

# NeRF-VINS: A Real-time Neural Radiance Field Map-based Visual-Inertial Navigation System

Saimouli Katragadda<sup>1</sup>, Woosik Lee<sup>1</sup>, Yuxiang Peng<sup>1</sup>, Patrick Geneva<sup>1</sup>,  
Chuchu Chen<sup>1</sup>, Chao Guo<sup>2</sup>, Mingyang Li<sup>2</sup>, and Guoquan Huang<sup>1</sup>

**Abstract**— Achieving accurate, efficient, and consistent localization within an *a priori* environment map remains a fundamental challenge in robotics and computer vision. Conventional map-based keyframe localization often suffers from sub-optimal viewpoints due to limited field of view (FOV), thus degrading its performance. To address this issue, in this paper, we design a real-time tightly-coupled Neural Radiance Fields (NeRF)-aided visual-inertial navigation system (VINS), termed NeRF-VINS. By effectively leveraging NeRF’s potential to synthesize novel views, essential for addressing limited viewpoints, the proposed NeRF-VINS optimally fuses IMU and monocular image measurements along with synthetically rendered images within an efficient filter-based framework. This tightly coupled integration enables 3D motion tracking with bounded error. We extensively compare the proposed NeRF-VINS against the state-of-the-art methods that use prior map information, which is shown to achieve superior performance. We also demonstrate the proposed method is able to perform real-time estimation at 15 Hz, on a resource-constrained Jetson AGX Orin embedded platform with impressive accuracy<sup>1</sup>.

## I. INTRODUCTION

The ability to achieve centimeter-level localization accuracy is pivotal for resource-constrained edge devices which have become prevalent through computation miniaturization enabling AR/VR [1], [2], consumer drones [3], [4], and autonomous navigation [5]. The ubiquitous use of cameras and inertial measurement units (IMU) due to their low cost, low power, and small size makes the Visual-Inertial Navigation System (VINS) a critical foundational component for the aforementioned applications [6]. VINS, without knowing global information, e.g., GPS, loop-closures, or a prior map, can only provide ego-motion information whose errors may grow unbounded. Over the past two decades, a particular focus has been placed on leveraging *a priori* map as additional costly sensors are not required [7]–[13].

A crucial component of successful map-based localization is an accurate place retrieval algorithm such as DBOW [14], placeless [15], or NetVLAD [16], which allows for recovery of correspondence information to construct constraints to historical information. However, these methods may be vulnerable to viewpoint variations, poor viewpoint coverage limiting recall, scene ambiguities, and sensitivities to environmental changes after mapping [17].

To address the aforementioned challenges, in this work, we propose to avoid the need for place recognition via the ren-

<sup>1</sup>The authors are with Robot Perception and Navigation Group (RPNG), University of Delaware. Email: {saimouli, woosik, yxpeng, pgeneva, ccchu, ghuang}@udel.edu

<sup>2</sup>The authors are with Google ARCore. Email: {chaoguo, mingyangli} @google.com

<sup>1</sup>Demo: <https://www.youtube.com/watch?v=X7wDy1enkTs>.

dering of novel synthetic views adjacent to the current state estimate, enabling high-quality and informative loop-closure constraints that are not susceptible to these failure modes. Specifically, we introduce a new paradigm for map-based localization which leverages the recent Neural Radiance Fields (NeRF) advancements in deep learning to compress the collection of images, e.g. a prior keyframe image map, into a trained network, and then leverage during localization the high-fidelity image rendering of synthesize novel camera viewpoints. While NeRF is able to accurately reconstruct complex environments has encouraged researchers to build dense NeRF maps [18], [19], we focus on achieving real-time localization on edge devices with limited computational resources and thus look to leverage the comparably cheaper novel viewpoint rendering via hashing [20].

To this end, we effectively leverage NeRF as an *a priori* map while maintaining real-time drift-free VINS localization. The main contribution of this work includes:

- To the best of our knowledge, this is the first real-time NeRF-based VINS that tightly-couples *a priori* NeRF map to enable drift-free localization.
- We conduct extensive numerical studies to demonstrate the impact of different NeRF map construction methods and descriptor algorithms on rendered NeRF views, and environmental changes.
- The proposed NeRF-VINS is among the first to demonstrate centimeter-level drift-free pose estimates on an edge platform (Jetson AGX Orin rendering at over 15 Hz) and outperform existing state-of-the-art methods which leverage prior map information.

## II. RELATED WORK

In this section, we provide an overview of methods related to visual and visual-inertial and NeRF-based localization.

### A. Prior Map-based Classical Localization

*Single-View Visual Localization:* The classical structure-based method is the Perspective-n-Point (PnP) solver within a RANSAC loop for robustness [21], [22]. The 2D-3D correspondences between the query image and a map points are typically found through the matching of local feature descriptors [23]–[27]. To mitigate the complexity increase as the map size grows, image retrieval methods that narrow down the search space typically retrieve top similar matches (place recognition) and query keypoints in the region defined by these images for correspondences (local matching) [9], [28]. The quality of this approach heavily relies on the effectiveness of the image retrieval methods. DBOW

[14] has gained great popularity thanks to its efficiency, but recent deep learned-based HF-Net [9], which leverages NetVLAD [16] and SuperPoint [29] for global retrieval and local matching respectively, has demonstrated state-of-the-art performance in localization tasks. Although there are end-to-end deep learning methods available, their poor accuracy and complexity still make structure-based methods appealing [30]–[32]. Additionally, all discussed methods can suffer from global descriptor ambiguities, particularly in scenarios with sparse images or significant changes in viewpoint, and poor recall due to limited view coverage of the scene which we aim to address through the proposed NeRF-VINS rendering paradigm.

*Visual-Inertial Localization:* As compared to single-view visual localization, visual-inertial localization aims to continuously provide estimates against a prior map and can leverage historical information to reduce the search space and thus complexity. There is a rich literature, for which we refer the interested reader to the references in [10] for a summary. One which is of particular relevance to this work is the open-sourced ROVIOLI [12] extension of ROVIO [33], [34] which performs 2D-3D matches against an optimized global map commonly constructed using maplab [12], [13].

*SLAM Systems:* In contrast to previous approaches that construct maps offline for accurate localization, SLAM builds maps online and utilizes them via loop closures. A typical SLAM architecture includes a real-time thread for camera pose tracking using sparse keypoints [35], [36] or dense/semi-dense representations [37], [38], along with a non-real-time thread that optimizes and constructs the map. These methods use classical image retrieval techniques to query images for loop closure, which can be affected by limited viewpoint coverage and ambiguities.

## B. Neural Radiance Fields

The work [39] introduced the NeRF methodology and revolutionized scene representation, novel view generation, and high-fidelity rendering. Later works such as BARF [40] and Nerf [41] have shown that knowing the exact poses is not required, while iMAP [42] and NICE-SLAM [43] showed that the joint optimization of poses in respect the NeRF can further improve performance. There additionally have been works that have focused on map representation [18], and the integration within SLAM [19], [44].

As compared to the online generation of NeRF maps, we instead look to leverage a previously built NeRF to provide high-quality loop-closure information and bound estimator drift. Only a few works have focused on leveraging NeRF to provide prior environmental information for the betterment of visual tracking. iNeRF [45] proposed to localize camera poses by optimizing the photometric error between the real and NeRF-generated images within a small static environment context but remained sensitive to the initial pose guess and large computational cost. More recently, Loc-NeRF [46] was proposed to employ a particle filter to remove the need for an initial guess. While this method does not require any initial guess, it necessitates image rendering for each particle, which could easily become computationally prohibitive if

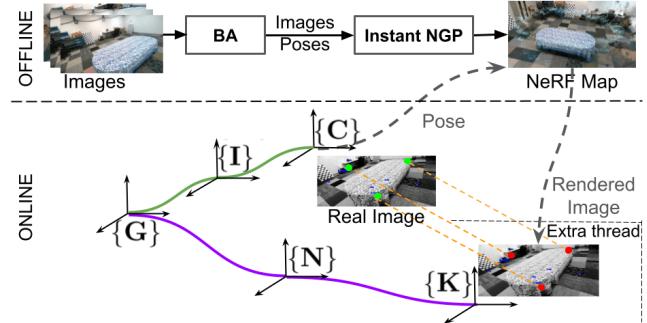


Fig. 1: Overview of the proposed NeRF-VINS, where  $\{G\}$  is the global VIO frame,  $\{N\}$  is the map frame,  $\{K\}$  denotes the NeRF rendered image.  $\{I\}$  and  $\{C\}$  are IMU and camera frame, respectively.

using a large number of particles. Another work in this family is by Adamkiewicz et al. [47] which leveraged a pre-trained NeRF map to localize and additionally optimize future trajectories. As compared to these works which are constrained by rendering speed and their alignment computational complexity, the proposed NeRF-VINS combines the novel viewpoint rendering strength with the efficient, accurate, and consistent MSCKF-based VINS.

## III. NERF-VINS ESTIMATOR DESIGN

Built on top of the MSCKF-based OpenVINS [48], [49], the proposed NeRF-VINS estimator extends the MSCKF to fuse the prior NeRF map in a tightly-couple manner (see Fig. 1). As such, for presentation brevity, in the following, we will primarily focus on visual measurement update.

In particular, at time  $t_k$ , the system state  $\mathbf{x}_k$  consists of the current inertial navigation states  $\mathbf{x}_{I_k}$ , historical IMU poses  $\mathbf{x}_{T_k}$ , and a subset of 3D environmental point features,  $\mathbf{x}_f$ :

$$\mathbf{x}_k = [\mathbf{x}_{I_k}^\top \mathbf{x}_{T_k}^\top \mathbf{x}_f^\top]^\top \quad (1)$$

$$\mathbf{x}_{I_k} = [{}^I_G \bar{q}^\top {}^G \mathbf{p}_{I_k}^\top {}^G \mathbf{v}_{I_k}^\top \mathbf{b}_g^\top \mathbf{b}_a^\top]^\top \quad (2)$$

$$\mathbf{x}_{T_k} = [{}^I_G \bar{q}^\top {}^G \mathbf{p}_{I_k}^\top \dots {}^{I_{k-c}}_G \bar{q}^\top {}^G \mathbf{p}_{I_{k-c}}^\top]^\top \quad (3)$$

$$\mathbf{x}_f = [{}^G \mathbf{p}_{f_1}^\top \dots {}^G \mathbf{p}_{f_i}^\top]^\top \quad (4)$$

where  ${}^I_G \bar{q}$  is the unit quaternion ( ${}^I_G \mathbf{R}$  in rotation matrix form) that represents the rotation from the global  $\{G\}$   ${}^G \mathbf{p}_I$ ,  ${}^G \mathbf{v}_I$ , and  ${}^G \mathbf{p}_{f_i}$  are the IMU position, velocity, and  $i$ 'th point feature position in  $\{G\}$ ;  $\mathbf{b}_g$  and  $\mathbf{b}_a$  are the gyroscope and accelerometer biases. Note that other state variables can be included, e.g., spacial-temporal calibration, but have been omitted for clarity.

The state is propagated over time based on the IMU measurements. A canonical three-axis IMU provides linear acceleration,  ${}^I \mathbf{a}_m$ , and angular velocity measurements,  ${}^I \boldsymbol{\omega}_m$ . The IMU nonlinear kinematics is generically given by [50]:

$$\mathbf{x}_{I_{k+1}} = \mathbf{f}(\mathbf{x}_{I_k}, {}^I \mathbf{a}_k, {}^I \boldsymbol{\omega}_k, \mathbf{n}_{I_k}) \quad (5)$$

where  $\mathbf{n}_I = [\mathbf{n}_g^\top \mathbf{n}_a^\top \mathbf{n}_{wg}^\top \mathbf{n}_{wa}^\top]^\top$ ;  $\mathbf{n}_g^\top$  and  $\mathbf{n}_a^\top$  are Gaussian white noises, and  $\mathbf{n}_{wg}$  and  $\mathbf{n}_{wa}$  are the random walk bias noises of gyroscope and accelerometer, respectively. With this model (5), we can perform EKF propagation of the state estimate and covariance [48].

### A. Measurement Update with Real Images

As in [49], the bearing measurements of detected features seen at time  $t_k$  are modeled as follows:

$$\mathbf{z}_{C_k} = \mathbf{h}_c(\mathbf{x}_{T_k}, {}^G\mathbf{p}_f) + \mathbf{n}_{C_k} =: \boldsymbol{\Lambda}({}^C\mathbf{p}_f) + \mathbf{n}_{C_k} \quad (6)$$

$${}^{C_k}\mathbf{p}_f = {}_I^C\mathbf{R} {}_G^I\mathbf{R}({}^G\mathbf{p}_f - {}^G\mathbf{p}_{I_k}) + {}^C\mathbf{p}_I \quad (7)$$

where  $\boldsymbol{\Lambda}([x \ y \ z]^\top) = [x/z \ y/z]^\top$  and  $\mathbf{n}_{C_k}$  is the white Gaussian noise. Linearizing Eq. (6) the following measurement residual:

$$\mathbf{r}_{C_k} = \mathbf{z}_{C_k} - \mathbf{h}_c(\hat{\mathbf{x}}_{T_k}, {}^G\hat{\mathbf{p}}_f) \simeq \mathbf{H}_T \hat{\mathbf{x}}_{T_k} + \mathbf{H}_f {}^G\tilde{\mathbf{p}}_f + \mathbf{n}_{C_k} \quad (8)$$

where  $\mathbf{H}_T$  and  $\mathbf{H}_f$  are the Jacobian matrix of the measurement with respect to each state. We keep the long-tracked features in the state till lost in order to leverage their future observations, while the short-tracked features are updated via the efficient MSCKF nullspace projection [48].

### B. Measurement Update with NeRF Images

When a camera image reading is received, a NeRF render is triggered at a pose with a small horizontal positional offset (e.g., 10 cm, as in our experiments) to the current camera pose. This synthetic image should have a significant overlapping field of view (FOV) with the current real image, which facilitates high-quality feature matching. The small positional offset also enables robust triangulation and accurate feature matching between the real and synthetic images even when the camera is static.

Once the rendering is completed, descriptor-based feature matching is performed to the current image, where a 2D-to-2D prior keyframe measurement model is leveraged [51]. For example, consider that from the rendered image we get a bearing measurement,  $\mathbf{z}_{N_k}$ , which is related to the state as:

$$\mathbf{z}_{N_k} = \mathbf{h}_n({}^G\mathbf{p}_f) + \mathbf{n}_{N_k} =: \boldsymbol{\Lambda}({}^N\mathbf{p}_f) + \mathbf{n}_{N_k} \quad (9)$$

$${}^N\mathbf{p}_f = {}^K\mathbf{p}_N + s_N {}^N\mathbf{R}({}^N\mathbf{p}_G + {}_G^N\mathbf{R} {}^G\mathbf{p}_f) \quad (10)$$

where  $s$  is the scale factor of the map and  $\mathbf{n}_{N_k}$  is the zero mean Gaussian noise. Note that we model the bearing as only a function of the feature  ${}^G\mathbf{p}_f$ , and consider the map transform  $\{{s, {}_G^N\mathbf{R}, {}^N\mathbf{p}_G}\}$  (see Sec. IV-C) to be known and the rendered camera pose  $\{{}^K\mathbf{R}, {}^K\mathbf{p}_N\}$  to have some known orientation and position uncertainty  $\{{}_G^N\tilde{\theta}, {}^N\tilde{\mathbf{p}}_G\}$ . Thus, we have the following linearized model:

$$\mathbf{r}_{N_k} = \mathbf{z}_{N_k} - \mathbf{h}_n({}^G\hat{\mathbf{p}}_f) = s\mathbf{H}_N {}^K\mathbf{R} {}_G^N\mathbf{R} {}^G\tilde{\mathbf{p}}_f + \mathbf{n}'_{N_k} \quad (11)$$

where  $\lfloor \cdot \times \rfloor$  is the skew-symmetric matrix and

$$\mathbf{n}'_{N_k} = s\mathbf{H}_N {}^K\mathbf{R}(\lfloor {}_G^N\mathbf{R} {}^G\mathbf{p}_f \times \rfloor {}_G^N\tilde{\theta} + {}^N\tilde{\mathbf{p}}_G) + \mathbf{n}_{N_k} \quad (12)$$

The linearized model can be used to update the features in the state or can be stacked with the real image measurements (8) to perform (SLAM or MSCKF) EKF update.

## IV. NERF-VINS SYSTEM INTEGRATION

Armed with the NeRF-VINS estimation theory presented in the previous section, we now describe how to integrate the NeRF model and feature matching between synthetic and real images to form a tightly-coupled system.

In particular, our system leverages the open-source InstantNGP [20] for rendering and prior map training. The OpenVINS [49] frontend is modified to incorporate SuperPoint

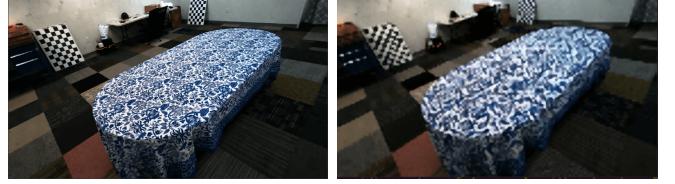


Fig. 2: Exemplary rendered images for testing matching methods. *Left*: Rendered image with full resolution. *Right*: Rendered image with 141x80 resolution and up-scaled with FSRCNN [52]

TABLE I: Average descriptor extraction time, number of matches, and ATE reported on the UD AR Table 1-8 dataset [59] for different matching methods.

	AKAZE	BRISK	ORB	KAZE	SP	SP Opt.
Time (ms)	31	88	13	140	15	7
Matches	55	85	20	117	31	30
ATE (deg/m)	2 FAIL	5 FAIL	6 FAIL	2.40 / 0.29	1.16 / 0.15	1.18 / 0.16

descriptors using Tensor-RT pipeline [53]. We used OpenCV [54] and CUDA to convert GPU-rendered images to a 32bit-float RGB image on the CPU. Additional care has been taken to convert the NeRF-rendered image coordinate system to a right-hand coordinate system by inverting the y and z axes of InstantNGP. The code is written in C++ and CUDA and runs on Jetson AGX Orin unless specified.

### A. Feature Descriptors

A crucial component is the ability to match features between the current frame and the rendered NeRF viewpoint. Thus significant effort has been spent to investigate the performance of various feature matching methods such as AKAZE [55], KAZE [56], BRISK [57], ORB [58], and the selected SuperPoint [29]. For this test, we choose a challenging scenario by rendering at 1/6 of the original size and is upscaled using FSRCNN [52], see Fig. 2. Shown in Tab. I, the average descriptor extraction time, number of matches between the rendered and current camera image, and Absolute Trajectory Error (ATE) of VINS for each method have been compared. It is clear that the handcrafted matching methods (AKAZE, BRISK, ORB, and KAZE) often fail and show large errors which is expected due to the limited fidelity in the up-sampled low-resolution NeRF image. On the other hand, SuperPoint (SP) and its optimized variant (SP Opt.) are shown to be robust to these conditions and report the highest accuracy and shortest descriptor extraction time. This leads us to select the optimized SuperPoint for its robustness and efficiency for synthetic NeRF to real image matching.

### B. Image Rendering and Feature Matching

Rendering NeRF images remains a computationally expensive operation, particularly on embedded devices like the Jetson AGX Orin. It takes approximately 660 ms (2Hz) to render an image with dimensions 424 x 240. To improve render speed and minimize loop-closure latency, we use a two-step process. Initially, we generate NeRF renders at half resolution (212 x 140). Then, we employ the lightweight FSRCNN [52] for up-sampling to the original size. This approach strikes a balance between computational speed

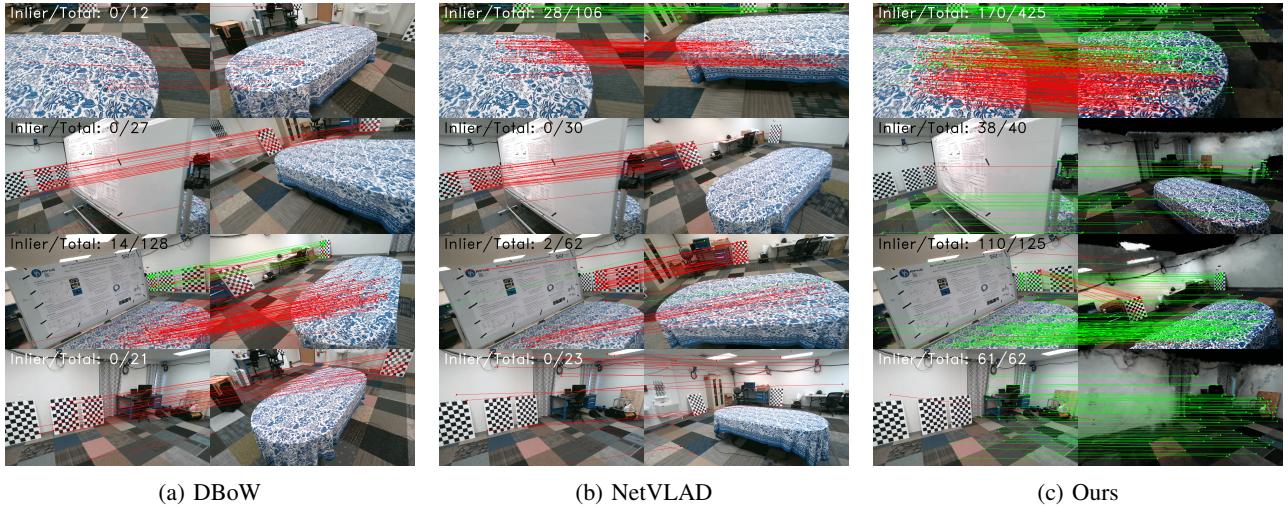


Fig. 3: Qualitative study of failure cases of classical place recognition method. Green and Red lines indicate inliers and outliers, respectively. Input image (left of each column) and retrieved, rendered for the NeRF case, image is shown (right of each column).

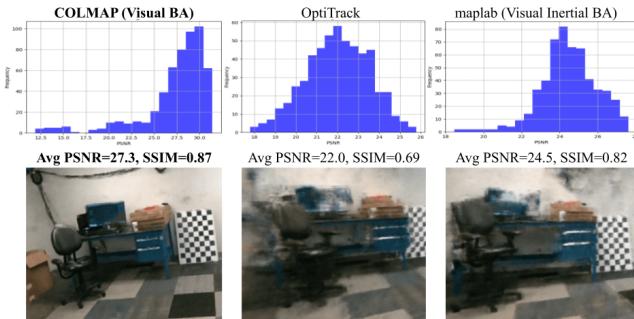


Fig. 4: Qualitative comparison of NeRF Map trained with different methods using 543 keyframe images. The top row shows the PSNR histograms and the bottom row shows exemplary images rendered from each method.

and image quality, as demonstrated in Fig. 2. We further enhance performance by reducing resolution levels and the hashing size of the model in InstantNGP [20]. Additionally, we minimize multiple copy times to the CPU by directly transferring GPU-rendered images.

The rendering is run on a separate thread to prevent blocking of the real-time VINS. The SuperPoint feature matching network has been modified to use a lightweight ResNet18 [60] and optimized to support a 16-bit floating point using the TensorRT pipeline [53] to further improve performance. This secondary thread which performs rendering and matching runs at an adequate speed of 15Hz on the Jetson to aid in real-time estimation.

### C. NeRF Map Generation

We now detail how we build the prior NeRF map to aid visual-inertial localization. We explored a variety of Bundle Adjustment (BA) methods to examine how they impact the map quality of the NeRF created by InstantNGP [20] on the Table 5 dataset (see Fig. 1 offline part). First, maplab [13], which optimizes both visual reprojection and inertial errors in

a visual-inertial BA, resulting 543 keyframes over the dataset was used. These keyframes were also fed into COLMAP [26], [27] which does up-to-scale bundle adjustment. Finally, the groundtruth motion capture pose and inertial information was processed using vicon2gt [61] to recover the pose of each keyframe. We use the Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM) to evaluate the quality of rendered images [39].

Most importantly, as shown in Fig. 4, we find that the vision-only BA of COLMAP (without scale) yields the superior map quality. It is our conjecture that the fusion of inertial information has introduced additional errors by requiring additional calibration and sensor synchronization. However, the groundtruth (motion capture) pose and inertial optimization do not intrinsically minimize any re-projection errors and thus the poses are likely not optimal in the desired geometric re-projection errors crucial for high quality NeRF creation. Nevertheless, this calls for future in-depth investigation. For this reason, we chose to use COLMAP poses to train our map. To correct the ambiguous scale, we aligned the COLMAP and groundtruth poses based on similarity transformation (sim3).

## V. EXPERIMENTAL VALIDATION

We validate the proposed NeRF-VINS and baseline methods on the recently released AR Table Dataset [62]. This dataset is ideal for NeRF reconstruction due to its object-centric trajectories which observe a table placed centrally. This dataset additionally enables us to evaluate the robustness of algorithms to changing environments (see Fig. 6), due to the addition of a whiteboard for the three datasets (Table 5-7) and the moving of the table to the side of the room in Table 8. Unless specifically noted, all prior map methods leverage Table 1 for datasets 1-4 and Table 5 for 5-7. The dataset has 30Hz RGB images and 400Hz IMU from an Intel Realsense D455 and groundtruth poses from a full-room OptiTrack rig.

In particular, for comprehensive validation, we evaluated the following state-of-the-art methods:

1) **Single-Shot Visual Localization**: The open-source Hierarchical Localization (HLoc) system [9] that used NetVLAD for image retrieval, and SuperPoint [29] descriptor establishes a baseline for expected state-of-the-art performance. In this system, local matching is performed using a nearest-neighbor search with a ratio test and geometric verification, which aligns with our pipeline. Notably, the use of Lightglue [63] matching remains computationally expensive (16 ms for a pair, thus 800 ms for top 50 on A3000 GPU) and did not yield substantially better results. The same images and poses that are used to train the NeRF are leveraged in its map. We evaluated the performance with the top 5 and 50 nearest neighbor matches: HLoc (top5) and HLoc (top50), respectively. Due to its single-shot nature, we found that for many image localization accuracy was poor, and thus in most results presented we select an inlier set of good quality success to provide a reasonable comparison. Note that the below map-based methods and proposed NeRF-VINS provide *continuous* estimation.

2) **Map-based Visual-Inertial Localization**: For map-based VINS, the filter-based ROVIO with additional re-localization module [34] (ROVIOLI) from maplab [13] provides one of the closest direct comparisons to the proposed method. We report the accuracy of both the odometry, ROVIOLI, and the map-aided, ROVIOLI+Map, which leverages the maplab optimized prior map with the same keyframes used to train the NeRF. VINS-Fusion [64], is additionally compared against as it has support in its secondary loop-closure thread for re-localization against a previous-built relative pose graph using DBoW2 [14]. Thus we run VINS-Fusion on the prior map dataset to generate a pose graph that is then leveraged for sequential datasets (e.g. the whole dataset Table 1 is used, as compared to the other approaches that use a small set of keyframes). Both the odometry, VF, the secondary pose graph without relocalization, VF+Loop, and then the secondary thread which is able to relocalize against the prior map pose graph, VF+Loop+Map.

#### A. Localization Accuracy

Table IV shows the Absolute Trajectory Error (ATE) of each state-of-the-art AR Table dataset including the proposed method on our desktop (termed Nerf-VINS (D)) equipped with an A4500 NVIDIA graphics card and on Jetson AGX Orin (termed Nerf-VINS (J)), and our VINS system (Open-VINS [49]). It is clear that our proposed method achieved one of the best accuracies over all algorithms while HLoc showed competing results (note we excluded large failures of HLoc from statistics). An interesting observation is that VF reported higher accuracy than VINS-Fusion+Loop which was due to multiple false loop closures induced by incorrect DBoW matching. It is extrapolated that other method that leverages DBoW showed lower performance than the proposed method for the same reason. Additionally, the Relative Pose Error (see Fig. 5) highlights the significant advantage of incorporating NeRF map features, which effectively mitigates drift and maintains error within bounds. We attribute

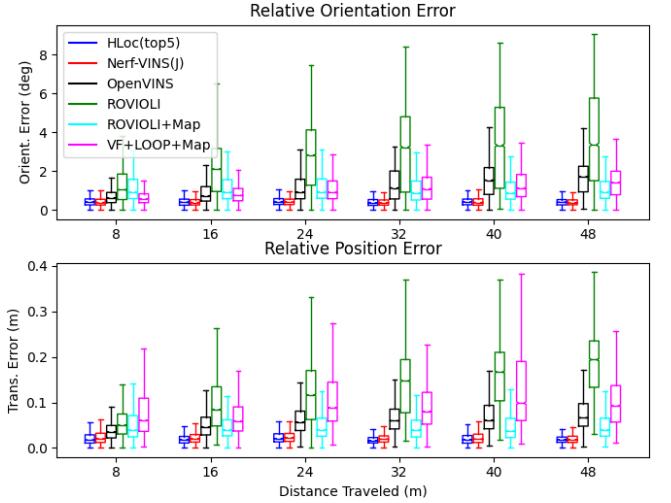


Fig. 5: Boxplot of the relative error trajectory error statistics. The middle box spans the first and third quartiles, while the whiskers are the upper and lower limits.

this performance gain is the ability to effectively utilize NeRF in our pipeline that can provide good novel scenes resulting in good viewpoints and a good number of quality measurements (Fig. 3). Though HLoc was able to provide good accuracy, there were many failures that were excluded from the statistics, and moreover, the classification of inliers and outliers for real-time estimation is challenging.

We additionally investigated the average timing of each function of our system and compared it with HLoc. Note that we disabled the multi-threading of our system for a fair comparison. The results reported in Table III show the total time of the proposed system takes 30 ms which is almost half of the total timing of HLoc with top 5 match results. Though the performance of HLoc can be improved by retrieving more images, however, this will introduce a significant computation burden for local matching and PnP, making it difficult to run in real-time (HLoc (top 50) pipelines take 331.9 ms per frame as shown in Table III). This clearly shows that our lightweight pipeline is capable of high-rate rendering of the NeRF images enabling real-time localization fully leveraging the NeRF map information.

#### B. Robustness to Environment Changes

To assess our system in generating favorable viewpoints enabling robust localization even when the environment is changed after mapping, we examined a more challenging scenario: employing Table 1 as the map and running on Table 5-8 each with distinct environments (refer to Fig. 6). Our system still shows one of the best records showing robust localization performance which is also competitive with HLoc (note that HLoc encounters numerous failures, which



Fig. 6: Environmental changes over different sequences of AR Table dataset.

TABLE II: The Absolute Trajectory Errors (ATE) of each state-of-the-art algorithm on the AR Table dataset (degree/cm). The best results are highlighted with a bold green color.

Algorithms	Table 1	Table 2	Table 3	Table 4	Table 5	Table 6	Table 7	Average
Map-based	Nerf-VINS (D)	0.51 / 1.8	0.27 / 1.0	0.50 / 1.0	0.35 / 1.5	0.43 / 1.4	0.59 / 1.9	0.46 / 1.6
	Nerf-VINS (J)	0.47 / 2.0	0.29 / 0.8	0.50 / 0.9	0.31 / 1.6	0.43 / 1.3	0.54 / 1.9	0.51 / 1.7
	VF+Loop+Map	0.93 / 4.1	1.27 / 7.1	0.88 / 6.1	1.39 / 5.2	0.72 / 3.2	0.93 / 3.7	1.68 / 5.3
	ROVIOLI+Map	0.54 / 2.1	1.30 / 3.6	0.67 / 2.2	1.15 / 4.3	0.86 / 3.7	2.33 / 17.9	2.42 / 13.6
	HLoc (top5)*	0.41 / 1.0	0.40 / 1.6	0.38 / 1.4	0.31 / 1.3	0.41 / 1.2	0.60 / 1.6	0.51 / 2.0
	HLoc (top50)*	0.41 / 1.0	0.33 / 1.4	0.35 / 1.2	0.30 / 1.2	0.40 / 1.2	0.57 / 1.6	0.51 / 2.0
VINS	OpenVINS	1.17 / 5.43	0.55 / 2.16	1.02 / 3.40	1.21 / 5.88	0.50 / 3.28	1.04 / 3.73	1.31 / 7.23
	ROVIOLI	2.05 / 7.08	1.11 / 4.08	2.63 / 7.86	1.48 / 11.07	2.50 / 12.08	1.10 / 4.25	3.12 / 15.92
	VF+Loop	1.25 / 6.7	1.18 / 9.2	0.95 / 6.5	1.10 / 5.7	0.88 / 2.8	0.98 / 11.2	1.57 / 10.1
	VF	1.62 / 5.8	1.32 / 3.0	1.47 / 7.6	1.75 / 5.6	1.12 / 3.4	0.98 / 5.3	1.67 / 9.3

\* Large failures (errors larger than 5 degrees or 10 cm) of HLoc (top5) and HLoc (top50) are excluded from statistics:

HLoc (top5) failure rates: Table 2 37%, Table 3 5.5%, Table 4 0.4%, Table 5 0.5%, Table 6 1%, Table 7 0.5%

HLoc (top50) failure rates: Table 2 39%, Table 3 2.4%, Table 4 0.4%

TABLE III: Average timing for proposed NeRF-VINS and HLoc pipeline in milliseconds. Recorded on a laptop with A3000 GPU and 11th Gen Intel(R) Core(TM) i7-11800H @ 2.30GHz CPU.

Step	Nerf-VINS (D)	HLoc (top 5)	HLoc (top 50)
Tracking	8.5	-	-
Rendering / NetVLAD	11.6	12.9	12.9
Superpoint Extraction	5.4	7.6	7.6
Local Matching	1.7	15.2	153.7
Update / PnP	2.5	21.3	157.7
Total	29.8	57.0	331.9

TABLE IV: AR table ATE (degree / cm) and Table 1 is used as a map for the following sequence. Blanks indicate failures.

Algorithm	Table 5	Table 6	Table 7	Table 8	Average
Nerf-VINS (J)	0.49 / 3.0	0.61 / 4.1	0.54 / 3.3	0.38 / 3.0	<b>0.50 / 3.4</b>
HLoc (top5)*	0.61 / 3.5	0.64 / 3.6	0.61 / 3.1	0.50 / 3.7	0.59 / 3.5
HLoc (top50)*	0.65 / 3.4	0.67 / 3.6	0.62 / 3.0	0.47 / 3.1	0.60 / <b>3.3</b>
VF+Loop+Map	0.95 / 12.4	0.82 / 3.3	1.60 / 9.3	2.44 / 9.9	1.45 / 8.7
ROVIOLI+Map	2.48 / 11.3	1.89 / 12.9	2.59 / 14.8	- / -	2.32 / 13.0

\* HLoc error larger than 5 degrees or 10 cm are removed to be presentable  
HLoc (top5) failure rates: Table 5 38.9%, Table 6 36.8%, Table 7 37.1%,  
Table 8 30.5%

HLoc (top50) failure rates: Table 5 23.9%, Table 6 29.4%, Table 7 21.8%,  
Table 8 10.7%

are omitted from consideration). In contrast, our system consistently delivers advantageous viewpoints, facilitating large inlier measurements (Fig. 3).

As can be seen from Fig. 7, around 80% percent of images for our pipeline are localized within a 2.5 cm high accuracy threshold, while HLoc is only around 70% when matching with the top 50 images. Our system is able to localize almost all the images within a 7.5 cm position error, while HLoc using the top 5 images and top 50 can only localize 80.9% and 89.3% images within a 20 cm error bound, respectively.

### C. Discussion and Limitations

While we have demonstrated that the proposed method exhibits superior localization performance, similar to other NeRF methods, our map is also object-centric. To train the map effectively, we require images that surround the object, which may not be scalable to larger environments. We argue that this limitation can be addressed by leveraging

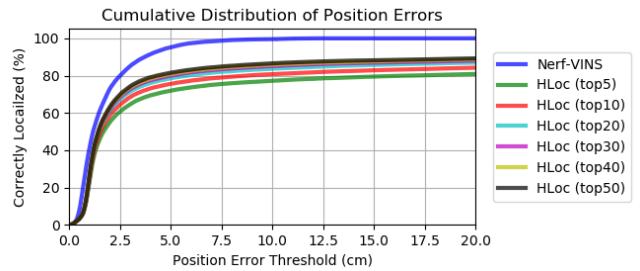


Fig. 7: The percentage of images successfully localized under a certain position error threshold using Table 1 as a map to evaluate Table 1-8.

the combination of F2-Nerf [65] and Block-Nerf [66], which does not assume to have bounded camera trajectory, and the other one trains large scale NeRF maps. Our pipeline can benefit from faster-rendering speed as we can leverage more viewpoints for stronger constraints. This is a quickly changing landscape with recent works such as Kerbl et al. [67] offering greater rendering speed and opening a new avenue for exploration. We leave these as future work.

## VI. CONCLUSIONS AND FUTURE WORK

In this paper, we have developed a real-time tightly-coupled NeRF-VINS algorithm. Built on top of the MSCKF-VINS, the proposed NeRF-VINS extends to efficiently and accurately fuses the NeRF synthetic images to overcome the limited viewpoint challenges commonly encountered by the keyframe map-based localization methods. In particular, as NeRF is able to generate novel views from any viewpoint, we exploit this advantage to synthesize better views to provide higher inlier matches that allow for full utilization of the map information, resulting in performance gain. In the future, we will investigate NeRF map-based initialization i.e., initializing the transform between the IMU and map frames.

## ACKNOWLEDGMENT

This work is partially supported by the University of Delaware (UD) College of Engineering, and Google ARCore. The authors thank Nathaniel Merrill for his fruitful initial discussions about the project.

## REFERENCES

- [1] Google, “ARCore,” <https://developers.google.com/ar>.
- [2] Apple, “ARKit,” <https://developer.apple.com/augmented-reality/>.
- [3] S. Katragadda, B. A. Mondal, and A. Deane, “Stereoscopic mixed reality in unmanned aerial vehicle search and rescue,” in *AIAA Scitech 2019 Forum*, 2019, p. 0154.
- [4] C. Chen, Y. Yang, P. Geneva, W. Lee, and G. Huang, “Visual-inertial-aided online mav system identification,” in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2022.
- [5] C. Häne, L. Heng, G. H. Lee, F. Fraundorfer, P. Furgale, T. Sattler, and M. Pollefeys, “3d visual perception for self-driving cars using a multi-camera system: Calibration, mapping, localization, and obstacle detection,” *Image and Vision Computing*, 2017.
- [6] G. Huang, “Visual-inertial navigation: A concise review,” in *Proc. International Conference on Robotics and Automation*, Montreal, Canada, May 2019.
- [7] S. Lynen, T. Sattler, M. Bosse, J. A. Hesch, M. Pollefeys, and R. Siegwart, “Get out of my lab: Large-scale, real-time visual-inertial localization,” in *Robotics: Science and Systems*, vol. 1, 2015, p. 1.
- [8] A. Kasyanov, F. Engelmann, J. Stückler, and B. Leibe, “Keyframe-based visual-inertial online slam with relocalization,” in *2017 IEEE/RSJ international conference on intelligent robots and systems (IROS)*. IEEE, 2017, pp. 6662–6669.
- [9] P.-E. Sarlin, C. Cadena, R. Siegwart, and M. Dymczyk, “From coarse to fine: Robust hierarchical localization at large scale,” in *CVPR*, 2019.
- [10] P. Geneva and G. Huang, “Map-based visual-inertial localization: A numerical study,” in *Proc. International Conference on Robotics and Automation*, Philadelphia, USA, May 2022.
- [11] 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*, 2009.
- [12] T. Schneider, M. T. Dymczyk, M. Fehr, K. Egger, S. Lynen, I. Gilitschenski, and R. Siegwart, “maplab: An Open Framework for Research in Visual-inertial Mapping and Localization,” *IEEE Robotics and Automation Letters*, vol. 3, no. 3, pp. 1418–1425, 2018.
- [13] A. Cramariuc, L. Bernreiter, F. Tschopp, M. Fehr, V. Reijgwart, J. Nieto, R. Siegwart, and C. Cadena, “maplab 2.0 – A Modular and Multi-Modal Mapping Framework,” *IEEE Robotics and Automation Letters*, vol. 8, no. 2, pp. 520–527, 2023.
- [14] D. Gálvez-López and J. D. Tardós, “Bags of binary words for fast place recognition in image sequences,” *IEEE Transactions on Robotics*, 2012.
- [15] S. Lynen, M. Bosse, P. Furgale, and R. Siegwart, “Placeless place-recognition,” in *2014 2nd International Conference on 3D Vision*, vol. 1. IEEE, 2014, pp. 303–310.
- [16] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, “Netvlad: Cnn architecture for weakly supervised place recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.
- [17] K. Li, Y. Ma, X. Wang, L. Ji, N. Geng, et al., “Evaluation of global descriptor methods for appearance-based visual place recognition,” *Journal of Robotics*, vol. 2023, 2023.
- [18] C. Jiang, H. Zhang, P. Liu, Z. Yu, H. Cheng, B. Zhou, and S. Shen, “H2-mapping: Real-time dense mapping using hierarchical hybrid representation,” 2023.
- [19] C.-M. Chung, Y.-C. Tseng, Y.-C. Hsu, X.-Q. Shi, Y.-H. Hua, J.-F. Yeh, W.-C. Chen, Y.-T. Chen, and W. H. Hsu, “Orbeez-slam: A real-time monocular visual slam with orb features and nerf-realized mapping,” in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 9400–9406.
- [20] T. Müller, A. Evans, C. Schied, and A. Keller, “Instant neural graphics primitives with a multiresolution hash encoding,” *ACM Trans. Graph.*, 2022.
- [21] O. Chum, J. Matas, and J. Kittler, “Locally optimized ransac,” in *Pattern Recognition: 25th DAGM Symposium, Magdeburg, Germany, September 10-12, 2003. Proceedings 25*. Springer, 2003.
- [22] O. Chum and J. Matas, “Optimal randomized ransac,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2008.
- [23] T. Sattler, B. Leibe, and L. Kobbelt, “Fast image-based localization using direct 2d-to-3d matching,” in *2011 International Conference on Computer Vision*, 2011.
- [24] L. Liu, H. Li, and Y. Dai, “Efficient global 2d-3d matching for camera localization in a large-scale 3d map,” in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017.
- [25] T. Sattler, B. Leibe, and L. Kobbelt, “Efficient & effective prioritized matching for large-scale image-based localization,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.
- [26] J. L. Schönberger and J.-M. Frahm, “Structure-from-motion revisited,” in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- [27] J. L. Schönberger, E. Zheng, M. Pollefeys, and J.-M. Frahm, “Pixel-wise view selection for unstructured multi-view stereo,” in *European Conference on Computer Vision (ECCV)*, 2016.
- [28] M. Humenberger, Y. Cabon, N. Guerin, J. Morat, V. Leroy, J. Revaud, P. Rerole, N. Pion, C. de Souza, and G. Csurka, “Robust image retrieval-based visual localization using kapture,” 2020.
- [29] D. DeTone, T. Malisiewicz, and A. Rabinovich, “Superpoint: Self-supervised interest point detection and description,” in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2018.
- [30] A. Kendall, M. Grimes, and R. Cipolla, “Posenet: A convolutional network for real-time 6-dof camera relocalization,” in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2938–2946.
- [31] A. Kendall and R. Cipolla, “Geometric loss functions for camera pose regression with deep learning,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 5974–5983.
- [32] L. Yang, Z. Bai, C. Tang, H. Li, Y. Furukawa, and P. Tan, “Sanet: Scene agnostic network for camera localization,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 42–51.
- [33] M. Bloesch, S. Omari, M. Hutter, and R. Siegwart, “Robust visual inertial odometry using a direct ekf-based approach,” in *2015 IEEE/RSJ international conference on intelligent robots and systems (IROS)*. IEEE, 2015, pp. 298–304.
- [34] 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*, 2017.
- [35] G. Klein and D. Murray, “Parallel tracking and mapping for small AR workspaces,” in *Proc. Sixth IEEE and ACM International Symposium Mixed and Augmented Reality*, 2007.
- [36] S. Leutenegger, S. Lynen, M. Bosse, R. Siegwart, and P. Furgale, “Keyframe-based visual-inertial odometry using nonlinear optimization,” *The International Journal of Robotics Research*, 2015.
- [37] J. Engel, T. Schöps, and D. Cremers, “Lsd-slam: Large-scale direct monocular slam,” in *European conference on computer vision*. Springer, 2014, pp. 834–849.
- [38] C. Forster, M. Pizzoli, and D. Scaramuzza, “SVO: Fast semi-direct monocular visual odometry,” in *IEEE International Conference on Robotics and Automation (ICRA)*, 2014.
- [39] 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,” *ECCV*, 2020.
- [40] C.-H. Lin, W.-C. Ma, A. Torralba, and S. Lucey, “Barf: Bundle-adjusting neural radiance fields,” in *IEEE International Conference on Computer Vision (ICCV)*, 2021.
- [41] Z. Wang, S. Wu, W. Xie, M. Chen, and V. Adrian Prisacariu, “Nerf—: Neural radiance fields without known camera parameters,” 2021.
- [42] E. Sucar, S. Liu, J. Ortiz, and A. Davison, “iMAP: Implicit mapping and positioning in real-time,” in *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021.
- [43] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.
- [44] A. Rosinol, J. J. Leonard, and L. Carlone, “NeRF-SLAM: Real-time dense monocular slam with neural radiance fields,” 2022.
- [45] 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 International Conference on Intelligent Robots and Systems (IROS)*, 2021.
- [46] D. Maggio, M. Abate, J. Shi, C. Mario, and L. Carlone, “Loc-nerf: Monte carlo localization using neural radiance fields,” in *IEEE International Conference on Robotics and Automation*. IEEE, 2023.
- [47] M. Adamkiewicz, T. Chen, A. Caccavale, R. Gardner, P. Culbertson, J. Bohg, and M. Schwager, “Vision-only robot navigation in a neural radiance world,” *IEEE Robotics and Automation Letters*, 2022.

- [48] A. I. Mourikis and S. I. Roumeliotis, “A multi-state constraint kalman filter for vision-aided inertial navigation,” in *Proceedings 2007 IEEE International Conference on Robotics and Automation*, 2007.
- [49] P. Geneva, K. Eckenhoff, W. Lee, Y. Yang, and G. Huang, “OpenVINS: a research platform for visual-inertial estimation,” in *Proc. of the IEEE International Conference on Robotics and Automation*, Paris, France, 2020. [Online]. Available: [https://github.com/rpng/open\\_vins](https://github.com/rpng/open_vins)
- [50] A. B. Chatfield, *Fundamentals of High Accuracy Inertial Navigation*. Reston, VA: American Institute of Aeronautics and Astronautics, Inc., 1997.
- [51] P. Geneva, K. Eckenhoff, and G. Huang, “A linear-complexity EKF for visual-inertial navigation with loop closures,” in *Proc. International Conference on Robotics and Automation*, Montreal, Canada, May 2019.
- [52] C. Dong, C. C. Loy, and X. Tang, “Accelerating the super-resolution convolutional neural network,” in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part II 14*. Springer, 2016.
- [53] NVIDIA, “TensorRT,” <https://github.com/NVIDIA/TensorRT>.
- [54] G. Bradski, “The OpenCV Library,” *Dr. Dobb’s Journal of Software Tools*, 2000.
- [55] P. F. Alcantarilla and T. Solutions, “Fast explicit diffusion for accelerated features in nonlinear scale spaces.” *IEEE Trans. Patt. Anal. Mach. Intell.*, vol. 34, no. 7, pp. 1281–1298, 2011.
- [56] P. F. Alcantarilla, A. Bartoli, and A. J. Davison, “Kaze features,” in *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7–13, 2012, Proceedings, Part VI 12*. Springer, 2012, pp. 214–227.
- [57] S. Leutenegger, M. Chli, and R. Y. Siegwart, “Brisk: Binary robust invariant scalable keypoints,” in *2011 International conference on computer vision*. Ieee, 2011, pp. 2548–2555.
- [58] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, “Orb: An efficient alternative to sift or surf,” in *2011 International conference on computer vision*. Ieee, 2011, pp. 2564–2571.
- [59] C. Chen, Y. Yang, P. Geneva, and G. Huang, “FEJ2: a consistent visual-inertial state estimator design,” in *Proc. International Conference on Robotics and Automation*, Philadelphia, USA, May 2022.
- [60] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.
- [61] P. Geneva and G. Huang, “vicon2gt: Derivations and analysis,” University of Delaware, Tech. Rep. RPNG-2020-VICON2GT, 2020, available: [http://udel.edu/~ghuang/papers/tr\\_vicon2gt.pdf](http://udel.edu/~ghuang/papers/tr_vicon2gt.pdf).
- [62] C. Chen, P. Geneva, Y. Peng, W. Lee, and G. Huang, “Monocular visual-inertial odometry with planar regularities,” in *Proc. of the IEEE International Conference on Robotics and Automation*, London, UK., 2023. [Online]. Available: [https://github.com/rpng/ar\\_table\\_dataset](https://github.com/rpng/ar_table_dataset)
- [63] P. Lindenberger, P.-E. Sarlin, and M. Pollefeys, “LightGlue: Local Feature Matching at Light Speed,” in *ICCV*, 2023.
- [64] T. Qin, S. Cao, J. Pan, and S. Shen, “A general optimization-based framework for global pose estimation with multiple sensors,” *arXiv preprint arXiv:1901.03642*, 2019.
- [65] P. Wang, Y. Liu, Z. Chen, L. Liu, Z. Liu, T. Komura, C. Theobalt, and W. Wang, “F2-nerf: Fast neural radiance field training with free camera trajectories,” in *CVPR*, 2023.
- [66] M. Tancik, V. Casser, X. Yan, S. Pradhan, B. Mildenhall, P. Srinivasan, J. T. Barron, and H. Kretzschmar, “Block-NeRF: Scalable large scene neural view synthesis,” *arXiv*, 2022.
- [67] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, “3d gaussian splatting for real-time radiance field rendering,” *ACM Transactions on Graphics*, 2023.