# Supplementary Material: Robust Neural Visual Inertial Odometry with Deep Velocity Constraint

Pengfei Gu, Pengkun Zhou and Ziyang Meng

Tsinghua University

## 1 Simulation Method

We formulate the low-light image synthesis process in Section V-C of the paper as a combination of brightness/contrast reduction and noise injection as done in [1]. In [1], a noise-free synthetic under-exposed image is firstly generated based on the original image by applying the linear and gamma transformations. The low-light image is generated as follows:

$$I_u = 255 \times \beta \times (\alpha \times I_i/255)^{\gamma}, \tag{1}$$

where $I_u$ and $I_i$ represent the synthetic image and the original image (both are grayscale images), respectively. $\alpha$ and $\beta$ are the linear transformation. $(\cdot)^{\gamma}$ represents the gamma transformation. In the paper, the values of $\alpha$, $\beta$ and $\gamma$ are set as 0.9, 0.5 and 5, respectively. After generating the noise-free low-light image $I_u$, the Gaussian-Poisson mixed noise is added in the in-camera image processing pipeline to simulate real low-light noise in [1]. For simplicity, we skip the in-camera processing module and directly add the Gaussian-Poisson mixed noise $n(I_u)$ on the grayscale image $I_u$. The noise $n(I_u)$ obeys a Gaussian distribution $N(0, \sigma^2(I_u))$ with its covariance defined as follows:

$$\sigma^2(I_u) = I_u \times \sigma_s^2 + \sigma_c^2, \tag{2}$$

where the values of $\sigma_s$ and $\sigma_c$ are set as 0.1 and 1, respectively.

Table 1 reports some statistics of the visual features tracked by the VINS-Mono [2] algorithm on the original and synthetic TUM_VI *outdoors*3 sequence, including the feature count per frame, the feature tracking length per feature (unit: frame count) and the feature tracking rate per frame (i.e. the percentage

**Table 1:** Statistics of visual features on the original and synthetic TUM_VI *outdoors*3 sequence. *Min.* and *std.* represent the minimum value and the stand deviation, respectively.

| | feature count | | | tracking length | | | tracking rate | | |
|---|---|---|---|---|---|---|---|---|---|
| | mean | min. | std. | mean | min. | std. | mean | min | std. |
| original sequence | 120 | 17 | 37 | 4.0 | 1 | 15 | 0.99 | 0.39 | 0.02 |
| low-light sequence | 80 | 5 | 34 | 2.3 | 1 | 8.9 | 0.98 | 0.10 | 0.06 |

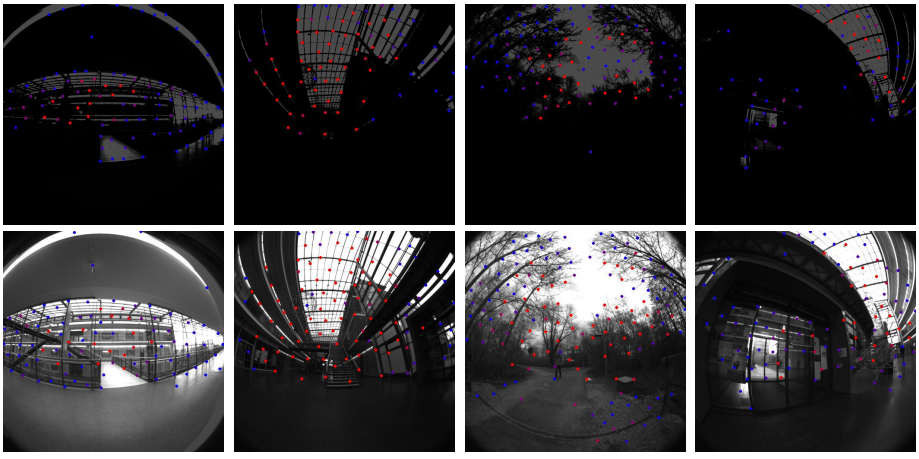

**Fig. 1:** Synthetic low-light image examples (top row) and their corresponding original images (bottom row). Best viewed in screen.

of the successfully tracked features in all features of last frame). As shown in Table 1, on the average, the feature count and the tracking length obtained on the low-light sequence are 33% and 43% lower than those obtained on the original sequence, respectively, and at worst, the minimum feature count and the tracking rate per frame are reduced to 5 and 0.10 after the low-light simulation, respectively. Fig. 1 presents some synthetic low-light image examples and their corresponding original images. The points overlaid on the images with color from blue to red indicate the visual features with the tracking length from short to long. It is shown that the distribution of the visual features obtained on the synthetic sequence is more uneven than those obtained on the original sequence, which negatively affects the accuracy and the robustness of the VIO algorithm.

## 2 Evaluation of the network on the unseen dataset

In order to better access the generalization capability of the proposed network, we test the proposed network on the RNIN and IDOL datasets. In particular, the networks trained on the IDOL or RNIN dataset are tested on another dataset to evaluate the network's performance on the unseen dataset. Table 2 lists the localization error of the proposed network on the unseen datasets and the trained datasets. It is observed that the RTE-1s on the unseen dataset is three times larger than that on the trained dataset, which quantitatively shows the proposed network's ability to generalize. Fig. 2 plots the estimated trajectory and the velocity of the proposed network on the unseen dataset. It is shown that despite the velocity predictions of the proposed network roughly follow the true velocity, there are still large localization errors between the estimated trajectories and the ground truth.

**Table 2:** Localization errors of the proposed network on the unseen datasets and the trained datasets (unit: meter).

| Training dataset | RNIN | | IDOL | |
|---|---|---|---|---|
| Testing dataset | RNIN | IDOL | IDOL | RNIN |
| ATE | **1.28** | 12.74 | **2.84** | 6.66 |
| RTE-1s | **0.091** | 0.335 | **0.118** | 0.318 |
| RTE-2s | **0.167** | 0.633 | **0.215** | 0.605 |
| RTE-4s | **0.296** | 1.173 | **0.359** | 1.144 |
| RTE-8s | **0.481** | 2.089 | **0.548** | 2.051 |

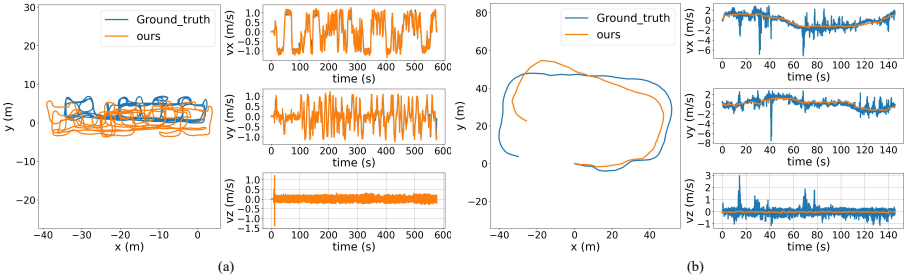


**Fig. 2:** The estimated trajectory and velocity of the proposed network on the unseen dataset. (a) the network trained on the RNIN dataset is tested on the IDOL dataset. (b) the network trained on the IDOL dataset is tested on the RNIN dataset.

## 3 Evaluation of the VIO on the unseen IMU and user

As discussed in Section V-C of the paper, we evaluate the proposed VIO algorithm, DV-VIO on an unseen dataset, the UMA-VI dataset [3] to test the generalizability. Since the UMA-VI dataset is also a pedestrian dataset, it has similar motion patterns as the training dataset (i.e. the TUM_VI and VCU-RVI dataset), which ensures the knowledge learned by our network has potential to be safely transferred without re-training. Fig. 3 shows the estimated trajectory of DV-VIO and VINS-Mono on the *long-corridor23-csc1* sequence of the UMA-VI dataset. After the VIO initialization completes, VINS-Mono encounters severe feature number reduction due to the over-exposure and then drifts. Thanks to the velocity prediction provided by the network, the proposed VIO algorithm obtains much better localization accuracy than VINS-Mono does. Note that in this experiment, we use the network trained on the TUM_VI and VCU-RVI datasets (i.e. the same network trained in Section V-C of the paper) and directly test it on the UMA-VI dataset. Thus, the above experiment demonstrates that the proposed method can to some extent generalize to an unseen pedestrian dataset with unseen IMU sensors and users which are not presented in the training dataset with a degraded accuracy.

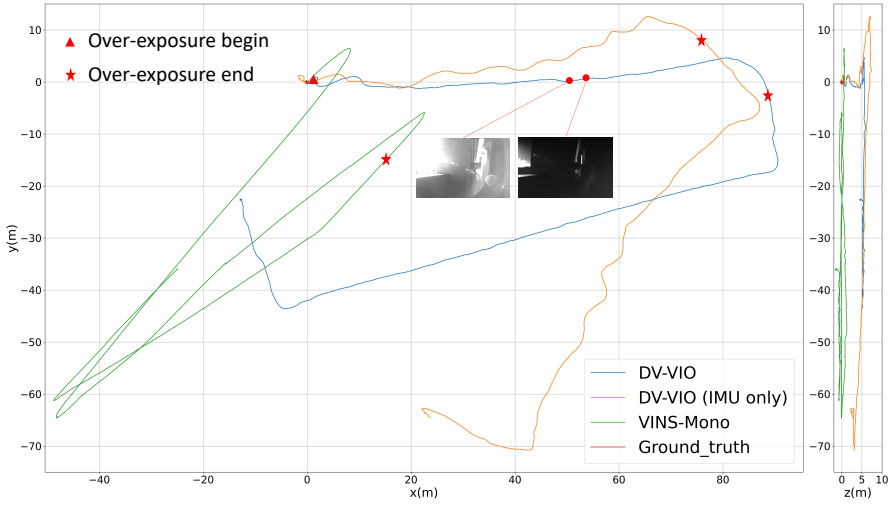


**Fig. 3:** Trajectory of the VIO algorithms on the UMA-VI dataset. Note that the ground truth is only available for the beginning and the ending part of the trajectory (very limited part around the *triangle* mark), and the true trajectory should start and end at the same place. The *triangle* and the *star* marks identify the beginning and the end of the period where the VIO algorithms encounter severe visual challenges due to the over-exposure and frequent brightness change. The orange trajectory is estimated by DV-VIO without the image input, which only uses the network velocity prediction and the raw IMU measurement for localization.

## References

1. Lv, F., Li, Y., Lu, F.: Attention guided low-light image enhancement with a large scale low-light simulation dataset. International Journal of Computer Vision **129**(7), 2175–2193 (2021) 1
2. Qin, T., Li, P., Shen, S.: Vins-mono: A robust and versatile monocular visual-inertial state estimator. IEEE Transactions on Robotics **34**(4), 1004–1020 (2018). https://doi.org/10.1109/TRO.2018.2853729 1
3. Zuñiga-Noël, D., Jaenal, A., Gomez-Ojeda, R., Gonzalez-Jimenez, J.: The uma-vi dataset: Visual–inertial odometry in low-textured and dynamic illumination environments. The International Journal of Robotics Research **39**(9), 1052–1060 (2020) 3