

U-ARE-ME: Uncertainty-Aware Rotation Estimation in Manhattan Environments

Supplementary Material

7. Training details for surface normal estimator

Our surface normal estimator is trained on the meta-dataset introduced by [2]. DSINE [2] uses per-pixel ray direction as input and thus requires camera intrinsics. We removed this dependency and warped the training images such that the principal point is at the center and the field-of-view (FOV) is 60° . Nonetheless, U-ARE-ME generalises well to images taken with different intrinsics (e.g. the video game sequence in the attached video has FOV of 86°). DSINE [2] also proposed to improve the piece-wise smoothness and crispness of the prediction by recasting surface normal estimation as iterative rotation estimation. We also remove this iterative process and hence improve the efficiency to give real-time estimates. While this degrades the quality of the surface normal prediction, the rotation estimates from our framework stay robust. U-ARE-ME robustly fuses the per-pixel predictions by weighting them with a confidence κ and is thus robust to mild inaccuracies in the surface normal prediction.

Our network estimates two quantities: the per-pixel surface normal vector \mathbf{n} and the corresponding confidence κ . \mathbf{n} is supervised with the angular loss, and κ with the negative log-likelihood defined in Eq. 2 (in the main manuscript). All other training protocols (e.g. batch size, number of epochs, data augmentation, optimiser, and learning-rate schedulers) are the same as [2]. The training only takes 9 hours on a single NVIDIA 4090 GPU. After training, κ is capped at 100 to prevent the over-confident predictions from dominating the optimisation.

8. Robustness to out-of-distribution camera intrinsics

To show that our algorithm is robust to a wide variety of out-of-distribution images, a brief study is performed with varying camera intrinsics on the same sequence. We re-evaluate on the ICL-NUIM living room 2 sequence but change the focal length and principal point of the input images, as shown in Fig. 5. By performing an increasingly aggressive central crop and then resizing back to the original resolution, a narrowing of the FOV is achieved while keeping a constant principal point. In the shift tests, both the focal length and the principal point vary by cropping from the bottom right corner and resizing.

Tab. 3 shows the accuracy comparison between U-ARE-ME, H-VP and ORB-SLAM (intrinsics given to ORB-

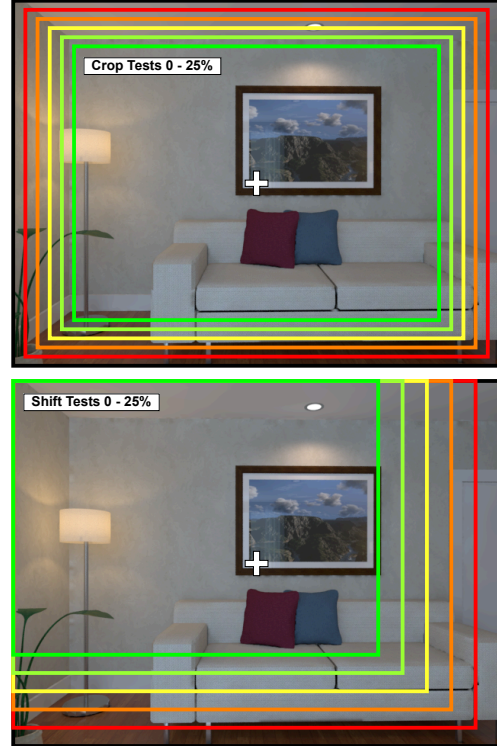


Figure 5. Robustness to out-of-distribution images. Crop tests simulate a narrowing focal length with a constant principal point and shift tests simulate both focal length and principal point change. Images are resized to their original resolution after cropping.

Table 3. Results from modified intrinsics tests.

Crop	Ours	H-VP	ORB-SLAM
Original	2.39	4.17	0.57
5% crop	2.29	4.57	2.64
10% crop	2.26	5.80	4.89
15% crop	2.26	6.58	5.64
20% crop	2.34	7.40	8.59
25% crop	2.62	8.50	13.34
Shift	Ours	H-VP	ORB-SLAM
original	2.39	4.17	0.57
5% shift	2.12	6.17	4.35
10% shift	2.50	7.51	10.27
15% shift	3.53	9.25	17.01
20% shift	4.99	10.65	26.73
25% shift	6.41	12.17	×



Figure 6. IMU sensors are prone to drift especially in non-inertial frames of reference (e.g. inside a moving vehicle). The horizon line in each image represents the *up-vector* inferred from our proposed method (middle) and an IMU sensor (right).

SLAM are kept constant throughout). Since our normal network does not need specific camera intrinsics, we are extremely robust to both image cropping and image shifting – more so for the former. Due to the network being trained on $\sim 60^\circ$ FOV images, a minor improvement is observed at ~ 10 -15% cropping as the ICL-NUIM images have a slightly wider than ideal 67° FOV. As expected, ORB-SLAM rapidly deteriorates with incorrect intrinsics. Despite H-VP not requiring known intrinsics, a narrower FOV in general reduces the number of sparse line features and therefore accuracy degrades proportionally (a major upside to dense predictors).

9. Horizon estimation in Non-inertial Frames of Reference

Motivated by the limitations of using Inertial Measurement Units (IMUs) when operating within a non-inertial reference frame, we apply U-ARE-ME to a real-world video sequence captured from a bus with longitudinally and laterally accelerating motions. We perform multi-frame rotation estimation for the sequence of RGB video frames, and perform a comparison against IMU data. For each frame, we compute the ARE between the estimated rotation and the initial rotation, and plot the evolution over time in Fig. 6.

The camera is stationary with respect to the bus throughout the sequence so there should be no relative rotation. The Intel Realsense D435i camera was used to take synchronised RGB and IMU measurements which were integrated using the complementary filter provided by the Intel Realsense SDK [22].

As the bus turns harshly, our visual-only approach maintains a steady rotation estimate, while the IMU suffers from strong accelerations causing a large error in its measured rotation. This is seen in Fig. 6 where the green horizon lines for each method have been calculated using the *up-vector* derived from the estimated rotation. We envisage that

U-ARE-ME can be used to complement traditional IMU-based applications in non-inertial reference frames.

10. Ground segmentation with U-ARE-ME

The *up-vector* can be used to perform real-time ground segmentation from RGB images, which is an important pre-processing step in many applications including robotics [36], autonomous driving [12], and 3D object tracking for augmented reality [46], where it can be used to seamlessly integrate virtual objects with the real-world environment.

As U-ARE-ME produces per-pixel surface normal estimates, we can directly use the result of the rotation esti-



Figure 7. Even in non-Manhattan scenes, our framework can optimise the rotation such that the global *up* direction is aligned with the surface normal of the ground plane. It can thus be used to accurately segment the ground plane, shown here in green, which can be useful for many applications in robotics.

Table 4. Timestamps for the contents of the demo video

Section		Timestamp
U-ARE-ME Demo	ICL-NUIM	0:13
	TUM-RGBD	0:30
	Tokyo walking sequence	0:47
	Video game sequence	1:05
Robustness to Bad Frames		1:29
Applications	Non-inertial Reference Frame	1:58
	Ground Segmentation	2:27

mation to segment areas of an input image corresponding to the ground, assuming that this is aligned with the world up-vector. Our method can be applied to real-world indoor and outdoor images to segment the ground even when the scene is non-Manhattan, as seen in Fig. 7.

11. Demo video

We encourage the reader to view the accompanying video file in the supplementary material for a visual overview of our method. The video can also be found on our project page².

For the reader’s convenience, the timestamps of each section in the video are summarised in Table 4. We provide additional detailed explanations for each section below.

11.1. U-ARE-ME demo

We first demonstrate the operation of U-ARE-ME on the ICL-NUIM and TUM-RGBD datasets discussed in our paper. The coordinate frame in the center of the video depicts the orientation of the global Manhattan frame. Throughout the demonstration, the video cycles through various visualisations of the scene:

- RGB image input to U-ARE-ME
- Predicted surface normals (using X-Y-Z to R-G-B colour mapping)
- Confidence of predicted surface normals (greyscale – white represents high confidence)

11.1.1 ICL-NUIM: living-room-2

This synthetic scene shows a camera moving through a living room which is generally Manhattan in structure, but does contain some features which do not agree with a Manhattan assumption (e.g. curtains, lamps, and sofa cushions). Of note is the textured wall painting at 0:13 which is predicted as having the same surface normal as the wall it is on, while the painting’s frame is predicted to have high uncertainty normals. The curtains seen at 0:21 (a large non-Manhattan area) are also shown with high uncertainty

normals as they have irregular geometry. As U-ARE-ME is uncertainty-aware, it estimates rotation with a greater weighting on those surfaces that agree with the Manhattan assumption, e.g., the walls and floor (shown in white on the confidence images), whilst down-weighting the non-Manhattan features.

11.1.2 TUM-RGBD: fr-3 large cabinet

This dataset contains videos captured with a real-world hand-held camera, exhibiting some pitch and roll with unsteady camera motion. The normals of objects in the background with fine structures and lots of occlusion are harder to accurately predict, and thus are estimated as having surface normals with higher uncertainty. As a result, our method can produce accurate estimates of rotation throughout the sequence by down-weighting such uncertain normals.

11.1.3 Tokyo Walking Sequence

We apply our method to an *in-the-wild* video taken directly from YouTube [49] showing a hand-held camera viewpoint walking through the streets of Tokyo. This is a challenging situation for rotation estimation as the camera intrinsics are not known. The environment is dynamic, with many pedestrians walking in the scene, and our method produces reliable rotation estimation despite the small quantity of static Manhattan-aligned objects and buildings.

11.1.4 Video Game sequence

In this example, our method is shown to estimate rotation on a synthetic sequence from the video game Star Citizen [50]. Once again the camera intrinsics are not known, and this is a relatively non-Manhattan environment. During the sequence, the camera switches between 1st person and 3rd person (during which the player character takes up a significant portion of the screen) yet rotations remain consistent with the game world. U-ARE-ME takes this into consideration by estimating a high uncertainty on the player model.

²Project Page: <https://callum-rhodes.github.io/U-ARE-ME>

11.2. Robustness to dropped frames

We compare U-ARE-ME operating **with** (bottom videos) and **without** (top videos) multi-frame optimisation, when black frames are randomly injected into the sequence, simulating dropped frames.

For example, at 01:50 the addition of a black frame cause the non-multi-frame rotation estimate (top) to differ significantly from its previous estimate. Using our proposed uncertainty-aware multi-frame optimisation however, U-ARE-ME is seen to produce temporally consistent rotation estimates, in the presence of dropped frames.

11.3. Applications

Finally, we demonstrate some applications of U-ARE-ME as discussed in the paper.

11.3.1 Operation in a Non-inertial Reference Frame

This section of the demo video shows the benefit of applying U-ARE-ME to a real-world video sequence captured in a non-inertial reference frame for estimating a horizon as discussed in section 9. The video shows an outward view demonstrating the motion of the bus (left), and the horizon lines estimated by U-ARE-ME (middle) and based on IMU measurements (right).

11.3.2 Ground Segmentation

Further to section 10, we apply U-ARE-ME to a video taken directly from YouTube [48]. Without knowledge of the video’s camera intrinsics, and in the presence of blurring noise, our method is able to estimate the up-vector and highlight the ground-aligned pixels in green.