VERF: Runtime Monitoring of Pose Estimation with Neural Radiance Fields

Dominic Maggio, Courtney Mario, Luca Carlone

***Abstract*—We present VERF, a collection of two methods (VERF-PnP and VERF-Light) for providing runtime assurance on the correctness of a camera pose estimate of a monocular camera without relying on direct depth measurements. We leverage the ability of NeRF (Neural Radiance Fields) to render novel RGB perspectives of a scene. We only require as input the camera image whose pose is being estimated, an estimate of the camera pose we want to monitor, and a NeRF model containing the scene pictured by the camera. We can then predict if the pose estimate is within a desired distance from the ground truth and justify our prediction with a level of confidence. VERF-Light does this by rendering a viewpoint with NeRF at the estimated pose and estimating its relative offset to the sensor image up to scale. Since scene scale is unknown, the approach renders another auxiliary image and reasons over the consistency of the optical flows across the three images. VERF-PnP takes a different approach by rendering a stereo pair of images with NeRF and utilizing the Perspective-n-Point (PnP) algorithm. We evaluate both methods on the LLFF dataset, on data from a Unitree A1 quadruped robot, and on data collected from Blue Origin's sub-orbital New Shepard rocket to demonstrate the effectiveness of the proposed pose monitoring method across a range of scene scales. We also show monitoring can be completed in under half a second on a 3090 GPU.**

I. Introduction

Estimating the pose of a camera from a monocular image is a fundamental problem in computer vision. However, limited work has been done to independently monitor the accuracy of the estimated pose and detect incorrect estimates without having direct access to depth information of the scene. This need is motivated by the growing use of monocular camera localization in high-stakes scenarios such as self driving [1], spacecraft entry decent and landing [2], [3], [4], and robotics tasks [5]. For instance, the detection of repeatedly incorrect estimates can be used to decide when to alert the user or trigger mitigation measures (*e.g.,* performing a safety landing for a drone, or disengaging the autopilot of a self-driving car).

Recent works such as [6], [7], [8], [9], [10], [11] explore the use of NeRF [12] (Neural Radiance Fields) for camera pose estimation. NeRFs are fully connected networks trained on sequences of RGB images to learn an implicit representation of a scene, from which the network can be used to generate RGB images at novel viewpoints. As an example, Loc-NeRF [8] uses NeRF as a map of an environment and utilizes a particle filter backbone to output a pose estimate of a provided sensor image. However, there is no clear and reliable measure to determine if the outputted pose is correct —where we define correct as being within a distance ϵ of the true pose— and existing approaches can fail without notice.

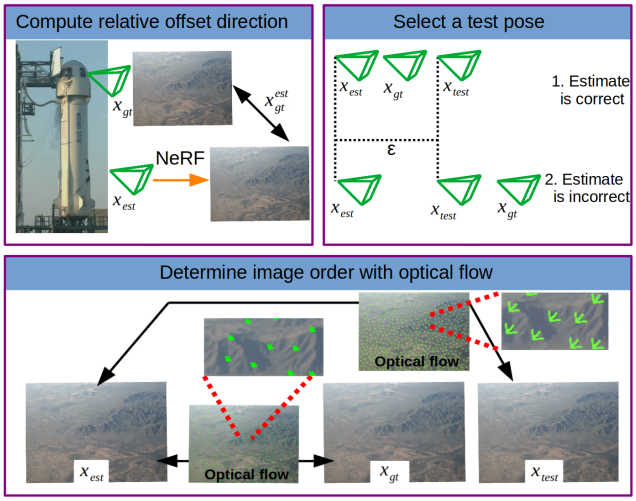


Fig. 1. Three main phases of VERF-Light. First, the relative error of a pose estimate up to scale is found by comparing a sensor image (collected at the ground truth pose, $\boldsymbol{x}_{gt}$,) to a NeRF image rendered at the pose estimate, $\boldsymbol{x}_{est}$. Next, a test pose, $\boldsymbol{x}_{test}$, is selected at an ϵ distance from the estimated pose such that all three poses are co-linear. Determining if the pose estimate is correct is lastly done by estimating the order of the three poses by comparing optical flow between the three corresponding images.

To overcome this limitation, we propose VERF, a collection of two approaches coined VERF-PnP and VERF-Light. VERF uses the sensor image already present in the pose optimization phase to provide assurance that the pose estimate is correct. We additionally require a NeRF model of the scene, but NeRF does not need to be used to produce the pose estimate being monitored, which allows VERF to be used for pose monitoring regardless of the pose estimation method. VERF-PnP renders a stereo pair of images with NeRF, one of which is at the estimated pose and the other at a given baseline, and uses the Perspective-n-Point (PnP) solver with RANSAC [13] to estimate the relative offset to the sensor image. VERF-Light uses a different methodology which can be stated concisely as follows. We first render an image with NeRF at the estimated pose, $\boldsymbol{x}_{est}$, and use it to determine the relative translation up to scale between the estimated pose and the ground truth pose, $\boldsymbol{x}_{gt}$. To overcome scale ambiguity we render a test image at a pose $\boldsymbol{x}_{test}$ which is at a distance ϵ from the estimate pose in the direction of the

D. Maggio is with the Laboratory for Information & Decision Systems, Massachusetts Institute of Technology, Cambridge, MA, USA, and is a Draper Scholar with the Perception and Embedded ML Group, Draper, Cambridge, MA, USA, drmaggio@mit.edu

C. Mario is with the Draper Perception and Embedded ML Group, Draper, Cambridge, MA, USA, cmario@draper.com

L. Carlone is with the Laboratory for Information & Decision Systems, Massachusetts Institute of Technology, Cambridge, MA, USA, lcarlone@mit.edu

This work was partially funded by the NASA Flight Opportunities under grant Nos 80NSSC21K0348 and 80NSSC20K0104.

sensor image. If the camera origin of these three images are co-linear with no rotation, then we show that we can compare optical flow fields between the three images to determine the order of the camera centers and hence the correctness of the pose estimate (Fig. 1). To enable assurance in the presence of noise, we incorporate an estimate of optical flow error and add outlier rejection using geometric constraints to compute a measure of confidence instead of a binary decision. We remark that as the rotation error can be directly observed between the sensor image and the image rendered at the estimated pose, we only focus our attention on determining the quality of the position estimate. We provide results on the publicly available LLFF dataset [14], on real data collected by an A1 quadruped, and on data collected onboard Blue Origin's sub-orbital New Shepard rocket at heights up to 8 km above the ground and at speeds over 800 km/hr. The results showcase the potential of VERF to perform in challenging real-world conditions. Our runtime monitoring approach runs in less than half a second on a 3090 GPU.

The rest of the paper is organized as follows. Section II discusses related work. Section III provides notation and preliminary concepts. Our two approaches are presented in Sections IV and V. Section VI evaluates the methods on three types of experiments: LLFF, A1 robot, and sub-orbital rocket. Finally, Section VII concludes the paper. Extra results and studies are included in the appendix (Section VIII).

II. Related Work

Neural Radiance Fields. NeRF was introduced by Mildenhall *et al.* [12] and represents a 3D scene with a neural implicit encoding that can be used to render novel viewpoints of the scene. Several extensions are beginning to leverage NeRF for robotic tasks such as localization. Yen *et al.* [6] develop iNeRF which inverted the NeRF paradigm by solving for a pose given an image. Adamkiewicz *et al.* [7] develop NeRF-Navigation which uses NeRF for a full autonomy pipeline of localization, planning, and control. Zhu *et al.* [10] propose LATITUDE to perform pose estimation with large-scale scenes. Maggio *et al.* [8] develop Loc-NeRF which uses a particle filter backbone and performs localization while using NeRF as a map. Lin *et al.* [9] use parallelized Monte Carlo Sampling to estimate camera poses. Rosinol *et al.* [15] develop NeRF-SLAM which builds a NeRF as images and poses become available. Sucar *et al.* [16] proposes iMAP and Zhu *et al.* [17] develop NICE-SLAM which use depth from a stereo camera along with RGB to create a neural implicit map of room-size scenes. Alignment accuracy of NeRF is studied and improved by Jiang el al. [18]. Moreau *et al.* [19] develop CROSSFIRE which uses PnP for localization with NeRF by training self-supervised feature descriptors and rendering depth directly from a neural renderer. Li *et al.* [20] develop NeRF-Pose which uses PnP with NeRF for object pose estimation by training a pose regression network to predict 2D-3D correspondences.

Visual Localization. Since VERF can monitor the accuracy of a pose estimate independent of the estimation method, we also include a brief mention of visual localization methods outside the scope of NeRF. Classical methods for robotic localization typically use either matching of sparse keypoints [21], [22], [23] or a dense representation [24]. Visual terrain relative navigation is the problem of estimating the pose of a camera given a terrain map (oftentimes built with satellite or aerial imagery and elevation data) [25], [2], [26], [4]. Absolute Pose Regressors [27] use Convolutional Neural Networks to predict poses by learning the end-to-end localization pipeline. We refer to Piasco *et al.* [28] for a more in-depth review of visual localization.

Certifiable Perception and Runtime Monitoring. Carlone and Dellaert [29] and Rosen *et al.* [30] develop optimality certification techniques for pose synchronization problems. Yang *et al.* [31], [32] develop certifiable algorithms for outlier robust estimation. Garcia-Salguero *et al.* [33], [34] certify the optimality of a relative pose estimate. Zhao *et al.* [35] present a certifiably optimal approach to estimate the generalized essential matrix. Here, we instead focus on monitoring the correctness of the pose estimate, rather than optimality of the estimation backend. Yang *et al.* [31] and Carlone [36] develop estimation contracts which certify the correctness of a geometric perception problem given conditions are met on the inputs. Talak *et al.* [37] extend certification of correctness for learning-based object pose estimation. Yang and Pavone [38] provide statistical bounds on object pose estimation given a heatmap predictions of object keypoints. Other works provide confidence metrics to monitor the correctness of perception algorithms without providing a certificate of correctness. Hu and Mordohai [39] provide a survey on confidence metrics for stereo matching. Rahman *et al.* [40] provide a survey on monitoring the correctness of learning-based methods for robotic perception. Antonante *et al.* [41] use a diagnostic graph to formalize detecting and identifying faults in a perception system.

III. Notation and Preliminaries

Notation. We use lowercase symbols (e.g., ϵ) to represent scalars, bold lowercase letters (e.g., $\boldsymbol{x}$) for vectors, and bold uppercase letters (e.g., $\boldsymbol{E}$) for matrices. Sets are represented with capital calligraphic fonts (e.g., $\mathcal{R}$). Unit vectors and homogeneous vectors are denoted with a bar and tilde (e.g., $\bar{\boldsymbol{x}}$ and $\tilde{\boldsymbol{x}}$) respectively. Estimated quantities are shown with a caret (e.g., $\hat{\boldsymbol{x}}$, $\hat{\boldsymbol{E}}$). We express the 2-norm of a vector as $\|\cdot\|$.

Let $\boldsymbol{r}_i = (x, y)$ be a coordinate in an image I_i. The sensor image will be referred to as I_{gt} as it is taken by a camera at the true pose. The estimated and test images will be referenced as I_{est} and I_{test}. Let $\boldsymbol{v}(\boldsymbol{r}_i)_{I_i,I_j}$ be the optical flow vector at point $\boldsymbol{r}$ in some image i to the corresponding point in some image j such that $\boldsymbol{r}_i + \boldsymbol{v}(\boldsymbol{r}_i)_{I_i,I_j} = \boldsymbol{r}_j$. $[\boldsymbol{a}]_\times$ is the skew-symmetric matrix such that $\boldsymbol{a} \times \boldsymbol{b} = [\boldsymbol{a}]_\times \boldsymbol{b}$.

The Essential Matrix. Assuming points have been calibrated using the camera intrinsics, the essential matrix $\boldsymbol{E}_{i,j}$ relates corresponding homogeneous coordinates $\tilde{\boldsymbol{r}}_i$, $\tilde{\boldsymbol{r}}_j$ in two images with the following constraint:

$$(\tilde{\boldsymbol{r}}_j)^T \boldsymbol{E}_{i,j} \tilde{\boldsymbol{r}}_i = 0. \tag{1}$$

The matrix $\boldsymbol{E}_{i,j}$ describes the relative pose transform between two cameras defined with a rotation matrix $\boldsymbol{R}$ and translation $\boldsymbol{t}$ up to scale as:

$$\boldsymbol{E}_{i,j} = \boldsymbol{R}[\boldsymbol{t}]_\times. \tag{2}$$

Decomposing $\boldsymbol{E}$ to recover $\boldsymbol{t}$ and $\boldsymbol{R}$ yields four solutions, of which only one satisfies the cheiral inequalities [42] which in summary state that triangulated points must lie in front of the two cameras. Since eq. (1) does not restrict scale, $\boldsymbol{E}_{i,j}$ along with a point $\boldsymbol{r}_i$ constrains a corresponding point $\boldsymbol{r}_j$ in I_j to a line known as the *epipolar* line.

Problem statement. Our objective is to determine if a given position estimate $\boldsymbol{x}_{est}$ is within some acceptable error bound, ϵ, from the true position $\boldsymbol{x}_{gt}$:

$$\|\boldsymbol{x}_{est} - \boldsymbol{x}_{gt}\| < \epsilon. \tag{3}$$

All we assume are available is the position estimate $\boldsymbol{x}_{est}$, the sensor image I_{gt}, and a NeRF model whose weights are trained on a scene containing I_{gt}.

IV. VERF-PnP

Here we present a simple yet effective method to estimate the correctness of a pose estimate using NeRF. We leverage NeRF to render a pair of stereo images to perform PnP. We first render an image I_{est} at the estimated pose $\boldsymbol{x}_{est}$. Since the true pose $\boldsymbol{x}_{gt}$ is by definition the camera position corresponding to I_{gt}, the correctness constraint in eq. (3) can be satisfied by showing that the metric offset between $\boldsymbol{x}_{gt}$ and $\boldsymbol{x}_{est}$ is less than ϵ. Towards this goal, we render a second image I_{right} at $\boldsymbol{x}_{right}$ by translating 2ϵ to the right with respect to $\boldsymbol{x}_{est}$. The image pair I_{est} and I_{right} whose poses are both known can then be used as a classical stereo pair of images. We compute the optical flow between these two images using RAFT [43] and use good features to track [44] to get sparse optical flow from RAFT's dense optical flow field. Likewise, we find the correspondences between I_{est} and I_{gt} for the same sparse points with RAFT. We then triangulate the 3D location of the sparse points by knowing $\boldsymbol{x}_{est}$ and $\boldsymbol{x}_{right}$ and finally apply PnP with RANSAC [13] to estimate the transform $\hat{\boldsymbol{x}}_{gt}^{est}$ between $\boldsymbol{x}_{est}$ and the unknown $\boldsymbol{x}_{gt}$. Our level of confidence in the accuracy of $\boldsymbol{x}_{est}$ is then estimated as follows:

$$\mathbb{P}(\|\hat{\boldsymbol{x}}_{gt}^{est}\| < \epsilon). \tag{4}$$

We model $\|\hat{\boldsymbol{x}}_{gt}^{est}\|$ as a random variable whose mean value is the estimated position from PnP and standard deviation is manually selected. We will show in Section VI the effectiveness of VERF-PnP despite its simplicity.

V. VERF-Light

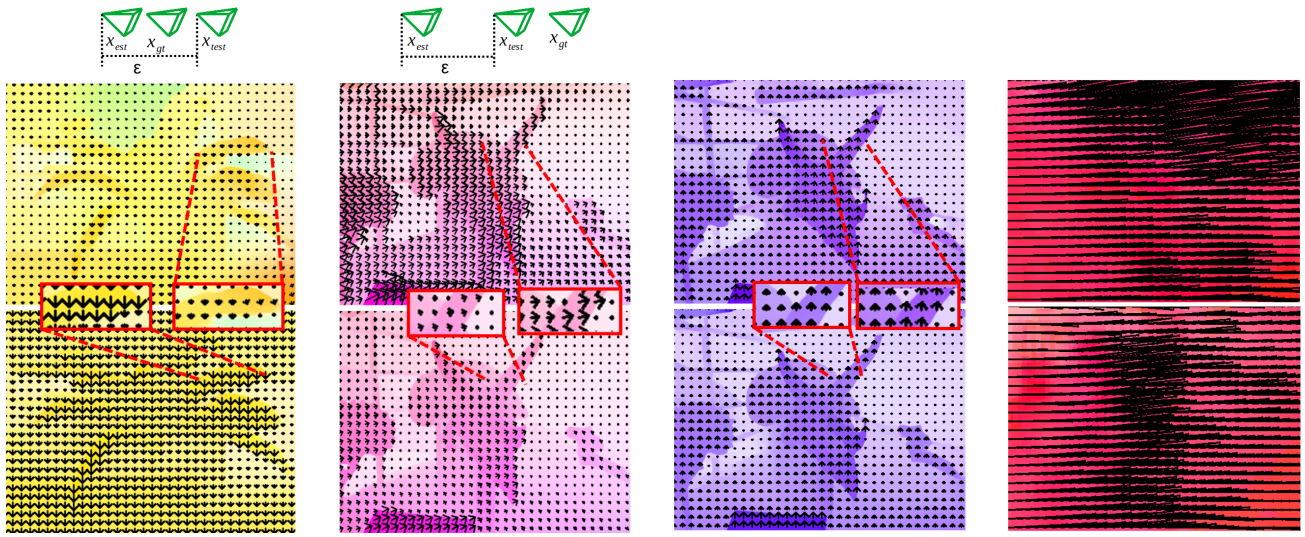
VERF-Light can be divided into three phases (Fig. 1): computing the relative offset between $\boldsymbol{x}_{est}$ and $\boldsymbol{x}_{gt}$ up to scale, selecting a test position $\boldsymbol{x}_{test}$ distance ϵ from $\boldsymbol{x}_{est}$ and co-linear with the latter two poses, and computing a quality of assurance that eq. (3) is met by using an application of the cheiral constraint. In particular, we leverage the fact that given three images from camera poses that are co-linear and with the same rotation, their order along the line they belong to can be determined by comparing the optical flow fields between them. For this arrangement, the flow fields between I_{est} and I_{gt} will be in the same direction as the flow field between I_{est} and I_{test}, and the order of the three positions $\boldsymbol{x}_{est}$, $\boldsymbol{x}_{gt}$, and $\boldsymbol{x}_{test}$ can be estimated by comparing the magnitude of corresponding vectors between the two flow fields. If $\boldsymbol{x}_{gt}$ falls between $\boldsymbol{x}_{est}$ and $\boldsymbol{x}_{test}$ in such ordering, we can conclude that the error of $\boldsymbol{x}_{est}$ is less than ϵ.

Examples and intuition. Figure 2 shows four example conditions that VERF-Light could potentially encounter. In Fig. 2a the flow field should provide confidence that the estimated pose is correct. First, the two optical flow fields have similar directions (and hence the same epipole) which validates our assumption of $\boldsymbol{x}_{est}$, $\boldsymbol{x}_{gt}$, and $\boldsymbol{x}_{test}$ being co-linear. Secondly, the magnitude of the optical flow between I_{est} and I_{test} (which have camera centers ϵ apart) is significantly greater than the corresponding flow between I_{est} and I_{gt} meaning that $\boldsymbol{x}_{gt}$ falls between $\boldsymbol{x}_{est}$ and $\boldsymbol{x}_{test}$ and hence the estimated pose is within ϵ of the true pose. In Figure 2b, the estimate can safely be labeled as incorrect as there is clear evidence from the flow field that the flow between I_{est} and I_{test} is less in magnitude than the flow field between I_{est} and I_{gt} and again that the three perspectives are co-linear. Figure 2c on the other hand does not allow drawing strong conclusions. In this case there should be reduced confidence in the correctness assessment as the flow field is roughly the same and differences may be only the result of noise. Figure 2d should be determined to be an incorrect pose but because of a different reason than Fig. 2b — here a cue that the pose is wrong is because no clear correspondences can be found between I_{est} and I_{gt}.

A. Computing Relative Error Direction of Position Estimate

We use NeRF along with $\boldsymbol{x}_{est}$ to render an image I_{est} which is the image that the camera would see if its center were at $\boldsymbol{x}_{est}$. We use RAFT to compute the dense optical flow between I_{gt} and I_{est}, and use good features to track [44] to extract a set $\mathcal{R}$ of n pixel coordinates $\boldsymbol{r}_{est}$ to get sparse optical flow between the two images.

We use the 5-point algorithm [45] with RANSAC to determine the essential matrix $\boldsymbol{E}_{est,gt}$. RANSAC attempts to search for a $\hat{\boldsymbol{E}}_{est,gt}$ such that a maximum number of points in $\mathcal{R}$ have sampson distance (a geometric constraint related to eq. (1)) less than δ. In short, the sampson distance [46] is an approximation of error to the epipolar line for two corresponding points. The unique solution to extracting the relative position $\hat{\hat{\boldsymbol{x}}}_{gt}^{est}$ up to scale from $\hat{\boldsymbol{E}}_{est,gt}$ is found using the cheiral constraints with maximum consensus. Any points whose correspondence are not part of the maximum consensus or whose sampson distance is larger than δ are removed from the set of inliers $\mathcal{R}$ reducing the set of points to $\boldsymbol{r}_{est} \in \mathcal{R}' \in \mathcal{R}$ where n' is the number of points currently labeled as inliers.



(a) True position error is 1.9 cm. The flow field should allow for concluding that the pose estimate is correct with high confidence. Order of camera positions shown above.

(b) True position error is 10 cm. The flow field should allow for concluding that the pose estimate is incorrect with high confidence. Order of camera positions shown above.

(c) True position error is 5.6 cm. Pose can potentially be verified as correct, but should be done with low confidence.

(d) True position error is much larger than ϵ such that there are no clear correspondences between images.

Fig. 2. Example of optical flow between I_{est}, I_{gt}, and I_{test} for pose estimates with a correctness condition of $\epsilon = 5cm$. **Top row**: optical flow between I_{est} and I_{gt}. **Bottom row**: optical flow between I_{est} and I_{test}.

B. Computing Location of Test Position

We now calculate a test position, $\boldsymbol{x}_{test}$, that is distance ϵ from $\boldsymbol{x}_{est}$ and co-linear with I_{est} and I_{gt}:

$$\boldsymbol{x}_{test} = \boldsymbol{x}_{est} + \epsilon \hat{\vec{\boldsymbol{x}}}_{gt}^{est}. \tag{5}$$

The correctness condition in (3) can now be stated as:

$$\|\boldsymbol{x}_{est} - \boldsymbol{x}_{gt}\| < \|\boldsymbol{x}_{est} - \boldsymbol{x}_{test}\| = \epsilon \tag{6}$$

where the exact pose of $\boldsymbol{x}_{est}$ and $\boldsymbol{x}_{test}$ are known and chosen to be ϵ apart. Note that since the positions are collinear by construction, the condition $\|\boldsymbol{x}_{est} - \boldsymbol{x}_{gt}\| < \|\boldsymbol{x}_{est} - \boldsymbol{x}_{test}\|$ is the same as requiring that these positions are ordered as $\boldsymbol{x}_{est}, \boldsymbol{x}_{gt}, \boldsymbol{x}_{test}$ along the line they belong to. We render a new image I_{test} at $\boldsymbol{x}_{test}$ using NeRF.

C. Determining the Confidence Score

We again use RAFT to compute the dense optical flow, this time between I_{est} and I_{test} and get sparse optical flow $\hat{\boldsymbol{v}}(\boldsymbol{r}_{est})_{I_{est},I_{test}}$ for coordinates $\boldsymbol{r}_{est} \in \mathcal{R}'$.

We now consider several properties given our particular choice of $\boldsymbol{x}_{test}$. The first is that it is unnecessary to compute $\boldsymbol{E}_{est,test}$ as we directly know it without error from the true poses of $\boldsymbol{x}_{est}$ and $\boldsymbol{x}_{test}$. Furthermore, it is simply the same as our estimate of $\boldsymbol{E}_{est,gt}$ since $\boldsymbol{x}_{est}$, $\boldsymbol{x}_{gt}$, and $\boldsymbol{x}_{test}$ are aligned and co-linear. This is summarized in the following relation:

$$\hat{\boldsymbol{E}}_{est,gt} = \boldsymbol{E}_{est,test}. \tag{7}$$

Determining whether eq. (6) is satisfied now reduces to solving an image ordering problem for $I_{est}, I_{gt}, I_{test}$ outlined visually in Fig. 1. If $\mathcal{R}'$ contains only true, noiseless inliers, the image ordering problem could now be solved using an application of the cheiral constraint:

$$\begin{aligned} &\|\boldsymbol{x}_{est} - \boldsymbol{x}_{gt}\| < \epsilon \iff \\ &\forall \boldsymbol{r}_{est} \in \mathcal{R}', \|\boldsymbol{v}(\boldsymbol{r}_{est})_{I_{est},I_{gt}}\| < \|\boldsymbol{v}(\boldsymbol{r}_{est})_{I_{est},I_{test}}\| \end{aligned} \tag{8}$$

Equation (8) states that for noiseless optical flow fields, the condition of correctness in (6) implies the optical flow vector relating a point $\boldsymbol{r}_{est}$ to its corresponding point in I_{gt} should be of less magnitude than the flow vector relating $\boldsymbol{r}_{est}$ to its corresponding point in I_{test}. The two corresponding vectors are in the same direction since the three poses are co-linear and hence the points $\boldsymbol{r}_{gt}$ and $\boldsymbol{r}_{test}$ corresponding to $\boldsymbol{r}_{est}$ are bound to the same epipolar line.

However, in the presence of noise and false inliers, we must consider the possibility that the epipolar constraint in eq. (1) is not exactly satisfied and hence $\mathcal{R}'$ may contain false inliers, the location of points $\boldsymbol{r}_{gt}$ and $\boldsymbol{r}_{test}$ along the epipolar line $\hat{\boldsymbol{E}}_{est,gt}\tilde{\boldsymbol{r}}_{I_{est}}$ are perturbed by noise, and that $\hat{\boldsymbol{E}}_{est,gt}$ differs from $\boldsymbol{E}_{est,gt}$. A primary source of error in our proposed monitoring method is the calculation of optical flow. Our estimate of the optical flow for any particular point can be expressed as follows:

$$\hat{\boldsymbol{v}}(\boldsymbol{r}_i)_{ij} = \boldsymbol{v}(\boldsymbol{r}_i)_{ij} + \boldsymbol{o}_{ij} + \boldsymbol{\gamma}_{ij} \tag{9}$$

where $\|\boldsymbol{\gamma}_{ij}\| \leq \delta$ and $\boldsymbol{o}_{ij}$ is 0 if $\hat{\boldsymbol{v}}(\boldsymbol{r}_i)_{ij}$ is an inlier with sampson distance less than δ. Otherwise, in the case of an outlier, $\boldsymbol{o}_{ij}$ is any arbitrary value such that $\hat{\boldsymbol{v}}(\boldsymbol{r}_i)_{ij}$ can exist at any location in the image. By computing the sampson distance of each $\hat{\boldsymbol{v}}(\boldsymbol{r}_{est})_{I_{est},I_{test}}$ w.r.t. $\hat{\boldsymbol{E}}_{est,gt}$, we can filter out points with error larger than δ. Note this does not check for error along the epipolar line. We additionally

filter out points which are not part of the cheiral set of maximum consensus. We again prune out any points whose correspondences have been labeled as outliers from a set of size n' to a set of n'', i.e., $\boldsymbol{r}_{est} \in \mathcal{R}'' \in \mathcal{R}'$. Lastly, we project all of $\hat{\boldsymbol{r}}_{I_{gt}}$ and $\hat{\boldsymbol{r}}_{test}$ to the epipolar line defined by $\hat{\boldsymbol{E}}_{est,gt}\tilde{\boldsymbol{r}}_{I_{est}}$ yielding $\hat{\hat{\boldsymbol{r}}}_{gt}$ and $\hat{\hat{\boldsymbol{r}}}_{test}$ such that pairs of corresponding points satisfy eq. (1).

Computing the confidence score. Now we must estimate the confidence, q, that the optical flow for corresponding points between I_{est} to I_{test} is greater than the ones between I_{est} to I_{gt}, i.e., $\|\boldsymbol{v}(\boldsymbol{r}_{est})_{I_{est},I_{gt}}\| < \|\boldsymbol{v}(\boldsymbol{r}_{est})_{I_{est},I_{test}}\|$. Using the optical flow vectors from $\boldsymbol{r}_{est}$ to the projected points $\hat{\hat{\boldsymbol{r}}}_{gt}$ and $\hat{\hat{\boldsymbol{r}}}_{test}$ we define the following confidence score:

$$q = \frac{1}{n''}\sum_{i=1}^{n''} \mathbb{P}(\|\hat{\boldsymbol{v}}(\boldsymbol{r}_{est})_{I_{est},I_{gt}}\| < \|\hat{\boldsymbol{v}}(\boldsymbol{r}_{est})_{I_{est},I_{test}}\|). \quad (10)$$

Explicitly, the confidence score in (10) is computed using the Normal CDF with a user-specified variance V. Standard deviation is set to a reasonable value of pixel error (e.g., 0.5). As a results, we rewrite (10) as:

$$q = \frac{1}{n''}\sum_{i=1}^{n''} \Phi\left(\frac{\hat{\boldsymbol{v}}(\boldsymbol{r}_{est})_{I_{est},I_{test}} - \hat{\boldsymbol{v}}(\boldsymbol{r}_{est})(\boldsymbol{r}_{est})_{I_{est},I_{gt}}}{\sqrt{V[\hat{\boldsymbol{v}}(\boldsymbol{r}_{est})_{I_{gt}}]}}\right) \quad (11)$$

where Φ is the Normal CDF. The confidence score mimics a probability, however due to simplifying assumptions such as approximating optical flow uncertainty and potential errors in computing the essential matrix, we do not claim it to be a true probability.

VI. Experiments

We now present results of running VERF-PnP and VERF-Light on three types of environments ranging from small-scale indoor scenes to a rocket trajectory spanning 8 km. For all experiments, we use torch-ngp [47] as our NeRF model. To get experimental sensor images we use randomly selected images from the NeRF training set. For each image, we generate a pose estimate to be checked for correctness by adding a random offset to the corresponding ground-truth position. To get a diverse distribution of correct and incorrect poses, we randomly selected either a low or high error regime when generating offsets.

In addition to comparing the two proposed methods, we include a simple baseline method that we refer to as Disparity Check. For this, we simply compute the optical flow between I_{est} and I_{gt} and determine the mean disparity from sparse flow. A naive approach is to assume low disparity means a correct pose estimation whereas a high disparity points to an incorrect pose. We use a folded normal distribution which computes a confidence level of correctness given a mean disparity. All experiments use a standard deviation of 4 pixels for the folded normal distribution. Since this method makes no efforts to handle scale ambiguity, we will show that it does not generalize well across varying scene size.

We pick a 0.5 cutoff confidence level for each method to estimate if the pose is correct or not. To show the generalizability of VERF, for all experiments we use the same standard deviation in (11) for VERF-Light (0.5 pixels) and the same standard deviation for VERF-PnP in (4). Likewise, the same RANSAC, RAFT, and good features to track parameters are used for all experiments.

A. LLFF dataset

Setup. We first evaluate VERF on 4 scenes (Fern, Fortress, Horns, and Room) from the LLFF dataset [14]. We pick 250 randomly selected views from the training set of images for each scene to serve as the sensor image I_{gt} and for each image randomly generate a choice for $\boldsymbol{x}_{est}$. We downscale I_{gt} to 504×378 and render the same resolution images when using NeRF. For these 1000 tests, we set ϵ to be 5 cm.

Results. In Fig. 3 we show the level of confidence VERF computed that the position error is less than ϵ compared to the actual position error for each test. As expected, confidence levels approach 1 as the position error is well less than ϵ and approach 0 when the position error is much greater than ϵ. On a 3090 GPU, total time to produce a confidence score from VERF-Light is on average 0.4 seconds with 0.25 seconds of that used for NeRF rendering, and is on average 0.35 seconds for VERF-PnP with the same time used for rendering since each method renders two NeRF images.

A summary of results is provided in Table I. Similar performance is observed by VERF-Light and VERF-PnP with most misclassifications occurring for pose estimates with errors near epsilon. Both methods outperform the Disparity Check baseline by a vast margin.

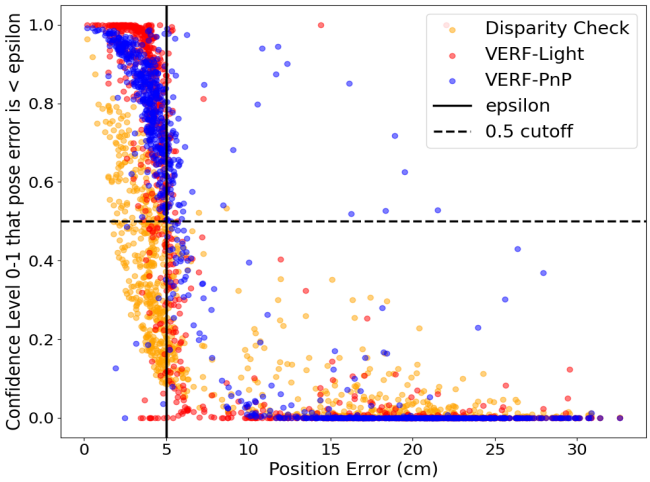


Fig. 3. VERF confidence level that for 1000 randomly sampled position estimates for LLFF scenes error is less than $\epsilon = 5$ cm

	Disparity Check	VERF-PnP	VERF-Light
True Positives (423)	146	415	381
True Negatives (577)	572	494	545
False Positives	5	83	32
False Negatives	277	8	42
Total Correct	**72%**	**91%**	**93%**

TABLE I. Summary of results for all proposed methods on 1000 tests on LLFF dataset. Classification is made with a 0.5 confidence score cutoff.

B. A1 Quadruped

Setup. We train a NeRF (Fig. 4) using RGB images collected with a Realsense D455 camera mounted on a Unitree A1 quadruped robot (Fig. 4). The robot transverses around a table at varying distances to the table in a motion capture room. Training images and sensor images are downscaled to 640×360. Ground-truth poses are estimated with COLMAP [48]. To correct from the ambiguous scale from COLMAP, we use vicon odometry to add metric scale to the poses. We again randomly select 1000 images with replacement from the dataset as sensor images and generate a random pose estimate for each image to be verified.

Fig. 4. A1 quadruped robot collecting monocular RGB data for NeRF training and VERF evaluation (top left). Three example NeRF-rendered views using weights trained by camera data collected onboard an A1 robot.

Results. We pick epsilon to be 5 cm and observe similar results as with the LLFF experiment with nearly all VERF mistakes occurring for position errors near the value of epsilon. Results are summarized in Table II and shown visually in Fig. 5. The Disparity Check baseline is shown to generalize poorly for different scale scenes as most of its errors are false negatives for the LLFF experiment whereas most of its errors are false positives for the A1 experiment.

	Disparity Check	VERF-PnP	VERF-Light
True Positives (421)	304	418	411
True Negatives (579)	513	561	551
False Positives	66	18	28
False Negatives	117	3	10
Total Correct	**82%**	**98%**	**96%**

TABLE II. Summary of results for all proposed methods on 1000 tests of A1 robot dataset. Classification is made with a 0.5 confidence score cutoff.

C. Sub-Orbital Rocket

Setup. Here we demonstrate the potential for VERF to be used in a highly complex scenario such as for precision spacecraft navigation. This experiment uses data we collected for [4] in which we mounted two cameras inside the capsule of Blue Origin's New Shepard rocket which point out the capsule windows towards the terrain (Fig. 6).

We train on 140 images collected during the rocket's ascent from an altitude range of approximately 0.2 to 8 km above

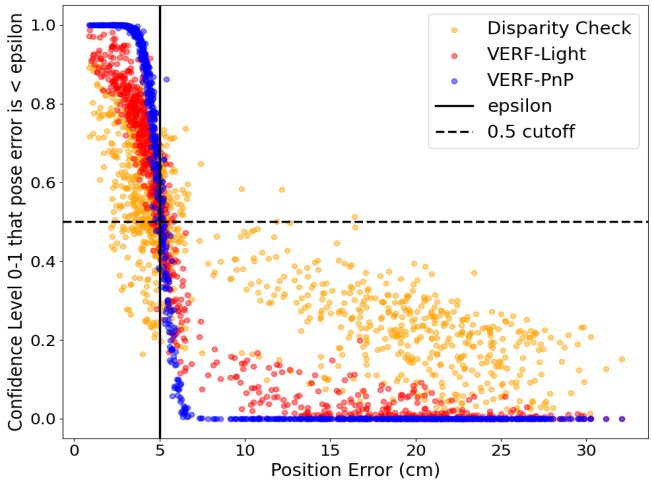


Fig. 5. VERF confidence level that for 1000 randomly sampled position estimates of the A1's pose, error is less than $\epsilon = 5$ cm

ground level during which the rocket reaches a speed up to 880 km/hr. We do not include data at higher altitudes as there was a mishap during flight NS-23 which triggered the capsule escape system. The curious reader can refer to [4] for more details of our flight data collection.

For simplicity, we train on images collected during the flight and use estimated poses from COLMAP as ground truth. In practice, a NeRF could be trained from prior satellite maps as was done in [49]. Again, similar to the A1 experiment, VERF is run on a scaled NeRF model and we provide metric scale to the COLMAP reconstruction from ground truth poses of the training images — in this case from GPS inside the rocket's capsule.

Fig. 6. Sub-orbital rocket on launch pad used for data collection (left). Close view of camera mounted inside the capsule window (bottom right) and a view of both cameras inside the capsule before launch (top right). Images courtesy of Blue Origin.

Results. We pick 40 m for epsilon since this is on the order of typical spacecraft landing accuracy for planetary exploration [2]. A summary of results is shown in Table III and visually in Fig. 8. VERF-PnP performs notably better than VERF-Light on this dataset which we believe to be caused by inaccuracies in the essential matrix estimation due to the scene being approximately planar at high altitudes.

Fig. 7. Example of four NeRF rendered views from sub-orbital rocket ascent from an altitude range of approximately 1 km to 8 km.

Section VIII provides a study on the effects of error in the essential matrix on VERF-Light.

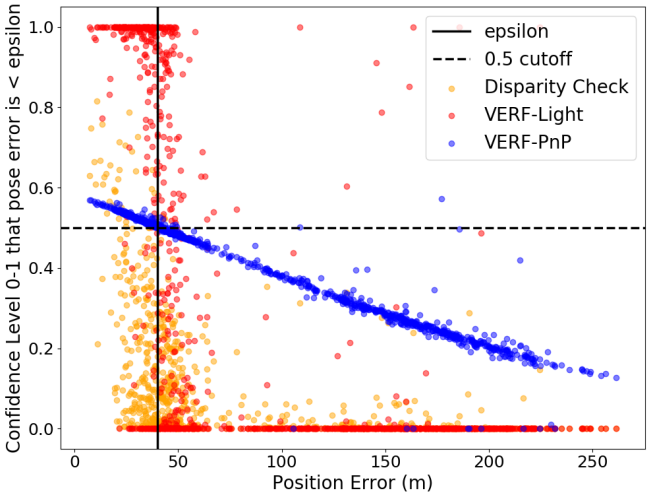


Fig. 8. VERF confidence level that for 1000 randomly sampled position estimates of the rocket's pose, error is less than $\epsilon = 40$ m

	Disparity Check	VERF-PnP	VERF-Light
True Positives (260)	38	259	214
True Negatives (740)	738	710	640
False Positives	2	30	100
False Negatives	222	1	46
Total Correct	**78%**	**97%**	**85%**

TABLE III. Summary of results for all proposed methods on 1000 tests of rocket dataset. Classification is made with a 0.5 confidence score cutoff.

Additionally, as VERF must perform well across a wide range of altitudes for the rocket dataset, we show that VERF-PnP and VERF-Light perform well across all altitudes while Disparity Check does not generalize well. To further demonstrate this point, in Fig. 9 we pick estimated poses with error 15 m (with epsilon again set to 40 m) and run all three methods on sequential images during launch. Figure 9 shows that the performance of the Disparity Check is dependent on altitude (switching its decision from incorrect to correct after 4 km) while VERF-PnP and VERF-Light perform consistently throughout the rocket's accent.

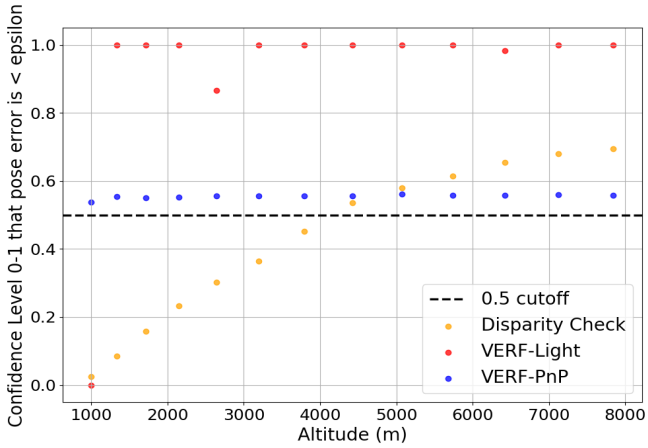


Fig. 9. VERF confidence level vs altitude for rocket dataset with fixed 15 m of position error. Epsilon is selected as 40 m. Disparity Check is shown to not generalize with varying scene scale while VERF-PnP and VERF-Light performance are independent of the rocket's altitude.

VII. Conclusion

We have presented two approaches (VERF-PnP and VERF-Light) to leverage Neural Radiance Fields to monitor the correctness of a pose estimate acquired from a monocular camera. Our methods functions independently of how the pose is estimated (i.e., NeRF does not have to be used for pose estimation) and can provide a level of assurance in under half a second. Experiments have shown the effectiveness of VERF on scene scales ranging from small rooms to kilometer-scale outdoor scenes. As a limitation, we remark that VERF is intended to be a local pose monitoring approach in the sense that if an arbitrarily large epsilon were used, it is possible for NeRF rendered images to be outside the range of the trained NeRF or fail to match features to the sensor image leading to a false assumption of an incorrect pose.

Acknowledgement

The authors gratefully acknowledge Jingnan Shi, Brett Streetman, and Ted Steiner for their help collecting data for our experiments.

References

[1] P. Wang, X. Huang, X. Cheng, D. Zhou, Q. Geng, and R. Yang, "The ApolloScape open dataset for autonomous driving and its application," *IEEE Trans. Pattern Anal. Machine Intell.*, 2019.

[2] A. E. Johnson, S. B. Aaron, J. Chang, Y. Cheng, J. F. Montgomery, S. Mohan, S. Schroeder, B. E. Tweddle, N. Trawny, and J. X. Zheng, "The lander vision system for mars 2020 entry descent and landing," 2017.

[3] D. A. Lorenz, R. D. Olds, A. May, C. Mario, M. E. Perry, E. E. Palmer, and M. G. Daly, "Lessons learned from osiris-rex autonomous navigation using natural feature tracking," *IEEE Aerospace Conference*, pp. 1–12, 2017.

[4] D. Maggio, C. Mario, B. Streetman, T. Steiner, and L. Carlone, "Vision-based terrain relative navigation on high altitude balloon and sub-orbital rocket," in *AIAA SciTech Forum*, 2023.

[5] L. Manuelli, W. Gao, P. Florence, and R. Tedrake, "kpam: Keypoint affordances for category-level robotic manipulation," in *Proc. of the Intl. Symp. of Robotics Research (ISRR)*, 2019.

[6] L. Yen-Chen, P. Florence, J. T. Barron, A. Rodriguez, P. Isola, and T.-Y. Lin, "iNeRF: Inverting neural radiance fields for pose estimation," in *IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, 2021.

[7] M. Adamkiewicz, T. Chen, A. Caccavale, R. Gardner, P. Culbertson, J. Bohg, and M. Schwager, "Vision-only robot navigation in a neural radiance world," *CoRR*, vol. abs/2110.00168, 2021. [Online]. Available: https://arxiv.org/abs/2110.00168

[8] D. Maggio, M. Abate, J. Shi, C. Mario, and L. Carlone, "Loc-NeRF: Monte carlo localization using neural radiance fields," in *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, 2022, (pdf) (video).

[9] Y. Lin, T. Müller, J. Tremblay, B. Wen, S. Tyree, A. Evans, P. A. Vela, and S. Birchfield, "Parallel inversion of neural radiance fields for robust pose estimation," 2022.

[10] Z. Zhu, Y. Chen, Z. Wu, C. Hou, Y. Shi, C. Li, P. Li, H. Zhao, and G. Zhou, "Latitude: Robotic global localization with truncated dynamic low-pass filter in city-scale nerf," 2022. [Online]. Available: https://arxiv.org/abs/2209.08498

[11] G. Avraham, J. Straub, T. Shen, T.-Y. Yang, H. Germain, C. Sweeney, V. Balntas, D. Novotny, D. DeTone, and R. Newcombe, "Nerfels: Renderable neural codes for improved camera pose estimation," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 5061–5070.

[12] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," *arXiv preprint arXiv:2003.08934*, 2020.

[13] M. Fischler and R. Bolles, "Random sample consensus: a paradigm for model fitting with application to image analysis and automated cartography," *Commun. ACM*, vol. 24, pp. 381–395, 1981.

[14] B. Mildenhall, P. P. Srinivasan, R. Ortiz-Cayon, N. K. Kalantari, R. Ramamoorthi, R. Ng, and A. Kar, "Local light field fusion: Practical view synthesis with prescriptive sampling guidelines," *ACM Transactions on Graphics (TOG)*, 2019.

[15] A. Rosinol, J. Leonard, and L. Carlone, "NeRF-SLAM: Real-time dense monocular SLAM with neural radiance fields," *arXiv preprint: 2210.13641*, 2022, (pdf).

[16] E. Sucar, S. Liu, J. Ortiz, and A. Davison, "iMAP: Implicit mapping and positioning in real-time," in *Intl. Conf. on Computer Vision (ICCV)*, 2021.

[17] Z. Zhu, S. Peng, V. Larsson, W. Xu, H. Bao, Z. Cui, M. R. Oswald, and M. Pollefeys, "NICE-SLAM: Neural implicit scalable encoding for slam," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, June 2022.

[18] Y. Jiang, P. Hedman, B. Mildenhall, D. Xu, J. T. Barron, Z. Wang, and T. Xue, "Alignerf: High-fidelity neural radiance fields via alignment-aware training," *arXiv preprint arXiv:2211.09682*, 2022.

[19] A. Moreau, N. Piasco, M. Bennehar, D. Tsishkou, B. Stanciulescu, and A. de La Fortelle, "Crossfire: Camera relocalization on self-supervised features from an implicit representation," 2023. [Online]. Available: https://arxiv.org/abs/2303.04869

[20] F. Li, H. Yu, I. Shugurov, B. Busam, S. Yang, and S. Ilic, "Nerf-pose: A first-reconstruct-then-regress approach for weakly-supervised 6d object pose estimation," 2022.

[21] G. Klein and D. Murray, "Parallel tracking and mapping for small ar workspaces," in *2007 6th IEEE and ACM international symposium on mixed and augmented reality*. IEEE, 2007, pp. 225–234.

[22] T. Qin, J. Pan, S. Cao, and S. Shen, "A general optimization-based framework for local odometry estimation with multiple sensors," *arXiv preprint: 1901.03638*, 2019.

[23] C. Cadena, L. Carlone, H. Carrillo, Y. Latif, D. Scaramuzza, J. Neira, I. Reid, and J. Leonard, "Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age," *IEEE Trans. Robotics*, vol. 32, no. 6, pp. 1309–1332, 2016, arxiv preprint: 1606.05830, (pdf).

[24] R. A. Newcombe, S. J. Lovegrove, and A. J. Davison, "Dtam: Dense tracking and mapping in real-time," in *Intl. Conf. on Computer Vision (ICCV)*. IEEE, 2011, pp. 2320–2327.

[25] A. I. Mourikis, N. Trawny, S. I. Roumeliotis, A. E. Johnson, A. Ansar, and L. Matthies, "Vision-aided inertial navigation for spacecraft entry, descent, and landing," *IEEE Transactions on Robotics*, vol. 25, no. 2, pp. 264–280, 2009.

[26] C. D. N. et al., "Autonomous navigation performance using natural feature tracking during the osiris-rex touch-and-go sample collection event," *The Planetary Science Journal*, vol. 3, no. 5, p. 101, may 2022. [Online]. Available: https://dx.doi.org/10.3847/PSJ/ac5183

[27] T. Sattler, Q. Zhou, M. Pollefeys, and L. Leal-Taixe, "Understanding the Limitations of CNN-based Absolute Camera Pose Regression," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 3302–3312.

[28] N. Piasco, D. Sidibé, C. Demonceaux, and V. Gouet-Brunet, "A survey on visual-based localization: On the benefit of heterogeneous data," *Pattern Recognition*, vol. 74, pp. 90–109, 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0031320317303448

[29] L. Carlone and F. Dellaert, "Duality-based verification techniques for 2D SLAM," in *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, 2015, pp. 4589–4596, (pdf) (code)

[30] D. Rosen, L. Carlone, A. Bandeira, and J. Leonard, "SE-Sync: a certifiably correct algorithm for synchronization over the Special Euclidean group," *Intl. J. of Robotics Research*, 2018, arxiv preprint: 1611.00128, (pdf).

[31] H. Yang, J. Shi, and L. Carlone, "TEASER: Fast and Certifiable Point Cloud Registration," *arXiv preprint: 2001.07715*, 2020, (pdf).

[32] H. Yang and L. Carlone, "Certifiably optimal outlier-robust geometric perception: Semidefinite relaxations and scalable global optimization," *IEEE Trans. Pattern Anal. Machine Intell.*, 2022, (pdf).

[33] M. Garcia-Salguero and J. Gonzalez-Jimenez, "Fast and robust certifiable estimation of the relative pose between two calibrated cameras," *arXiv preprint arXiv:2101.08524*, 2021.

[34] M. Garcia-Salguero, J. Briales, and J. Gonzalez-Jimenez, "Certifiable relative pose estimation," *Image and Vision Computing*, vol. 109, p. 104142, 2021.

[35] J. Zhao, W. Xu, and L. Kneip, "A certifiably globally optimal solution to generalized essential matrix estimation," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[36] L. Carlone, "Estimation contracts for outlier-robust geometric perception," *arXiv preprint: 2208.10521*, 2022, (pdf).

[37] R. Talak, L. Peng, and L. Carlone, "Certifiable 3D object pose estimation: Foundations, learning models, and self-training," *arXiv preprint: 2206.11215*, Jan. 2023, (pdf).

[38] H. Yang and M. Pavone, "Object pose estimation with statistical guarantees: Conformal keypoint detection and geometric uncertainty propagation," 2023.

[39] X. Hu and P. Mordohai, "A quantitative evaluation of confidence measures for stereo vision," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 34, no. 11, pp. 2121–2133, 2012.

[40] Q. Rahman, P. Corke, and F. Dayoub, "Run-time monitoring of machine learning for robotic perception: A survey of emerging trends," *IEEE Access*, vol. PP, pp. 1–1, 01 2021.

[41] P. Antonante, H. Nilsen, and L. Carlone, "Monitoring of perception systems: Deterministic, probabilistic, and learning-based fault detection and identification," *arXiv preprint: 2205.10906*, 2022, (pdf).

[42] R. I. Hartley, "Chirality," *Intl. J. of Computer Vision*, vol. 26, pp. 41–61, 1998.

[43] Z. Teed and J. Deng, "Raft: Recurrent all-pairs field transforms for optical flow," *ArXiv*, vol. abs/2003.12039, 2020.

[44] J. Shi and C. Tomasi, "Good Features to track," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 1994, pp. 593–600.

[45] D. Nistér, "An efficient solution to the five-point relative pose problem," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 26, no. 6, pp. 756–770, 2004.

[46] P. D. Sampson, "Fitting conic sections to "very scattered" data: An iterative refinement of the bookstein algorithm," *Computer Graphics and Image Processing*, vol. 18, no. 1, pp. 97–108, 1982. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0146664X82901010

[47] J. Tang, "Torch-ngp: a pytorch implementation of instant-ngp," 2022, https://github.com/ashawkey/torch-ngp.

[48] J. L. Schönberger and J.-M. Frahm, "Structure-from-motion revisited," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[49] Y. Xiangli, L. Xu, X. Pan, N. Zhao, A. Rao, C. Theobalt, B. Dai, and D. Lin, "BungeeNeRF: Progressive neural radiance field for extreme multi-scale scene rendering," in *European Conf. on Computer Vision (ECCV)*, 2022.

VIII. Appendices

A. Effects of Essential Matrix Error

Here we study the effects of the accuracy of the essential matrix estimation for VERF-Light. We repeat each of the three experiments with the same setup except we now provide VERF-Light with the true essential matrix $\boldsymbol{E}_{est,gt}$. Table IV shows notable improvements on all experiments with VERF-Light correctly classifying 99% of pose estimates. The most significant improvement is on the rocket dataset. Not only does accuracy go from 86% to 99% but as shown in Fig. 10, the confidence levels follow a cleaner distribution. We believe this to be caused by the approximate planar scene from high altitudes. This study thus shows the potential to improve VERF-Light with a more effective essential matrix estimation method.

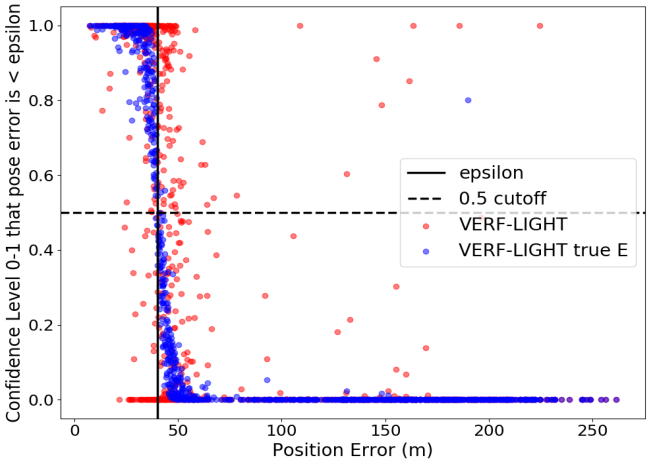


Fig. 10. VERF confidence level that for 1000 randomly sampled position estimates for rocket data that their error is less than $\epsilon = 40m$. VERF-Light shown with and without the true essential matrix.

	LLFF	A1	Rocket
True Positives	415	420	255
True Negatives	574	567	736
False Positives	3	12	4
False Negatives	8	1	5
Total Correct	**99%**	**99%**	**99%**

TABLE IV. Summary of results on running VERF-Light on all experiments using the true essential matrix.

B. Estimating Metric Error with PnP

A logical question to pose is since PnP can estimate metric distance, how well can VERF-PnP estimate the true error instead of just estimating correctness with respect to an epsilon threshold. With an estimate of the true error, the pose estimate can then be corrected. Figures 11 and 12 show position errors before an after being corrected in this fashion by VERF-PnP with errors decreasing by an order of magnitude. For each experiment there were a small number of pose estimates omitted (1 for Fig. 11 and 9 for Fig. 12) as PnP diverged. One potential option to automatically check and prevent this is to only accept the updated pose if VERF-Light predicts that the corrected pose is less than epsilon.

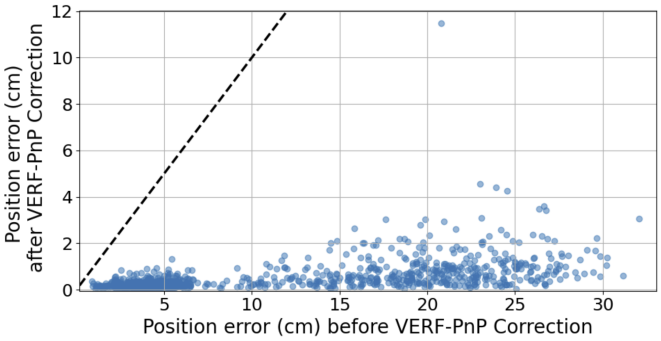


Fig. 11. Position errors before and after being corrected using the position error estimate from VERF-PnP. Results shown for 1000 tests on the A1 dataset.

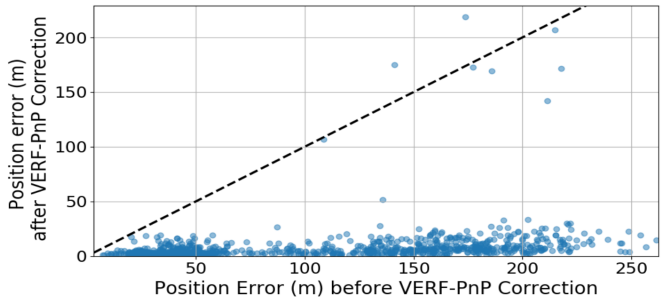


Fig. 12. Position errors before and after being corrected using the position error estimate from VERF-PnP. Results shown for 1000 tests on the Rocket dataset.