ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-7605-5/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICASSP39728.2021.9413912

ATVIO: ATTENTION GUIDED VISUAL-INERTIAL ODOMETRY

Li Liu[1], *Ge Li*[1], *Thomas H Li*[*2]

[1]School of Electronic and Computer Engineering, Peking University Shenzhen Graduate School
[2]Advanced Institute of Information Technology, Peking University

ABSTRACT

Visual-inertial odometry (VIO) aims to predict trajectory by ego-motion estimation. In recent years, end-to-end VIO has made great progress. However, how to handle visual and inertial measurements and make full use of the complementarity of cameras and inertial sensors remains a challenge. In the paper, we propose a novel attention guided deep framework for visual-inertial odometry (ATVIO) to improve the performance of VIO. Specifically, we extraordinarily concentrate on the effective utilization of the Inertial Measurement Unit (IMU) information. Therefore, we carefully design a one-dimension inertial feature encoder for IMU data processing. The network can extract inertial features quickly and effectively. Meanwhile, we should prevent the inconsistency problem when fusing inertial and visual features. Hence, we explore a novel cross-domain channel attention block to combine the extracted features in a more adaptive manner. Extensive experiments demonstrate that our method achieves competitive performance against state-of-the-art VIO methods.

Index Terms— Visual-Inertial Odometry, Attention, Feature Fusion, IMU

1. INTRODUCTION

Visual-inertial odometry (VIO) gains significant popularity in robotics, aiming to estimate ego-motion with camera sensors and Inertial Measurement Unit (IMU) sensors. For the reason that camera sensors and IMU sensors are rather low-priced, power-efficient in mobile agents, VIO enjoys extensive applications in the field of robotics, autonomous driving, and AR/VR. Moreover, visual and inertial sensors are complementary to each other. Compared to odometry by a single sensor such as visual odometry (VO) [1, 2] and inertial odometry (IO) [3, 4], VIO demonstrates a better performance on the accuracy and robustness in heterogeneous scenes.

There develop traditional methods and learning-based methods to solve the VIO problem. Over the past decade, many traditional VIO methods have been proposed. They can be divided into loosely coupled methods [5] and tightly coupled methods [6, 7, 8] by the sensor fusion type.

The above-mentioned conventional methods resort to hand-crafted design that heavily relies on human experience and domain knowledge. Therefore, there are increasing learning-based and task-specific VIO frameworks. In the end-to-end VIO frameworks, the core issues are the visual feature extraction, IMU feature extraction, and the fusion of visual features and IMU features. For the first issue, there already exist many VO methods [9, 10, 2]. With the powerful feature extraction capability of Deep Convolutional Neural Networks (CNNs) [11, 12, 13], pure visual odometry has achieved acceptable results. Thus the problem of visual feature extraction is easier to solve. We utilize the current advanced encoder based on the neural architecture search technology for visual feature extraction.

However, for the second issue, the inertial encoders in VIO frameworks are rarely investigated. The main reason is that IMU data has less information compared to images, even though it is sampled at a high frequency. At each sampling timestamp, the information is only contained in six numbers. Besides, due to the strong temporal dependency of IMU data, most existing VIO frameworks [14, 15, 16] adopt LSTM to process IMU information. However, the LSTM-based inertial encoders are low-efficient and unsatisfactory. Therefore, in our paper, we design a more effective inertial encoder to extract inertial features.

The last thorny issue is the integration of visual features and inertial features. Pure VO has been able to achieve favorable performance. After adding the inertial features, if the combination is irrational, the IMU data may be regarded as noise and cannot help to improve the performance of the trajectory estimation. In fact, an irrational fusion of inertial and visual features will lead to worse results. Clark et al. [14] directly concatenate the two kinds of representations. Chen et al. [15] explore two fusion strategies to integrate the observations from cameras and IMU sensors. However, all the mentioned methods do not build an effective correlation between IMU data and visual information. It is because there exists a gap between the two distributions, and some crucial information cannot be sufficiently exploited. To address the problem, we present a novel attention guided visual-inertial feature fusion framework to enhance the performance of the VIO system.

In conclusion, we propose a novel attention guided deep framework for visual-inertial odometry (ATVIO), and our main contributions are as follows:

- We propose a fast and efficient inertial encoder to extract features from raw IMU data.
- We explore a new attention module to solve the inconsistency of data distributions in the fusion stage.
- Extensive experiments are conducted to compare with state-of-the-art VIO methods and show our method has competitive performance.

2. METHOD

The framework of ATVIO is presented in Fig. 1. It contains three stages, i.e. a feature extraction stage, a feature fusion stage, and a pose regression stage. In the feature extraction stage, the inputs are sequential stacked images and IMU measurements. We use two encoders to extract vision features and inertial features respectively. In the feature fusion stage, two kinds of extracted features are inte-

*Corresponding author: tli@aiit.org

This project was supported by Hangzhou Science and Technology Development Program (No.20182014B09) and Key-Area Research and Development Program of Guangdong Province (No.2019B121204008).

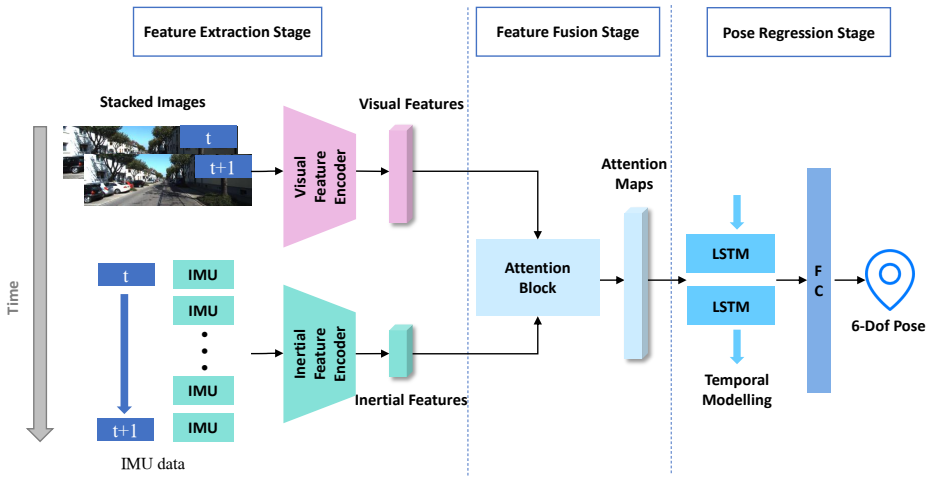


Fig. 1: Architecture of the proposed ATVIO. It includes (1) the feature extraction stage for visual and inertial features extraction, (2) the feature fusion stage to get attention maps, (3) the pose regression stage to predict 6-DoF poses.

grated by an attention block to output attention maps. Finally, there are LSTM layers to model temporal dependence, then the 6-DoF (six degrees of freedom, including the translation and the rotation) pose is output by a fully connected layer.

2.1. Visual Feature Encoder

Deep Convolutional Neural Networks (CNNs) have exhibited strong feature extraction capability in computer vision. In particular, VGG [11] and ResNet [12] are the most widely used structures. However, in the VIO task, geometric information is critical. It cannot be easily captured by conventional CNNs. In order to learn geometric features, we utilize the first two modules (the stem module used to process the input data and multiple blocks used to extract features) of VONAS-A architecture in [2] as our visual feature encoder, which is designed by neural architecture search technology. It is appropriate for VIO due to the circumstantial architecture search for a specific task. It also achieves a more preferable performance than FlowNet [13] in other VIO approaches.

2.2. Inertial Feature Encoder

Most of the existing learning-based VIO methods adopt LSTM to process IMU data, but it is time-consuming. To enhance the performance of the network, we propose an innovative inertial feature encoder based on 1D Convolutional Neural Network (Conv1d) to process IMU data. Our model not only has fewer parameters but also outperforms the LSTM-based inertial encoders.

As presented in Fig. 2, the input of our inertial feature encoder contains $3\times N$ angular velocity $\boldsymbol{\omega}$ and $3\times N$ acceleration $\boldsymbol{a}$. The size of the window is determined by the sampling frequency of images and IMU data. In our experiments, N is 11. Firstly, we use Conv1d whose kernel size is 3 and whose output channel is 128 to process gyroscope data $\boldsymbol{\omega}$ and acceleration $\boldsymbol{a}$ separately. Then a max-pooling layer (kernel size $=$ 3, and padding $=$ 1) following two convolutional layers is adopted. The output of these layers is concatenated into a 256×7 feature map. In order to perform the following feature fusion stage, the width and the height of the output should be identical to features extracted by the visual feature encoder. Therefore, we reshape the inertial feature map into $128\times 2\times 7$ and then interpolate it into $128\times 4\times 13$.

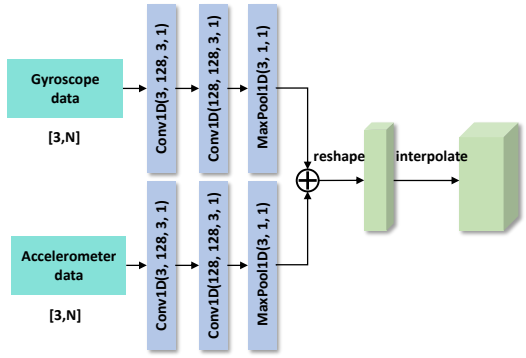


Fig. 2: Architecture for the inertial encoder, and $\oplus$ denotes concatenation operation.

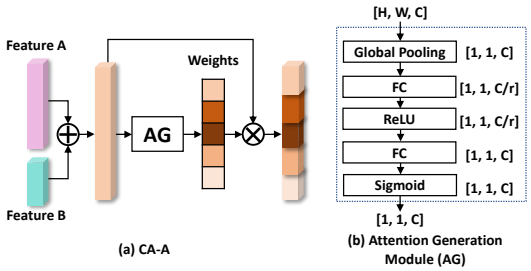


Fig. 3: An overview of the attention block, and $\otimes$ denotes element-wise production.

2.3. Attention Guided Feature Fusion

Conventional feature fusion strategies usually take channel-wise feature concatenation. But it is not robust for data with different distributions. In other words, simply concatenating two kinds of features lacks flexibility for our feature representations. Because static and dynamic input scenes have large deviations in motion descriptions which immensely influence our final predictions. To fuse the inertial and visual features in a more effective way, we attempt to design a

cross-domain channel attention structure. To realize it, four kinds of attention blocks are taken into consideration. One of them with the best performance CA-A is shown in Fig. 3, while the three other (CA-B, CA-C, CA-D) are introduced in the ablation study.

The attention block has two inputs, feature maps A and feature maps B. They are generated by the visual encoder and the inertial encoder in the feature extraction stage. A and B have the same shape of feature maps and a different number of channels. Let $A = [a_1, a_2, ..., a_{C_1}]$, $B = [b_1, b_2, ..., b_{C_2}]$, where $a_i, b_i \in \mathbb{R}^{H \times W}$ are one of feature maps, C_1 and C_2 is the number of channels in A and B respectively. The pipeline of our attention model is as follows. Firstly, we concatenate A and B and get the combined feature representation $M = [m_1, m_2, ..., m_C]$, where $C = C_1 + C_2$ and $m_i \in \mathbb{R}^{H \times W}$. Take F_{ag} to be the operator of Attention Generation Module (AG), then we have transformation $F_{ag} : M \rightarrow W$, where $W = [w_1, w_2, ..., w_C]$ and $w_i \in \mathbb{R}$ is the weight of channels. Ultimately, we assign these weights to M and output attention maps $Y = [w_1 m_1, w_2 m_2, .., w_C m_C]$.

In the attention block, there is an attention generation module (AG) aiming to generate a weight for each channel. Inspired by [17], we introduce Squeeze-and-Excitation (SE) block to our attention generation module. To recalibrate channel-wise feature responses adaptively, SE block models interdependencies between channels explicitly. By this means, CA-A is able to build correlations of input features, and allocates a group of appropriate weights for features involving large gaps.

2.4. Pose Regression

LSTM is capable of modeling dependencies in a sequence and VIO is a sequence-to-sequence problem. Thus in the pose regression stage, we choose a 2-layer LSTM that has 128 hidden states to do temporal modeling. Finally, we utilize a fully connected layer to regress the 6-DoF relative pose of two adjacent frames.

To make the model performance in ego-motion estimation better, the adaptive loss in [18] is applied to our training process. It is defined as:

$$\mathcal{L}(x, \alpha, c) = \frac{|2-\alpha|}{\alpha} \left(\left(\frac{(x/c)^2}{|2-\alpha|} + 1 \right)^{(\alpha/2)} - 1 \right) \quad (1)$$

where $c>0$ is a scale parameter that controls the size of the loss's quadratic bowl near x = 0 and $\alpha \in \mathbb{R}$ is a shape parameter that controls the robustness of the loss.

Given the ground truth pose (p, φ) and the predicted pose (p', φ'), we define the final loss function as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}(p - \hat{p}) + \lambda \cdot \mathcal{L}(\varphi - \hat{\varphi}) \quad (2)$$

where p is the translation, φ is the rotation (Euler angle), and λ is a scale factor to balance the weights of translation and rotation.

3. EXPERIMENTAL RESULTS

3.1. Datasets and Metrics

We evaluate our proposed approaches on KITTI Odometry benchmark [21]. We conduct experiments on sequences 00-10 with ground truth, expect for sequence 03 without corresponding raw IMU file. We use monocular images and raw IMU data as the input of our model. The sampling frequency of the IMU data is 10 times as the monocular images and ground truth, where the former is 100 Hz and the latter is 10Hz. The input images are scaled to 416×128, and the size of IMU data between two consecutive frames is 11×6. In the training phase, the batch size is set to 16; the total epoch is set to 100; λ is set to 100; the sequence length is 5. The Adam optimizer is applied to optimize our model, where $\beta_1 = 0.9$, $\beta_2 = 0.999$, the weight decay is 3.5×10^{-4} and the learning rate is 0.001.

We use $t_{rel}(\%)$ (The average translational RMSE drift (%) on lengths of 100m-800m) and $r_{rel}(\%)$ (the average rotational RMSE drift (°/100m) on lengths of 100m-800m) as the evaluation metrics in our experiments. They are widely used to evaluate the accuracy of the estimated trajectory in KITTI Odometry benchmark.

3.2. Comparison with State-of-the-Art Approaches

We compare the performance of our model ATVIO against a number of VO/VIO approaches: DeepVIO [16], VIOLearner [19], ORB-SLAM [1], VINS-Mono [8], and SFMLearner [20]. The results are given in Table 1. We highlight the top-two algorithms with the best experimental results. It can be seen that deep models outperform traditional methods ORB-SLAM and VINS-Mono. ORB-SLAM is without loop closure here, and it uses monocular images as input. VINS-Mono is a tightly coupled VIO method. Obviously, if without tight time synchronization between the camera and the IMU, traditional methods have poor results on trajectory estimation than learning-based methods. Compared to state-of-the-art [16, 19], our model achieves a better performance on the average error of translation in all sequences. Notably, we pay more attention to $t_{rel}(\%)$, because it reflects the accuracy of localization.

3.3. Ablation Study

As shown in Table 2, we demonstrate the comparisons with different feature fusion approaches, including state-of-the-art sensor fusion work [15], DirectCat (directly concatenating two kinds of features) and LSTM-VONAS (LSTM-based inertial encoder + VONAS-based visual encoder). We also compare with VONAS-A [2], which is a pure VO method. VONAS-A is our baseline. In order to make the results more clear, we visualize them in Fig. 4.

3.3.1. Ablation on the inertial encoder

In order to validate the effectiveness of our inertial encoder based on Conv1d, we replace our inertial encoder with 2-layer LSTM (LSTM-VONAS). And we remove the inertial encoder (VONAS-A). As shown in Fig. 4 (a), the performance of VIO methods are better than VONAS-A. In particular, our CA-A outperforms the LSTM-VONAS. The results prove the effectiveness of our inertial encoder. Besides, we compare our Conv1d-based inertial encoder with LSTM-based inertial encoder, both of which are followed by a fully connected layer to regress the relative pose. The former is 0.11 MB and the latter is 2.54 MB. Furthermore, we verify the efficiency from the perspective of time. We conduct experiments on sequence 00, 05, 07 on Geforce RTX 2080 Ti. Our model CA-A can process 235 frames per second on average, while LSTM-VONAS is 231.

3.3.2. Ablation on the attention block

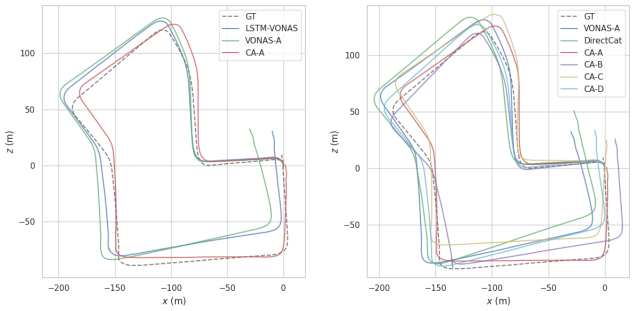
We compare with SelectFusion (denoted as SF) [22], because we both concentrate on integrating the observations from different sensors. The results of SF (soft) and SF (hard) are reproduced from [22]. Even though SF utilizes the information of lidar and vision instead of IMU and vision, and the image size is larger than ours, our fusion method CA-A still outperforms SF in the translation error of sequence 07, sequence 10 and the average. According to the

Table 1: Comparison to VO/VIO approaches. † are VO methods. Our model train and test on the same sequences as [16].

	ATVIO (Ours)		*DeepVIO [16]*		*VIOLearner [19]*		*ORB-SLAM*† *[1]*		*VINS-Mono [8]*		*SFMLearner*† *[20]*	
Seq	t_{rel}	r_{rel}	t_{rel}	r_{rel}	t_{rel}	r_{rel}	t_{rel}	r_{rel}	t_{rel}	r_{rel}	t_{rel}	r_{rel}
00	**2.53**	1.01	**11.62**	2.45	14.27	5.29	25.29	7.37	18.83	2.49	65.27	6.23
02	**1.41**	0.60	4.52	1.44	**4.07**	1.48	26.30	3.10	21.03	2.61	57.59	4.09
05	**1.97**	0.98	**2.86**	2.32	3.00	1.40	26.01	10.62	21.90	2.72	16.76	4.06
07	**1.64**	1.17	**2.71**	1.66	3.60	2.06	24.53	10.83	15.39	2.42	17.52	5.38
08	**2.04**	0.88	**2.13**	1.02	2.93	1.32	32.40	12.13	32.66	3.09	24.02	3.05
09	3.99	1.63	**1.38**	1.12	**1.51**	0.90	45.52	3.10	41.47	2.41	21.63	3.57
10	4.06	1.81	**0.85**	1.03	**2.04**	1.37	6.39	3.20	20.35	2.73	20.54	10.93
Avg	**2.52**	1.16	**3.72**	1.58	4.49	1.97	25.72	7.19	21.61	2.64	31.90	5.33

Table 2: Comparison to different feature fusion work. † is a VO method. We use sequences 00, 02, 04, 06, 08, 09 for training and 05, 07, 10 for testing. The image size of SF is 512×256, others are 416×128.

Method	Seq.05		Seq.07		Seq.10		Avg	
	t_{rel}	r_{rel}	t_{rel}	r_{rel}	t_{rel}	r_{rel}	t_{rel}	r_{rel}
SF(soft) [22]	**4.65**	1.83	**4.36**	2.19	8.35	2.01	5.78	2.01
SF(hard) [22]	**4.25**	1.67	4.46	2.17	**5.81**	1.55	**4.84**	1.80
VONAS-A † [2]	5.92	2.66	5.00	3.16	6.16	3.19	5.70	3.00
Ours(CA-A)	4.93	2.40	**3.78**	2.59	**5.71**	2.96	**4.80**	2.65
DirectCat	4.20	1.74	5.32	3.17	8.25	4.19	5.92	3.03
LSTM+VONAS	5.44	2.24	4.30	3.19	6.08	2.92	5.27	2.78
Ours(CA-B)	4.32	1.87	4.36	2.86	7.03	3.54	5.24	2.76
Ours(CA-C)	4.48	1.74	6.15	3.37	6.79	2.39	5.81	2.50
Ours(CA-D)	5.13	2.34	4.08	2.57	6.62	3.49	5.28	2.80



(a) Ablation of the inertial encoder (b) Ablation of the attention block

Fig. 4: Ablation Study on Sequence 07. The gray dotted line is the ground truth. The red line is our model ATVIO.

results, our model has a better representation of models in translation, which is what we are concerned with. In general, our proposed model achieves excellent performance in feature fusion.

Besides CA-A, we consider verifying whether the visual features and inertial features have potential relations. So, we have thought of an investigation that learns intra-class weights for two kinds of features respectively, and then allocates weights for the inter-class features, namely CA-B, CB-C, CA-D, as shown in Fig. 5. Firstly, They separately processes the two kinds of original feature maps to generate internal weights (the weights presenting intra-class relations), obtaining two kinds of attention maps. Then CA-B adopts two kinds of attention maps to generate the external weights (the weights presenting inter-class relations), while CA-C and CA-D use original feature maps to produce the external weights. CA-C and CA-D differ in the way to produce the weight α and β. CA-C feeds the feature maps to an operating group of $\{CNN, Pooling\}$, while CA-D utilizes the operating group of $\{Pooling, FC\}$.

We evaluate the importance of our proposed attention blocks. The results are shown in Fig. 4 (b). We firstly compare with vision-only odometry VONAS-A [2]. It is obvious that CA-A (red line) achieves better performance than VONAS-A (dark-blue). Compared to DirectCat, it demonstrates that the proposed attention-guided fusion strategy CA-A is superior to directly feature concatenation. We notice CA-A has better performance than CA-B, CA-C and CA-D. It is because CA-A concatenates channel-wise features and deals with two kinds of features via a global manner, which simultaneously learns the intra-class and inter-class relationships of visual features and inertial features. However, other modules just rely on α and β to represent external weights, and the internal correlations of the two kinds of features are not sufficiently exploited. The results also manifest visual information and inertial information have a strong internal relationship, which reflects that our fusion strategy is effective.

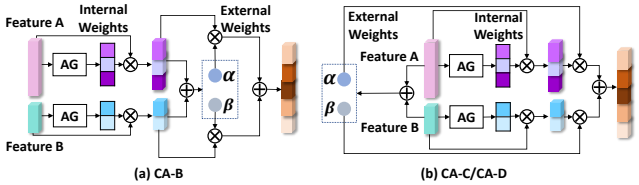


Fig. 5: The proposed variations of other attention blocks: (a) CA-B, (b) CA-C and CA-D share the same framework but generate α and β in different ways.

4. CONCLUSION AND FUTURE WORK

We propose a deep-learning framework based on the channel attention mechanism for visual-inertial odometry. We adopt the current advanced encoder based on a neural architecture search technology for visual feature extraction. To process IMU measurements more quickly and effectively, we design a sophisticated network for inertial feature extraction. Besides, we propose a novel cross-domain channel attention block to combine the extracted features in a more adaptive manner. The extensive ablation studies demonstrate the effectiveness of the proposed inertial encoder and the attention block. The experimental results show the proposed method is competitive against state-of-the-art methods.

5. REFERENCES

[1] Raul Mur-Artal, J. M. M. Montiel, and Juan D. Tardós, "ORB-SLAM: A versatile and accurate monocular SLAM system," *IEEE Trans. Robotics*, vol. 31, no. 5, pp. 1147–1163, 2015.

[2] Xing Cai, Lanqing Zhang, Chengyuan Li, Ge Li, and Thomas H. Li, "VONAS: network design in visual odometry using neural architecture search," in *ACM MM*. 2020, pp. 727–735, ACM.

[3] Changhao Chen, Xiaoxuan Lu, Andrew Markham, and Niki Trigoni, "Ionet: Learning to cure the curse of drift in inertial odometry," in *AAAI*, 2018, pp. 6468–6476.

[4] Mahdi Abolfazli Esfahani, Han Wang, Keyu Wu, and Shenghai Yuan, "Aboldeepio: A novel deep inertial odometry network for autonomous vehicles," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 5, pp. 1941–1950, 2019.

[5] Stephan Weiss and Roland Siegwart, "Real-time metric state estimation for modular vision-inertial systems," in *2011 IEEE international conference on robotics and automation*. IEEE, 2011, pp. 4531–4537.

[6] Ke Sun, Kartik Mohta, Bernd Pfrommer, Michael Watterson, Sikang Liu, Yash Mulgaonkar, Camillo J Taylor, and Vijay Kumar, "Robust stereo visual inertial odometry for fast autonomous flight," *IEEE Robotics and Automation Letters*, vol. 3, no. 2, pp. 965–972, 2018.

[7] Mingyang Li and Anastasios I Mourikis, "Improving the accuracy of ekf-based visual-inertial odometry," in *2012 IEEE International Conference on Robotics and Automation*. IEEE, 2012, pp. 828–835.

[8] Tong Qin, Peiliang Li, and Shaojie Shen, "Vins-mono: A robust and versatile monocular visual-inertial state estimator," *IEEE Transactions on Robotics*, vol. 34, no. 4, pp. 1004–1020, 2018.

[9] Sen Wang, Ronald Clark, Hongkai Wen, and Niki Trigoni, "Deepvo: Towards end-to-end visual odometry with deep recurrent convolutional neural networks," in *2017 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2017, pp. 2043–2050.

[10] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G Lowe, "Unsupervised learning of depth and ego-motion from video," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1851–1858.

[11] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," in *ICLR*, 2015.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[13] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox, "Flownet: Learning optical flow with convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2758–2766.

[14] Ronald Clark, Sen Wang, Hongkai Wen, Andrew Markham, and Niki Trigoni, "Vinet: Visual-inertial odometry as a sequence-to-sequence learning problem," in *AAAI*, 2017, pp. 3995–4001.

[15] Changhao Chen, Stefano Rosa, Yishu Miao, Chris Xiaoxuan Lu, Wei Wu, Andrew Markham, and Niki Trigoni, "Selective sensor fusion for neural visual-inertial odometry," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10542–10551.

[16] Liming Han, Yimin Lin, Guoguang Du, and Shiguo Lian, "Deepvio: Self-supervised deep learning of monocular visual inertial odometry using 3d geometric constraints," in *IROS*. 2019, pp. 6906–6913, IEEE.

[17] Jie Hu, Li Shen, and Gang Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[18] Jonathan T Barron, "A general and adaptive robust loss function," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4331–4339.

[19] E. Jared Shamwell, Sarah Leung, and William D. Nothwang, "Vision-aided absolute trajectory estimation using an unsupervised deep network with online error correction," in *IROS*. 2018, pp. 2524–2531, IEEE.

[20] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G. Lowe, "Unsupervised learning of depth and ego-motion from video," in *CVPR*, 2017, pp. 6612–6619.

[21] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun, "Vision meets robotics: The kitti dataset," *The International Journal of Robotics Research*, vol. 32, no. 11, pp. 1231–1237, 2013.

[22] Changhao Chen, Stefano Rosa, Chris Xiaoxuan Lu, Niki Trigoni, and Andrew Markham, "Selectfusion: A generic framework to selectively learn multisensory fusion," *CoRR*, vol. abs/1912.13077, 2019.