. 最后将 Y 展平为融合向量 Y'.

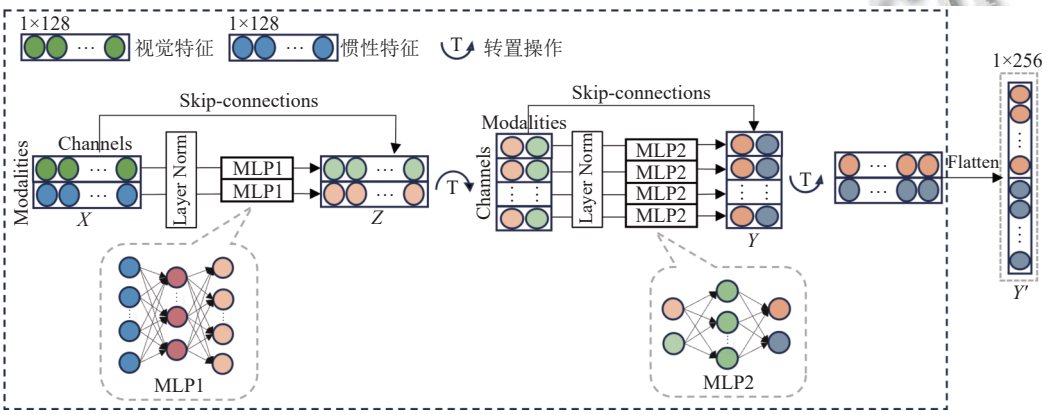


图 8 全局 MLP-Mixer 层

2.4 输出层

输出层由时间建模和位姿回归组成. 该层核心在于建模序列特征的时间依赖性, 以确保对连续的位姿变化进行准确的预测. 自我运动估计不仅需要考虑单一时刻的特征, 还需要捕捉不同时间之间的特征关联. 在时间建模中, 采用两层 LSTM 学习复杂的运动模型并推导出序列特征之间的关联, 同样两层结构可以捕获更复杂的时间依赖关系, 能够得到深层次的特征关联. 将最终的融合向量 Y'输入到两层 LSTM 模型中, 来捕捉序列数据的动态变化. LSTM 会接收时间 T–1 到 T 的输入特征 $Y_{T-1,T}$ 以及前一个时刻的隐藏状态 $h_{T-2,T-1}$:

$$P_{T-1,T} = f_{\text{LSTM}}(Y_{T-1,T}, h_{T-2,T-1}) \tag{27}$$

其中, f_{LSTM} 表示两层 LSTM 的递归函数, 通过学习序列中特征的时间依赖性, 输出与当前时刻相对位姿相关的特征 $P_{T-1,T}$. 时间建模之后, 通过两层全连接层对特征进一步处理, 以回归得到最终的 6D Pose, 6D Pose 表示时间 T–1 到 T 的相对平移和相对旋转.

2.5 损失函数

损失函数表示真实值与预测值之间的差距, 模型训练的目标是最小化损失函数值. 本文模型的训练损失函数由预测相对位姿 P 和真实相对位姿 P'之间的均方误差之和构成, 公式如下:

$$Loss = \| P_{tr} - P'_{tr} \|_2^2 + \beta \| P_r - P'_r \|_2^2 \tag{28}$$

其中, P_{tr} 和 P_r 表示模型预测的相对平移和由欧拉角表示的相对旋转, P'_{tr} 和 P'_r 表示对应真实的相对平移和由欧拉角表示的相对旋转. β 是尺度因子, 用于平衡位置和平移的权重, 使得在训练过程中平移和旋转得到类似的重视程度.

3 实验分析

3.1 数据集

(1) KITTI 里程计数据集

本文使用 KITTI 里程计数据集. KITTI 里程计数据集展示了城市道路上的室外驾驶环境. 该数据集包含丰富的传感器数据, 包括图像、IMU 和 GPS 数据. 图像由安装在车辆上的相机采集, 提供前视立体图像对, 分辨率为 1241×376 像素. IMU 数据主要使用加速度和角速度, 陀螺仪测量角速度的单位是弧度每秒 (rad/s), 而加速度计测量的单位是 m/s^2. 其中提供的图像和地真值采集频率为 10 Hz, IMU 数据采集频率为

100 Hz. KITTI 里程计数据集有 22 个序列, 但只有 00–10 有用于训练的地面实况轨迹, 而其中序列 03 由于缺失 IMU 数据, 相应的原始文件不可用, 因此选取其中的序列 00、01、02、04、05、06、07、08 进行训练, 10% 作为验证集, 在序列 09 和 10 上进行测试[24].

(2) EuRoC 数据集

EuRoC 数据集通过 AscTec Firefly 微型无人机在 3 种不同的室内环境中执行 6 自由度运动时手动操控录制, 共包含 11 个序列. 在每个环境中, 序列的难度随着序列编号的增加逐渐提升. 所有 EuRoC 序列均由一个前置视觉-惯性传感器单元录制, 立体相机和 IMU 的时间戳紧密同步, 数据采集由 MAV 完成. 本文仅使用左相机图像作为单目图像, 采样频率为 20 Hz, 同时 Skybotix VI IMU 传感器捕获的加速度和角速度数据, 采样频率为 200 Hz. 在 Vicon Room 序列中, 真实轨迹由 Vicon 运动捕捉系统提供, 在机器大厅序列中, 真实轨迹由 Leica MS50 激光跟踪仪提供. 与 KITTI 数据集不同, EuRoC 数据集是在室内非结构化路径上录制, 具有运动模糊, 轨迹呈现高度不规则的路径. 在实验中, 将图像和 IMU 数据分别降采样到 10 Hz 和 100 Hz, 并使用 MH_03 和 MH_05 序列进行测试, 其余序列用于训练, 且 10% 的训练数据用于验证集.

3.2 数据低质量处理

在许多现实世界的场景中, 传感器因为各种原因, 会导致数据损坏. 对于相机来说, 这些损坏可能以遮挡、噪声、光线变化、模糊图像以及缺失图像帧的形式出现. IMU 可能在陀螺仪数据中存在偏差, 也可能在加速度计数据中存在噪声. 为了对传感器低质量数据的影响进行广泛的研究, 并评估所提出方法的性能, 本文通过在原始数据中添加各种类型的噪声和遮挡来模拟现实情况, 主要有以下 3 大类低质量数据.

(1) 视觉数据处理

遮挡: 为模拟由于传感器上的灰尘或污垢、或传感器附近的静止物体可能导致的遮挡, 本文在样本图像的顶部随机覆盖一个尺寸为 128×128 像素的遮罩.

模糊+噪声: 为模拟相机移动或光线条件变化引起的运动模糊和噪点, 本文对样本图像应用高斯模糊 ($\sigma = 15$ 像素) 和添加噪声.

曝光/过暗: 为模拟不良的光照条件 (如过曝或过暗), 本文根据随机选择的曝光类型调整样本图像亮度. 如果是过曝, 则将图像亮度提高 (×8), 如果是过暗, 则降低图像亮度 (×0.5), 最后将图像值范围限制在 0–255 之间.

缺失数据: 为模拟数据传输中可能的数据包丢失, 本文随机删除 10% 的样本图像. 这种情况通常发生在数据包由于网络过载或临时传感器断开连接而从传输总线上丢失时.

(2) IMU 数据处理

噪声+偏置: 考虑到传感器可能因温度变化或机械冲击而引入噪声和偏置, 本文在加速度计和陀螺仪数据中添加了额外的白噪声和固定偏置, 模拟传感器在现实使用中遭受的环境和物理影响.

缺失数据: 为模拟 IMU 测量不稳定或者数据传输中可能的数据包丢失, 本文随机移除两个连续的视觉帧之间的惯性样本窗口.

(3) 跨传感器数据处理

空间错位: 通过随机改变相机与 IMU 之间的相对旋转角度, 模拟由于设备安装不当或校准错误导致的空间错位. 本文通过旋转矩阵实现, 每个 IMU 数据点都会应用旋转矩阵, 模拟最高 10°的错位.

时间错位: 通过在视觉和惯性数据流之间引入时间偏移, 模拟独立传感器子系统间的时钟漂移. 本文在每个 IMU 数据点的时间戳上随机应用小幅度偏移, 模拟最高 500 ms 的时间错位.

在后续实验中, 除使用公开的原始 KITTI、 EuRoC 数据集外, 本文使用这 3 种低质量数据场景来系统性地测试所提出方法的鲁棒性和有效性. 首先, 独立评估视觉数据的低质量情况, 模拟现实环境中的遮挡、模糊加噪声、不良光照 (过曝或过暗) 以及数据缺失; 其次, 测试仅包含 IMU 数据低质量的场景, 干扰包括白噪声、偏置以及数据丢失; 最后, 设计了一种包含视觉和 IMU 两种传感器低质量数据的综合场景, 进一步验证模型的性能.

3.3 实验设置

网络架构是用 PyTorch 实现的, 并在 RTX 4090D GPU 上进行了训练. 为了提取这两帧图像之间的光流信息, 本文采用了预训练的 FlowNetS 模型, 该模型在 FlyingChairs 数据集上进行训练, 能够预测真实场景中的光流. 同时, 为了适应 GPU 的内存大小, KITTI、EuRoC 数据集的图像分别被调整为 512×256、376×240. 为了生成更多训练数据以提高模型的泛化能力并防止过拟合, 序列被分割成长度一样的多条轨迹, 长度设置为 3.

包括基线在内的所有网络都使用 Adam 优化器以 8 个批处理大小进行训练, epoch 设置为 200, 学习率 lr 设置为 0.000 1, LSTM 网络中 dropout 设置为 0.2, 隐藏层设置为 512. 由文献[25]证明, 在融合模块中固定通道数最佳值可以由 $\min\{C_V, C_I\}/2$ 得到, 因此将通道数设置为 128.

3.4 结果分析

对于 KITTI 数据集, 我们使用官方评价工具进行评估. 评估指标为在长度为 100–800 m 的子序列上计算的平均平移 RMSE t_{rel} (%) 和平均旋转 RMSE r_{rel} (°/100 m). t_{rel} 和 r_{rel} 衡量的是模型预测的值与实际值之间的差距, 越低表示模型的预测越准确.

本文选择了 4 种基线模型: 仅使用视觉估计位姿的模型 (VO)、直接融合视觉与惯性数据的模型 (VIO-D)、基于软掩码融合模型 (VIO-soft), 以及基于多头注意力机制的融合模型 (VIO-A). 其中, VO 模型仅依赖视觉特征进行位姿估计, 包含视觉特征提取、时间建模和位姿回归模块; VIO-D 模型则通过简单的通道拼接实现视觉与惯性数据的融合; VIO-soft[13]通过软掩码进行特征融合, 将视觉和惯性特征与它们对应的软掩模相乘作为新的重加权向量, 这一过程与注意力机制有相似之处; 而 VIO-A[26]则采用多头注意力机制进行融合, 除此之外的部分采用与本文相同的框架. 实验结果如表 1 所示.

表 1 基线模型与本文模型相对轨迹误差对比

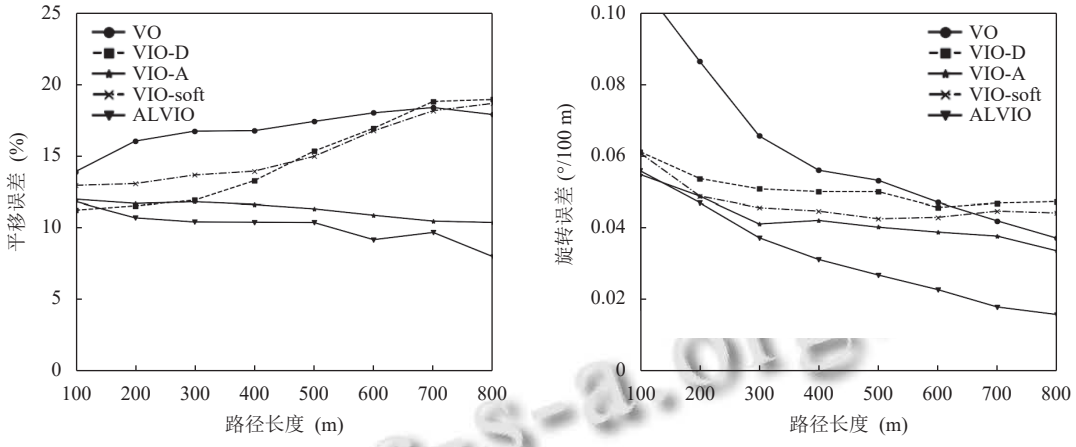
模型	融合模块参数量	N		I		V		ALL	
		t_{rel} (%)	r_{rel} (°/100 m)	t_{rel} (%)	r_{rel} (°/100 m)	t_{rel} (%)	r_{rel} (°/100 m)	t_{rel} (%)	r_{rel} (°/100 m)
VO	—	27.28	5.91	31.68	16.59	35.08	11.76	56.07	15.09
VIO-D	—	14.86	4.21	17.66	5.90	16.80	4.76	18.35	5.98
VIO-A	1 313 280	11.37	3.93	12.52	4.29	12.87	3.83	13.68	3.93
VIO-soft	**262 656**	11.44	3.86	12.58	4.70	14.03	4.21	14.76	4.13
ALVIO	265 570	**8.35**	**2.08**	**8.55**	**2.19**	**9.44**	**2.43**	**9.19**	**2.91**

注: 加粗为最优值; N表示原始KITTI数据, I表示仅低质量惯性数据, V表示仅低质量视觉数据, ALL表示包含所有低质量数据.

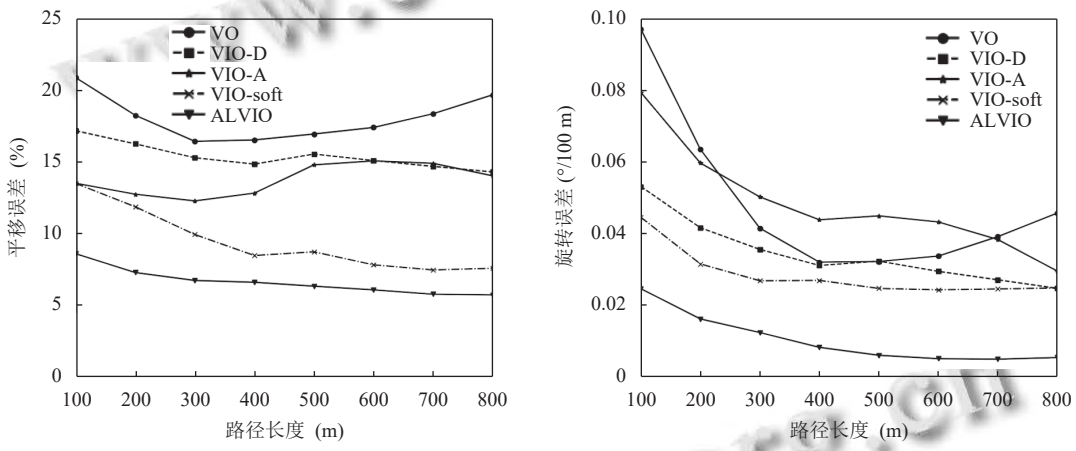
可以看出, 本文模型的平移和旋转平均 RMSE 值均低于基线模型, 在多种低质量数据场景下表现最佳. 对于 VO 模型, 原始 KITTI 数据中的平移和旋转平均 RMSE 值为 27.28 和 5.91, 其位姿估计误差较差, 当惯性和视觉都进行低质量处理之后, 平移和旋转的平均 RMSE 值为 56.07 和 15.09, 可见当数据质量下降时, 平移和旋转估计的性能大幅下降. 相比之下, 加入惯性数据的模型在面对低质量数据时依然表现稳定, 比如 VIO-D 在原始数据下的平移和旋转的平均 RMSE 值为 14.86 和 4.21, 在低质量数据下的平移和旋转的平均 RMSE 值为 18.35 和 5.98, 说明多模态融合有效地增强了系统的鲁棒性. 值得注意的是, 当视觉数据质量下降时, 各方法的平移误差增幅较大, 而旋转误差相对稳定; 相反, 在惯性数据质量下降时, 旋转误差增加显著, 而平移误差变化较小, 比如 VIO-A 在原始数据下的平移和旋转的平均 RMSE 值为 11.37 和 3.93, 在视觉数据质量下降时的平移和旋转的平均 RMSE 值为 12.52 和 4.29, 在惯性数据质量下降时的平移和旋转的平均 RMSE 值为 12.87 和 3.83. 这表明视觉特征主要有助于平移估计, 而惯性数据更适合旋转估计. 在融合方式的比较中, 本文将视觉和惯性特征充分融合, 有效实现了模态间的互补性, 并且在低质量数据场景下表现突出. 与基线融合模块的参数量相比, VIO-D 直接拼接的方式简单, 但未能发挥不同模态的作用, 在低质量下 ALVIO 相较于 VIO-D 平移误差和旋转误差分别减少了 49.92% 和 51.34%. 参数量大的模型如 VIO-A 虽然位姿估计的效果好, 但代价是更多的计算资源消耗. 而 VIO-soft 和 ALVIO 在保持低参数量的同时也能够取得接近甚至更好的性能. 特别是 ALVIO, 在原始数据集的平移误差分别比 VIO-soft 和 VIO-A 降低了 27.01% 和 26.56%, 旋转误差则降低了 46.11% 和 47.07%. 当视觉和惯性数据质量同时下降时, ALVIO 的平移误差分别比 VIO-soft 和 VIO-A 降低了 37.74% 和 32.82%, 旋转误差则降低了 29.54% 和 25.95%. 因此, 本文模型在各种场景下展现了最佳的综合表现.

除此之外, 在序列 09 和序列 10 上, 我们对提出的模型进行了平移和旋转误差的评估, 覆盖了 100–800 m 的各个分段, 结果如图 9 所示. 分段误差的分析能够更加细致地展示各个方法之间的性能差异. 从结果中可以明显看出, VO 的分段误差始终是所有方法中最高的, 单模态方法由于缺乏多模态之间的互补性, 鲁棒性较差, 导致性能较弱. 相比之下, VIO-soft 和 VIO-A 的分段误差相对接近. 而我们的模型 ALVIO, 无论是在序列 09 还是序列 10 上, 其分段误差始终低于所有基

线模型, 显示了不错的表现. 而且在长距离评估中, 分段误差在逐渐下降, 表明模型在训练过程中逐步逼近真实轨迹, 表现出越来越高的稳定性, 模型能够更好地捕捉位姿变化趋势, 与真实轨迹的偏差逐渐缩小.



(a) 序列09上不同路径长度的平移和旋转的分段误差



(b) 序列10上不同路径长度的平移和旋转的分段误差

图 9 基线模型与本文模型在序列 09 和序列 10 的分段误差图

为了更全面地评估所提出的融合方法, 本文除了与提出的基线比较, 还将提出模型与其他基于深度学习的 VIO 模型 (如 EMA-VIO) 以及传统 SLAM 方法 (如 ORB-SLAM、VINS-Mono、OKVIS) 进行对比. 考虑到传统模型在低质量情况下会有严重漂移, 因此仅采用原始 KITTI 数据集进行实验对比. EMA-VIO 是由外部注意力机制进行融合视觉与惯性, 在惯性特征提取时采用 WaveNet 模块. 在传统的 SLAM 方法中, ORB-SLAM[27]是一种基于特征点的 SLAM 系统, 使用 ORB 特征进行定位和建图, VINS-Mono 采用紧耦合的方式, 将单目相机和 IMU 传感器数据结合在一起, 通过滑动窗口优化模型实现视觉和惯性数据的融合, OKVIS 则通过非线性优化方法, 将关键帧的视觉特征和 IMU 数据融合在一起. 如表 2 所示.

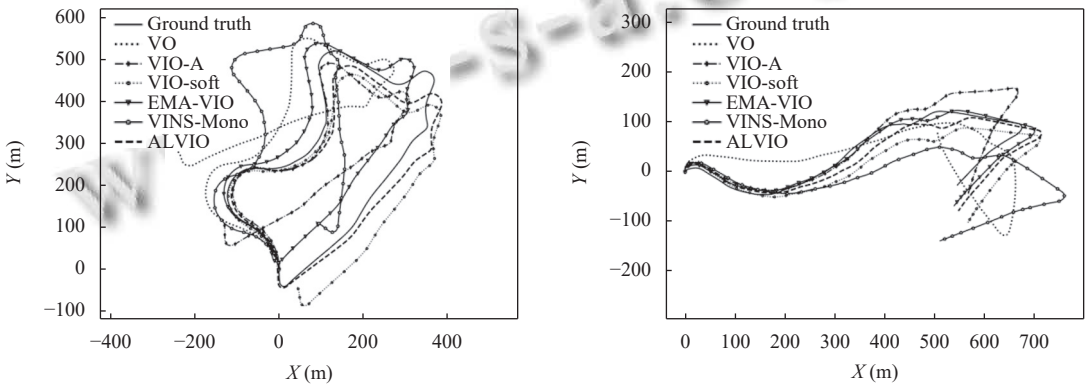
表 2 各算法相对轨迹误差对比

模型	序列09		序列10		平均	
	t_{rel} (%)	r_{rel} (°/100 m)	t_{rel} (%)	r_{rel} (°/100 m)	t_{rel} (%)	r_{rel} (°/100 m)
VO	26.69	6.53	27.87	5.28	27.28	5.91
VIO-D	14.17	4.73	15.54	3.68	14.86	4.21
VIO-A	11.47	3.53	11.26	4.33	11.37	3.93
VIO-soft	12.77	4.74	10.10	2.97	11.44	3.86
EMA-VIO	9.68	2.54	**8.46**	2.26	9.07	2.40
ORB-SLAM	23.19	7.62	28.35	6.69	25.77	7.16
VINS-Mono	21.30	7.52	19.59	4.14	20.45	5.83
OKVIS	13.38	4.00	15.01	3.68	14.20	3.84
ALVIO	**7.84**	**1.96**	8.85	**2.20**	**8.35**	**2.08**

可以看出, 基于深度学习的方法相较于传统 VIO 系统表现出优势, 可能是由于 KITTI 数据集中相机和 IMU 没有严格校准和时间同步, 这对手工设计的 VIO

系统构成挑战. 在所有模型中, ALVIO 的平移和旋转平均误差均优于其他方法. 具体来讲, 与传统方法 ORB-SLAM、VINS-Mono、OKVIS 相比, ALVIO 的平移误差分别降低了 67.5%、59.2%、41.8%, 旋转误差分别降低了 70.9%、64.3%、45.4%, 与误差最小的 EMA-VIO 相比, 平移误差降低了 7.94%, 旋转误差降低了 13.33%, 证明了本文模型的有效性. 外部注意力机制主要特点是引入外部存储, 以增强模型的记忆和信息处理能力, 但外部记忆存储会增加额外的内存需求. 而本文模型通过网络的逐步融合机制进行学习, 将不同模态特征组合起来, 以保持和充分利用各模态的特征信息, 确保每个模态对最终预测的贡献, 无需依赖外部记忆, 在实际应用中更为灵活. 图 10 展示了各模型在序列 09 和序列 10 上的轨迹估计对比结果, 进一步验证了本文模型的性能较好.

在 EuRoC 数据集中, 同样使用基线模型 VO、VIO-D、VIO-A、VIO-soft 以及传统模型 VINS-Mono、OKVIS 进行实验对比. 为了验证本文模型的鲁棒性和有效性, 将原始数据集进行低质量处理并进行实验对比, 实验结果如表 3 所示. 同时, 与传统模型实验对比如表 4 所示. 在评价指标方面, 本文计算了序列的绝对轨迹误差 ATE, 利用官方评价工具 EVO 进行评估.



(a) 序列 09　　(b) 序列 10

图 10　各模型在序列 09 和序列 10 的轨迹图

表 3　本文模型与基线模型绝对轨迹误差对比 ATE (m)

模型	N	I	V	ALL
VO	1.34	2.14	2.83	3.12
VIO-D	0.99	1.30	1.55	2.56
VIO-A	0.58	1.05	1.28	1.72
VIO-soft	0.68	1.18	1.33	2.05
ALVIO	**0.42**	**0.62**	**0.65**	**1.18**

注: N表示原始EuRoC数据, I表示仅低质量惯性数据, V表示仅低质量视觉数据, ALL表示包含所有低质量数据.

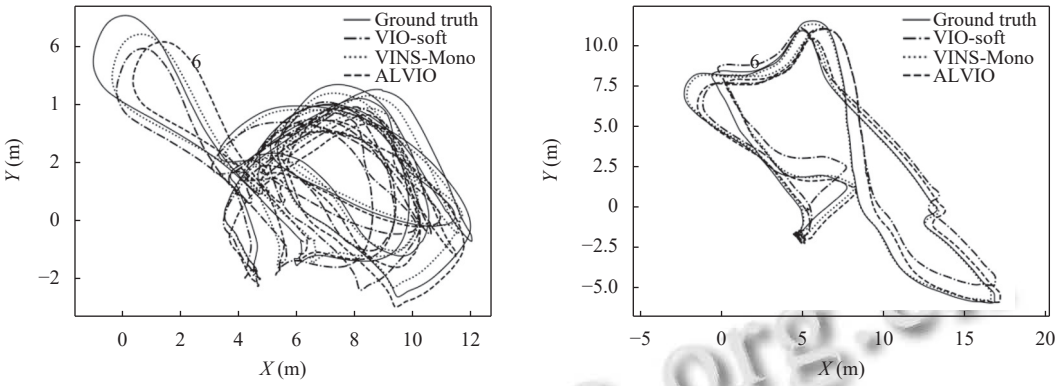
表 4　本文模型与传统模型绝对轨迹误差对比 ATE (m)

模型	MH_03	MH_05	Mean
VINS-Mono	**0.37**	**0.32**	**0.35**
OKVIS	0.51	0.46	0.49
ALVIO	0.49	0.35	0.42

在表 3 中可以看出, 无论是原始数据还是低质量的情况下, 本文模型优于其他的基线模型. 在表 4 中, 与传统模型相比, 本文模型在序列 MH_03 和序列 MH_05 上都优于 OKVIS, 误差分别降低了 3.92% 和 23.91%. 由于 EuRoC 数据集中相机和 IMU 数据紧密同步, 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, 因此其性能高于本文模型. 此外, EuRoC 数据集中的 IMU 数据包含较大的噪声和偏置, 对视觉信息产生干扰, 在 IMU 数据质量较差且与视觉信息难以相互补充的情况下, 传统方法会不断修正测量值, 以优化位姿估计, 因此传统模型在原始数据集中的表现优于深度学习模型. 然而, 在现实环境中, 传感器数据同步存在局限性, 且数据质量常常不稳定. 相比之下, 本文模型具有较强的鲁棒性, 能够在复杂和不确定的环境中保持较为稳定的性能, 在实际应用中更具实用性.

图 11 展示了各模型在序列 MH_03 和序列 MH_05 上的轨迹估计对比结果, 进一步验证了本文模型的性能较好. 根据以上实验对比展示, 本文模型展现出较好的稳定性和性能优势. 与其他融合方式相比, 局部交互与全局融合模块通过由局部到全局的特征融合策略, 能够更精确地调整和优化各个模态特征的贡献, 使得

视觉和惯性特征的互补性得到了充分发挥. 这样在低质量数据的情况下, 自动识别和关注这两种数据模态中最相关的信息, 实现各模态的互补优势, 并提升整体的预测性能.



(a) MH_03 (b) MH_05

图 11 各模型在序列 MH_03 和序列 MH_05 的轨迹图

3.5 消融实验

为了确认框架中各模块的重要性, 我们使用 KITTI 数据集进行了消融实验, 实验结果如表 5 所示. 首先, 我们将去掉基于 DCT 的通道注意力层, 使用按通道拼接的方式进行直接融合作为基线模型. 基线模型没有使用任何网络进行特征融合, 因此其位移和旋转误差相对较大. 可以推测, 这是因为直接拼接的方式无法充分捕捉不同模态之间的信息关联, 导致信息融合效果较弱. 在加入基于 DCT 的通道注意力后, 位移和旋转误差都有所下降. 这表明通道注意力能够更好地捕捉不同模态中特定通道的重要性, 强调重要特征, 抑制不重要的特征. 特征融合层由改进的分散注意力层和全局 MLP-Mixer 层两部分组成. 为了进一步分析这两个子模块的重要性, 我们进行消融实验时, 分别去掉其中一个部分, 保留另一个部分进行测试. 可以看出, 即使只有局部交互与全局融合的一部分, 也比通道拼接的方式效果好, 但与完整的模型相比, 性能仍有一定差距. 这表明每个子模块在框架中的作用是相互补充、不可或缺的.

表 5 消融实验对比

VIO	通道注意力层	改进的分散注意力层	全局MLP-Mixer层	序列09		序列10		平均	
				t_{rel} (%)	r_{rel} (°/100 m)	t_{rel} (%)	r_{rel} (°/100 m)	t_{rel} (%)	r_{rel} (°/100 m)
√	—	—	—	14.17	4.73	15.54	3.68	14.86	4.21
√	√	—	—	13.56	3.86	14.50	3.23	14.03	3.55
√	√	√	—	9.99	2.24	10.70	**1.97**	10.35	2.11
√	√	—	√	10.72	2.04	**8.52**	2.58	9.62	2.31
√	√	√	√	**7.84**	**1.96**	8.85	2.20	**8.35**	**2.08**

4 结论与展望

本文针对视觉惯性里程计领域中如何有效融合多模态数据以提升位姿估计准确性的问题, 提出了一种基于注意力机制与局部交互的视觉惯性里程计模型, 通过实验表明, 该方法在各种数据低质量的场景下展现出了较好的性能和鲁棒性, 与其他深度学习模型和传统方法的对比也验证了该方法的有效性. 未来的工作可以进一步优化融合策略, 包括探索无监督的方式、结合传统方法, 开发出更高效的 VIO 系统. 此外, 为了进一步验证本文融合方式的通用性, 可以考虑将其扩展到更多领域, 从而提升其他复杂多模态任务的性能和效率.

参考文献

1 陈妍妍, 田大新, 林椿眄, 等. 端到端自动驾驶系统研究综述. 中国图象图形学报, 2024, 29(11): 3216–3237.

2 孙弋, 张雪丽. 基于改进 MSCKF 算法的室内机器人定位方法. 计算机系统应用, 2020, 29(2): 238–243. [doi: 10.15888/j.cnki.csa.007263]

3 Dong RF, Le XX, Quan MX, *et al*. Robust initialization and high-accuracy state estimation for filtering-based visual-inertial system. IEEE Transactions on Instrumentation and Measurement, 2023, 72: 7505612. [doi: 10.1109/TIM.2023.3298420]

4 von Stumberg L, Cremers D. DM-VIO: Delayed marginalization visual-inertial odometry. IEEE Robotics and Automation Letters, 2022, 7(2): 1408–1415. [doi: 10.1109/

LRA.2021.3140129]
5 Wang X, Pan YQ, Yan ZK, *et al*. Visual-inertial odometry based on kinematic constraints in IMU frames. IEEE Robotics and Automation Letters, 2022, 7(3): 6550–6557. [doi: 10.1109/LRA.2022.3173040]
6 Qin T, Li PL, Shen SJ. VINS-Mono: A robust and versatile monocular visual-inertial state estimator. IEEE Transactions on Robotics, 2018, 34(4): 1004–1020. [doi: 10.1109/TRO.2018.2853729]
7 Campos C, Elvira R, Rodriguez JJG, *et al*. ORB-SLAM3: An accurate open-source library for visual, visual-inertial, and multimap slam. IEEE Transactions on Robotics, 2021, 37(6): 1874–1890. [doi: 10.1109/TRO.2021.3075644]
8 王文森, 黄凤荣, 王旭, 等. 基于深度学习的视觉惯性里程计技术综述. 计算机科学与探索, 2023, 17(3): 549–560. [doi: 10.3778/j.issn.1673-9418.2209014]
9 Wang S, Clark R, Wen HK, *et al*. DeepVO: Towards end-to-end visual odometry with deep recurrent convolutional neural networks. Proceedings of the 2017 IEEE International Conference on Robotics and Automation (ICRA). Singapore: IEEE, 2017. 2043–2050. [doi: 10.1109/ICRA.2017.7989236]
10 张震宇, 杨小冈, 卢瑞涛, 等. VFE-VO: 视觉特征增强的光流法视觉里程计算法. 激光与光电子学进展, 2025, 62(6): 0612006.
11 Chen DP, Wang N, Xu RS, *et al*. RNIN-VIO: Robust neural inertial navigation aided visual-inertial odometry in challenging scenes. Proceedings of the 2021 IEEE International Symposium on Mixed and Augmented Reality (ISMAR). Bari: IEEE, 2021. 275–283. [doi: 10.1109/ISMAR52148.2021.00043]
12 Clark R, Wang S, Wen HK, *et al*. VINet: Visual-Inertial odometry as a sequence-to-sequence learning problem. Proceedings of the 31st AAAI Conference on Artificial Intelligence. San Francisco: AAAI, 2017. 3995–4001. [doi: 10.1609/aaai.v31i1.11215]
13 Chen CH, Rosa S, Miao YS, *et al*. Selective sensor fusion for neural visual-inertial odometry. Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach: IEEE, 2019. 10534–10543. [doi: 10.1109/CVPR.2019.01079]
14 Qu H, Zhang LL, Hu XP, *et al*. SelfOdom: Self-supervised egomotion and depth learning via bi-directional coarse-to-fine scale recovery. IEEE Transactions on Intelligent Vehicles, 2024, 9(5): 4962–4978. [doi: 10.1109/TIV.2023.3342803]
15 Shinde K, Lee J, Humt M, *et al*. Learning multiplicative interactions with Bayesian neural networks for visual-inertial odometry. arXiv:2007.07630, 2020.
16 Tu ZM, Chen CH, Pan XF, *et al*. EMA-VIO: Deep visual-inertial odometry with external memory attention. IEEE Sensors Journal, 2022, 22(21): 20877–20885. [doi: 10.1109/JSEN.2022.3208200]
17 刘海涛, 戴娟, 朱胜涛, 等. 基于深度学习的移动机器人视觉里程计. https://link.cnki.net/urlid/21.1476.TP.20231124.1204.002. (2023-11-27).
18 Chen CH, Lu XX, Markham A, *et al*. IONet: Learning to cure the curse of drift in inertial odometry. Proceedings of the 32nd AAAI Conference on Artificial Intelligence. New Orleans: AAAI, 2018. 6468–6476. [doi: 10.1609/aaai.v32i1.12102]
19 Jiang MW, Zeng PY, Wang K, *et al*. FECAM: Frequency enhanced channel attention mechanism for time series forecasting. Advanced Engineering Informatics, 2023, 58: 102158. [doi: 10.1016/j.aei.2023.102158]
20 Zhang H, Wu CR, Zhang ZY, *et al*. ResNeSt: Split-attention networks. Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. New Orleans: IEEE, 2022. 2735–2745. [doi: 10.1109/CVPRW56347.2022.00309]
21 Tolstikhin I, Houlsby N, Kolesnikov A, *et al*. MLP-Mixer: An all-MLP architecture for vision. Proceedings of the 35th International Conference on Neural Information Processing Systems. Curran Associates Inc., 2021. 1857.
22 Sun H, Wang HY, Liu JQ, *et al*. CubeMLP: An MLP-based model for multimodal sentiment analysis and depression estimation. Proceedings of the 30th ACM International Conference on Multimedia. Lisboa: ACM, 2022. 3722–3729. [doi: 10.1145/3503161.3548025]
23 Shi JH, Zhang Q, Tang YH, *et al*. Polyp-Mixer: An efficient context-aware MLP-based paradigm for polyp segmentation. IEEE Transactions on Circuits and Systems for Video Technology, 2023, 33(1): 30–42. [doi: 10.1109/TCSVT.2022.3197643]
24 Almalioglu Y, Turan M, Saputra MRU, *et al*. SelfVIO: Self-supervised deep monocular visual-inertial odometry and depth estimation. Neural Networks, 2022, 150: 119–136. [doi: 10.1016/j.neunet.2022.03.005]
25 Boulahia SY, Amamra A, Madi MR, *et al*. Early, intermediate and late fusion strategies for robust deep learning-based multimodal action recognition. Machine Vision and Applications, 2021, 32(6): 121. [doi: 10.1007/s00138-021-01249-8]
26 Shinde K, Lee J, Humt M, *et al*. Learning multiplicative interactions with bayesian neural networks for visual-inertial odometry. Proceedings of the 37th International Conference on Machine Learning (ICML). 2020.
27 Mur-Artal R, Montiel JMM, Tardos JD. ORB-SLAM: A versatile and accurate monocular SLAM system. IEEE Transactions on Robotics, 2015, 31(5): 1147–1163. [doi: 10.1109/TRO.2015.2463671]

(校对责编: 张重毅)