IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 15 April 2023, accepted 14 May 2023, date of publication 23 May 2023, date of current version 22 June 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3279292

METHODS

VIUNet: Deep Visual–Inertial–UWB Fusion for Indoor UAV Localization

PENG-YUAN KAO[1], **HSIU-JUI CHANG**[2], **KUAN-WEI TSENG**[3], **(Student Member, IEEE)**, **TIMOTHY CHEN**[2], **HE-LIN LUO**[4], **AND YI-PING HUNG**[1], **(Member, IEEE)**

[1]Graduate Institute of Networking and Multimedia, National Taiwan University, Taipei 10617, Taiwan
[2]Department of Computer Science and Information Engineering, National Taiwan University, Taipei 10617, Taiwan
[3]Department of Computer Science, Tokyo Institute of Technology, Tokyo 152-8550, Japan
[4]Graduate Institute of Animation and Film Art, Tainan National University of the Arts, Tainan 72045, Taiwan

Corresponding author: Yi-Ping Hung (hung@csie.ntu.edu.tw)

This work was supported in part by the Ministry of Science and Technology, Taiwan, under Grant MOST 109-2221-E-002-207-MY3.

ABSTRACT Camera, inertial measurement unit (IMU), and ultra-wideband (UWB) sensors are commonplace solutions to unmanned aerial vehicle (UAV) localization problems. The performance of a localization system can be improved by integrating observations from different sensors. In this paper, we propose a learning-based UAV localization method using the fusion of vision, IMU, and UWB sensors. Our model consists of visual–inertial (VI) and UWB branches. We combine the estimation results of both branches to predict global poses. To evaluate our method, we augment a public VI dataset with UWB simulations and conduct a real-world experiment. The experimental results show that our method provides more robust and accurate results than VI/UWB-only localization. Our codes and data are available at https://imlabntu.github.io/VIUNet/.

INDEX TERMS Visual-inertial odometry, ultra-wideband, sensor fusion, deep learning.

I. INTRODUCTION

With the increasing development of techniques such as smart cities and the Internet of Things (IoT), precise localization results have become more important. Although GPS applications are well-developed and widely available in global localization techniques, GPS signals are unreliable in indoor scenes. Moreover, the system provides only meter-level accuracy, which is insufficient for unmanned aerial vehicle (UAV) flight. Therefore, many localization techniques have been designed to provide accurate positioning.

A localization system uses sensors to capture environmental information and estimate agent positions. However, indoor scenes contain states and environments that a single sensor modality cannot always observe. For instance, visual sensors provide color and texture information but are defeated by changes in illumination, by motion blur, by dynamic objects, or by textureless scenes [1], [2]. Inertial sensors provide acceleration and angular rate, which environmental changes generally do not affect. However, inertial sensors used in localization still suffer from noisy measurements and accumulated error [3], [4]. Ultra-wideband (UWB) sensors are also often used for localization. In a UWB localization system, UWB sensors estimate distances between targets and anchors with known locations to calculate the coordinates of the targets. A schematic diagram of a UWB localization system is shown in Figure 1.

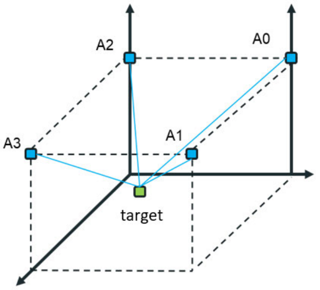


FIGURE 1. Schematic diagram of UWB localization system. A0, A1, A2, and A3 are UWB anchors with known locations. Target coordinates are calculated after estimating the distance between the target and each anchor.

The associate editor coordinating the review of this manuscript and approving it for publication was Wenbing Zhao.

Multimodal sensor fusion combines different sensors to increase the robustness of the system. In visual–inertial odometry (VIO) systems [5], [6], visual and inertial measurements are combined to provide 6DoF camera poses in real-time. In VIO systems, vision provides rich information that is not available in an inertial measurement unit (IMU), whereas IMU makes vision more robust to environmental changes. However, the odometry results of VIO systems lack global information, and the initialization process largely determines their performance. Therefore, we propose a VIO with UWB-aided relocalization, a learning-based method combining visual, inertial, and UWB sensors to estimate the global 6DoF poses of the target, including translation and rotation. UWB sensors provide a good initialization for the VIO system and offer global information that the VIO system can reference.

We propose a real-time[1] learning-based Visual-Inertial-UWB fusion (VIU-Net) that contains a visual–inertial (VI) branch and a UWB branch. The VI branch takes image sequences and IMU measurements as input and computes relative poses between consecutive images, and the UWB branch utilizes UWB measurements to regress the global position. We integrate the outputs of the two branches to predict accurate global poses. By using deep learning, our method are robust to data corruption and noisy sensor measurements. Since the model tends to *learn* those biases in the training process, the proposed method does not depend on accurate calibration of UWB beacon location. However, with conventional methods, those biases will easily drift the result of estimated position. To evaluate the proposed method, we add UWB simulations on EuRoC [7], a public VI dataset, and conduct a real-world experiment. The experimental results reveal that compared to VI/UWB-only positioning, the fusion of VI and UWB improves global localization accuracy.

To summarize, this work has three contributions:

1) Our method is the first deep learning method combining Vision, IMU, UWB measurements for localization.
2) We collected the first dataset for Vision, IMU, and UWB localization.
3) We improved the localization accuracy of VI branch and VIU-Net with loss function.

II. RELATED WORK

The proposed system is a learning-based framework that contains visual, inertial, and UWB sensors. This section introduces deep learning methods for localization and discusses sensor fusion, including traditional and learning-based approaches.

A. DEEP LEARNING FOR LOCALIZATION

1) VISUAL SENSORS

Deep learning has achieved great success in visual odometry (VO) [8], [9], [10], [11], [12], [13]. Learning-based methods have shown results comparable to those of traditional methods, but do not always require modules in the classic VO pipeline, such as camera calibration and outlier rejection.

2) INERTIAL SENSORS

Yan et al. [3] were the first research to integrate sophisticated machine learning techniques with inertial navigation. Chen et al. [4] propose a neural network framework to learn inertial odometry directly from IMU raw data. Both studies show that it is possible to use deep learning to estimate inertial odometry.

3) ULTRA-WIDEBAND SENSORS

Deep learning methods have been developed for UWB localization, including long short-term memory (LSTM) networks [14], CNNs [15], and deep neural networks (DNNs) [16]. Deep learning can also be used to correct UWB sensor measurement errors [17]. In this work, we use DNN to predict positions, and further integrate this with our VI branch.

B. SENSOR FUSION

1) TRADITIONAL METHODS

Sensor measurements can be combined in either a loosely-coupled or a tightly-coupled manner. Loosely-coupled sensor fusion treats the estimation of different sensor units as independent [18], whereas the tightly-coupled approach integrates raw sensor data at a lower processing level [19]. Traditional methods, in turn, can be divided into filtering-based and optimization-based approaches, mainly according to the backend optimization type [20]. Filtering-based methods such as the Kalman filter and the particle filter use a linear or nonlinear model to estimate the state of a dynamic system and predict results by finding the most similar target to the model. When the measurement is received, the filtering-based method executes propagation and update steps to update the state. Optimization-based algorithms typically combine the error terms into a cost function and optimize the system to minimize the cost function. Yang et al. [21] propose R-UVIS, a tightly-coupled UWB–visual–inertial indoor localization system, along with an optimation-based algorithm. Xu et al. [22] use tightly-coupled VIO results and the UWB-based distance measurements of a pair of drones to accomplish optimization-based decentralized visual–inertial–UWB fusion for relative state estimation. However, with conventional methods, biases of UWB and IMU will easily drift the result of estimated position if the accurate calibration is not applied. On the contrary, we use the deep learning method to fuse VIO results and UWB results. By using deep learning methods, our method benefits from some traits of deep learning, such as toleration of bias in input datas. Since the model tends to *learn* those biases in the training process, the proposed method does not depend on accurate calibration of UWB beacon location.

[1] The inference speed reaches 30+ fps on a RTX3090 GPU.

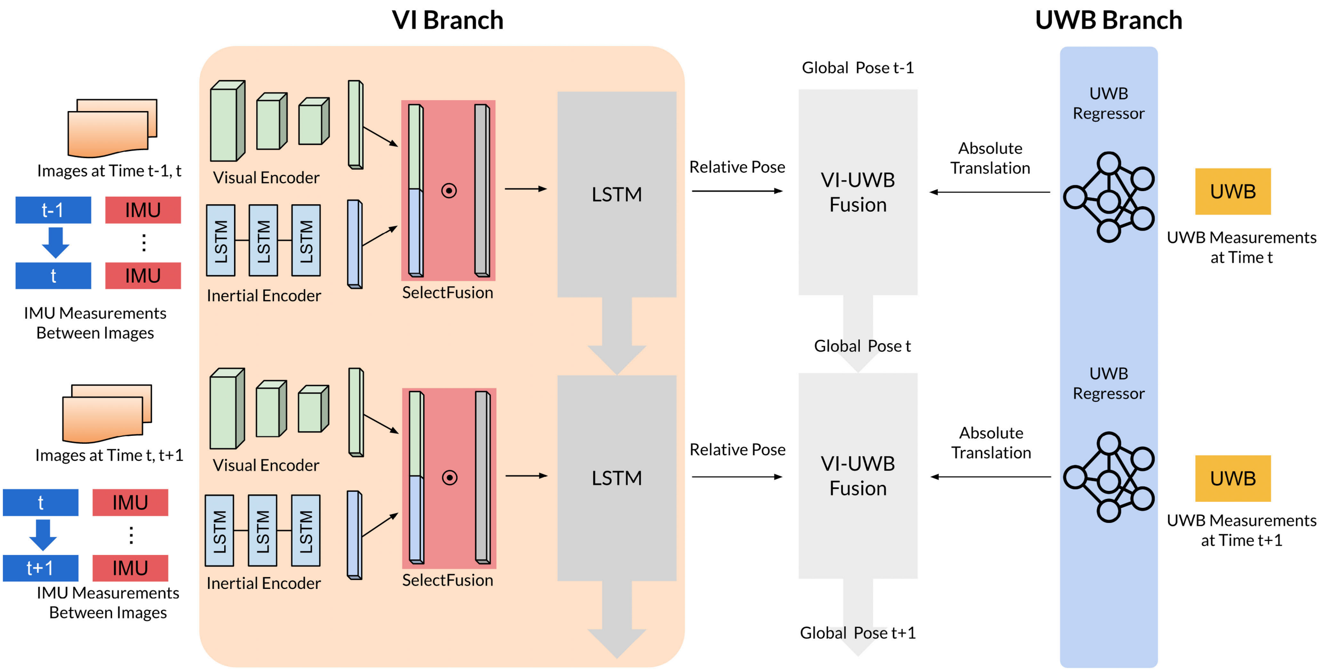


FIGURE 2. **Network architecture of proposed method. The VI branch (with the orange background) takes images and IMU measurements as input and predicts the relative pose between each image frame. The UWB branch (blue background) regresses the absolute position according to the UWB measurements.**

2) LEARNING-BASED METHODS

The goal of multimodal machine learning [23] is to build models that can process heterogeneous sources from multiple modalities. Multimodal machine learning has recently been used in language and vision applications to provide more robust predictions, since multimodal systems can still operate when one of the modalities is unavailable. For localization, applications fuse visual–LiDAR [24], [25], and visual–inertial [25], [26], [27], [28], [29] data. Among these, SelectFusion [25] proposes fusion modules based on attention strategies. Such ideas have shown success in vision–depth, vision–LiDAR, and visual–inertial fusion. SelectFusion includes two selective fusion modules—one using deterministic soft fusion and the other using Gumbel-softmax-based hard fusion—to integrate different modality features. However, none of the abovementioned methods utilize the deep fusion of visual, inertial, and UWB information. Our approach adds the UWB sensors and enhances the model's performance.

III. PROPOSED METHOD

In this section, we describe our method in detail. The system comprises the VI branch, the UWB branch, and the VI-UWB fusion mechanism. The VI branch extracts visual and inertial features to estimate the relative pose between each frame, and the UWB branch regresses the absolute position by considering the anchor positions and the distances between the tag and each anchor. The poses estimated by these two branches are then fused in an adaptive manner to estimate the global pose of the agent. In addition, a multi-task loss function is used to handle translation and rotation loss. The network architecture of the proposed method is shown in Figure 2.

A. VISUAL–INERTIAL (VI) BRANCH

The VI branch is constructed following [25] in the VIO task. In the visual encoder, we calculate the optical flow of two consecutive image frames by FlowNetSimple [30], a popular convolutional neural network (CNN), to estimate optical flow. We convert the optical flow to a 256-dimensional visual feature by a fully connected layer. The inertial encoder is a two-layer bi-directional long short-term memory (LSTM) network with 128 hidden states that takes as input the triaxial acceleration and triaxial angular velocity between two consecutive image frames, and outputs inertial features.

After the visual and inertial features are computed, they are concatenated and the soft fusion method [25] is used to generate a weighted feature, which is passed through a two-layer unidirectional LSTM to model temporal dependencies. The output of the unidirectional LSTM is then used to estimate the relative translation and rotation between two frames at time steps $t-1$ and t. We denote the output as $\Delta T^{\mathrm{VI}}_{t-1,t}$ and $\Delta R^{\mathrm{VI}}_{t-1,t}$, where $\Delta T^{\mathrm{VI}}_{t-1,t} \in \mathbb{R}^3$ is a translation vector, and $\Delta R^{\mathrm{VI}}_{t-1,t} \in \mathbb{R}^4$ is a quaternion rotation vector.

B. ULTRA-WIDEBAND (UWB) BRANCH

We propose a learning-based model to estimate global poses based on UWB measurements. The proposed model can be roughly viewed as a traditional trilateration method to calculate positions, the difference being that our model can be

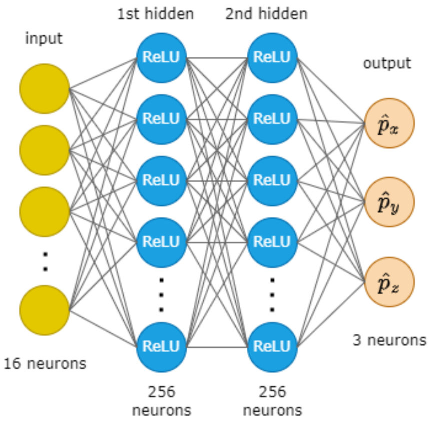


FIGURE 3. **Deep neural network model for UWB branch.**

fine-tuned for better performance, since the model can learn the UWB measurement bias in a specific environment.

The model consists of three fully connected layers. The first and the second hidden states contain 256 neurons, whereas the input contains 16 neurons and the output contains three neurons. Each fully connected layer, except for the last one, is followed by a ReLU activation function. The model architecture is shown in Figure 3. Given N UWB anchors, the input should contain the distances from the target to the anchors, $d_i, i = 1, \ldots, N$, and the anchor positions $p_i^A = (p_{ix}^A, p_{iy}^A, p_{iz}^A), i = 1, \ldots, N$. In our experiments, we set N to 4, and the output is the estimated global translation $\widehat{T}^{\text{UWB}} = (\widehat{p}_x, \widehat{p}_y, \widehat{p}_z)$.

C. VI-UWB FUSION

After computing the VI and UWB branches, we integrate their outputs. As mentioned above, the VI branch provides the relative pose estimations $\Delta T_{t-1,t}^{\text{VI}}$ and $\Delta R_{t-1,t}^{\text{VI}}$. This facilitates the calculations of the VI predictions of the global pose, $\widehat{T}_t^{\text{VI}}$ and $\widehat{R}_t^{\text{VI}}$, which is based on the global pose estimation of time step $t-1$, $\widehat{T}_{t-1}$ and $\widehat{R}_{t-1}$, by

$$\widehat{T}_t^{\text{VI}} = \widehat{T}_{t-1} + \widehat{R}_{t-1}\Delta T_{t-1,t}^{\text{VI}} \tag{1}$$

$$\widehat{R}_t^{\text{VI}} = \widehat{R}_{t-1}\Delta R_{t-1,t}^{\text{VI}}, \tag{2}$$

where $\widehat{T}_t^{\text{VI}}, \widehat{T}_{t-1}$ are 3-dimensional translation vectors; $\widehat{R}_t^{\text{VI}}, \widehat{R}_{t-1}$ are 3×3 rotation matrices.

The output of the UWB branch at time step t is denoted by $\widehat{T}_t^{\text{UWB}}$, as it contains translation but no rotation. Then, the final estimation of our model is calculated as

$$\widehat{T}_t = \begin{cases} \widehat{T}_t^{\text{UWB}}, & t = 0 \\ \alpha \widehat{T}_t^{\text{VI}} + (1-\alpha)\widehat{T}_t^{\text{UWB}}, & t \geq 1 \end{cases}, \tag{3}$$

$$\widehat{R}_t = \begin{cases} I_3, & t = 0 \\ \widehat{R}_t^{\text{VI}}, & t \geq 1 \end{cases}, \tag{4}$$

where α is a learnable parameter to decide the weight of VI and UWB estimation. Note that when $t = 0$, there are no consecutive image frames for the VI branch. In this case, we use $\widehat{T}_t^{\text{UWB}}$ to initialize the global translation. Although UWB can produce inaccurate measurements, our method corrects the position error after several frames. We initialize the global rotation using the identity matrix.

D. LOSS FUNCTIONS

For the loss function, we first define the loss for relative translation, relative rotation, global translation, and global rotation.

$$\mathcal{L}_{\text{relative}T}^{\phi} = \left\| \Delta T_{t-1,t}^{\phi} - \Delta T_{t-1,t} \right\|_2 \tag{5}$$

$$\mathcal{L}_{\text{relative}R}^{\phi} = \min(\|\Delta R_{t-1,t}^{\phi} - \Delta R_{t-1,t}\|_1, \|\Delta R_{t-1,t}^{\phi} + \Delta R_{t-1,t}\|_1) \tag{6}$$

$$\mathcal{L}_{\text{global}T}^{\phi} = \left\| \widehat{T}_t^{\phi} - T_t \right\|_2 \tag{7}$$

$$\mathcal{L}_{\text{global}R}^{\phi} = \min(\|\widehat{R}_t^{\phi} - R_t\|_1, \|\widehat{R}_t^{\phi} + R_t\|_1) \tag{8}$$

Here, $\phi \in \{all, \text{VI}, \text{UWB}\}$, where *all* indicates the total result loss, *VI* the VI branch loss, and *UWB* the UWB branch loss. Note that since UWB provides only global translation results, $\mathcal{L}_{\text{relative}T}^{\text{UWB}}$, $\mathcal{L}_{\text{relative}R}^{\text{UWB}}$, and $\mathcal{L}_{\text{global}R}^{\text{UWB}}$ are not defined. We empirically chose norm-2 and norm-1 in the loss functions for translation and rotation, respectively that achieved the best localization accuracy by the experimental results of several trials.

In equations (5)–(8), $\Delta T_{t-1,t}$, $\Delta R_{t-1,t}$ are the ground-truth relative translation and rotation, and T_t and R_t are the ground-truth global translation and rotation. Since we use quaternions, and since quaternions q and $-q$ represent the same rotation, we resolve the antipodal problem in the loss function as equations (6) and (8). We also use the mean absolute error between the true and predicted quaternions, whereas the translation error is defined as the mean square error.

E. TRAINING PROCESS AND IMPLEMENTATION

We propose two-stage training for both the VI and UWB branches. In the first stage, we train the VI branch and UWB branch separately. However, position initialization is not possible with just the VI branch. To train the VI branch, we modify equation (3) to not use the UWB branch information:

$$\widehat{T}_t = \begin{cases} T_t, & t = 0 \\ \widehat{T}_t^{\text{VI}}, & t \geq 1. \end{cases} \tag{9}$$

In addition, we use multi-task loss functions [31], [32] to combine the translation and rotation loss:

$$\mathcal{L}_{\text{comb}} = \frac{1}{2\sigma_1^2}(\mathcal{L}_{\text{relative}T}^{\text{VI}} + \mathcal{L}_{\text{global}T}^{\text{VI}}) + \ln(1+\sigma_1^2) + \frac{1}{2\sigma_2^2}(\mathcal{L}_{\text{relative}R}^{\text{VI}} + \mathcal{L}_{\text{global}R}^{\text{VI}}) + \ln(1+\sigma_2^2), \tag{10}$$

where σ_1 and σ_2 are learnable parameters for the weight of translation and rotation. Using multi-task loss functions is more robust when training the model, since manually defining loss weights is time-consuming and easily lead to models not converging. After completing the VI and UWB branch training, we choose the best epoch according to the validation

TABLE 1. Loss in each stage.

Target	Stage 1 VI	Stage 1 UWB	Stage 2 Full network
Relative translation	✓	–	–
Relative rotation	✓	–	✓
Global translation	✓	✓	✓
Global rotation	✓	–	✓
Combined	✓	–	✓

results. Then, in the second stage, we fix the UWB branch and train the whole network by

$$\mathcal{L} = \frac{1}{2\sigma_3^2}(\mathcal{L}_{\text{global}T}^{\text{all}}) + \ln(1+\sigma_3^2) + \frac{1}{2\sigma_4^2}(\mathcal{L}_{\text{relative}R}^{\text{all}} + \mathcal{L}_{\text{global}R}^{\text{all}}) + \ln(1+\sigma_4^2), \qquad (11)$$

where σ_3 and σ_4 are learnable parameters for the translation and rotation weights. Note that we do not use relative translation loss since the VI branch is trained with fixing the parameters of the UWB branch instead of predicting the great relative pose by itself. The loss used to train each branch is shown in Table 1. Moreover, the VI and UWB estimation weight α in equation (3) is also trained in the second stage. Learnable parameters: α, σ_1, σ_2, σ_3, and σ_4, which controls the importance of different UWB/VI ratio and the balance between relative and absolute poses. These parameters are fully differentiable. We jointly optimize these parameters along with deep neural networks' parameters by simply adding these parameters into the optimizers.

The proposed networks were implemented with PyTorch and trained on an NVIDIA GeForce RTX 3090 GPU. For the VI branch, we set learning rate $\text{lr} = 1\times 10^{-4}$ and trained for 80 epochs with a batch size of 4 using the Adam optimizer. We also set the training sequence length to 3. For the UWB branch, we set $\text{lr} = 1\times 10^{-4}$, the dropout rate to 0.2, and trained for 100 epochs with a batch size of 10 using the Adam optimizer. In the second stage of the training process, the settings were similar to the VI branch, but with 16 epochs and a training sequence length of 5.

IV. EXPERIMENTS

We conducted experiments to evaluate the proposed VI-UWB fusion method.

A. EVALUATION METRICS

To evaluate the resulting output trajectories, we calculated the root-mean-square error (RMSE) of the relative translation and rotation as

$$\text{RMSE}_{\text{relative}T} = \sqrt{\frac{1}{n}\sum_{t=1}^{n}\left\|\Delta\widehat{T}_{t-1,t} - \Delta T_{t-1,t}\right\|^2} \qquad (12)$$

$$\text{RMSE}_{\text{relative}R} = \sqrt{\frac{1}{n}\sum_{t=1}^{n}\left\|\text{Euler}(\Delta\widehat{R}_{t-1,t}^{-1}\Delta R_{t-1,t})\right\|^2}, \qquad (13)$$

where $\Delta\widehat{T}_{t-1,t} \in \mathbb{R}^3$ is the predicted relative translation vector, $\Delta T_{t-1,t} \in \mathbb{R}^3$ is the ground-truth relative translation vector, $\Delta\widehat{R}_{t-1,t}$ is the predicted 3×3 relative rotation matrix, and $\Delta R_{t-1,t}$ is the ground-truth 3×3 relative rotation matrix.

Similarly, the RMSE of the global translation and rotation is calculated as

$$\text{RMSE}_{\text{global}T} = \sqrt{\frac{1}{n+1}\sum_{t=0}^{n}\left\|\widehat{T}_t - T_t\right\|^2} \qquad (14)$$

$$\text{RMSE}_{\text{global}R} = \sqrt{\frac{1}{n+1}\sum_{t=0}^{n}\left\|\text{Euler}(\widehat{R}_t^{-1}R_t)\right\|^2}, \qquad (15)$$

where $\widehat{T}_t \in \mathbb{R}^3$ is the predicted global translation vector, $T_t \in \mathbb{R}^3$ is the ground-truth global translation vector, $\widehat{R}_t$ is the predicted 3×3 global rotation matrix, and R_t is the ground-truth 3×3 global rotation matrix.

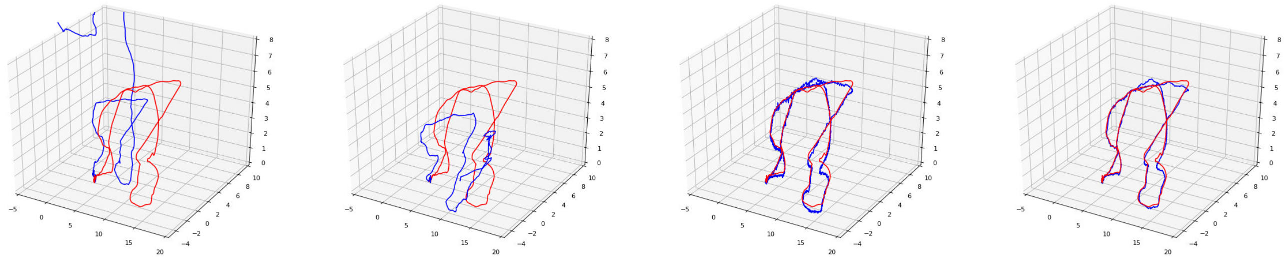
TABLE 2. Characteristics of each sequence in EuRoC dataset [7].

Name	Length / Duration	Average Velocity / Angular Velocity	Note
MH_01_easy	80.6 m 182 s	0.44 m/s 0.22 rad/s	Good texture, bright scene
MH_02_easy	73.5 m 150 s	0.49 m/s 0.21 rad/s	Good texture, bright scene
MH_03_medium	130.9 m 132 s	0.99 m/s 0.22 rad/s	Fast motion, bright scene
MH_04_difficult	91.7 m 99 s	0.93 m/s 0.24 rad/s	Fast motion, dark scene
MH_05_difficult	97.6 m 111 s	0.88 m/s 0.21 rad/s	Fast motion, dark scene
V1_01_easy	58.6 m 144 s	0.41 m/s 0.28 rad/s	Slow motion, bright scene
V1_02_medium	75.9 m 83.5 s	0.91 m/s 0.56 rad/s	Fast motion, bright scene
V1_03_difficult	79.0 m 105 s	0.75 m/s 0.62 rad/s	Fast motion, motion blur
V2_01_easy	36.5 m 112 s	0.33 m/s 0.28 rad/s	Slow motion, bright scene
V2_02_medium	83.2 m 115 s	0.72 m/s 0.59 rad/s	Fast motion, bright scene
V2_03_difficult	86.1 m 115 s	0.75 m/s 0.66 rad/s	Fast motion, motion blur

B. EuRoC DATASET

The EuRoC dataset [7] is a public benchmark containing 11 sequences with synchronized stereo images, IMU measurements, and ground-truth poses. The EuRoC dataset was recorded by a micro aerial vehicle in two indoor scenes: Machine Hall (MH) and Vicon Room (V). The characteristics of each sequence are shown in Table 2. We used sequence *V1_01_easy* for model validation, sequence *MH_04_difficult* for testing, and the other sequences for training.

Since the EuRoC dataset contains no UWB data, we simulated UWB measurements by calculating the ground-truth distances from target to anchors and adding Gaussian noise with standard deviation σ to each distance. In our experiments, we set $\sigma = 0.03$ m and 0.1 m in both scenes. To choose the proper anchor coordinates, we plotted all the trajectories and determined the bounding box that contained all the sequences. The positions of the anchors were the vertices on top of the bounding box.



(a) SelectFusion (b) VI branch (c) UWB branch (d) VI-UWB

FIGURE 4. Trajectories of different localization methods on *MH_04_difficult*. Ground-truth and estimated trajectories are marked in red and blue, respectively.

TABLE 3. Global translation 3D RMSE (unit: m) on each sequence in EuRoC dataset using *V1_01_easy* as the validation sequence. Bold numbers indicate the lowest error value. VIUWB*: Same structure, but with higher error ($\sigma = 0.10$) on UWB simulation.

Sequence	SelectFusion [25]	VI branch	UWB branch	VI-UWB	VI-UWB*
MH_01_easy	11.69	9.48	0.10	**0.10**	0.17
MH_02_easy	10.50	6.98	0.13	**0.09**	0.14
MH_03_medium	6.41	4.41	**0.25**	**0.25**	0.36
MH_04_difficult	7.06	3.57	0.26	**0.17**	0.37
MH_05_difficult	6.06	3.36	0.18	**0.16**	0.35
V1_02_medium	4.60	3.12	0.09	**0.08**	0.10
V1_03_difficult	9.58	8.58	0.11	**0.09**	0.14
V2_01_easy	6.58	9.03	0.12	**0.07**	0.07
V2_02_medium	5.26	4.74	0.14	**0.08**	0.18
V2_03_difficult	1.27	1.5	0.11	**0.09**	0.15
Average	6.90	5.48	0.15	**0.12**	0.18

TABLE 4. Results of different localization methods on EuRoC dataset sequence *MH_04_difficult*. Bold numbers indicate the lowest error value.

	Relative 3D RMSE		Global 3D RMSE	
	Trans (m)	Rot (°)	Trans (m)	Rot (°)
SelectFusion	**0.03**	**0.06**	7.06	14.81
Proposed (VI branch)	0.04	0.13	3.57	**7.85**
Proposed (UWB branch)	0.09	N/A	0.26	N/A
Proposed (VI-UWB)	0.04	0.08	**0.17**	8.29

The quantitative results on the *MH_04_difficult* sequence are shown in Table 4. In terms of global RMSE, the proposed VI-UWB methods exhibit superior performance on translation, with rotation close to the results of the VI branch. In terms of relative RMSE, SelectFusion achieves the best results because it focuses solely on relative performance. However, in real applications such as flight control and navigation systems, global pose is more important than relative pose. Since we have added global loss, $\mathcal{L}^{\phi}_{\text{global}T}$ and $\mathcal{L}^{\phi}_{\text{global}R}$, in our loss function, our VI branch and VI-UWB outperform in global pose estimation.

The output trajectories of different localization methods are shown in Figure 4. Ground-truth and estimated trajectories are marked in red and blue, respectively. In Figures 4a and 4b, the trajectories show that compared to SelectFusion, the VI branch prevents the predicted positions from drifting too far from the ground truth. Although the VI branch sometimes compromises the relative transformation performance, it contributes to a better global pose after integration with the UWB branch. Figure 4c shows the output trajectories of the UWB branch. Compared to VI methods, the UWB branch performs much better on global translation, but does not provide a rotation estimation. Moreover, with the time-dependent localization method of VI, the trajectory of UWB can be relatively smooth. In addition, the UWB branch has the most significant relative translation error, since UWB measurements are time-independent (time-independent means the measurements of the specific sampling is not related to the previous measurements of sampling.). In Figure 4c, the trajectory is clearly close to the ground truth, but a closer look reveals that it fluctuates. Our fusion method (Figure 4d) reduces UWB's global translation error and produces smoother trajectories, making our model more practical in real-world applications.

The global translation error is the most important metric in UAV localization. Therefore, we recorded and integrated the global translation RMSE of each sequence in Table 3. To evaluate each sequence, we used *V1_01_easy* as the validation sequence and the other sequences for training to ensure the training data did not contain the testing data. This table illustrates that our method is stable and robust in different environments.

C. REAL-WORLD DATASET

Since few datasets contain vision, IMU, and UWB, we recorded a dataset in our environment. We used a RealSense D435i camera to capture RGB images and IMU data, the Nooploop LinkTrack S system to record UWB measurements, and the Vicon motion capture system to record the 6DoF pose ground truth. As the Vicon motion capture system supports millimeter-level measurement precision, it is often used in localization tasks as a ground-truth system.

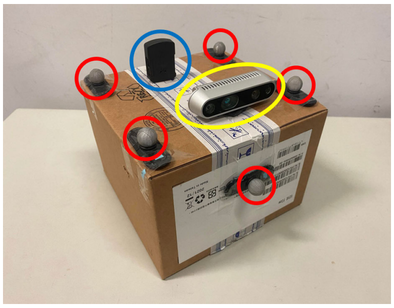


FIGURE 5. Positioning target in a real-world dataset. The yellow and blue circles mark the camera and UWB tag, respectively, and the five red circles indicate the Vicon ball for ground-truth estimation.

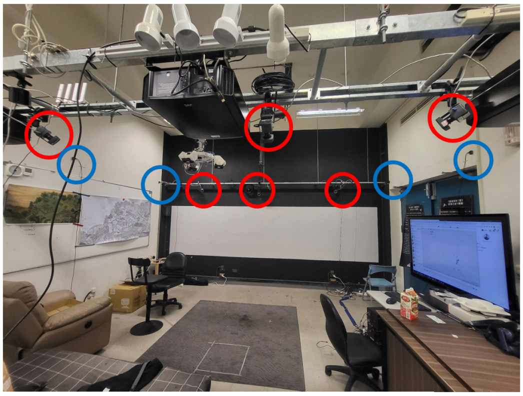


FIGURE 6. Environment for real-world dataset. The red and blue circles indicate the position of the Vicon cameras and UWB anchors, respectively.

In Figure 5, we illustrate the target used for localization. The yellow and blue circles mark the camera and UWB tag, respectively, and the five red circles indicate the Vicon ball used for ground-truth estimation. The experimental setup of the room is shown in Figure 6. The red and blue circles indicate the position of the Vicon cameras and UWB anchors, respectively. We held the object in Figure 5 and moved arbitrarily around the room to simulate a UAV flight. We recorded eleven sequences in total, including eight training sequences, one validation sequence, and two testing sequences. The characteristics of each sequence are shown in Table 5. Note that *Sequence_11* is a special case in which the trajectory is a rectangle at the same height instead of arbitrary movement; we used this sequence as a testing sequence.

Tables 6 and 7 show the quantitative results on real-world testing sequences; the output trajectories of different localization methods are shown in Figures 7 and 8. We also recorded the results of the Nooploop LinkTrack S, a positioning system based on IMU and UWB.

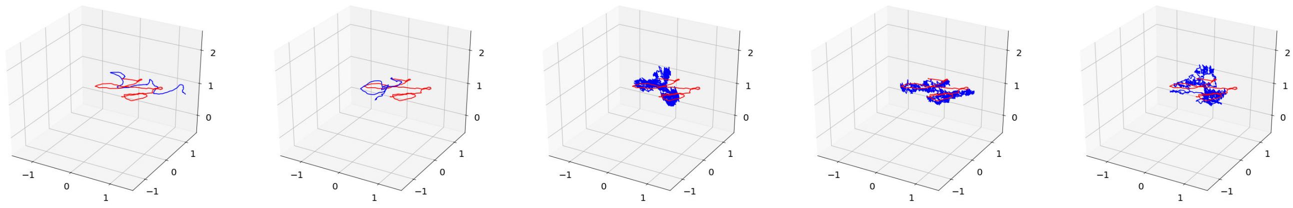
Comparing SelectFusion and the VI branch in Table 6, the VI branch reduce the global translation RMSE as expected. In addition, VI-UWB performs the best on global translation. However, our global rotation is worse than that of SelectFusion because our data contains more rotation. As shown in Table 5, in the real-world dataset, the numerical value of the angular velocity is greater than the numerical value of the average velocity, whereas in the EuRoC dataset, the opposite is true. For this reason, none of the methods predict rotation well on the real-world dataset. In this case, the model with a lower relative rotation error yields a lower global rotation error.

TABLE 5. Characteristics of each sequence in real-world dataset.

Name	Length / Duration	Average Velocity / Angular Velocity
Sequence_01	12.2 m 65 s	0.19 m/s 0.35 rad/s
Sequence_02	10.4 m 62 s	0.17 m/s 0.38 rad/s
Sequence_03	10.3 m 76 s	0.14 m/s 0.35 rad/s
Sequence_04	8.4 m 60 s	0.14 m/s 0.29 rad/s
Sequence_05	8.7 m 66 s	0.13 m/s 0.35 rad/s
Sequence_06	7.1 m 52 s	0.14 m/s 0.23 rad/s
Sequence_07	8.6 m 54 s	0.16 m/s 0.36 rad/s
Sequence_08	9.6 m 53 s	0.18 m/s 0.41 rad/s
Sequence_09	6.8 m 64 s	0.10 m/s 0.26 rad/s
Sequence_10	7.9 m 69 s	0.11 m/s 0.28 rad/s
Sequence_11	4.0 m 64 s	0.06 m/s 0.05 rad/s

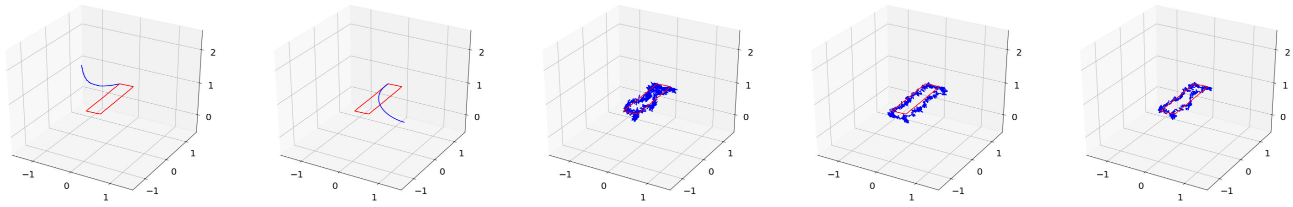
In *Sequence_11*, we find that the VI branch outperforms SelectFusion in terms of global RMSE due to the smaller drift in the Z-axis. However, SelectFusion and the VI branch both fail to match the ground-truth trajectory, as shown in Figures 8a and 8b. Utilizing UWB measurements solves this problem because UWB provides global information that can be referenced by the VI branch. Figures 8c, 8d, and 8e show the trajectories of the UWB branch, Noopploop LinkTrack S, and VI-UWB, respectively, all of which outperform the VI methods. To compare the trajectories of the methods more clearly, we also provide a three-view drawing of UWB branch, Nooploop LinkTrack S, and VI-UWB on *Sequence_11* in Figures 9, 10, and 11, respectively. From the figures and the values in Table 7, we find that adding UWB improves the performance of global pose estimation significantly. Although orientation is not observable in the UWB system, it helps the VI branch to learn and better predict global rotation.

Taking into account Tables 6 and 7, the results show that the proposed UWB branch yields accuracy similar to that of the Nooploop LinkTrack S method. At the same time, our VI-UWB method fuses the results of the UWB and VI branches to obtain the best global translation accuracy. However, in Figure 7, we find that the shape of Nooploop LinkTrack S's trajectory better matches the shape of the ground-truth trajectory. This is because Nooploop LinkTrack S's algorithm contains a filter to smooth the trajectories and UWB measurements, which results in an offset to match two trajectories. The offset in each trajectory is different, and may be related to the starting point of the trajectory or to environmental factors. This explains the poor RMSE of Nooploop LinkTrack S's global translation.



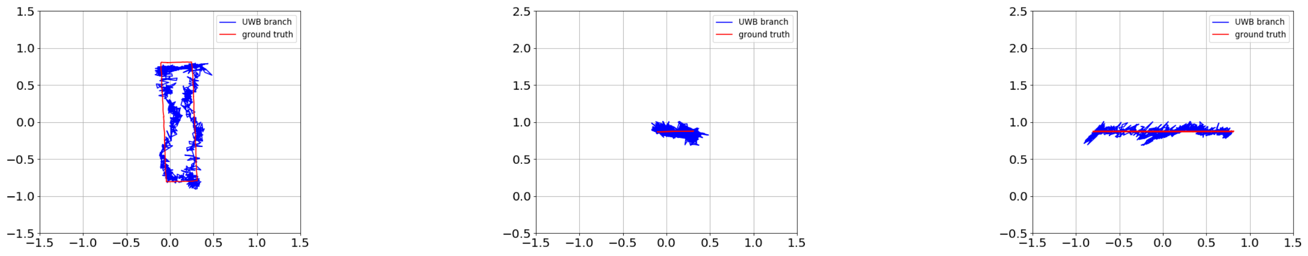
(a) SelectFusion (b) VI branch (c) UWB branch (d) Noopploop (e) VI-UWB

FIGURE 7. Trajectories of different methods on real-world dataset *Sequence_10*. Ground-truth and estimated trajectories are marked in red and blue, respectively.



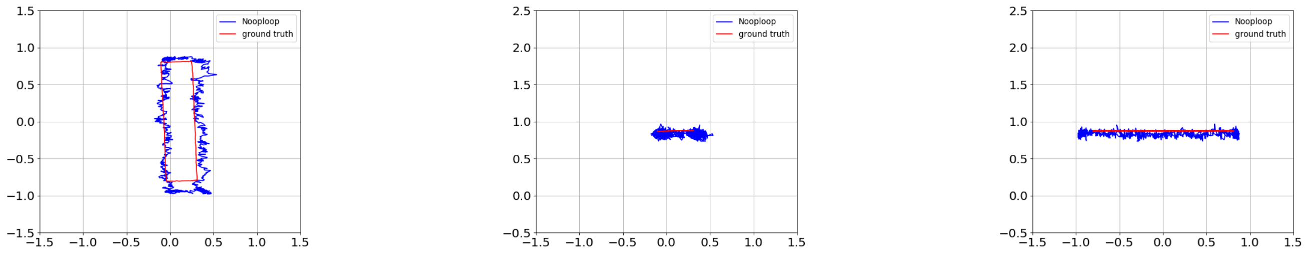
(a) SelectFusion (b) VI branch (c) UWB branch (d) Noopploop (e) VI-UWB

FIGURE 8. Trajectories of different methods on real-world dataset *Sequence_11*. Ground-truth and estimated trajectories are marked in red and blue, respectively.



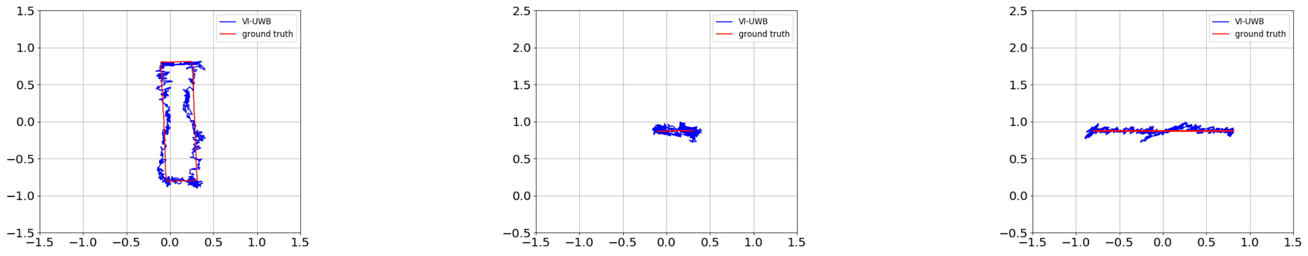
(a) top view (b) front view (c) side view

FIGURE 9. Top, front, and side views of UWB branch on real-world dataset *Sequence_11*.



(a) top view (b) front view (c) side view

FIGURE 10. Top, front, and side views of Noopploop LinkTrack S on real-world dataset *Sequence_11*.



(a) top view (b) front view (c) side view

FIGURE 11. Top, front, and side views of VI-UWB on real-world dataset *Sequence_11*.

TABLE 6. Results on real-world dataset *Sequence_10*. Bold numbers indicate the lowest error value.

	Relative 3D RMSE		Global 3D RMSE	
	Trans (m)	Rot (°)	Trans (m)	Rot (°)
SelectFusion	**0.004**	**0.16**	1.18	**23.05**
Proposed (VI branch)	**0.004**	0.45	0.73	38.75
Proposed (UWB branch)	0.072	N/A	0.45	N/A
Nooploop (IMU+UWB)	0.045	N/A	0.43	N/A
Proposed (VI-UWB)	0.035	0.37	**0.42**	33.43

TABLE 7. Results on real-world dataset *Sequence_11*. Bold numbers indicate the lowest error value.

	Relative 3D RMSE		Global 3D RMSE	
	Trans (m)	Rot (°)	Trans (m)	Rot (°)
SelectFusion	**0.004**	0.20	1.00	79.35
Proposed (VI branch)	**0.004**	0.15	0.78	69.68
Proposed (UWB branch)	0.067	N/A	0.12	N/A
Nooploop (IMU+UWB)	0.040	N/A	0.13	N/A
Proposed (VI-UWB)	0.028	**0.13**	**0.07**	**12.09**

In sum, VI-UWB, the proposed method, effectively leverages both the VI and UWB branches. The VI branch provides relative results, and the UWB branch has good global localization results and can perform relocalization for the VI branch. The results on the EuRoC and real-world datasets show that VI-UWB is the best method for global pose estimation, and exhibits the most consistent performance.

V. CONCLUSION

In this paper, we propose a learning-based method for vision, inertial, and UWB sensor fusion to improve global localization results. In the VI task, the proposed loss function restricts predicted positions from drifting too far from the ground truth. Also, the UWB model provides good global information for the proposed VI–UWB network. To train our models, we present a two-stage training process and multi-task loss function that combines translation and rotation loss. Experimental results on the EuRoC dataset and real-world experiments illustrate that our method achieves high global pose accuracy and outperforms other methods that use only UWB or VI with initialization.

REFERENCES

[1] M. O. A. Aqel, M. H. Marhaban, M. I. Saripan, and N. B. Ismail, "Review of visual odometry: Types, approaches, challenges, and applications," *SpringerPlus*, vol. 5, no. 1, pp. 1–26, Dec. 2016.

[2] K. Li Lim and T. Bräunl, "A review of visual odometry methods and its applications for autonomous driving," 2020, *arXiv:2009.09193*.

[3] H. Yan, Q. Shan, and Y. Furukawa, "RIDI: Robust IMU double integration," in *Computer Vision—ECCV 2018* (Lecture Notes in Computer Science), vol. 11217, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds. Munich, Germany: Springer, Sep. 2018, pp. 641–656.

[4] C. Chen, X. Lu, A. Markham, and N. Trigoni, "IONet: Learning to cure the curse of drift in inertial odometry," in *Proc. AAAI Conf. Artif. Intell.*, Apr. 2018, vol. 32, no. 1, pp. 1–9.

[5] S. Leutenegger, S. Lynen, M. Bosse, R. Siegwart, and P. Furgale, "Keyframe-based visual–inertial odometry using nonlinear optimization," *Int. J. Robot. Res.*, vol. 34, no. 3, pp. 314–334, 2015.

[6] P. Geneva, K. Eckenhoff, W. Lee, Y. Yang, and G. Huang, "OpenVINS: A research platform for visual-inertial estimation," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2020, pp. 4666–4672.

[7] M. Burri, J. Nikolic, P. Gohl, T. Schneider, J. Rehder, S. Omari, M. W. Achtelik, and R. Siegwart, "The EuRoC micro aerial vehicle datasets," *Int. J. Robot. Res.*, vol. 35, no. 10, pp. 1157–1163, Sep. 2016.

[8] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe, "Unsupervised learning of depth and ego-motion from video," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6612–6619.

[9] A. Kendall, M. Grimes, and R. Cipolla, "PoseNet: A convolutional network for real-time 6-DOF camera relocalization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2938–2946.

[10] S. Wang, R. Clark, H. Wen, and N. Trigoni, "DeepVO: Towards end-to-end visual odometry with deep recurrent convolutional neural networks," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2017, pp. 2043–2050.

[11] A. Ranjan, V. Jampani, L. Balles, K. Kim, D. Sun, J. Wulff, and M. J. Black, "Competitive collaboration: Joint unsupervised learning of depth, camera motion, optical flow and motion segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*. Los Alamitos, CA, USA: IEEE Computer Society, Jun. 2019, pp. 12232–12241.

[12] N. Yang, L. von Stumberg, R. Wang, and D. Cremers, "D3VO: Deep depth, deep pose and deep uncertainty for monocular visual odometry," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 1278–1289.

[13] L. Lin, W. Luo, Z. Yan, and W. Zhou, "Rigid-aware self-supervised GAN for camera ego-motion estimation," *Digit. Signal Process.*, vol. 126, Jun. 2022, Art. no. 103471.

[14] A. Poulose and D. S. Han, "UWB indoor localization using deep learning LSTM networks," *Appl. Sci.*, vol. 10, no. 18, p. 6290, Sep. 2020.

[15] D. T. A. Nguyen, H.-G. Lee, E.-R. Jeong, H. L. Lee, and J. Joung, "Deep learning-based localization for UWB systems," *Electronics*, vol. 9, no. 10, p. 1712, Oct. 2020.

[16] Y. Lu, J. Sheu, and Y. Kuo, "Deep learning for ultra-wideband indoor positioning," in *Proc. IEEE 32nd Annu. Int. Symp. Pers., Indoor Mobile Radio Commun. (PIMRC)*, Sep. 2021, pp. 1260–1266.

[17] P. Krapež, M. Vidmar, and M. Munih, "Distance measurements in UWB-radio localization systems corrected with a feedforward neural network model," *Sensors*, vol. 21, no. 7, p. 2294, Mar. 2021.

[18] L. Kneip, S. Weiss, and R. Siegwart, "Deterministic initialization of metric state estimation filters for loosely-coupled monocular vision-inertial systems," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Sep. 2011, pp. 2235–2241.

[19] A. R. J. Ruiz, F. S. Granja, J. C. P. Honorato, and J. I. G. Rosas, "Accurate pedestrian indoor navigation by tightly coupling foot-mounted IMU and RFID measurements," *IEEE Trans. Instrum. Meas.*, vol. 61, no. 1, pp. 178–189, Jan. 2012.

[20] C. Chen, H. Zhu, M. Li, and S. You, "A review of visual-inertial simultaneous localization and mapping from filtering-based and optimization-based perspectives," *Robotics*, vol. 7, no. 3, p. 45, Aug. 2018.

[21] B. Yang, J. Li, and H. Zhang, "Resilient indoor localization system based on UWB and visual–inertial sensors for complex environments," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–14, 2021.

[22] H. Xu, L. Wang, Y. Zhang, K. Qiu, and S. Shen, "Decentralized visual-inertial-UWB fusion for relative state estimation of aerial swarm," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2020, pp. 8776–8782.

[23] T. Baltrusaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 423–443, Feb. 2019.

[24] C. Debeunne and D. Vivet, "A review of visual-LiDAR fusion based simultaneous localization and mapping," *Sensors*, vol. 20, no. 7, p. 2068, Apr. 2020.

[25] C. Chen, S. Rosa, Y. Miao, C. X. Lu, W. Wu, A. Markham, and N. Trigoni, "Selective sensor fusion for neural visual-inertial odometry," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 10534–10543.

[26] R. Clark, S. Wang, H. Wen, A. Markham, and N. Trigoni, "VINet: Visual-inertial odometry as a sequence-to-sequence learning problem," in *Proc. AAAI Conf. Artif. Intell.*, Feb. 2017, vol. 31, no. 1, pp. 1–7.

[27] L. Han, Y. Lin, G. Du, and S. Lian, "DeepVIO: Self-supervised deep learning of monocular visual inertial odometry using 3D geometric constraints," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Nov. 2019, pp. 6906–6913.

[28] E. J. Shamwell, K. Lindgren, S. Leung, and W. D. Nothwang, "Unsupervised deep visual-inertial odometry with online error correction for RGB-D imagery," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 10, pp. 2478–2493, Oct. 2020.

[29] Y. Almalioglu, M. Turan, M. R. U. Saputra, P. P. B. de Gusmão, A. Markham, and N. Trigoni, "SelfVIO: Self-supervised deep monocular visual–inertial odometry and depth estimation," *Neural Netw.*, vol. 150, pp. 119–136, Jun. 2022.

[30] A. Dosovitskiy, P. Fischer, E. Ilg, P. Häusser, C. Hazirbas, V. Golkov, P. van der Smagt, D. Cremers, and T. Brox, "FlowNet: Learning optical flow with convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2758–2766.

[31] L. Liebel and M. Körner, "Auxiliary tasks in multi-task learning," 2018, *arXiv:1805.06334*.

[32] R. Cipolla, Y. Gal, and A. Kendall, "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7482–7491.

PENG-YUAN KAO received the B.S. degree in computer science and information engineering from the National Taiwan University of Science and Technology, in 2014, and the Ph.D. degree from the Graduate Institute of Networking and Multimedia, National Taiwan University, in 2023. His research interests include computer vision, deep learning, and multimedia.

HSIU-JUI CHANG received the bachelor's degree in computer science and information engineering and a minor in economics from National Taiwan University, in 2020, and the master's degree from the Image and Vision Laboratory (imLab), National Taiwan University, in 2022. He is currently a Software Engineer with Perfect Corporation. He works on developing innovative features for the company's miniprogram products.

KUAN-WEI TSENG (Student Member, IEEE) is currently pursuing the M.S. degree in artificial intelligence with the Department of Computer Science, Tokyo Institute of Technology. Prior to his graduate study, he was a full-time Research Associate with the AI Applications and Integration Laboratory (AI^2Lab) and the Image and Vision Laboratory (imLab), National Taiwan University. His research interests include visual computing, focusing on 3D computer vision, deep learning for computer vision, and their applications to augmented and virtual reality. He is also a Student Member of ACM.

TIMOTHY CHEN received the B.S. degree in CS from National Taiwan University, where he is currently pursuing the master's degree majoring in computer science and information engineering. His research interests include computer vision, augmented reality, and blockchain technologies.

HE-LIN LUO received the master's degree from the Graduate School of Arts and Technology, Taipei National University of the Arts, and the Ph.D. degree from the Graduate Institute of Networking and Multimedia, National Taiwan University. He specializes in creating interdisciplinary works using art and technology. For his creations, he draws from his personal experience of being extremely addicted to online games during middle and upper school to explore "the power of virtual worlds," "the thrill of speed," and other variations during this era of technology. Furthermore, his works are centered around the concept of "immigrant illness" amidst this generation of digital immigrants. His works have been recognized at many electronic and contemporary art festivals both domestically and internationally. These include the Digital Arts Award Taipei, in 2008, 2010, 2011, and 2015, FILE—Electronic Language International Festival, in 2009, 2010, and 2013, and other competitions. He has also participated in many major international exhibitions, such as SIGGRAPH, SIGGRAPH ASIA, FILERIO, PIXILERATIONS (v.8), Asian Students and Young Artists Art Festival (ASYAAF), ACM Multimedia Art Exhibition, Ars Electronica Festival, in 2014 and 2017, Click Festival in Denmark, and other exhibitions, and biennales, such as 2015 WRO Media Art Biennale, 2017 Asian Art Biennial, and more. He is the first Taiwanese artist to create works using a four-axis drone. His works have won first prize for the Performance Award at Digital Art Festival Taipei. He was also specially invited to Ars Electronica Festival to present his inter-disciplinary works made through drones.

YI-PING HUNG (Member, IEEE) received the B.S. degree in EE from National Taiwan University, in 1982, and the Ph.D. degree from Brown University, in 1990.

Then, he was with the Institute of Information Science, Academia Sinica, for 12 years, while he was an Adjunct Professor with the Department of Computer Science and Information Engineering, National Taiwan University. During this period, he was the Deputy Director of the Institute of Information Science and awarded the Young Researcher Award from Academia Sinica, in 1997. He moved to the Department of Computer Science and Information Engineering, National Taiwan University, in 2002, as a full-time Professor. He was the Director of the Graduate Institute of Networking and Multimedia, from 2007 to 2013. From February 2017 to July 2020, he was with the Tainan National University of the Arts as the Dean of Research and Development and then as the Provost. He was also a Distinguished Professor with the Institute of Animation and Film Art. He was the Founding President of the Taiwan Society of Human-Computer Interaction, from 2016 to 2018. He is currently the President of the Taiwan Art and Technology Association. He has also served as the Executive Director for the Chinese Society Image Processing and Pattern Recognition. His research interests include computer vision, image processing, pattern recognition, and interactive multimedia. His current research interests include virtual reality, augmented reality, non-fungible tokens, and their applications in the metaverse, digital art, and in culture and museums.