Federated Meta Learning for Visual Navigation in GPS-denied Urban Airspace

Burak Yuksek[1], Zhengxin Yu[2], Neeraj Suri[2], Gokhan Inalhan[1]
[1]School of Aerospace, Transport and Manufacturing, Cranfield University, United Kingdom
[2]School of Computing and Communications, Lancaster University, United Kingdom
Email: {Burak.Yuksek, Inalhan} @cranfield.ac.uk {z.yu8, neeraj.suri} @lancaster.ac.uk

***Abstract*—Urban air mobility (UAM) is one of the most critical research areas which combines vehicle technology, infrastructure, communication, and air traffic management topics within its identical and novel requirement set. Navigation system requirements have become much more important to perform safe operations in urban environments in which these systems are vulnerable to cyber-attacks. Although the global navigation satellite system (GNSS) is a state-of-the-art solution to obtain position, navigation, and timing (PNT) information, it is necessary to design a redundant and GNSS-independent navigation system to support the localization process in GNSS-denied conditions. Recently, Artificial intelligence (AI)-based visual navigation solutions are widely used because of their robustness against challenging conditions such as low-texture and low-illumination situations. However, they have weak adaptability to new environments if the size of the dataset is not sufficient to train and validate the system. To address these problems, federated meta learning can help fast adaptation to new operation conditions with small dataset, but different visual sensor characteristics and adversarial attacks add considerable complexity in utilizing federated meta learning for navigation. Therefore, we proposed a robust-by-design Federated Meta Learning based visual odometry algorithm to improve pose estimation accuracy, dynamically adapt to various environments by using differentiable meta models and tunning its architecture to defense against cyber-attacks on the image data. In this proposed method, multiple learning loops (inner-loop and outer-loop) are dynamically generated. Each vehicle utilizes its collected visual data in different flight conditions to train its own neural network locally for a particular condition in the inner loops. Then, vehicles collaboratively train a global model in the outer loop which has generalizability across heterogeneous vehicles to enable lifelong learning. The inner loop is used to train a task-specific model based on local data, and the outer loop is to extract common features from similar tasks and optimize meta-model adaptability of similar tasks in navigation. Moreover, a detection model is designed by utilizing key characteristics in trained neural network model parameters to identify attacks.**

***Index Terms*—Federated Learning, Meta Learning, Visual Odometry, Navigation**

I. Introduction

Utilization of urban air mobility (UAM) operations has been increasing in recent years due to improvements in vehicle technology, infrastructure and autonomy. All of these improvements constitutes a basis for reliable, safe, effective and sustainable operations even in challenging conditions in urban canyons.

One of the key technologies for safe and effective UAM applications is positioning, navigation and timing (PNT) solutions to obtain the current vehicle location within an operational environment with minimal error. Currently, GNSS is state-of-the-art technology to obtain PNT information of the vehicle, however it is quite vulnerable to multi-path error within urban canyons and cyber-attacks such as GNSS-spoofing and GNSS-jamming. Such attacks may result in catastrophic accidents if there is no precaution against these situations. Visual Odometry (VO) is one of the most effective techniques to complement the GNSS in which image sequences are utilised to estimate the camera pose in the presence of these abovementioned error sources. Moreover, it is suitable to be integrated with different kind of sensors such as inertial measurement unit (IMU) and light detection and ranging (LIDAR) to mitigate the estimation error.

VO methods are quite promising for manned and unmanned applications in indoor and GNSS-denied outdoor environments. Over the past 35 years, many researchers have studied these methods to improve its estimation performance against challenging operational conditions such as textureless environments, low-illumination situations and motion blur due to fast movement. To handle these problems, researchers have focused on two approaches; a) Geometry-based VO, and b) Learning-based VO. Geometry based VO applications, which are further divided into two sub-categories as feature-based methods and direct methods. This class of approaches are well-defined and recognized as benchmarks for further improvements. However, to obtain a superior performance, they require significant effort to calibrate and fine-tune the system parameters. In addition, scale ambiguity is a fundamental problem in monocular VO as the depth of the scene cannot be calculated directly and this results in drift in pose estimation.

In the last decade, as a result of the improvements in computation power, Artificial Intelligence (AI) has been successfully applied for many computer vision tasks such as object detection, segmentation, tracking, etc. This also has enabled the application of AI, especially Deep Learning (DL), on VO problems and prompted the researchers to create suitable neural-network (NN) structures for VO applications. It has been shown that DL-based VO algorithms are able to learn effective and robust feature representation by using large datasets without extensive tuning requirements unlike in the geometry-based VO applications. This makes the DL one of the most effective candidate to solve VO problems in challenging conditions [1].

A multitude of methods have been proposed to address nav-

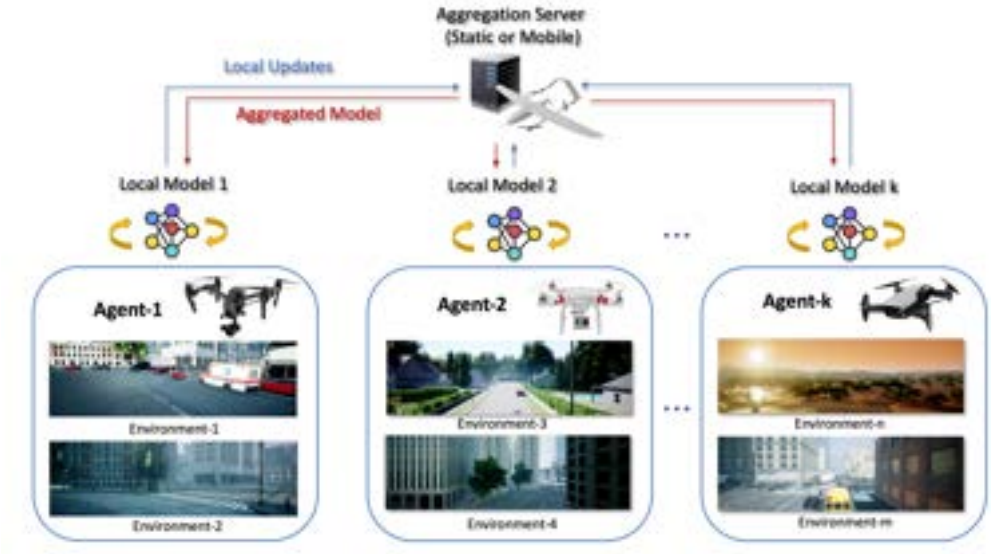

Fig. 1. General overview of the FL-VO.

igation problems in a learning-enabled framework. However, these learning-based methods are not able to rapidly adapt to new environments and/or conditions. Meta learning [2] is a promising method to address the aforementioned issues by leveraging previous experiences across a range of learning tasks to significantly accelerate learning of new tasks.

In this study, we propose a Federated meta Learning-based Visual Odometry (FLVO) to improve pose estimation accuracy and rapidly adapt to various environments by using differentiable models. The proposed method learns multiple meta models for different groups of vehicles and rapidly obtains an effective model for each vehicle based on the meta model and local data, which also combined with Federated Learning (FL). In this way, the training efficiency in learning new tasks can be improved and the navigation algorithm becomes more adaptive to the dynamic environment. Moreover, the proposed method can leverage resources from different groups of vehicles. More specifically, the proposed FLVO conducts two learning loops: outer loop and inner loop. The outer loop is to train for the meta model which runs on the FL server that can be a roadside unit, a base station or a moving vehicle. It also uses its experiences over many task contexts to gradually adjust parameters of the meta model that governs the operation of an inner loop. The inner loop is processed on vehicles, which trains for the specific environment and condition. Normally, the inner loop can adapt faster to new tasks through several gradient updates with a small amount of sampling data [3].

The major contributions of this paper can be summarized as follows:

- FLVO has high sample efficiency towards new learning tasks, thus it enables vehicles to have the lifelong learning capability and run the training process by using their own data even with limited computation resources.
- FLVO dynamically generates multiple learning loops to train meta visual odometry model which can speed up the whole training process and improve pose estimation accuracy.
- An robust-by-design federated meta learning framework is proposed to achieve fast adaption to new environments and conditions. Adversarial attacks can be detected by learning multiple meta models and optimize a global meta model generalizability across vehicles that operate in different environments.

The rest of the paper is organized as follows. A brief introduction to navigation, FL and meta Learning is given in Section II. The problem formulation for navigation is presented in Section III. The details of the proposed FLVO are described in Section IV. Section V-C provides evaluation results and performance analysis. Obtained results of FLVO are discussed in Section VI. Finally, Section VII gives concluding remarks and future works.

II. Related Works

As mentioned before, visual navigation has been studied for longer than 3 decades and significant outcomes have been obtained so far. From a general point of view, visual navigation solutions in the literature could be divided into two sub-groups as; a) Geometry-based methods and, b) Learning-based methods. In order to have a clear overview about each group and to emphasize the differencies of these approaches, a literature survey is given in this section.

A. Geometry-based Visual Odometry

Over the past decades, significant developments and improvements have been demonstrated in both feature-based and appearance-based methods. In [4], ORB-SLAM framework is introduced which is a feature-based monocular simultaneous localization and mapping (SLAM) algorithm that operates in

real-time in both indoor and outdoor environments. As an extension study of [4], ORM-SLAM3 is given in [5] which contains an open-source library for visual, visual-inertial and multimap SLAM applications and suitable for monocular, stereo, RGB-D cameras. Appearance-based approaches observe changes in intensity of pixel information instead of performing feature detection and matching processes. So, these methods skip the pre-computation step and directly use the pixel intensity values received directly from the environment. In [6], a direct monocular SLAM algorithm is proposed which allows to build large scale and consistent maps. This study is extended for stereo and omni-directional cameras in [7], [8]. In [9], direct sparse odometry (DSO) is proposed which combines benefits of direct methods with the flexibility of sparse approaches.

B. Learning-based Navigation

When comparing to the geometry-based visual odometry approaches, DL-based methods have promising results as they are able to learn more effective and robust feature representation when sufficient size of dataset is provided. In addition, they don't require manual tuning and calibration process which makes them more applicable in real-world conditions. Although their pose estimation accuracy is not as high as geometry-based approaches, they could learn further to adapt to new environments thanks to their learning-enabled structure.

In [10], end-to-end, sequence-to-sequence probabilistic visual odometry (ESP-VO) is proposed for monocular camera setting based on deep recurrent convolutional neural networks (RCNN). In [11], Deep Depth, Deep Pose, and Deep Uncertainty for Visual Odometry (D3VO) method is proposed for monocular camera setting. This framework utilizes deep neural networks on three levels for depth (with DepthNet [12]), pose (with PoseNet [13]) and uncertainty estimation purposes. In [14], DeepVO application is presented which is an end-to-end framework applied for monocular VO and utilizes Recurrent Convolutional Neural Network (RCNN) structure. Main aim of the DeepVO is not only learning the feature representations on the images by utilising Convolutional Neural Networks (CNNs), but also modeling the sequential dynamics and relations with the help of Recurrent Neural Networks (RNNs) [14].

C. Federated Learning Applications in Navigation

Privacy and security issues have been become important issues as the data size and number of users are increased. Hence, several learning-based navigation approaches utilized Federated Learning (FL) to reduce privacy and security risks. McMahan *et al.* [15] first introduced FL which enables multiple vehicles collaboratively train a global model without sharing vehicles' local data. Kong *et al.* [16] proposed a privacy-preserving aggregation for FL based navigation. Liu *et al.* [17] developed a lifelong federated reinforcement learning architecture for navigation by using a knowledge fusion algorithm and transfer learning. However, these learning-based methods are not able to fast adapt to new environments, because they need full retraining to learn updated model for new environments, which is time-consuming. Furthermore, they cannot handle heterogeneous vehicles that contain varied data distributions.

In recent years, federated learning has been used for odometry applications on autonomous vehicles because of its unique properties such as minimum data sharing requirements and user privacy. In [18], it is proposed a learning-based cooperative SLAM (FC-SLAM) approach for cloud robotic applications to improve the performance of visual-LIDAR SLAM. To do so, a federated deep learning algorithm is developed for feature extraction and dynamic vocabulary designation. It is shown that the proposed method has better feature extraction perforance when compared to SIFT and ORB under different illumination and viewpoint conditions. In [19], a reconfigurable holographic surface (RHS)-aided SLAM application is developed in which federated learning method is utilised to enhance SLAM performance. Here, a multi-vehicle SLAM protocol is proposed to regulate data processing across multiple vehicles. Simulation results indicate that the localization error is reduced when compared to non-cooperative training schemes with the same hardware cost and radiation power by utilising the federated learning architecture.

III. System Model and Design Goals

In this section, we describe the assumptions, system model and identify design goals.

A. System model

We design a federated meta learning based visual navigation framework for UAVs, as shown in Fig.1. Specifically, the proposed FLVO utilizes a monocular camera to collect data to train a neural network based visual odometry navigation model. For ease of implementation, development and evaluation process of the FLVO, following assumption is made which is about the operation conditions;

Assumption 1: Monocular cameras are mounted on the aerial vehicles horizontally. Aerial vehicles operate at low and fixed-altitude and low-speed flight conditions.

Our proposed framework is based on a peer-to-peer FL framework, which consists of a lead vehicle and a group of vehicles which are defined below.

- *Lead vehicle (Aggregation Server)*: The vehicle that has sufficient computation and communication capacity can be selected as a lead vehicle. The lead vehicle is responsible for updating a navigation model via T training iterations. At the beginning of iteration t, the lead vehicle broadcasts the navigation model to all participating vehicles. At the end of iteration t, the outer loop is executed in the lead vehicle to conduct a global meta navigation model by aggregating model updates from inner loop.
- *Vehicles*: At iteration t, each participating vehicle first uses a camera to collect data in its local storage. Secondly, each vehicle initializes its local navigation by using the received model from the lead vehicle. Thirdly, each

vehicle utilizes its local data to obtain the updated navigation model by using Stochastic Gradient Descent (SGD). Position and attitude data are exploited to calibrate the predicted output. After the training process is completed, all participating vehicles upload all their updated model parameters to the lead vehicle.

Training an optimal navigation model is performed by multiple communication rounds and each communication round r consists of the above steps. Communication rounds are repeated until an optimal navigation model achieved at the lead vehicle. Here, the assumption about the communication system is given below.

Assumption 2: The connections between the lead vehicle and participating vehicles are via wireless links with low transmission delay and high bandwidth.

B. Design Goals

The goal of the proposed FLVO framework is to design a robust-by-design and dynamic FL framework for navigation to improve pose estimation accuracy, handle heterogeneous vehicles and rapidly adapt to new environments.

- *Robustness*: Navigation system can be vulnerable to various kinds of attacks, such as pixel attacks on images and GPS spoofing. These attacks can corrupt the navigation model training process. The proposed FLVO can detect these attacks and remove them from model aggregation in the lead vehicle to resist such attacks and improve navigation system robustness.
- *Adaptation*: Vehicles operate in different environments (i.e. city area, rural area, etc) and conditions (i.e. sunny, rainy, cloudy, etc.). These environments are frequently changing and the FLVO aims to fast adapt to new environments by using multiple learning loops. Different learning loops train different models that can be applied to different environments. Moreover, FL framework supports flexible joining and leaving of participating vehicles. The proposed FLVO also need to be adaptive to this situation.
- *Accuracy & Efficiency*: Vehicles have varied data distributions. Training a navigation model over these data distribution is difficult, which needs more training steps and training samples. FLVO can extract common features from multiple vehicles to improve model accuracy and efficiency.

IV. A robust-by-design Federated Meta Learning Based Navigation Solution

In this section, the proposed federated meta learning based visual odometry navigation is introduced in details. It consists of two parts: a) federated meta learning framework and, b) visual odometry.

A. End-to-end Visual Odometry Architecture

In this study, each vehicle performs DeepVO framework which has a Recurrent Convolutional Neural Network (RCNN) structure to estimate the pose of the vehicle by using monocular camera image sequences. Here, a tensor is created by stacking two consecutive images and fed into the RCNN structure. The RCNN structure contains a CNN for feature extraction and RNN for sequential learning. This is one of the most important aspect of the RCNN, as it allows simultaneous feature extraction and sequential modeling of the visual odometry by using CNN and RNN structures. For more information about the RCNN architecture and its application in VO, readers may refer to [14].

The main aim of the end-to-end visual odometry is to compute the conditional probability of the vehicle poses $\mathbf{Y} = \{y_0, y_1, ..., y_t\}$ for a given sequence of camera images $\mathbf{X} = \{x_0, x_1, ..., x_t\}$ in the time interval $T = [0, 1, ..., t]$. Mathematical representation of this is given in Eq. (1).

$$p(\mathbf{Y}_t|\mathbf{X}_t) = p(y_0, y_1, ..., y_t|x_0, x_1, ..., x_t) \tag{1}$$

During the training process of the neural network structure, optimal network parameters θ^* are calculated to minimize the Euclidean distance between the ground truth pose $(\mathbf{p}_k, \psi_k)$ and estimated pose $(\hat{\mathbf{p}}_k, \hat{\psi}_k)$ as shown in Eq. (2).

$$\theta^* = \arg\min_{\theta} \frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{t} \|\hat{\mathbf{p}}_k - \mathbf{p}_k\|_2^2 + \kappa \|\hat{\psi}_k - \psi_k\|_2^2 \tag{2}$$

where $\|.\|$ is 2-norm, κ is a scale factor to provide balance between position and orientation errors, and N is number of samples [14].

B. Federated meta learning for navigation

We combine FL with meta learning in the proposed FLVO. It conducts two learning loops (inner loop and outer loop) to train the navigation model. The outer loop is used to generate meta model by extracting common features from inner loops, while inner loop is based on the meta model from the outer loop to fast train a personalized model [20].

Inner loop: Each vehicle uses its local dataset $\mathcal{D}_i$ to train an navigation model $f(\theta_i, \mathcal{D}_i)$ parameterized by θ [21]. The target function of the inner loop is to minimize loss function $\mathcal{L}_{in}$.

$$\theta_i^* = \arg\min_{\theta_i \in R} \mathcal{L}_{in}(\mathcal{D}_i; \theta_i, w), \tag{3}$$

where θ^* is the updated local model. w is the parameters of meta model in the outer loop. The loss function $\mathcal{L}_{in}$ is defined as [22]:

$$\mathcal{L}_{in} = f_i(\theta_i) + \frac{\lambda}{2} \|\theta_i - w\|^2, \tag{4}$$

where λ is the weight of w to the trained model.

FLVO optimizes this loss through SGD [23] to achieve θ_i^*:

$$\begin{aligned} \theta_i^t &= \theta_i^{t-1} - \alpha \nabla \mathcal{L}_{in}\left(\theta_i^{t-1}\right), \\ \text{s.t. } \theta_i^0 &= w, t = 1, 2, 3, ...T, \end{aligned} \tag{5}$$

where t is the current iteration in FL training and the total number of iterations is T. θ_i^0 is the initial inner loop model in the first iteration.

Outer loop: The outer loop is to conduct a meta-model w that governs the operation of an inner loop. The loss function of the outer loop is as follows:

$$\mathcal{L}_{ou} = \frac{1}{N}\sum_{i=1}^{N} F_i(w), \quad \text{where } F_i(w) = \min_{\theta_i \in R}\left\{ f_i(\theta_i) + \frac{\lambda}{2}\|\theta_i - w\|^2 \right\}. \tag{6}$$

The outer loop to update the meta-model is as:

$$w^t = (1-\beta)\, w^{t-1} + \beta \sum_{i=1}^{N} \frac{\theta_i^{t-1}}{|\mathcal{D}_i|}, \tag{7}$$

where β controls the weight of w from the last iteration. The data size of the vehicle also influences w in the new iteration.

Detection model: We utilize a Variational AutoEncoder (VAE) based detection model to identify attacks in the navigation system and remove them from the FL training process. VAE, an unsupervised learning model, copies its inputs to outputs by extracting features from low-dimensional embeddings in the latent space. In VAE, essential features from inputs can be kept, and irrelevant and noisy features are removed [24], since attackers have large reconstruction errors. During the FL training process, we compute a dynamic detection threshold (the average value of all reconstruction errors) to detect attacks in each communication round. Compared with this threshold, the model updates with larger reconstruction errors are identified as attackers. These attackers are removed from the outer loops.

V. Training and Experimentation Results

A. Dataset and Simulation Environment

In supervised learning applications, it is crucial to have a large and labeled dataset to train and test the agent properly. However, there is a limited number of data within the open literature that is collected from an aerial vehicle for learning-based visual odometry purposes. Hence, in this study, a RCNN model is typically initially trained for ground vehicle applications by using KITTI dataset [25], and then transferred to the aerial vehicle application and re-trained by using the synthetic data.

As a synthetic environment, AirSim [26] is used which is an open-source simulator for drones and cars. It contains sensor models such as RGB cameras, LIDAR, inertial measurement unit (IMU), barometer, GPS and magnetometer. It is possible to simulate different weather and environmental conditions (wind, rain, road wetness, leaves, etc.). Day time can also be set which affects the light conditions and, of course, pose estimation performance. In addition, it is possible to model the noise on the RGB camera image which could be considered to model the cyber-attacks on camera sensors.

TABLE I
Agents and Data Sequences

Agents	Training Sequences	Validation Sequences
Agent-1	K00, K02, K08, K09	K03, K05
Agent-2	A23, A27, A29, A30	A24, A28

TABLE II
Hyperparameters for Training

Parameters	Value
Epochs	80
Batch Size	12
Learning Rate (LR)	0.001
Decay Rate	5e-6
LR Decay Factor	0.1
Sequence Length	7
Image Overlap	1

Agents and data sequences that are used for the training phase are given in Table I. Here, "K" and "A" stand for the sequence from KITTI dataset and AirSim, respectively.

B. Training of the RCNN Structure

Training process of the RCNN for KITTI and AirSim datasets are performed by using the hyperparameters given in Table II.

Training and validation loss values for KITTI and AirSim environments are given in Fig.(2).

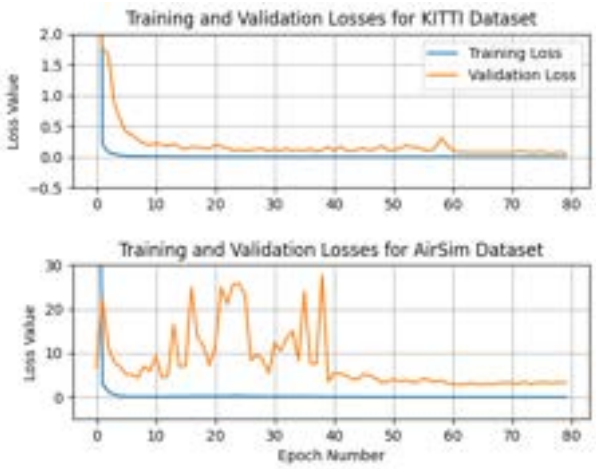

Fig. 2. Train and validation loss for KITTI and AirSim datasets.

C. Performance Evaluation

To evaluate the effeciveness of our federated learning framework on visual odometry applications, several aggregated agents are created by using different scaling factors k_n as shown in Eq. (8).

$$k_n = c\,\frac{\textit{User-1 \# of samples}}{\textit{User-2 \# of samples}} \tag{8}$$

TABLE III
EXPERIMENT RESULTS OF AGENTS IN DIFFERENT ENVIRONMENTS.

	Env-1 (K03)		Env-2 (K04)		Env-3 (A22)		Env-4 (A31)	
	t_{rel}	r_{rel}	t_{rel}	r_{rel}	t_{rel}	r_{rel}	t_{rel}	r_{rel}
Agent-1	5.44	0.035	7.95	0.024	85.95	0.6453	89.99	0.473
Agent-2	107.93	0.5817	107.89	0.1146	5.94	0.038	5.32	0.038
Agent-3	77.23	0.403	87.17	0.201	80.59	0.796	78.85	0.438
Agent-4	69.02	0.453	75.27	0.299	69.02	0.853	80.71	0.457
Agent-5	37.84	0.235	48.23	0.184	63.94	0.787	87.76	0.450
Agent-6	20.71	0.101	24.32	0.09	60.01	0.761	95.98	0.460

where c is a scale factor to adjust the overall weight of the user $n \in \{1, 2\}$. Here, c is selected as $c = \{1, 2, 3, 5\}$ for analysis purposes and this means relying more on the User-1 dataset while creating the aggregated visual navigation agent.

After defining the federation policy of the algorithm, the proposed FLVO framework is applied to obtain aggregated visual odometry agents with different data scaling factors, c. Description of the agents that are trained and tested in this study are described below.

- Agent-1: VO agent trained by utilising the KITTI dataset.
- Agent-2: VO agent trained by utilsing the synthetic AirSim dataset.
- Agent-3: Aggregated (Agent-1 + Agent-2) with $c = 1$.
- Agent-4: Aggregated (Agent-1 + Agent-2) with $c = 2$.
- Agent-5: Aggregated (Agent-1 + Agent-2) with $c = 3$.
- Agent-6: Aggregated (Agent-1 + Agent-2) with $c = 5$.

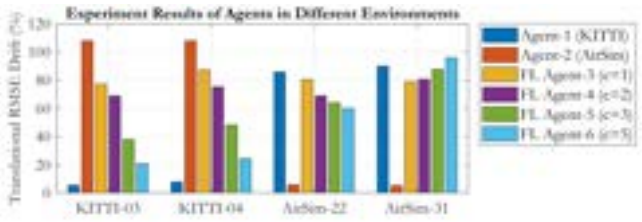

Fig. 3. Experiment results of agents in different environments.

VI. DISCUSSION

To evaluate the pose estimation performance of each agent, translational root mean square error (RMSE) drift percentage and rotational RMSE drift ($deg/100m$) are used as evaluation metrics. Results are given in Table III where t_{rel} is translational RMSE drift percentage and r_{rel} is rotational RMSE drift in $deg/100m$. These results are also represented graphically in Fig. 3.

- Agent-1, which is trained by utilizing KITTI dataset, has better performance in KITTI sequences 03 and 04, as expected. However, in AirSim-22 and AirSim-31 sequences, its translational RMSE drift is over 80%. A similar situation can be observed for Agent-2 which is trained by using AirSim dataset (i.e. better results in AirSim sequences but increased RMSE drift in KITTI sequences). The main reason of this result is operating the agent in different environments other than it is trained and validated.
- The proposed FLVO framework reduces translational RMSE drift in both KITTI-03, 04 and AirSim-22 environments. For example, Agent-2 is trained in AirSim environment and tested in KITTI-03 and KITTI-04 sequences. Its translational RMSE drift is higher than 100% in KITTI environment because any image data from KITTI dataset is not used in its training phase of Agent-2. However, as proposed in the FLVO framework, if the Agent-2 is aggregated with the Agent-1 that is trained by using KITTI dataset, aggregated agents (Agents 3, 4, 5, 6) have lower translational RMSE drift in the KITTI environment as a result of federation process. Similar situation could be observed in AirSim sequence 22. Agent-1 has higher translational RMSE drift in this sequence. However, pose estimation error is reduced by using the proposed FLVO framework as shown in Agents 3, 4, 5 and 6 in AirSim-22 sequence in Fig. 3.
- In the environment AirSim-31, translational RMSE drift of the FLVO Agent-3 is lower than that of the Agent-1. However, Agents 4, 5, and 6 have higher translational RMSE drift in this environment. This might be as a result of lack of sufficient number data collected from synthetic AirSim environment.
- In rotational RMSE drift results in Table III, it is shown that FLVO can reduce angular pose estimation error in KITTI-03, KITTI-04 and AirSim-22 environments (see aggregated agents). In AirSim-31 environment, although there is not an improvement in angular RMSE drift, it is observed that the angular RMSE drift remains constant for the aggregated agents 3, 4, 5, and 6.
- Even though a decreasing trend is observed in translational RMSE drift by adjusting scaling factor c, it should be further reduced by increasing the number of agents and the size of the dataset. This helps to increase robustness and adaptation capability of the FLVO in various conditions.

VII. CONCLUSION

In this paper, we have proposed a novel FLVO framework which can improve pose estimation accuracy in terms of translational and rotational RMSE drift while reducing security and privacy risks. It also enables fast adaptation to new conditions thanks to the aggregation process of the local agents which operate in different environments.

In addition, we have shown that it is possible to transfer an end-to-end visual odometry agent that is trained by using ground vehicle dataset (i.e. KITTI dataset) to an aerial vehicle pose estimation problem for low-altitude and low-speed operating conditions.

Dataset size is an important topic that should be considered in both AI-based end-to-end visual odometry applications and federated learning approaches. Although it is demonstrated that federated learning could be applied for visual odometry applications to aggregate the agents that are trained in different

environments, more data should be collected to improve the translational and rotational pose estimation performance of the aggregated agents.

In our future work, we will evaluate cyber-attack detection performance of the proposed FLVO framework by utilizing multiple learning loops. In addition, dataset size will be expanded by utilizing real flight tests to increase the realm of the training data and to improve the robustness of the proposed federated learning based end-to-end visual odometry algorithm.

ACKNOWLEDGMENTS

Research supported by the UKRI Trustworthy Autonomous Systems Node in Security/EPSRC Grant EP/V026763/1.

REFERENCES

[1] K. Wang, S. Ma, J. Chen, F. Ren, and J. Lu, "Approaches, challenges, and applications for deep visual odometry: Toward complicated and emerging areas," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 14, no. 1, pp. 35–49, 2020.

[2] T. Hospedales, A. Antoniou, P. Micaelli, and A. Storkey, "Meta-learning in neural networks: A survey," *arXiv preprint arXiv:2004.05439*, 2020.

[3] M. Botvinick, S. Ritter, J. X. Wang, Z. Kurth-Nelson, C. Blundell, and D. Hassabis, "Reinforcement learning, fast and slow," *Trends in cognitive sciences*, vol. 23, no. 5, pp. 408–422, 2019.

[4] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardos, "Orb-slam: a versatile and accurate monocular slam system," *IEEE transactions on robotics*, vol. 31, no. 5, pp. 1147–1163, 2015.

[5] C. Campos, R. Elvira, J. J. G. Rodríguez, J. M. Montiel, and J. D. Tardós, "Orb-slam3: An accurate open-source library for visual, visual–inertial, and multimap slam," *IEEE Transactions on Robotics*, vol. 37, no. 6, pp. 1874–1890, 2021.

[6] J. Engel, T. Schöps, and D. Cremers, "Lsd-slam: Large-scale direct monocular slam," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part II 13*. Springer, 2014, pp. 834–849.

[7] J. Engel, J. Stückler, and D. Cremers, "Large-scale direct slam with stereo cameras," in *2015 IEEE/RSJ international conference on intelligent robots and systems (IROS)*. IEEE, 2015, pp. 1935–1942.

[8] D. Caruso, J. Engel, and D. Cremers, "Large-scale direct slam for omnidirectional cameras," in *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2015, pp. 141–148.

[9] J. Engel, V. Koltun, and D. Cremers, "Direct sparse odometry," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 3, pp. 611–625, 2017.

[10] S. Wang, R. Clark, H. Wen, and N. Trigoni, "End-to-end, sequence-to-sequence probabilistic visual odometry through deep neural networks," *The International Journal of Robotics Research*, vol. 37, no. 4-5, pp. 513–542, 2018.

[11] N. Yang, L. v. Stumberg, R. Wang, and D. Cremers, "D3vo: Deep depth, deep pose and deep uncertainty for monocular visual odometry," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 1281–1292.

[12] A. CS Kumar, S. M. Bhandarkar, and M. Prasad, "Depthnet: A recurrent neural network architecture for monocular depth prediction," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 283–291.

[13] A. Kendall, M. Grimes, and R. Cipolla, "Posenet: A convolutional network for real-time 6-dof camera relocalization," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2938–2946.

[14] S. Wang, R. Clark, H. Wen, and N. 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.

[15] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proceedings of the International Conference on Artificial intelligence and statistics (AISTATS)*. PMLR, 2017, pp. 1273–1282.

[16] Q. Kong, F. Yin, R. Lu, B. Li, X. Wang, S. Cui, and P. Zhang, "Privacy-preserving aggregation for federated learning-based navigation in vehicular fog," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 12, pp. 8453–8463, 2021.

[17] B. Liu, L. Wang, and M. Liu, "Lifelong federated reinforcement learning: a learning architecture for navigation in cloud robotic systems," *IEEE Robotics and Automation Letters*, vol. 4, no. 4, pp. 4555–4562, 2019.

[18] Z. Li, L. Wang, L. Jiang, and C.-Z. Xu, "Fc-slam: Federated learning enhanced distributed visual-lidar slam in cloud robotic system," in *2019 IEEE International Conference on Robotics and Biomimetics (ROBIO)*. IEEE, 2019, pp. 1995–2000.

[19] H. Zhang, Z. Yang, Y. Tian, H. Zhang, B. Di, and L. Song, "Reconfigurable holographic surface aided collaborative wireless slam using federated learning for autonomous driving," *IEEE Transactions on Intelligent Vehicles*, 2023.

[20] Z. Yu, Y. Lu, P. Angelov, and N. Suri, "Ppfm: An adaptive and hierarchical peer-to-peer federated meta-learning framework," in *2022 18th International Conference on Mobility, Sensing and Networking (MSN)*, 2022, pp. 502–509.

[21] T. Hospedales *et al.*, "Meta-learning in neural networks: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[22] C. T Dinh *et al.*, "Personalized federated learning with moreau envelopes," *Advances in Neural Information Processing Systems*, vol. 33, pp. 21 394–21 405, 2020.

[23] L. Bottou, "Stochastic gradient descent tricks," in *Neural Networks: Tricks of the Trade*. Springer, 2012, pp. 421–436.

[24] S. Li *et al.*, "Learning to detect malicious clients for robust federated learning," *arXiv preprint arXiv:2002.00211*, 2020.

[25] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[26] S. Shah, D. Dey, C. Lovett, and A. Kapoor, "Airsim: High-fidelity visual and physical simulation for autonomous vehicles," in *Field and Service Robotics*, 2017. [Online]. Available: https://arxiv.org/abs/1705.05065