

CUAHN-VIO: Content-and-Uncertainty-Aware Homography Network for Visual-Inertial Odometry

Supplementary Document

Yingfu Xu and Guido C. H. E. de Croon

I. EXPERIMENTS IN MULTI-PLANE ENVIRONMENTS

Review 2, Comment 1

Based on the reviewer's remark, we test our approach (6-3 of Table 7 in the main article) in a scene with two planes with around a 0.5-meter height difference. The camera was first moved around above the lower plane and then moved to the higher plane. Our approach suffered from big estimation errors in height and the horizontal moving direction, as shown in Fig. 1. At around the 25th second, the camera moved to the higher plane and its field of view was mainly occupied by the higher plane. Since the speed of the camera was almost constant, moving above the higher plane led to bigger homography transformations. According to the design of the simple EKF backend of our approach, it assumes that there is only one flat plane. So the reason for observing bigger homography transformations under the assumption is that either the speed increased or the camera was suddenly closer to the plane (the height became lower). Both phenomena can be observed in Fig. 1 at around 25th second. At around the 50th second, the camera was moved back above the lower plane. The magnitude of homography transformation suddenly decreased. Under the single plane assumption, the reason could be that the speed slowed or the camera moved further from the plane. Both phenomena can be observed in Fig. 1 at around 50th second. In another experiment, when the camera shuttled frequently between the two planes, the estimation error grew fast over time. When the two planes were the flat ground and a wall, our estimation also drifted a lot.

The reason for our approach's failure is that the EKF assumes a single plane in the field of view, which applies to most indoor environments. When a plane other than the ground occupies the field of view, the VIO frontend (homography network) would predict the planar homography transformation of the new plane instead of the ground. The EKF backend would wrongly take the network prediction as the planar homography transformation of the ground to estimate egomotion. Therefore, our simple EKF backend with single plane assumption cannot handle scenes with more than one plane. Scenes with more than one plane are out of the scope of this work because we mainly focus on the learning-based visual frontend and use a toy-like EKF backend to build the VIO system.

The above-introduced experiments inspire us to potential improvements in our approach to handling multi-plane cases, which is interesting to explore in future research. A distance

sensor (e.g. an ultrasound rangefinder) could be integrated to provide the distance to the planar surface. When the camera moves above a new plane, the distance sensor can measure the step-like change in the distance to the plane in the field of view and accordingly update the position estimation in the z -axis of the backend. This potential solution requires an additional sensor and a more complicated backend design. In the cases of more complicated multi-plane scenes, it may be possible to have an additional network that performs segmentation and finds the dominant plane. The homography network is then supposed to predict the homography transformation of the dominant plane. Improvements are possible to enable our approach for scenes with a slope, as discussed in Subsection 7.6 of the main article. When the dominant plane changes, the VIO backend should increase the variance of the distance to the plane and the covariance of the normal vector of the plane to get a fast converge state estimation.

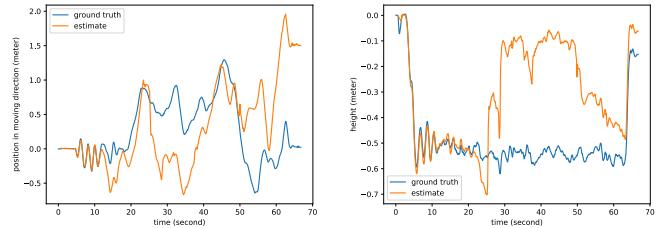


Fig. 1. The estimated and ground-truth position of the testing sequence where the camera was moved from above one plane to another. The left subplot shows the position in the moving direction of the horizontal plane. The right subplot shows the minus of height (because the z -axis of the IMU points to the ground).

II. EXPERIMENTS ON DENSE 3-D OBJECTS ON THE GROUND

Review 2, Comment 2

To explore the role of the content awareness training of the homography network of our approach, we test in an environment with much denser 3-d objects than the datasets in the main article. 37,608 image pairs are collected in the environment for training. Two networks without and with content awareness were trained and tested on a sequence filming dense 3-d objects as shown in Fig. 2. Their RMSEs of ATEs (absolute trajectory errors) are respectively 0.6717 and 0.6475. The VIO with the homography network that has content awareness is slightly more accurate than its counterpart

without content awareness. In Fig. 2 and 3, it is hard to find a visible advantage of the homography network that has content awareness. We have observed that the homography network trained without content awareness is already rather robust to non-homography image content. The content-aware training brings small improvement. To some extent, this explains the good generalization capacity of the homography networks without content awareness (e.g. 6-3 of Table 7 of the main article) in the real-world environments with non-homography image content, e.g. the UZH-FPV dataset and our MYNT CyberZoo dataset.

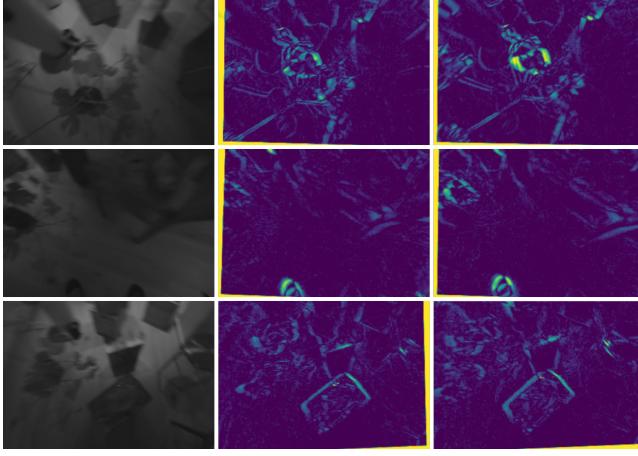


Fig. 2. Example images of the testing scene with dense 3-d object on the ground. The column on the left shows the raw images. The column in the middle shows the photometric error of image matching using the prediction of the homography network without content awareness. The column on the right shows the photometric error of image matching using the prediction of the homography network with content awareness. In the photometric error maps, yellow means big photometric error, blue means small photometric error.

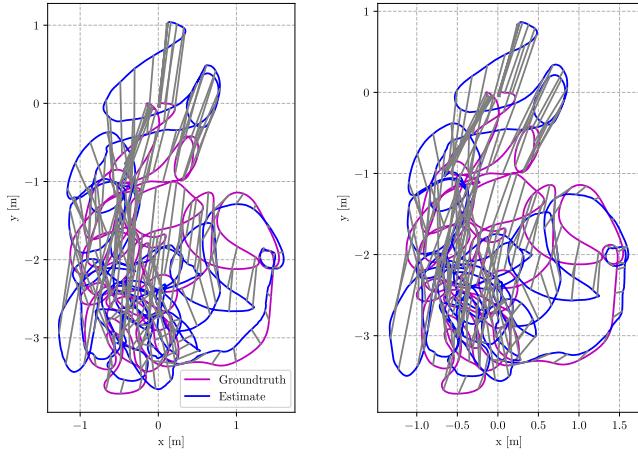


Fig. 3. Estimated trajectory aligned with the ground-truth trajectory. The left subplot shows the result of the VIO variant without content awareness. The right subplot shows the result of the VIO variant with content awareness.

III. COMPARISON BETWEEN NETWORK-ESTIMATED UNCERTAINTY AND CHI-SQUARE OUTLIER REJECTION

Review 2, Comment 3

We implement Chi-square outlier rejection for our VIO approach in the same way as the feature point outlier rejection

mechanism of ORB-SLAM. Same as ORB-SLAM, We select $\alpha = 0.05$. According to the Chi-square distribution table *, the threshold for a 8-DoF independent standard normal random variables is 15.507 when $\alpha = 0.05$. Therefore, assuming the measurement noise of each dimension of homography transformation is identical, a 8-d homography transformation z will be rejected if $(z - f)^T \cdot (z - f) > 15.507 \cdot R_{\text{meas}}$. f is the *a priori* homography transformation propagated by the EKF. The scalar R_{meas} is the hand-tuned fixed variance of the measurement noise of one dimension of z .

We apply Chi-square outlier rejection to the network homography transformation prediction of the Basic network, shown in Table 7 of the main article with index 1-3. We tune the value of R_{meas} based on the estimation accuracy of the 45-degree downward-facing Seq. 2 of the UZH-FPV dataset. The best R_{meas} on Seq. 2 is applied to other sequences. In Table I, we show the results of $\alpha = 0.05$ and $\alpha = 0.025$ (the corresponding threshold is 17.535). We also tried $\alpha = 0.1$ but it led to worse accuracy. For $\alpha = 0.05$, $R_{\text{meas}} = 29.5$. For $\alpha = 0.025$, $R_{\text{meas}} = 20.0$. From Table I, we notice that after Chi-square outlier rejection, the accuracy of Seq. 2, 4, and 12 is better than without outlier rejection. But outlier rejection led to the problem that too many measurements were rejected, the EKF states did not get enough updates and drifted a lot over time, marked by “-” in the table. To get better accuracy on the sequence with big drifts, it may be possible by tuning R_{meas} on the specific sequence. In contrast, our approach that uses the network-estimated uncertainty of measurement is tuned only on Seq. 2 and achieves decent accuracy on all sequences. Furthermore, as shown in Table 6 of the main article, our approach is not sensitive to the single tunable hyperparameter, making it convenient to deploy to a new environment, requiring little parameter tuning on-site. Overall, the VIO accuracy when using Chi-square outlier rejection and fixed hand-tuned measurement noise variance is lower than UAHN (6-3 of Table 7 in the main article) which uses network-estimated measurement noise variance.

TABLE I
ESTIMATION ACCURACY OF WITH AND WITHOUT CHI-SQUARE OUTLIER REJECTION. **BOLD** REPRESENTS THE BEST AND UNDERLINE REPRESENTS THE SECOND BEST.

VIO	RMSE (meter) of Absolute Translation Errors (ATE) ↓					
	Seq. 2	Seq. 4	Seq. 9	Seq. 12	Seq. 13	Seq. 14
Basic (1-3) (without outlier rejection)	1.1526	1.0022	<u>0.5777</u>	0.9379	<u>1.2572</u>	1.6455
Basic (1-3) $\alpha = 0.05$	1.0895	0.7494	-	-	1.9469	1.8514
Basic (1-3) $\alpha = 0.025$	<u>0.9344</u>	<u>0.6303</u>	-	<u>0.7392</u>	1.9463	-
UAHN (6-3)	0.3371	0.3139	0.3392	0.5837	0.4066	<u>1.7905</u>

IV. EXPERIMENTS ON MSCKF WITH DIFFERENT NUMBERS OF FEATURE POINTS

Review 2, Comment 6

*<https://www.scribbr.com/statistics/chi-square-distribution-table/>

TABLE II
TIME CONSUMPTION INDICATORS MEASURED ON TX2 PROCESSOR BY RUNNING SEQ. 2 OF UZH-FPV. **BOLD** REPRESENTS THE BEST. UNDERLINE MARKS THE FRAME RATE AND ATE VALID FOR ACCURACY EVALUATION.

VIO	Image Resolution (pixels)	Num. of Pts / Network Blocks	Histogram Equalization (HE)	Visual Processing Time (ms)	IMU Propagation Time (ms)	EKF Updating Time (ms)	Total Mean (ms) ↓	Total Variance (ms ²) ↓	Ratio of Long Processing Time (%) ↓	Avg. Processed Frame Rate (fps) ↑	RMSE (m) of ATE ↓
MSCKF	640×480	300	✓	38.37	19.38	8.43	66.19	3.95e4	66.30	18.06	0.3142
MSCKF	320×240	180	✓	24.47	2.29	6.04	32.80	471.61	25.94	23.78	0.4058
MSCKF	640×480	100	✓	26.14	2.47	3.94	32.57	549.43	29.88	23.14	0.3386
MSCKF	640×480	100	✗	22.85	2.09	3.80	28.74	460.15	24.93	24.24	0.3178
MSCKF	640×480	75	✓	24.29	2.26	3.30	29.86	503.42	27.94	23.70	0.3395
MSCKF	640×480	50	✓	21.81	2.04	2.64	26.49	467.78	23.88	24.35	0.3637
MSCKF	640×480	25	✓	16.96	1.68	1.83	20.48	336.94	16.27	25.90	0.3557
MSCKF	640×480	20	✓	14.71	1.67	1.65	18.03	258.57	12.08	<u>26.23</u>	0.3414
MSCKF	640×480	15	✓	15.46	1.64	1.42	18.52	239.88	12.47	<u>26.23</u>	0.4047
MSCKF	640×480	10	✓	15.81	1.66	1.28	18.75	233.38	12.21	<u>26.23</u>	0.4112
MSCKF	320×240	10	✓	11.57	1.78	1.48	14.83	157.05	7.01	<u>26.23</u>	0.6000
UAHN-VIO	320×224	2	✗	24.00	1.48	0.12	25.61	3.32	0.44	<u>26.11</u>	0.3544
UAHN-VIO	320×224	3	✗	27.30	1.46	0.12	28.89	2.66	0.78	<u>26.11</u>	0.3380

In this section, we show an extended version of Table 8 in the main article by adding the experimental results of MSCKF processing 15, 20, 25, 50, and 75 points in an image frame. From Table II, we notice that when the number of feature points to process is set below twenty, MSCKF processes every frame without frame dropping. However, there are more than 12% frames whose processing time is more than the time interval of temporally consecutive frames, which means the delay is more than 33.3 milliseconds. For the test where the number of points was set to twenty for MSCKF, the trajectory estimation accuracy (0.3414) is slightly higher than our approach with two network blocks (0.3544) but slightly lower than our approach with three network blocks (0.3380).

V. EXPERIMENTS ON FORWARD-FACING CAMERA

Although our approach is designed for planar scenes and thus forward-facing camera is out of the design scope of this work, because of the reviewer's remark and potential inspiration to future research, we test our approach UAHN-VIO (6-3 of Table 7) on two forward-facing camera sequences of UZH-FPV, indoor_forward_5_snapdragon (Seq. 5) and indoor_forward_7_snapdragon (Seq. 7). The MAV is relatively slow when recording Seq. 5, so the camera is straightly forward facing, which means that less than 50% of pixels are on the planar ground. More pixels are on the wall and other 3D objects. In contrast, Seq. 7 is a fast flight sequence. The big acceleration forced the MAV to "nose down", thus in many images, the ground covers almost 70% of the whole frame. According to our observation, due to too much non-planar content, many predicted homography transformations in Seq. 5 are not accurate. An example is shown in the left part of Fig. 4. From the photometric error map after image matching using the predicted homography transformation (the second from left in Fig. 4), we can see yellow patches on the ground, meaning big photometric errors. Relatively, the case of Seq. 7 is better.

Fig. 5 shows the estimated trajectory aligned with the ground truth. The estimated trajectory of Seq. 5 (left of Fig. 5) has a big drift. The estimated trajectory of Seq. 7 is relatively

accurate. Its RMSE of ATE (absolute trajectory error) of our UAHN-VIO (6-3 of Table 7) on Seq. 7 is 1.1277, not much worse than MSCKF (0.5125). This experiment demonstrates that our approach has the chance to work on a forward-facing camera if most of the pixels are on the planar ground, a bigger percentage of the image frame filming the planar ground can lead to higher accuracy. If the percentage of planar ground is small, our approach suffers from big estimation errors due to the inaccurate homography transformation deteriorated by a big number of non-planar pixels.

VI. NETWORK ARCHITECTURE AND IMPLEMENTATION

A. Network Architecture

A network block is made of several convolutional layers followed by one or two (only the last block has two) fully-connected layer(s). All the layers except for the output layer of each block are followed by a Leaky ReLU activation with a negative slope of 0.1. There is no normalization layer.

B. Model Size

The Basic Model has 5,228,280 parameters. The Master Model is an enlarged version of the Basic Model. It has 6,897,840 parameters. For the student model that outputs predictive uncertainty, a subnetwork of two fully-connected layers with 1,313,032 parameters is added to the 4th block. The total number of parameters of a student model is 6,541,312, among which 3,612,312 belong to the 4th block. These parameters are trained in uncertainty-aware learning. The convolutional decoder for mask prediction has 1,090,681 parameters that are trained in content-aware learning.

C. Implementation and Training

We implement the networks by PyTorch [1] v1.7.1. The training is executed on a server with GTX1080ti GPU and CUDA v10.1. The AdamW [2] optimizer with $\beta = (0.9, 0.999)$ and weight decay $\lambda = 0.01$ is utilized during the 50 training epochs. The batch size is 16. The initial learning rate is 2e-4 and it is divided by 2 after 10, 20, 30, 35, 40, and

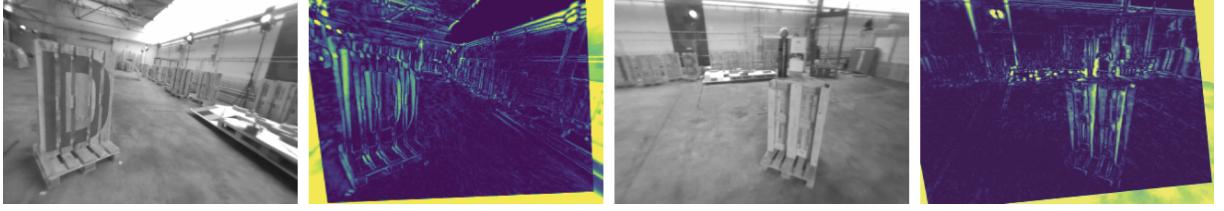


Fig. 4. Example images and photometric error maps after image matching by the network prediction of our approach. The image on the left is from `indoor_forward_5_snapdragon` and the image on the right is from `indoor_forward_7_snapdragon`. Yellow means big photometric error and blue means small photometric error.

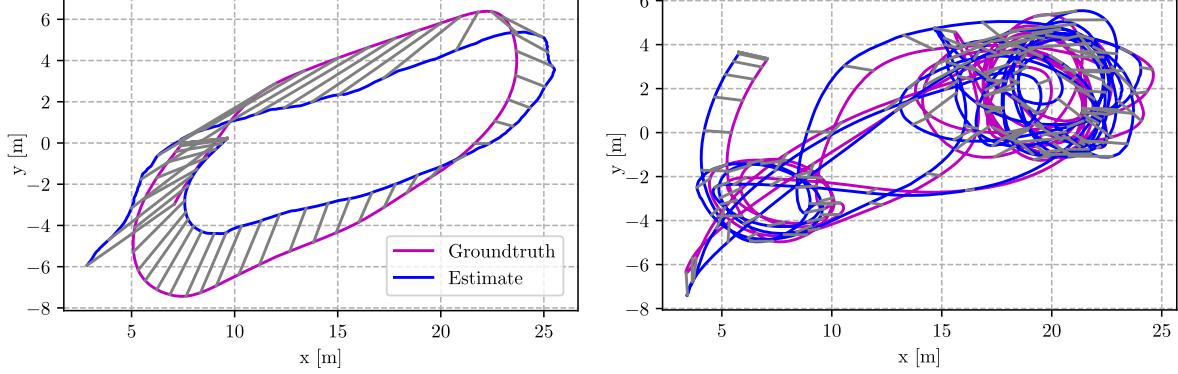


Fig. 5. The estimated trajectory and ground-truth trajectory of `indoor_forward_5_snapdragon` (left) and `indoor_forward_7_snapdragon` (right).

45 epochs. The weight parameters are initialized by Kaiming initialization [3] and bias parameters are initialized to zero. The network performs inference on the validation set after each training epoch. The set of parameters that achieve the smallest average validation error are saved as the best model and evaluated on the testing set. Most networks in this article follow the above training procedure, except that Master Model has a smaller initial learning rate (5e-5).

To run a network in a C++ environment, we utilize PyTorch C++ API. We use TorchScript to trace the trained network model in the Python environment to generate a file that can be loaded by C++ code. Further, we use LibTorch in the C++ environment to load the traced model and run the network.

The visual processing time of VIO shown in Table 8 of the main article is higher than pure network inference (Table 7 of the main article). The reason is that visual processing includes network inference and other processes supporting it. The processes are image undistortion and resize, and data type conversion between C++ double, OpenCV Mat, and LibTorch tensor. Their total time consumption is around 4.0 to 4.5 milliseconds.

In the training of the content-aware mask based on Laplacian probability distribution, we tried to train the network to directly output $\log b_k$ (Eq. (7) of the main article) in order to avoid zero values of b_k . However, the training became unstable. Good results were achieved when b_k is the sum of the square of the network output and a small value (2e-6).

D. Comparison of Basic Homography Networks

We compared different network architectures and downsample methods in the early stage of this work. The first architecture utilizes the feature pyramid extractor (FPE) network

inspired by the PWC-Net [4]. The FPE is made of three convolutional layers. It processes both images independently and outputs pyramidal feature maps. When the pyramid level is one higher, the height and width of the feature map are halved and the number of channels is doubled. This architecture has 5,466,320 parameters. The second architecture gets the image pyramid by downsampling the images to half/one-fourth/one-eighth of their height and width by average pooling or bilinear interpolation. To achieve a similar model size and inference speed to the first architecture, there are more channels in some of the intermediate tensors in the 3rd block and one more convolutional layer in the 4th block. This architecture has 5,228,280 parameters. We compare these two architectures because they achieved similar performance in translational motion prediction and outperformed others in [5].

TABLE III
COMPARISON OF BASIC HOMOGRAPHY NETWORKS

Network Input	Downsample	Avg. Error (pixel)		Time Cons. (ms)	
		Model 1	Model 2	Python	C++
Feature Maps	FPE	0.409	0.388	28.66	21.66
Pyramidal Images	Avg. Pooling	0.275	0.285	28.20	21.16
Pyramidal Images	Bilinear Interp.	0.280	0.281	29.06	21.65
Pyramidal Images	Bilinear Interp.	0.498*	0.501*	21.24*	18.27*

* Directly regress to homography matrix

We trained two models of each architecture to exclude the influence of individual circumstances. The time consumption of network inference is measured on the GPU of an Nvidia Jetson TX2 mobile processor in Max-P ARM power mode. The shown values in the Python column are the averages of one thousand times of inference on the validation set

of the Basic Dataset when the batch size is one. The C++ time consumption values are measured when processing the downward-facing Seq. 2 of the UZH-FPV dataset. From Table III we can conclude that the architecture using pyramidal images is more accurate and faster than its peer utilizing FPE. Downsampling by average pooling is slightly faster than bilinear interpolation. Their accuracy is very close. Besides corner flow, directly regressing to the eight elements of \mathbf{H} (the last diagonal element is 1.0) is also implemented for the first three blocks (bottom row of Table III). It does not require DLT solving so it runs faster. But obviously, it is less accurate.

Model 1 of the network using pyramidal images and average pooling in Table III is selected as the Basic Model in the main article.

E. Direct Linear Transformation (DLT) Solver

DLT solver aims at solving the homography matrix \mathbf{H} that has 8 degrees of freedom from at least four pairs of corresponding points. As (1) shows, \mathbf{H} relates the pixel coordinates in two images of a point lying on a planar surface.

$$\lambda \begin{bmatrix} u_2 \\ v_2 \\ 1 \end{bmatrix} = \mathbf{H} \begin{bmatrix} u_1 \\ v_1 \\ 1 \end{bmatrix}, \mathbf{H} = \begin{bmatrix} h_1 & h_2 & h_3 \\ h_4 & h_5 & h_6 \\ h_7 & h_8 & 1 \end{bmatrix} \quad (1)$$

From (1), we have $\lambda = h_7u_1 + h_8v_1 + 1$, then we can form up a linear equation set $\mathbf{A}_i\mathbf{h} = \mathbf{b}_i$, where

$$\begin{aligned} \mathbf{A}_i &= \begin{bmatrix} u_1 & v_1 & 1 & 0 & 0 & 0 & -u_2u_1 & -u_2v_1 \\ 0 & 0 & 0 & u_1 & v_1 & 1 & -v_2u_1 & -v_2v_1 \end{bmatrix} \\ \mathbf{h} &= [h_1 \ h_2 \ h_3 \ h_4 \ h_5 \ h_6 \ h_7 \ h_8]^T \\ \mathbf{b}_i &= [u_2 \ v_2]^T \end{aligned} \quad (2)$$

Since we only have the predicted offsets of four corner pixels, i ranges from 1 to 4. Then by forming up the 8×8 matrix \mathbf{A} and 8-d vector \mathbf{b} from \mathbf{A}_i and \mathbf{b}_i , and calculating the inverse matrix of \mathbf{A} , \mathbf{h} can be solved as $\mathbf{h} = \mathbf{A}^{-1}\mathbf{b}$.

VII. PREDICTIVE UNCERTAINTY

A. Why Learning Requires a Teacher Network?

In Subsection 4.3 of the main article, we introduce how to perform content-aware learning by the predictive uncertainty of photometric matching. A reader may think that the error of homography transformation prediction can be reflected by the photometric matching uncertainty, which is true. Here we explain the reason why we do not use the photometric matching uncertainty map to calculate the uncertainty of homography transformation prediction.

The predictive uncertainty map b_k explains the uncertainty of photometric matching that is affected by both image content and the accuracy of homography transformation. For a well-trained content-aware network, b_k mainly reflects the content information in complicated scenes as shown in Fig. 6 (main article). Because the photometric error induced by the prediction error is much smaller compared to the non-homography content. Since it is unknown whether a pixel belongs to the single non-reflective plane, it is not tractable to

decouple the prediction error and the image content. Even if b_k is determined by the prediction error alone, e.g., the pixels on the plane are accurately semantically segmented or the input images have no unfavorable content, it is not clear to us how to calculate the prediction uncertainty from b_k .

Therefore, we follow the common practice of using the negative log-likelihood (NLL) loss to learn the predictive uncertainty. The teacher network is supposed to have high accuracy to provide the learning targets of mean values.

B. Comparison of Different Output Dimensions

In addition to predicting eight different values for each σ_n^2 (8-d variance), inspired by the concise approach utilized in [6], we also implement predicting a generalized single value that represents the overall uncertainty (1-d variance) and duplicating it eight times to fill in all dimensions. It assumes that noise in the input affects all dimensions of f_4 equally, neglecting the difference among them. Two types of variance representation under different supervisions are compared in Table IV. The 4th block of the student model is randomly initialized. The best model in one training attempt is the one achieving the smallest average imitation error on the validation set. In terms of the prediction accuracy of the mean values, 1-d variance is slightly better. While 8-d variance has obvious advantages in predictive uncertainty as evidenced by lower AUSE and higher inside rate.

TABLE IV
COMPARISON OF PREDICTIVE UNCERTAINTY

Super- vision	Dim. of Var.	Avg. Error (pixel)	Avg. Imitation Error (pixel)	Avg. Var.* (pixel)	AUSE	Inside Rate (3 σ)
GT	1	0.454	0.490	28.16	455.9	96.46
GT	8	0.454	0.489	16.66	370.0	97.63
master	1	0.419	0.329	11.80	445.7	80.83
master	8	0.428	0.336	7.69	357.3	86.96
self	1	0.406	0.210	2.59	756.7	76.93
self	8	0.408	0.208	1.83	565.9	79.08

* The networks predict rare unreasonably big variances. The averages are calculated after removing the 0.1% biggest values.

C. Difficult Testing Samples

Fig. 6 shows several difficult testing samples from which we explore what leads to high prediction error and high predictive uncertainty. Comparing the Basic Model and the Master Model, the Master Model achieves accurate predictions except for the 1st and 5th images. Significant blur (1st and 2nd columns) and lack of texture (4th and 5th columns) cause big prediction errors of the Basic Model that has less model capacity. The master model achieves high accuracy on the images of the 2nd and 4th columns, while the errors of the Master-Teach student model are big. It indicates that the student fails to imitate the teacher well. From the appearance of the images, we speculate that the difficulty in imitation is due to the unfavorable image contents from which extracting desired information is hard for a network. That is to say, the big predictive variances are caused by unfavorable image contents.

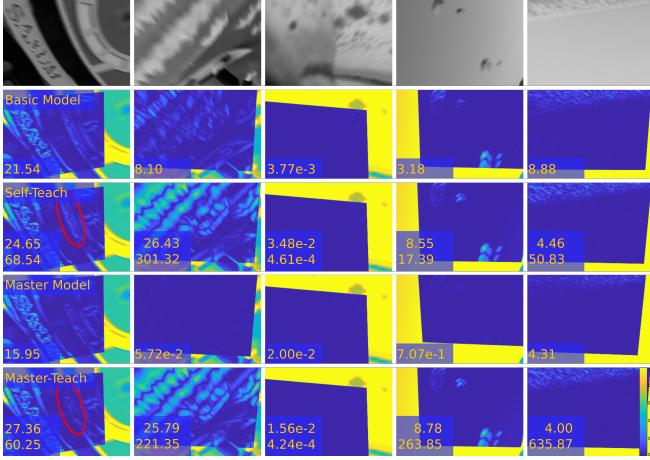


Fig. 6. Five testing samples from the Basic Dataset that contain unfavorable content: significant blur (1st to 4th) and lack of texture (4th and 5th). 1st row shows one of the input images \mathcal{I}_p . The 2nd to 5th rows show the photometric error map of $(\mathcal{I}_p, \mathcal{I}_c)$, where dark blue means a low error. The data shown at the bottom left corner of a photometric error map includes the error of the 1st element of $\mathbf{f}_{\text{total}}$ (*i.e.* the u component of the upper-left corner) and its predictive variance below it if the network is a student. The 2nd and 4th rows correspond to the teacher networks. The 3rd and 5th rows respectively correspond to the student networks in Self-Teach and Master-Teach. The red ellipses in the 1st column highlight one of their nuances.

It corroborates the definition of predictive uncertainty, which is that it captures the uncertainty in network output caused by the content-determined observation noise. Based on our observations, the noise is mainly caused by textureless image content and blur.

VIII. CORRELATION BETWEEN PREDICTIVE AND EMPIRICAL UNCERTAINTY

The predictive variance $\sigma_{\text{pred.}}^2$ and empirical variance $\sigma_{\text{emp.}}^2$ are logged and plotted in Fig. 7. The idea of estimating empirical uncertainty is to model the distribution of parameters by independent network samples. The randomly initialized networks are more independent from each other than the networks initialized by the same parameter set from the Basic Model. For the MC-Dropout networks, due to the initialization of convolutional layers being the same, the independence is compromised. Thus the randomly initialized ensemble has the best knowledge of the parameter distributions and captures the biggest $\sigma_{\text{emp.}}^2$. Comparing $\sigma_{\text{emp.}}^2$ and $\sigma_{\text{pred.}}^2$, in general $\sigma_{\text{emp.}}^2$ has smaller values. For all three models, $\sigma_{\text{emp.}}^2$ accounts for roughly one-third of the total variance for 90% of testing data as shown in the left three subplots.

In the right subplots, with $\sigma_{\text{pred.}}^2$ rapid grows, $\sigma_{\text{emp.}}^2$ also has the trend to increase. The correlation can also be observed in the left subplots. The reason behind the correlation is discussed as follows. A small $\sigma_{\text{pred.}}^2$ means that the network models in the ensemble “think” that their predictions are close to the teacher network with the same input. Thus the networks have similar predictions that lead to small $\sigma_{\text{emp.}}^2$. Similarly, a big $\sigma_{\text{pred.}}^2$ indicates the predictions are far different from the teacher’s. Different models may have very different predictions due to the randomness in training. Thus a bigger $\sigma_{\text{emp.}}^2$ is produced. In the perspective of data distribution of the dataset, the testing

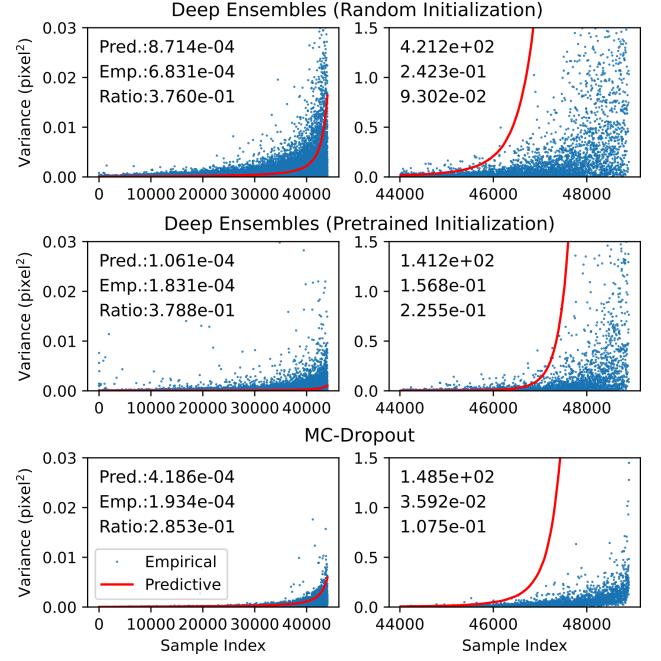


Fig. 7. Predictive variance $\sigma_{\text{pred.}}^2$ and empirical variance $\sigma_{\text{emp.}}^2$ of three network models in an ensemble. The $\sigma_{\text{pred.}}^2 - \sigma_{\text{emp.}}^2$ pairs of each model are sorted according to $\sigma_{\text{pred.}}^2$ and separated into two groups. 90% of the data that has smaller $\sigma_{\text{pred.}}^2$ is shown in the subplots on the left. The rest 10% that has big $\sigma_{\text{pred.}}^2$ is shown in the subplots on the right. Note that the scales of y -axes of the right subplots are much bigger than the left ones. The numbers around the top left corner of each plot are statistical values of the data in the group, which are the averages of $\sigma_{\text{pred.}}^2$, the averages of $\sigma_{\text{emp.}}^2$, and the average ratio of $\sigma_{\text{emp.}}^2$ in the total variance $\sigma_{\text{pred.}}^2 + \sigma_{\text{emp.}}^2$. 10% of the total data is plotted to avoid overly dense points.

data leading to big $\sigma_{\text{pred.}}^2$ is the minority. Such images are likely to have unfavorable content that is rare in the training set. Network’s unfamiliarity with such contents increases $\sigma_{\text{emp.}}^2$.

In order to gain more insight, we log the $\sigma_{\text{pred.}}^2$ and $\sigma_{\text{emp.}}^2$ on UZH-FPV and show the average values in Table V. The light yellow group shows the results of VIO 6-2 in Table 7 of the main article. The light green group shows the results of a variant of 4-4 with four blocks and without *a priori*. The three sequences on the left were recorded when there are crowded low objects on the ground. The other three sequences were filmed in the same flight arena but the ground is relatively cleaner from objects. We respectively sort the sequences of the two environments in ascending order of speed. Because the network input image can vary with *a priori* homography transformation, we run all four network blocks and the inputs

TABLE V
UNCERTAINTY ESTIMATION ON UZH-FPV.

Average Value	4 (6.55)	2 (6.97)	9 (11.23)	12 (4.33)	13 (7.92)	14 (9.54)
Pred. Var.	92.4	106.8	332.1	41.3	177.5	552.5
Emp. Var.	0.043	0.061	0.108	0.030	0.071	0.124
Ratio Emp. Var.	0.203	0.171	0.120	0.217	0.177	0.139
Pred. Var.	102.4	161.9	438.5	47.6	165.9	415.0
Emp. Var.	0.130	0.191	0.327	0.060	0.120	0.245
Ratio Emp. Var.	0.316	0.275	0.187	0.329	0.232	0.178

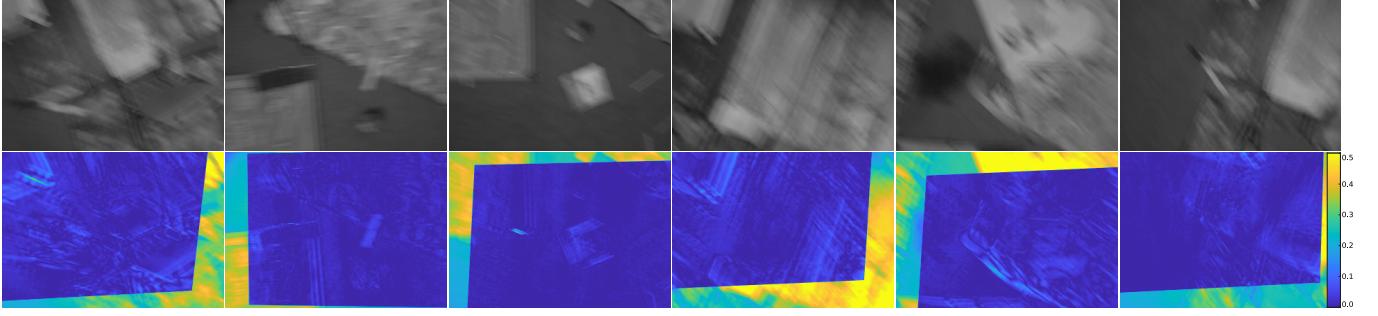


Fig. 8. Example images from the high-speed two-waypoint shuttle flight sequence. The top row shows the original images and the bottom row shows the photometric error maps of the image pairs warped according to the network predictions of homography transformation. Significant motion blur causes big predictive uncertainty. The average values of the variance estimations of the eight elements of the corner flow are respectively 394.99, 106.41, 59.80, 183.90, 190.22, 221.72 (pixel²) for the six examples.

are the original images. Thus different networks have identical inputs. The data in Table V tells that both predictive and empirical variances increase with speed. The growth in $\sigma_{\text{pred.}}^2$ is more significant and thus the ratio of $\sigma_{\text{emp.}}^2$ decreases with speed. It is related to the motion blur considered as observation noise by the network, as discussed before. The positive correlation between $\sigma_{\text{pred.}}^2$ and $\sigma_{\text{emp.}}^2$ is observable. For both $\sigma_{\text{pred.}}^2$ and $\sigma_{\text{emp.}}^2$, deep ensembles produce bigger values, implying more comprehensive uncertainty estimation that contributes to better AUSE and Inside Rate, as shown in Fig. 9 in the main article.

Snapshot Ensemble [7] learns multiple models faster than deep ensembles. Since it shares the disadvantage of high inference time cost with deep ensembles (limited by our current implementation), it is not separately discussed. The approach proposed in [8] that enhances model diversity in the ensemble by altering hyperparameters is not discussed for the same reason.

IX. NETWORK UNCERTAINTY AND VELOCITY

As supplementary to Fig. 13 in the main article, Fig. 9 illustrates the statistically positive correlation between the

magnitude of optical flow and network uncertainty estimation. We do not know the exact magnitude of optical flow but it is determined by the speed divided by the distance to the ground and the rotation rate. Since the rotation in the dataset is mostly slow, the height-scaled speed can serve as an approximate equivalent of the magnitude of optical flow. For both sequences, estimated variance is more likely to have a big value when the speed is high. In the slow sequence, most estimated variances have very small values thanks to the clear and sharp images. While in the fast sequence where blurry images are common, the positive correlation between the estimated variance and height-scaled speed is obvious. Fig. 8 shows cases where the network has big variance estimations.

X. UAHN-VIO FOR FEED-BACK CONTROL

Two of the common practices for VIO to obtain better performance in autonomous flights are 1) publishing the estimated states in higher frequency, *e.g.*, IMU measurement frequency, and 2) propagating the states to the current time using the IMU measurements that come after the latest image to compensate for the image processing latency. Since we aim to evaluate the VIO instead of pursuing the overall performance of autonomous flight, the above methods are not deployed in this article. Nevertheless, thanks to the short and almost constant processing time of UAHN-VIO, the MAV performed stable controlled flights. The RMSE of the ATE

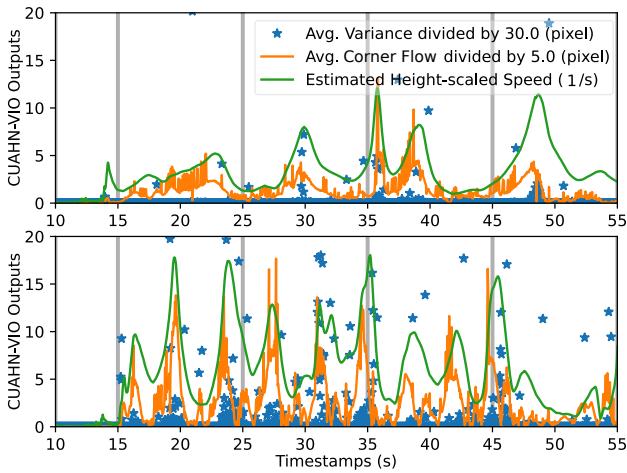


Fig. 9. The outputs of CUAHN-VIO (6-4 of Table 7 in the main article) evaluated on a slow sequence (Seq. 12) and a fast sequence (Seq. 14) of UZH-FPV dataset. Both sequences were collected in the same environment.

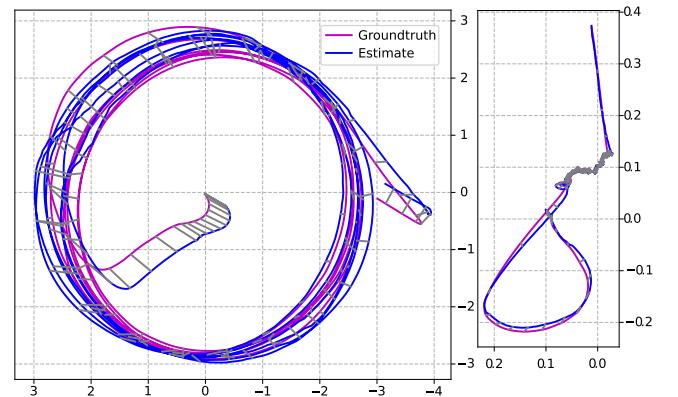


Fig. 10. Trajectories of circle trajectory tracking and fixed-point hovering.

of the estimated trajectory by UAHN-VIO (6-5 in Table 7) is 0.1521 in the two-waypoint shuttle flight (Fig. 14 of the main article). The trajectories of autonomous flights of tracking a circle trajectory and fixed-point hovering are shown in Fig. 10. The RMSEs of the ATEs are respectively 0.2775 and 0.0155, estimated by UAHN-VIO (6-5 in Table 7).

XI. VIO BACKEND

A. State Propagation

In our implementation, we use the simple zeroth-order integration (3). When propagating the states from t to $t + 1$, we use the average of the IMU measurements sampled at the two time points (4).

$$\begin{aligned} \mathbf{p}_{t+1} &= \mathbf{p}_t + \Delta t(-[\bar{\omega}]_{\times} \mathbf{p}_t + \mathbf{v}_t), \\ \mathbf{v}_{t+1} &= \mathbf{v}_t + \Delta t(-[\bar{\omega}]_{\times} \mathbf{v}_t + \bar{\mathbf{a}} + \mathbf{R}^{-1}(\mathbf{q}_t) \mathbf{g}), \\ \mathbf{q}_{t+1} &= \mathbf{q}_t \otimes \mathbf{q} \{ \bar{\omega} \Delta t \} = \mathbf{q}_t \otimes \begin{bmatrix} \cos(\|\bar{\omega}\| \Delta t / 2) \\ \frac{\bar{\omega}}{\|\bar{\omega}\|} \sin(\|\bar{\omega}\| \Delta t / 2) \end{bmatrix}, \quad (3) \\ \mathbf{b}_{a,t+1} &= \mathbf{b}_{a,t}, \quad \mathbf{b}_{g,t+1} = \mathbf{b}_{g,t}, \\ \mathbf{f}_{j,t+1} &= \mathbf{f}_{j,t} - \Delta t (\mathbf{I} - (\mathbf{c}_j + \mathbf{f}_{j,t}) \mathbf{e}_z^T) \mathbf{H} (\mathbf{c}_j + \mathbf{f}_{j,t}) \\ \bar{\mathbf{a}} &= 0.5 \cdot (\hat{\mathbf{a}}_t + \hat{\mathbf{a}}_{t+1}), \quad \bar{\omega} = 0.5 \cdot (\hat{\omega}_t + \hat{\omega}_{t+1}) \end{aligned} \quad (4)$$

B. a Priori Homography

As shown in Fig. 11 of the main article, the variance estimations for both image pairs are smaller when utilizing *a priori* homography. In the 2nd row, a noticeable phenomenon is that the photometric error map of $(\tilde{\mathcal{I}}_{t,2}, \mathcal{I}_{t-1})$ is bigger than $(\tilde{\mathcal{I}}_{t,\text{prior}}, \mathcal{I}_{t-1})$. The reason can be that, in training, the networks are trained to infer the whole homography transformation from the original image pairs. The 2nd block often handles bigger disparities in training than the ones between $\tilde{\mathcal{I}}_{t,\text{prior}}$ and \mathcal{I}_{t-1} and thus tends to output bigger values.

C. Iterative EKF

Inspired by [9], [10], we tried the iterative EKF scheme. After the first EKF update, the corner flow is updated. The image pair is then warped according to the updated corner flow and input to the network again to estimate the remaining visual disparities. And then EKF is updated once more according to the outputs of the second network inference. However, this scheme fails to improve accuracy.

XII. PARAMETER TUNING OF SOTA VIO APPROACHES

For OpenVINS, we use the open-sourced code of it and modified all the parameters mentioned in the report[†] from the original values in the launch file. For MSCKF, we use the same code as OpenVINS by setting the number of SLAM points to zero.

At the beginning of each sequence of UZH-FPV, the MAV was swung by the human operator and then put on the ground.

[†]<https://rpg.ifi.uzh.ch/uzh-fpv/ICRA2020/reports/open-vins.pdf>

For sequences No. 2, 9, and 12, when we started ROVIO at the very beginning of the sequences, the estimated trajectories showed huge drifts. By changing the starting time of ROVIO to after the MAV is swung and before taking off, we obtained much better results on these sequences. We tried the IMU noise parameters given by UZH-FPV and the parameters in the ROVIO's configuration file for EuRoC. After several attempts, we found that the combination of IMU noise densities for EuRoC and the IMU random walk parameters given by UZH-FPV yielded good results.

We modified the source code of VINS-Fusion to get odometry output at frame rate. We observed occasional big drifting in the estimated trajectories of VINS-Fusion. So we ran VINS-Fusion several times on each sequence until we got two or more good estimated trajectories, and used the best one in comparison. We used the IMU noise parameters given by UZH-FPV for VINS-Fusion. We observed that multiplying the given white noise standard deviation of the gyroscope and the accelerometer by 0.005 yields better results.

For LARVIO, we use the parameters mentioned in the report[‡]. The IMU noise parameters are the same as the ones in EuRoC.yaml of the open-sourced code. The VIO started at the beginning of Seq. 9. But for other sequences, the VIO started after the hand-swinging part.

ORB-SLAM3 initializes a map when the drone slowed down to make turns and the optical flow was relatively small. But it lost tracking very soon when the drone speeded up. We tried to lower the thresholds for the feature extraction but it made little difference.

REFERENCES

- [1] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, *et al.*, “Pytorch: An imperative style, high-performance deep learning library,” *Advances in neural information processing systems*, vol. 32, pp. 8026–8037, 2019.
- [2] I. Loshchilov and F. Hutter, “Decoupled weight decay regularization,” *arXiv preprint arXiv:1711.05101*, 2017.
- [3] K. He, X. Zhang, S. Ren, and J. Sun, “Delving deep into rectifiers: Surpassing human-level performance on imagenet classification,” in *Proceedings of the IEEE international conference on computer vision*, pp. 1026–1034, 2015.
- [4] D. Sun, X. Yang, M.-Y. Liu, and J. Kautz, “Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8934–8943, 2018.
- [5] Y. Xu and G. C. de Croon, “Cnn-based ego-motion estimation for fast mav maneuvers,” *arXiv preprint arXiv:2101.01841*, 2021.
- [6] M. Klodt and A. Vedaldi, “Supervising the new with the old: learning sfm from sfm,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 698–713, 2018.
- [7] G. Huang, Y. Li, G. Pleiss, Z. Liu, J. E. Hopcroft, and K. Q. Weinberger, “Snapshot ensembles: Train 1, get m for free,” *arXiv preprint arXiv:1704.00109*, 2017.
- [8] Z. Y. Ding, J. Y. Loo, V. M. Baskaran, S. G. Nurzaman, and C. P. Tan, “Predictive uncertainty estimation using deep learning for soft robot multimodal sensing,” *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 951–957, 2021.
- [9] M. Bloesch, M. Burri, S. Omari, M. Hutter, and R. Siegwart, “Iterated extended kalman filter based visual-inertial odometry using direct photometric feedback,” *The International Journal of Robotics Research*, vol. 36, no. 10, pp. 1053–1072, 2017.

[‡]<https://fpv.ifi.uzh.ch/wp-content/uploads/sourcenova/uni-comp/2019-2020-uzh-fpv-benchmark/submissions/20/details.pdf>

- [10] S. Zhong and P. Chirarattananon, “Direct visual-inertial ego-motion estimation via iterated extended kalman filter,” *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 1476–1483, 2020.
- [11] P. Geneva, K. Eckenhoff, W. Lee, Y. Yang, and G. Huang, “Openvins: A research platform for visual-inertial estimation,” in *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 4666–4672, IEEE, 2020.
- [12] X. Qiu, H. Zhang, and W. Fu, “Lightweight hybrid visual-inertial odometry with closed-form zero velocity update,” *Chinese Journal of Aeronautics*, 2020.
- [13] M. Li and A. I. Mourikis, “High-precision, consistent ekf-based visual-inertial odometry,” *The International Journal of Robotics Research*, vol. 32, no. 6, pp. 690–711, 2013.
- [14] T. Qin, P. Li, and S. Shen, “Vins-mono: A robust and versatile monocular visual-inertial state estimator,” *IEEE Transactions on Robotics*, vol. 34, no. 4, pp. 1004–1020, 2018.