2022 IEEE Intelligent Vehicles Symposium (IV)
June 5-9, 2022. Aachen, Germany

Attention Guided Unsupervised learning of Monocular Visual-inertial Odometry

Zhenke Wang[1], Yuan Zhu[1], Ke Lu[1], Daniel Freer[2],Hao Wu[2], Hui Chen[1]

2022 IEEE Intelligent Vehicles Symposium (IV) | 978-1-6654-8821-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/IV51971.2022.9827359

Abstract— **Visual-inertial Odometry (VIO) provides cars with position information by fusing data from a camera and inertial measurement unit (IMU) which are both widely equipped on intelligent vehicles. Recently, unsupervised VIO has made great progress. However, existing VIOs mainly concatenate features extracted from different domains (visual and inertial), leading to inconsistency during integration. These methods are also difficult to scale to longer sequences because absolute velocity is not available. Hence, we propose a novel network based on attention mechanism to fuse sensors in a self-motivated and meaningful manner. We design spatial and temporal branches that focus on pairwise images and a sequence of images respectively. Meanwhile, a tiny but effective module (referred to as "warm start") is introduced to produce velocity-related information for the IMU encoder. The proposed attention branches and warm start are shown to improve the robustness of the model in dynamic scenarios and in the case of rapid changes in vehicle velocity. Evaluation on KITTI and Malaga datasets shows that our method outperforms other recent state-of-the-art VO/VIO methods.**

I. INTRODUCTION

Ego-motion estimation is crucial for many tasks in the field of autonomous vehicles. Recently, visual-inertial odometry (VIO) has attracted extensive attention from the research community. Compared with visual odometry (VO) which suffers from dynamic motion and illumination variation, VIO incorporates information from an inertial measurement unit (IMU) and demonstrates better performance. Moreover, both camera and IMU are low-cost and power-efficient sensors, making VIO a more feasible solution to vehicle positioning.

In the existing VIO work, fusion strategy and sequence consistency are two commonly discussed problems. In terms of fusion strategy, traditional methods divide it into loosely coupled [1][2] and tightly coupled [3]-[5] models, then solve the problem of position estimation with a filter [1]-[4] or optimization [5] algorithm. However, the visual component of these methods relies on handcrafted features, (e.g. ORB [1][3], SIFT [4]),which leads to unsatisfactory performance in texture less environment. Meanwhile, learning-based approaches to sensor fusion [6]-[8] integrate RGB images and inertial measurements using deep networks. Long short-term memory (LSTM) networks [6], assigning weights to different sensors [7] and a selective strategy [8] have been explored to fuse visual and inertial features. However, these works do not learn the correlation between two sensors in a self-motivated and meaningful manner. Thus, we integrate IMU data and images in the form of *query and memory*, through migrating a transformer-like framework [9] from target detection to the localization problem.

To address sequence consistency, research in monocular VO has tended to predict scale-inconsistent results over different snippets [10]. Zou et al. [11] adopts a ConvLSTM with input size of 97 frames which causes high computational cost. In the presented paper, we build spatial and temporal attention branches to process inputs from a shared encoder, which effectively and efficiently improves sequence consistency.

Current mainstream VIO frameworks primarily employ LSTM as an encoder for IMU information, but the initial state of the LSTM is another important problem in VIO which has not yet been fully investigated. Hence, we designed a *warm start* module to extract motion features from the previously predicted pose and then convert it into constraints for the IMU encoder. In this way, the IMU encoder can have better knowledge of the vehicle's current kinematic state.

As shown in Fig. 1, we propose a novel unsupervised ego-motion estimation network. The framework takes sequences of image and inertial measurements as input, integrating them with transformer-like branches. Our major contributions are:

- An attention guided visual-inertial features fusion strategy is proposed to predict ego-motion in an unsupervised end-to-end fashion.
- Spatial and temporal branches are built to adopt separate attention mechanisms in spatial and temporal domains.
- A warm-up module is designed to build connections between the previously predicted pose and the initial state of inertial measurements.

The remainder of this paper is structured as follows. Existing work regarding VO and VIO is discussed in Section II. Section III introduces our proposal in detail, including network structure, loss functions and mathematical background. Section IV presents experimental results on datasets: KITTI [12] and Malaga [13], followed by conclusions in Section V.

II. RELATED WORK

A. Traditional visual-inertial odometry

As mentioned in Sec. I, loosely coupled and tightly coupled methods are the two main streams among traditional VIO works. The loosely-coupled approach treats visual and inertial subsystems as standalone pose estimators. Fusion is applied at the last stage to refine predicted pose from both

[1] School of Automotive Studies, Tongji University, Shanghai 201804, China. (corresponding author to provide e-mail: hui-chen@tongji.edu.cn).
[2] Antobot Ltd., BIC011, Arise Centre, Chelmsford CM1 1SQ, UK

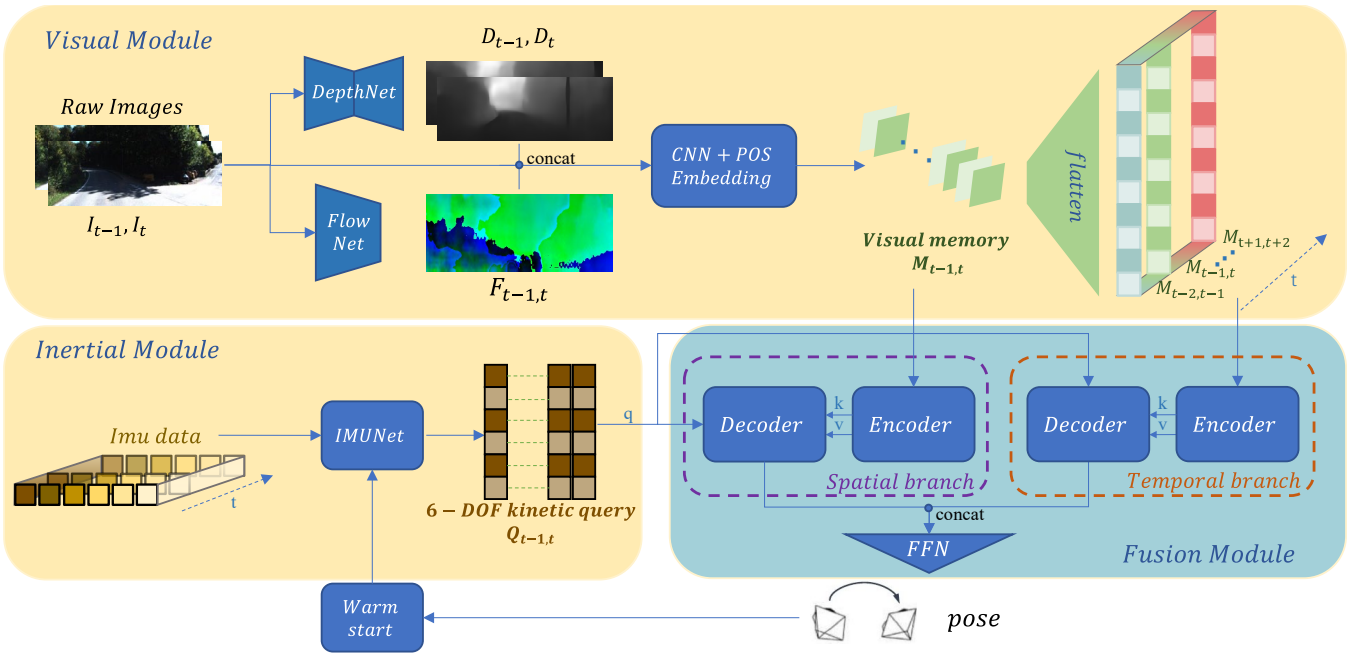


Fig. 1: **Proposed architecture overview.** Visual module takes image pairs (I_{t-1},I_t) as input. DepthNet and FlowNet predict depth map (D_{t-1},D_t) and 2D flow ($F_{t-1,t}$) separately. Raw images, depth map and 2D flow are then concatenated together and forwarded into a CNN based feature extractor. POS embedding records pixel coordinates which are required by transformer branches. The results of visual module are noted as *visual memory* (M). In parallel, inertial module extracts features from IMU data. Since these features describes 6-DOF motion states, they are noted as *kinetic query* (Q). There are two attention braches in fusion module, i.e. spatial branch and temporal branch. They take visual memory and kinetic query as input and their outputs are concatenated to predict poses by a followed feed forward network (FFN). Spatial branch estimates ego-motion from pairwise inputs (from t-1 to t) while Temporal branch focus on a sliding window of size n (n equals 5 in this work). A *warm start* module is proposed to build connections between previously predicted pose and current motion observation.

subsystems, where the conventional Kalman filter (KF) or derivative methods are commonly used. In [14], an indirect feedback KF is adopted to calculate relative position between consecutive frames. [2] uses Unscented Kalman filter (UKF) to avoid complex computation for Jacobian matrices. In contrast, tightly-coupled approaches fuse key features extracted from images with raw measurements of IMU at early stages and achieve better accuracy in general. An MSCKF-based monocular VIO system is presented in [4] where the 3D feature positions are not included in the filter, leading to relatively low computational complexity as well as inconsistency in performance. Compared with filter methods, optimization-based approaches achieve higher accuracy, but require more computational resources by carrying out iterative minimization of cost function, e.g., least square error. OKVIS [15] optimizes the position of landmarks and camera motion at the same time without loop closure constraint. VINS-Mono [5] fuses preintegrated IMU data and visual observations by adopting pose graph optimization to improve global consistency.

B. *Learning-based visual-inertial odometry*

To the best of our knowledge, VINet [6] is the first end-to-end VIO system which firstly extracts visual and inertial features and then learns to integrate data from the two sensors. The majority of later neural-network based VIO work follows a 3-part structure, containing a visual module, inertial module and fusion module.

Visual module. Large progress has been recently achieved in the development of single view depth estimation [10][16] and visual odometry. Learning-based VO systems can be further classified into supervised VO and unsupervised VO. Konda et al. [17] is one of the first supervised works. It treats visual odometry as a classification problem and predicts discrete changes of camera poses. DeepVO [18] combines a recurrent neural network (RNN) and convolutional neural network (CNN) to produce continuous motion estimation. A number of works [19][20] extend DeepVO and report impressive results. However, supervised approaches rely on labelled datasets and the ground truth is difficult and expensive to acquire. Thus, Zhou et al. [21] utilizes view synthesis loss as a supervisory signal so the framework can jointly learn depth and ego-motion from RGB frames. Their work has drawn growing interest in unsupervised multi-task learning of depth and motion estimation. Bian et al. [10] adopts geometric loss to produce scale-consistent prediction. GeoNet [22] generates 3D flow to have better knowledge of surroundings. Hu et al. [23] learns stable and repeatable pixel-level correspondences to improve performance when illumination changes. Compared with supervised VO, unsupervised VO shows more adaptability to different scenarios considering that labels are not always available. As a result, unsupervised VO is a more accessible solution to be tightly coupled with other sensors.

Inertial module. CNN and LSTM are common learning-based encoders for inertial measurements in VIO, but the form of IMU encoder largely depends on the fusion strategy. ATVIO [24] adopts a 1D CNN to extract the feature map from IMU data and then processes the concatenated visual-inertial feature maps with a pooling layer. Meanwhile, VINet [6] extracts IMU features via LSTM and then integrates the camera and IMU features with another LSTM module. In addition to VIO, LSTM is also widely used in inertial odometry (IO) for its ability to learn feature representation

from a sequence. IONet [25] adopts a 2-layer LSTM to track pedestrians with IMU data from smartphone. Aboldeepio [26] presents a triple-channel LSTM network that extracts features from accelerometer, gyroscope and time interval between IMU frames, and additionally improves the robustness to noise by adding noise to the training data. Because the LSTM has been used so widely by the research community, we have chosen to use it in our inertial encoder.

Fusion module. Data-driven methods of the fusion module integrate visual and inertial measurements in a tightly-coupled manner, and are more robust to calibration error than traditional methods. VINet [6] directly concatenates visual and inertial features together which are taken as input into a LSTM module to predict 6-DoF poses, and by including the history of system states via LSTM it shows robustness to bad time synchronization. Wei et al. [7] further extends VINet by learning a weight vector for concatenated visual-inertial features before feeding them into a RNN for pose estimation. Chen et al. [8] selectively uses soft and hard feature representations considering that not all features are useful. Aforementioned works perform data fusion by concatenating features together which are from different modalities, leading to sub-optimal results. Hence, we propose a self-motivated fusion framework based on attention mechanisms.

III. Proposed Method

Our framework is composed of three main blocks, i.e., visual module, inertial module and fusion module. The visual module extracts high-dimensional features (*visual memory,* M) from raw images, depth map and 2D flow. In parallel, the inertial module outputs *kinetic query* (Q) that describes 6-DOF motion states. Finally, two attention branches jointly estimate ego-motion from visual memory and kinetic query. Furthermore, a tiny *warm start* module is built to add connections between previously predicted poses and current motion states. The details of each module will be introduced in the following sections.

A. Input Encoding

Visual feature encoding. Since the whole model aims to estimate ego-motion, geometric and kinetic features should be learnt by the visual module. To tackle this problem, we concatenate adjacent frames (I_{t-1}, I_t), their depth map (D_{t-1}, D_t) and 2D optical flow ($F_{t-1,t}$) along the channel dimension as the input to the feature extractor. At this stage, depth map describes 3D surroundings and provides the model with scale information. The 2D optical flow predicts motion of pixels and creates geometric constraints. The raw image pair forwards original information in the form of short cuts. In addition, both depth map and optical flow are outputs from a neural network and no ground truth is used. In this work, we follow an unsupervised network of [10] for single-view depth estimation (DepthNet) while reducing the number of output channels for different resolutions. MaskFlownet [27] is adopted to estimate optical flow (FlowNet). We utilize ResNet-18 [28] as a feature extractor and trigonometric functions for positional encoding of pixels (POS embedding). The feature extractor and POS embedding can be expressed as function F_v:

$$M_{t-1,t} = F_v\left(\text{concat}\left(I_{t-1}, I_t, D_{t-1}, D_t, F_{t-1,t}\right)\right) \quad (1)$$

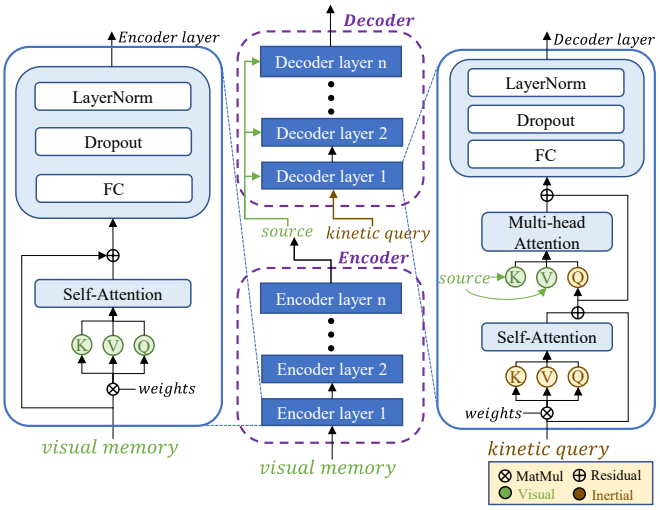


Fig. 2: Architecture of spatial branch

Inertial feature encoding. There are multiple groups (N) of inertial measurements between two adjacent image frames due to different sampling frequencies of sensors. Thus, the input size of the inertial module is 6×N where 6 denotes the linear acceleration (α) and angular velocity (β) in 3 dimensions. IMUNet consists of a two-layer LSTM and a fully connected layer. The hidden states (h) of the LSTM contain not only the values from last step but also the output of the warm start. We will discuss this further in Sec. III-D. IMUNet generates a 6-channel kinetic query, which conveys the trend of the motion in six degrees of freedom. This process can be noted as a function F_i:

$$Q_{t-1,t} = F_i\left((\alpha_{t-1,t}, \omega_{t-1,t}), h_t = h_{t-1} + warmstart\right) \quad (2)$$

B. Spatial Branch and Data Fusion

In the mainstream of VIO work, visual and inertial features are extracted from different modalities and domains, but are concatenated directly, leading to sub-optimal fusion results. Inspired by attention mechanism and its application in object detection DETR [9], we designed a spatial transformer branch to integrate two sensors in the form of query and memory. DETR learns 100 query slots to focus on different areas and box sizes of detected objects. As a result, up to 100 objects can be detected per frame. Since inertial measurements and the pose transformation have the same dimensions and higher-dimensional features are extracted from images, we use the encoded IMU data as the kinetic query. No pretrained model is used to obtain visual memory since odometry and object detection are different tasks. POS embedding injects pixel position encoding into visual memory so that spatial information is still available after the feature map is decomposed into vectors required by the attention mechanism.

Fig. 2 illustrates the structure of the spatial transformer branch, which helps to predict ego-motion between two adjacent images. First, visual memory is sent to the encoder to produce self-attention. Encoder layers share the same structure and are connected in series. The first layer multiplies the learned weights with the visual memory matrix to obtain key (K), value (V) and query (Q) respectively. Then, the self-attention is determined as:

$$SelfAtt = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V \quad (3)$$

where d_k denotes the dimension of key. After a fully connected layer (FC), drop out and layer normalization, the first encoder layer forwards self-attention to layer 2. The output of the final encoder layer is denoted as the *source*.

The decoder takes the visual source from the encoder and kinetic query as input. Each decoder layer is similar to the encoder layers, but additionally contain a multi-head attention module to integrate visual and inertial features. We can get multi-head attention in two steps. First, each head calculates self-attention as (3) where K and V come from visual source and Q is from kinetic query. In other words, the network focuses on the part of the visual feature that is related to 6-DOF motion with the help of query from inertial measurements. The second step is to concatenate self-attention of all heads together and multiply them by weight W^O to get multi-head attention that enhances the expressiveness of the model:

$$MultiAtt = \text{Concat}(SelfAtt_1, ..., SelfAtt_\tau)W^O \quad (4)$$

C. Temporal Branch and Sequence Consistency

The temporal branch is designed to maintain scale consistency over long sequences. Its structure is the same as the spatial branch, though the inputs are different. As shown in Fig. 3, the temporal branch looks at image frames within a sliding window rather than an image pair. Window size is set to 5 in this work. Take the relative pose estimation from t-1 to t as an example: we first flatten and concatenate visual memory between t-2 to t+2 as new visual inputs for the encoder of the temporal branch. With POS embedding, the spatial structure of feature map is reserved during this transformation. Meanwhile, the kinetic query from IMUNet at timestamp t is used as input to the decoder, in a similar way to the spatial branch. Then, a feed forward network (FFN) receives the concatenated results of two branches and predicts ego-motion. FFN is composed of two 2D convolutional layers (kernel size: 3) and an AvgPooling layer. We chose tanh as activation function to get both positive and negative translation and rotation.

The temporal branch matches kinetic query with the visual source that contains richer information. Landmarks existing in multiple frames are able to describe velocity and acceleration. Kinetic query at time step t indicates that the model should output frame-to-frame motion (t-1 to t). At last, the FFN learns to fuse local (spatial branch) and sub-global (temporal branch) estimation and refines the results.

D. Warm Start for IMUNet

Given only linear acceleration and no velocity, the inertial module tends to suffer from scale ambiguity. Hence, we extract velocity information from predicted poses and add them to the hidden states of IMUNet. Handcrafted and CNN-based methods are applied as:

$$handcraft = \frac{1}{2\Delta t}\left(pose_{t-2,t-1} + pose_{t-1,t}\right) \quad (5)$$

$$warmstart = handcraft + \text{CNN}\left(pose_{t-2,t-1}, pose_{t-1,t}\right) \quad (6)$$

where Δt denotes fixed time interval between two images

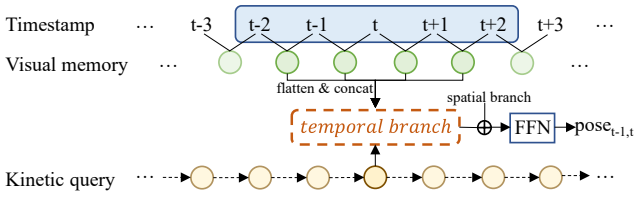


Fig. 3: Inputs and outputs of temporal branch

and CNN is a 1D conv layer. *Handcraft* directly computes velocity at t-1 while CNN maps relative poses to velocity features for IMUNet. As for initial steps where poses are not available, our model firstly estimates motion without a warm start and then repeats again with a warm start as in (2).

E. Loss Functions

Image reconstruction serves as the main supervisory signal in this work. Given camera intrinsic matrix (S), depth map and relative translation ($T_{t-1,t}$) between two images, we are able to synthesize image I'_t from I_{t-1} by

$$I'_t(p) = ST_{t-1,t}D_{t-1}S^{-1}I_{t-1}(p) \quad (7)$$

where p indexes over valid pixel coordinates that can be viewed at both t and t-1. In the same way, we can generate multiple I'_t with all other images within a time window and match each synthesized image with the original one. Image reconstruction loss is formulated as:

$$L_{IR} = \sum_{I'_t}\sum_{p}\left|I_t(p) - I'_t(p)\right| \quad (8)$$

Another loss function (L_D) comes from depth map. First, we project I_{t-1} into the 3D space, and then obtain the 3D points in the camera coordinate system at t with estimated relative pose. Then we can get the depth map in two ways. On the one hand, we calculate depth map D_t^c by projecting these 3D points to the camera plane at time t. On the other hand, the depth map can also be synthesized similar to (7), i.e., D_t^s . L_D is then defined as (9). In addition, pixels of different images but sampled from the same landmark are supposed to have same 3D coordinates $C(p)$ after being projected into space. We can therefore define 3D coordinates loss as (10).

$$L_D = \sum_{\left[D_t^c, D_t^s\right]}\sum_{p}\left|D_t^c(p) - D_t^s(p)\right| \quad (9)$$

$$L_C = \sum_{C'_t}\sum_{p}\left|C_t(p) - C'_t(p)\right| \quad (10)$$

The overall loss function can be summarized as

$$L_{all} = \lambda_1 L_{IR} + \lambda_2 L_{DR} + \lambda_3 L_C \quad (11)$$

IV. Experiments and Evaluation

A. Datasets

KITTI. The KITTI odometry split [12] contains 11 driving sequences with ground truth acquired by GPS. As in [7] and [22], we use sequences 00-08 as training set except for 03 where IMU data are not available and 09-10 as the test set. Images are recorded at 10 Hz and we scale them to 128×416 for the sake of computational cost. 100 Hz IMU data are collected from raw datasets. The size of inertial inputs between two consecutive frames is 10×6. We use linear interpolation to deal with missing IMU frames in original data.

TABLE I. COMPARISION OF ODOMETRY PERFORMANCE ON KITTI

Method	Type*	Metrics	Seq.09	Seq.10	Avg
ORB-SLAM[29]	VO(T)	t_{rel}(%)	45.52	6.39	25.96
		r_{rel}(°)	3.10	3.20	3.15
VINS[5]	VIO(T)	t_{rel}(%)	23.90	16.50	20.20
		r_{rel}(°)	2.47	2.34	2.41
SfMLearner[21]	VO(L)	t_{rel}(%)	21.63	20.54	21.09
		r_{rel}(°)	3.57	10.93	7.25
Bian[10]	VO(L)	t_{rel}(%)	12.43	11.86	12.15
		r_{rel}(°)	4.65	4.95	4.80
Wei et al.[7]	VIO(L)	t_{rel}(%)	4.13	5.51	4.82
		r_{rel}(°)	**0.89**	**0.53**	**0.71**
ATVIO†[24]	VIO(L)	t_{rel}(%)	3.99	5.61	4.80
		r_{rel}(°)	1.63	2.96	2.30
ours†	VIO(L)	t_{rel}(%)	**3.87**	**4.38**	**4.23**
		r_{rel}(°)	1.12	1.46	1.29

- †: 128×416 images are fed into our network and ATVIO. All other methods use higher resolution (256×832).
- * Type: T means traditional method and L denotes learning-based method.
- The best and second-best results of t_{rel} and r_{rel} are highlighted in **bold** and underline respectively.

Malaga. Malaga outdoor dataset [13] is employed to evaluate the robustness of the proposed framework under different scenarios. We collect 10 Hz images and 100 Hz inertial measurements from raw data. Sequences 01, 02, 04, 05, 06, 07 are used for training and 03 is applied for test.

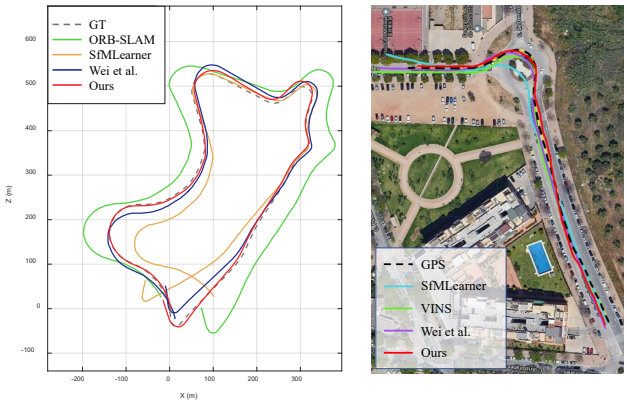
B. Training Details

We use ADAM optimizer to train the proposed network on both KITTI and Malaga dataset, with alpha=0.9, beta=0.999, and learning rate=1e-4. An exponential learning rate strategy is applied during the first 3,000 iterations. The network is trained using single-point precision on a NVIDIA GeForce 2080 Super GPU. The training process converges after about 35,000 iterations with a batch size of 4. The weights of the overall loss function (11) are set to λ_1=0.6 , λ_2=0.2 , λ_3=0.2 . Encoders and decoders of both temporal and spatial branches have 6 layers. Multi-head attention module has 6 headers.

C. Motion Estimation

Quantitative evaluation was carried out on the KITTI dataset and results are summarized in Table I. We chose translation error t_{rel} (%/100m) and rotation error r_{rel} (°/100m) as metrics. They compute average translation and rotation RMSE drift on lengths of 100m-800m. The methods involved in the evaluation for comparison are traditional VO [29], traditional VIO [5], learning-based VO [10][21] and learning-based VIO [7][24]. All works take single-view images as input and the predicted trajectories are aligned with ground truth in 7-DOF, i.e., 6-DOF motion and scale. The learning-based methods are all unsupervised and have the same training/test set partitioning as our work.

The results show that our framework outperforms other models in translation error and is the second-best performer in rotation error, although our approach uses lower resolution images. As with other VIO work, benefiting from extra IMU data as input, the proposed model achieves better localization performance than VO. In particular, compared with Bian [10] where the same depth map network is employed, which also means that we have the same supervisory signal, our network shows substantially better performance, indicating the effectiveness of the fusion strategy. Compared with other learning-based VIOs, our model can avoid cumulative errors more effectively due to the introduction of temporal and spatial branches, and thus has better accuracy on translation. In traffic scenarios, translation is directly related to vehicle position and is therefore a more important metric. The trajectory estimation of different methods is illustrated in Fig. 4 for qualitative evaluation. The two test sequences are selected from the KITTI (Seq.09) and Malaga datasets (Seq.03) respectively. Notably, our depth module is pre-trained on the KITTI dataset and not further optimized on Malaga. Nevertheless, the proposed framework can still predict more accurate positions than other traditional and learning-based methods on both datasets, demonstrating the adaptability of our model.



(a) Seq.09 (b) Seq.03

Fig. 4: Trajectory estimation of different methods on KITTI (a) and Malaga (b) test set.

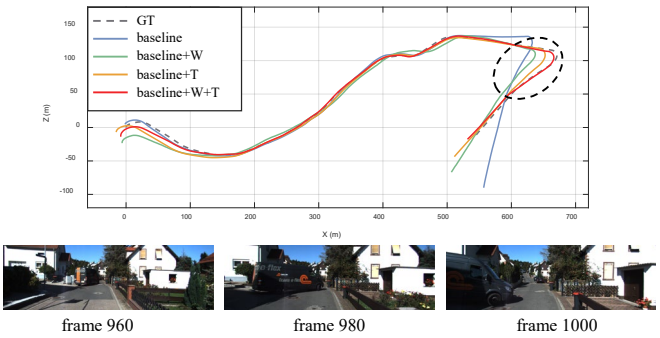


Fig. 5: Alblation study on KITTI sequence 10. At around frame 980 (circled area in the figure), a truck is reversing and ego car slows down rapidly, making it challenging for ego-motion estimation. *W* and *T* are abbreviations of warm start and temporal branch.

TABLE II. ALBLATION STUDY ON DIFFERENT PROPOSED MODULES

Method	Seq.09		Seq.10	
	t_{rel}(%)	r_{rel}(°)	t_{rel}(%)	r_{rel}(°)
baseline	8.42	2.21	11.73	4.53
baseline+warm start	7.45	1.56	9.92	3.08
baseline+T-branch	7.17	1.63	7.68	2.97
baseline+T-branch+warm start	**3.87**	**1.12**	**4.38**	**1.46**

- T-branch: temporal branch.

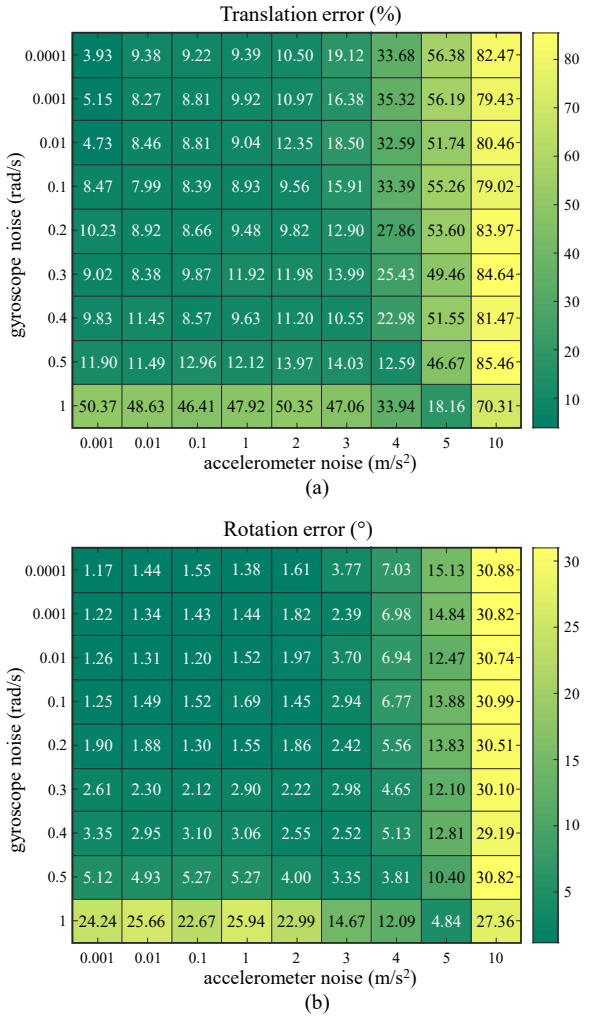


Fig 6. Robustness test on white noise added to accelerometer and gyroscope. Both translation (a) and rotation (b) error of our network are given.

TABLE III. ROBUSTNESS TEST ON BLURRED IMAGES AND TEMPORAL UNSYNCHRONIZATION

Method	Blur		Unsyn 30ms		Unsyn 60ms	
	t_{rel}(%)	r_{rel}(°)	t_{rel}(%)	r_{rel}(°)	t_{rel}(%)	r_{rel}(°)
VINS[5]	24.27	2.86	26.34	5.49	32.02	17.88
Ours	4.63	1.32	4.37	1.29	5.12	1.48

D. Ablation Study

An ablation study was conducted on the KITTI dataset to demonstrate the importance of each proposed component. The baseline takes two adjacent images and corresponding IMU data as input, and only the spatial branch is utilized in the fusion module. As shown in Table II, both translational and rotational performance of the network is improved after adding the temporal branch and warm start module. Fig. 5 illustrates the effectiveness of the two extra modules. Sequence 10 of the KITTI dataset was selected for evaluation since there is a challenging scenario near frame 980 where a truck is reversing in the field of the view. As a result, the ego car slows down rapidly. It can be seen in Fig. 5 that the baseline predicts a biased trajectory under this situation. The reason for improved performance after adding a warm start is that it provides velocity-related information to IMUNet, making our model more sensitive to acceleration and deceleration. The temporal branch contributes to lower translation and rotation error by introducing historical information from the RGB images. The visual part of the framework is able to extract features from the input sequence, similar to the inertial part, so that the model is no longer limited to estimating the pose between two frames. Meanwhile, the network learns to pay more attention to stable and static background landmarks and less attention to landmarks on dynamic targets. In general, the proposed temporal branch and warm start improve the robustness of our model in dynamic scenarios and in the case of rapid changes in vehicle velocity.

E. Robustness test

Due to human or environmental factors (e.g. vehicle vibration), the camera and IMU may generate noise, and the two sensors may also be out of time synchronization. Hence, we used polluted data (i.e. IMU noise, blurred image and unsynchronized image-IMU stream) to test the robustness of proposed method. Notably, our model was trained on a normal dataset and not finetuned for specific disruptions. We added white noise to the raw accelerometer and gyroscope data with magnitudes ranging from 0.001 to 10 m/s^2 and from 0.0001 to 1 rad/s respectively. Results are reported in Fig. 6. Our VIO still achieves competitive performance when accelerometer noise is up to 3 m/s^2 and gyroscope noise is up to 0.5 rad/s. Table III shows translation and rotation errors in the case of bad time synchronization and blurred images (Gaussian blur is adopted). Compared to VINS, the performance of our method is only slightly degraded due to the polluted data, indicating better robustness.

V. CONCLUSION AND FUTURE WORK

In this paper, we propose a brand-new method for end-to-end visual-inertial odometry. RGB images and IMU data are treated as word vectors and integrated based on an attention mechanism. We design spatial and temporal attention branches that focus on local and sub-global motion respectively. The temporal branch processes 5 consecutive images at the same time and improves the performance significantly. In addition, we introduce a warm start module for IMU embedding. Our method is shown to provide additional velocity-related information and improve the robustness of the framework to polluted data. We use different datasets collected in urban environments (KITTI and Malaga), and the superiority of our model is demonstrated in both qualitative and quantitative evaluation. For future work, we plan to explore longer image sequences as input with limited additional computational resource consumption.

REFERENCES

[1] Lin, H., and F. Defay, "Loosely coupled stereo inertial odometry on low-cost system." In *Proc. Int. Micro Air Vehicle Conf. Flight Competition (IMAV)*,2017, pp. 143-148.

[2] S. Shen, Y. Mulgaonkar, N. Michael and V. Kumar, "Multi-sensor fusion for robust autonomous flight in indoor and outdoor environments with a rotorcraft MAV," *2014 IEEE International Conference on Robotics and Automation (ICRA)*, 2014, pp. 4974-4981.

[3] F. Pang, Z. Chen, L. Pu and T. Wang, "Depth enhanced visual-inertial odometry based on Multi-State Constraint Kalman Filter," *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2017, pp. 1761-1767.

[4] A. I. Mourikis and S. I. Roumeliotis, "A Multi-State Constraint Kalman Filter for Vision-aided Inertial Navigation," *Proceedings 2007 IEEE International Conference on Robotics and Automation*, 2007, pp. 3565-3572.

[5] T. Qin, P. Li and S. Shen, "VINS-Mono: A Robust and Versatile Monocular Visual-Inertial State Estimator," in *IEEE Transactions on Robotics*, vol. 34, no. 4, pp. 1004-1020, Aug. 2018.

[6] R. Clark, S. Wang, H. Wen, A. Markham, and N. Trigoni, "VINet: Visual-Inertial Odometry as a Sequence-to-Sequence Learning Problem", *AAAI*, vol. 31, no. 1, Feb. 2017.

[7] P. Wei, G. Hua, W. Huang, F. Meng and H. Liu, "Unsupervised Monocular Visual-inertial Odometry Network." In *IJCAI*, 2020, pp. 2347-2354.

[8] C. Chen, S. Rosa, Y. Miao, C. Lu. X. Wu, A. Markham, and N. Trigoni, "Selective sensor fusion for neural visual-inertial odometry." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 10542-10551.

[9] N.Carion, F.Massa, G.Synnaeve, N.Usunier, A.Kirillov, S. Zagoruyko, "End-to-end object detection with transformers." In *European Conference on Computer Vision (ECCV)*, 2020, pp. 213-229.

[10] J.W. Bian, H. Zhan, N. Wang , Z. Li, L. Zhang, C. Shen, M. Cheng and I. Reid, "Unsupervised Scale-consistent Depth Learning from Video." *International Journal of Computer Vision*, 2021, pp. 1-17.

[11] Y. Zou, P. Ji, Q.H. Tran, J.B. Huang, and M. Chandraker, "Learning monocular visual odometry via self-supervised long-term modeling." In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings*, 2020, pp. 710-727.

[12] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset". *The International Journal of Robotics Research*, vol. 32, no. 11, pp.1231-1237, 2013

[13] J.L. Blanco-Claraco, F.A. Moreno-Duenas, and J. González-Jiménez, "The Málaga urban dataset: High-rate stereo and LiDAR in a realistic urban scenario. "*The International Journal of Robotics Research*, vol. 33, no. 2, pp.207-214, 2014.

[14] S. Sirtkaya, B. Seymen and A. A. Alatan, "Loosely coupled Kalman filtering for fusion of Visual Odometry and inertial navigation," *Proceedings of the 16th International Conference on Information Fusion*, 2013, pp. 219-226.

[15] S. Leutenegger, L. Simon, B. Michael, S. Roland, and P. Furgale, "Keyframe-based visual–inertial odometry using nonlinear optimization." *The International Journal of Robotics Research*, vol. 34, no. 3, pp. 314-334, 2015.

[16] Y. Chen, C. Schmid. and C. Sminchisescu, "Self-supervised learning with geometric constraints in monocular video: Connecting flow, depth, and camera." In *Proceedings of the IEEE/CVF International Conference on Computer Vision (CVPR)*, 2019, pp. 7063-7072.

[17] K. R. Konda, R. Memisevic, "Learning visual odometry with a convolutional network." In *VISAPP (1)*, 2015, pp. 486-490. 2015.

[18] S. Wang, R. Clark, H. Wen and N. Trigoni, "DeepVO: Towards end-to-end visual odometry with deep Recurrent Convolutional Neural Networks," *2017 IEEE International Conference on Robotics and Automation (ICRA)*, 2017, pp. 2043-2050.

[19] M.R.U. Saputra, P.P. De Gusmao, Y. Almalioglu, A. Markham and N. Trigoni, "Distilling knowledge from a deep pose regressor network." In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 263-272.

[20] F. Xue, X. Wang, S. Li, Q. Wang, J. Wang and H. Zha, "Beyond tracking: Selecting memory and refining poses for deep visual odometry." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8575-8583.

[21] T. Zhou, M. Brown, N. Snavely and D.G. Lowe, "Unsupervised learning of depth and ego-motion from video." In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2017, pp. 1851-1858.

[22] Z. Yin, and J. Shi, "Geonet: Unsupervised learning of dense depth, optical flow and camera pose." In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2018, pp. 1983-1992.

[23] H. Hu, L. Sackewitz and M. Lauer, "Joint Learning of Feature Detector and Descriptor for Visual SLAM," *2021 IEEE Intelligent Vehicles Symposium (IV)*, 2021, pp. 928-933.

[24] L. Liu, G. Li and T. H. Li, "ATVIO: Attention Guided Visual-Inertial Odometry," *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2021, pp. 4125-4129.

[25] Chen, C., Lu, X., Markham, A. and Trigoni, N. "Ionet: Learning to cure the curse of drift in inertial odometry." *In Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1. 2018.

[26] M. Abolfazli Esfahani, H. Wang, K. Wu and S. Yuan, "AbolDeepIO: A Novel Deep Inertial Odometry Network for Autonomous Vehicles," in *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 5, pp. 1941-1950, May. 2020.

[27] S. Zhao, Y. Sheng, Y. Dong, E.I. Chang and Y. Xu, "Maskflownet: Asymmetric feature matching with learnable occlusion mask." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 6278-6287.

[28] K. He, X. Zhang, S. Ren and J. Sun, "Deep residual learning for image recognition." In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2016, pp. 770-778.

[29] R. Mur-Artal, J. M. M. Montiel and J. D. Tardós, "ORB-SLAM: A Versatile and Accurate Monocular SLAM System," in *IEEE Transactions on Robotics*, vol. 31, no. 5, pp. 1147-1163, Oct. 2015.