VIII. Appendix

A. Comparison against VID-Fusion

VID-Fusion is a visual-inertial-model-based odometry method very similar to VIMO. The main difference between the two methods is the external force model. While in VIMO, the external force is modeled as a Gaussian random variable with zero mean, in VID-Fusion, the mean is obtained by integrating the difference between accelerometer and thrust measurements. According to the authors, this prior helps the estimation of continuous external forces. In the main paper, we did not include VID-Fusion [2] among the baselines for conciseness. Since the drone model is the same as the one used in VIMO, VID-Fusion has the same limitations. Consequently, including VID-Fusion in the baselines does not affect the conclusions drawn in the paper.

In this appendix, for the sake of completeness, we present the same experiments as in the main paper including VID-Fusion among the baselines. We refer the reader to the main paper for the detailed description of the experimental setup and focus here only on the results.

1) Blackbird Dataset: We present the ATE_T and ATE_R on the evaluation sequences of the Blackbird dataset in Table IV. Since VIMO and VID-Fusion use the same drone dynamics model and there are no external perturbances acting on the drone in this dataset, their trajectory estimation accuracy is similar. Our method outperforms both baselines as well as the VIO solution. Large improvements are in the fast sequences, where our learned aerodynamics model brings additional information to the VIO backend. However, the performance difference is small in slow sequences, where including aerodynamic effects in the drone model is less effective.

TABLE IV: Evaluation of the trajectory estimates in the Blackbird dataset. In bold are the best values, and the second-best values are underlined.

Trajectory Name	$v_{\max}$ [m/s]	Evaluation Metric: ATE_T [m] / ATE_R [deg]			
		VIO	VID	VIMO	HDVIO (ours)
Bent Dice	3	**0.20** / 1.78	0.25 / **1.18**	0.31 / 1.53	0.21 / 1.53
Clover	5	0.90 / 3.52	0.83 / 2.48	0.88 / 3.66	**0.60** / **2.08**
Egg	5	1.07 / 1.54	0.81 / 1.61	0.75 / 1.34	**0.59** / **1.21**
Egg	6	1.40 / 2.35	1.10 / 2.42	0.98 / 4.89	**0.83** / **1.62**
Egg	8	1.79 / 4.55	1.47 / 4.84	1.57 / 3.69	**1.06** / **2.89**
Mouse	5	1.10 / 4.54	0.54 / 2.10	0.76 / 2.14	**0.36** / **1.40**
Star	1	0.17 / 0.78	0.18 / **0.54**	0.18 / 1.05	**0.16** / 0.58
Star	3	0.62 / 3.50	0.50 / 2.93	0.43 / **1.38**	**0.38** / 1.40
Winter	4	0.97 / 2.92	0.66 / 2.05	0.69 / 2.46	**0.57** / **1.54**

2) VID Dataset: We evaluate the estimates of the external force in the *sequence 17* of the VID dataset in Fig 10. In this sequence, the quadrotor is attached to a rope. Ground-truth forces are available from a force sensor attached to the other end of the rope. The force estimates are aligned to the motion-capture reference frame using the *posyaw* alignment method [33]. Our hybrid drone model has learned to compensate for a systematic residual error affecting the thrust inputs.

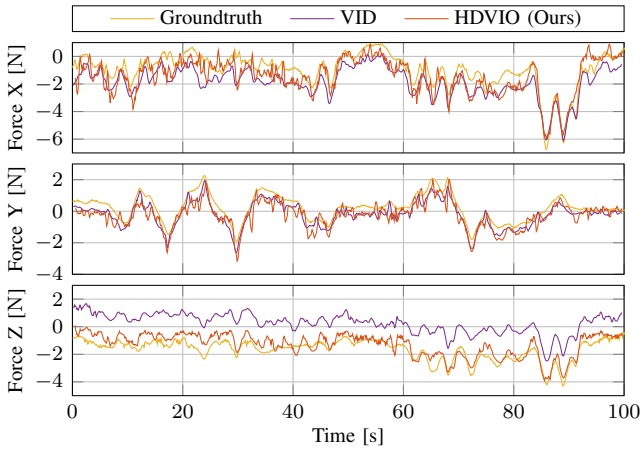
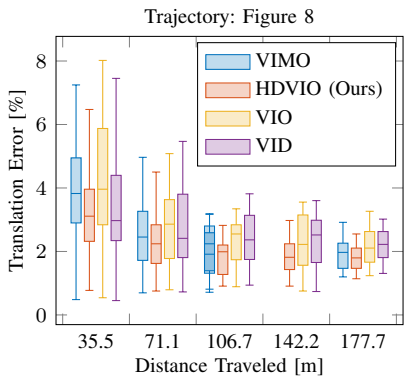


Fig. 10: Comparison of the external force estimate in *the sequence 17* of the VID dataset. Comparison of the external force estimate. HDVIO drastically improves the force estimation along the z-axis resulting in a 42% reduction of the RMSE compared to VID-Fusion.



	Method	Figure 8	Rectangle
ATE pos [m]	VIO	1.89	2.09
	VID	1.84	1.99
	VIMO	1.67	1.89
	HDVIO	**1.48**	**1.72**
ATE rot [deg]	VIO	4.02	2.01
	VID	**3.50**	2.49
	VIMO	3.70	1.92
	HDVIO	3.70	**1.86**

Fig. 11: The plot and accompanying table show how our HDVIO performs in a setting where the training data for the learning-based component is gathered from a vision-based SLAM system in the outdoor sequences of the VID dataset. The flown trajectories are at low speeds below 3 m/s, which is why all four methods show good performance, with HDVIO being the more accurate.

We believe that the cause of this error is inaccuracy in the rotor/thrust coefficients used to compute the collective thrust inputs from the rotor speed measurements. It is visible from the force estimates along the z-axis, that, VID-Fusion, similar to VIMO (see Fig. 6), is not able to compensate for this systematic error. The RMSE achieved by VID-Fusion along the z-axis is 1.95 N. The overall RMSE achieved by VID-Fusion is 1.12 N. The RMSE achieved by HDVIO along the z-axis is 0.55 N. The overall RMSE achieve by HDVIO is 0.65 N. We do not include the estimates of VIMO in Fig. 10 for the sake of readability. The ATE_T achieved by VID-Fusion is the same as the one achieved by VIO, VIMO, and, HDVIO namely 0.02 m.

For completeness, we evaluate VID-Fusion also on the outdoor sequences. We present the relative errors and the ATE_T and ATE_R in Fig. 11. In these experiments, our hybrid drone model has been trained without using an external motion-capture system. The flown trajectories are at low speeds below 3 m/s, which is why all four methods show good performance, with HDVIO being the more accurate.

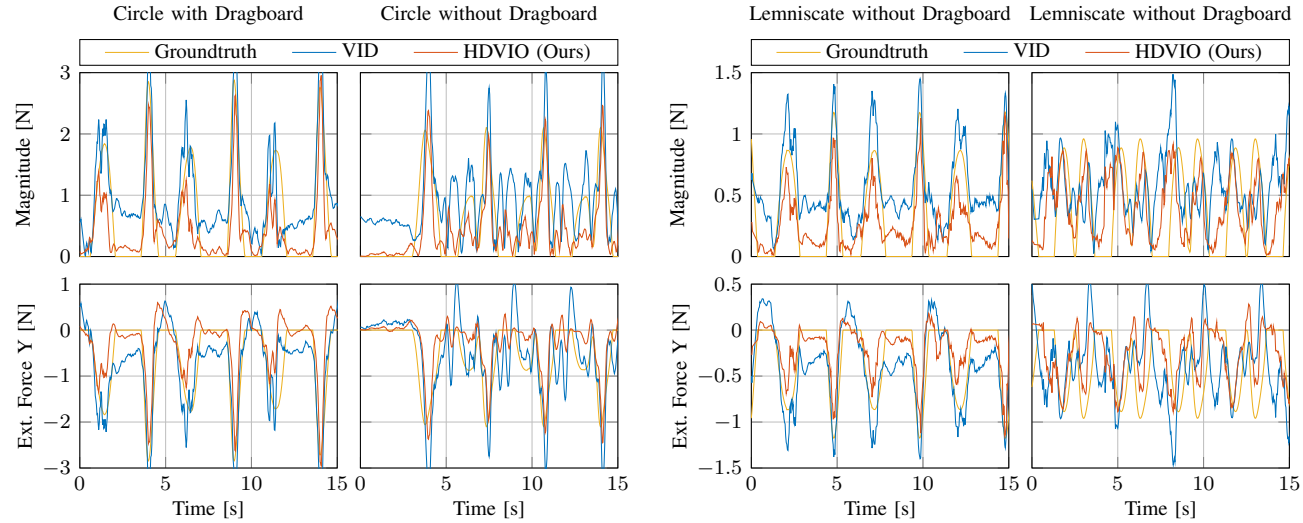


Fig. 12: Wind disturbance estimates in our real-world experiments. The magnitude and the y-axis component of the wind force estimated by HDVIO and VID-Fusion. Left: circle trajectory. Right: Lemniscate trajectory. In all the plots, it is visible that HDVIO achieves more accurate force estimates than VID-Fusion.

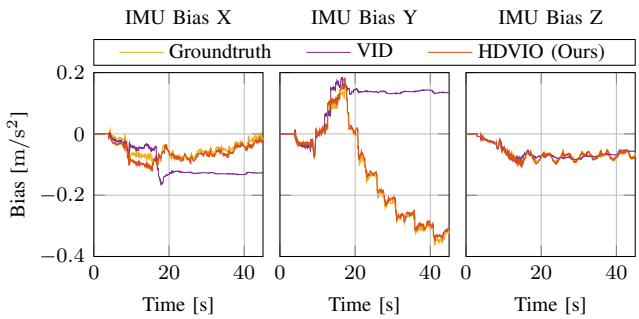


Fig. 13: Accelerometer bias estimates from our real-world experiments. The ground-truth bias is obtained from the VIO system. The estimates of our HDVIO match the ground-truth values, while, VID-Fusion estimates diverge along the x-axis.

TABLE V: Experimental results from our real-world experiments. We use (d) to indicate that a dragboard was attached to the drone.

	ATE Position [m]				ATE Rotation [deg]			
	VIO	VID	VIMO	HDVIO	VIO	VID	VIMO	HDVIO
Circle (d)	**0.07**	0.10	0.1	**0.07**	2.02	2.31	**1.80**	2.06
Circle	**0.06**	**0.06**	0.08	**0.06**	1.21	1.37	1.19	**1.17**
Lemniscate (d)	0.38	0.53	0.34	**0.30**	**2.39**	**2.39**	2.93	2.81
Lemniscate	0.27	0.28	0.32	**0.20**	2.44	2.05	1.93	**1.84**

3) Real-world Experiments: The estimates of the external force of VID-Fusion due to the wind gusts are in Fig. 12. We show the force along the y-axis, which is the direction of the wind gusts, after alignment to the world frame using the ground-truth orientations. From Fig. 12, it is clear that HDVIO outperforms VID-Fusion. Notably, VID-Fusion overestimates the wind force when the quadrotor enters the wind field. The reason for this behavior is that VID-Fusion jointly estimates the drag force with the external force.

Similar to VIMO, the measurement model in VID-Fusion introduces inconsistency in the estimates of the accelerometer bias resulting in decreased accuracy of the motion estimate. We show in Fig. 13 the accelerometer bias estimated by the VIO algorithm, by VID-Fusion, and by our method. This is done for a sequence where the quadrotor, equipped with the dragboard, flies a circle trajectory. Here, we consider the bias estimated by the VIO algorithm as a good approximation of the ground-truth values since the VIO achieves very high performance in this sequence thanks to the large number of visual features being tracked. The bias estimate of HDVIO closely matches the one of the VIO system, while the estimate of VID-Fusion converges to wrong values. We include the position and orientation absolute trajectory errors in Tab. V. In all 4 sequences, the rich texture environment simplifies the pose estimation problem. For this reason, the four algorithms achieve similar accuracy.

B. Clarification on Training with Vision-based SLAM Supervision

We show in Fig. 7, that our hybrid drone model can be trained using pose supervision from a vision-based SLAM system [34] on the VID dataset. However, the improvement in pose estimates is small compared to the baselines because these trajectories contain slow flights.

An alternative would have been to train with supervision from the vision-based SLAM system on the Blackbird dataset which contains faster flights. However, this was not possible because most of the train sequences did not include images at the time this work was carried out.