2024 IEEE International Conference on Robotics and Automation (ICRA 2024)
May 13-17, 2024. Yokohama, Japan

HERO-SLAM: Hybrid Enhanced Robust Optimization of Neural SLAM

Zhe Xin[1], Yufeng Yue[2], Liangjun Zhang[3], and Chenming Wu[3,†]

***Abstract*— Simultaneous Localization and Mapping (SLAM) is a fundamental task in robotics, driving numerous applications such as autonomous driving and virtual reality. Recent progress on neural implicit SLAM has shown encouraging and impressive results. However, the robustness of neural SLAM, particularly in challenging or data-limited situations, remains an unresolved issue. This paper presents HERO-SLAM, a Hybrid Enhanced Robust Optimization method for neural SLAM, which combines the benefits of neural implicit field and feature-metric optimization. This hybrid method optimizes a multi-resolution implicit field and enhances robustness in challenging environments with sudden viewpoint changes or sparse data collection. Our comprehensive experimental results on benchmarking datasets validate the effectiveness of our hybrid approach, demonstrating its superior performance over existing implicit field-based methods in challenging scenarios. HERO-SLAM provides a new pathway to enhance the stability, performance, and applicability of neural SLAM in real-world scenarios. Project page: *https://hero-slam.github.io*.**

I. Introduction

Visual Simultaneous Localization and Mapping (SLAM) is a fundamental task in robotics and computer vision that drives many applications, spanning from the intricacies of robot navigation and 3D scene reconstruction, to the cutting-edge fields of autonomous driving and virtual reality. The essence of visual SLAM lies in its ability to reconstruct the structure and visual details of a 3D environment, all while tracking the camera's position in real-time. The keys to its success in real-world applications are relying on runtime efficiency, scalability, and most importantly, robustness.

Visual SLAM can primarily be divided into two categories, sparse and dense, based on the nature of the map reconstruction. Specifically, sparse SLAM predominantly concentrates on deducing the camera trajectory from the sequential sensor data, generating sparse point clouds. In contrast, dense SLAM not only contemplates pose estimation but also initiates a detailed surface reconstruction. Conventional dense visual SLAM approaches heavily lean on manually engineered features and matching strategies. These methods often incur a significant computational expense to solve pre-established optimization issues. Recent advances in coordinate-based neural networks motivate many studies using implicit field representation in visual SLAM. A coordinate point can be encoded using sinusoidal positional or other formats of frequency encoding [2] to represent high-frequency details compactly. The benefits of using implicit field-based representation in dense visual SLAM tasks have been confirmed by pioneering work such as iMAP [3] and NICE-SLAM [4]. However, these methods are associated with high computational burdens and their running speeds are approximately 0.1 to 1 Hz, which restricts their applicability to a broader range of tasks. Recent methods like Co-SLAM [5] and E-SLAM [6] are designed to push forward the boundary of implicit field-based visual SLAM. Compared to iMAP [3] and NICE-SLAM [4], these methods have substantially improved the quality of dense reconstruction and pose estimation. Despite these improvements, an important issue that hinders the wider range application of neural SLAM is the robustness to tackle challenging scenes, for example, the circumstance when the number of provided frames falls below the standard camera frequency, which is very common in real-world applications due to the constraints such as limited bandwidth for data transferring or storage availability. Under these conditions, the success rate of existing methods is not satisfactory. In short, while recent advancements in neural implicit field-based visual SLAM have shown promise, there is still a need to improve their robustness and applicability in real-world applications.

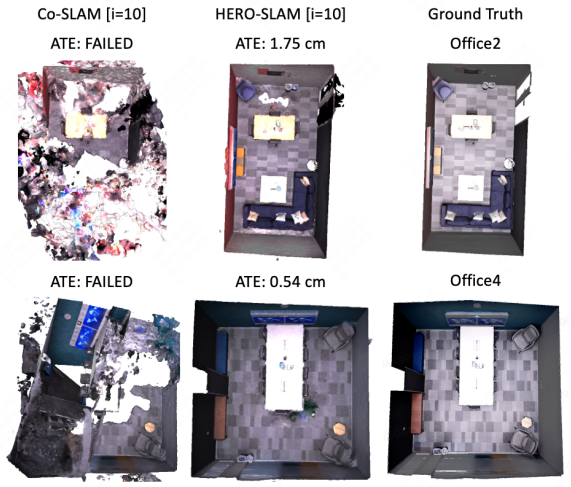

Fig. 1. The visualization of mapping and tracking errors on Replica [1] of challenging sparse inputs with large motion changes. This paper introduces a robust system for real-time dense 3D reconstruction, dubbed HERO-SLAM, which synergistically leverages the capabilities of neural implicit fields and feature-metric optimization, demonstrating exceptional resilience to large viewpoint changes and ensuring efficient runtime performance.

[1]Z. Xin is with the Institute of Automation, Chinese Academy of Sciences, Beijing, China. zhexin2015@gmail.com
[2]Y. Yue is with the School of Automation, Beijing Institute of Technology, Beijing, China. yueyufeng@bit.edu.cn
[3]L. Zhang and C. Wu are with the Robotics and Autonomous Driving Lab (RAL), Baidu Research. liangjun.zhang@gmail.com, wcm94@live.com, † denotes corresponding author.
This work was partially supported by National Natural Science Foundation of China under Grant No. NSFC 62233002, 92370203. The authors would like to thank F. Zhang for the suggestions.

The robustness issue that exists in neural SLAM approaches comes from the difficulty of optimizing neural networks. Despite the diverse underlying neural representations used to describe the implicit fields - including multilayer perceptron (MLP) [7], hash grid [8], codebook [9], triplane [10], and dense grid [11] - they all essentially function as large nonlinear optimization systems. Therefore, input images' quality, view coverage, and relevance are key determinants of the neural implicit fields. However, in situations where data is challenging or limited, the low relevance across all data frames can easily mislead the optimization process toward ambiguous local solutions. In light of these challenges, our work seeks to explore a new path, which in particular designs a ***hybrid*** representation that leverages both the capabilities of neural implicit field and feature-metric optimization. We aim to address the robustness problem for dense neural SLAM. This approach significantly improves the stability and performance of SLAM methods, particularly in challenging and data-limited situations, as shown in Fig. 1. The contributions of our work are summarized as follows.

- We propose a method that effectively leverages the advantages of neural implicit field and feature-metric optimization for visual SLAM. This results in increased robustness, especially in challenging environments involving abrupt view changes or sparse data collection.
- We propose a novel pipeline to optimize the hybrid feature-metric implicit fields using multiscale patch-based loss, which computes based on the warpings between feature points, feature maps, and RGB-D pixels.
- The comprehensive experiments on widely used benchmark datasets validate the effectiveness of our hybrid approach, particularly its superior performance compared to existing neural implicit field-based methods in challenging scenarios.

The outline of this paper is as follows. Section II provides a comprehensive literature review. We then present the detailed illustration of our proposed method, HERO-SLAM, in Section III. In Section IV, we extensively evaluate the performance of our method and validate the effectiveness of its various modules.

II. Related Work

A. Visual SLAM

Visual SLAM has emerged as a fundamental research area in the domains of robotics and computer vision. Traditional approaches for sparse/semi-dense visual SLAM include MonoSLAM [12], ORB-SLAM [13], VINS [14], LSD [15], DSO [16], where feature matching is widely used to recover the camera poses. In dense visual SLAM area, traditional approaches include KinectFusion [17], ElasticFusion [18], where RGB-D sensors are required and the scene completeness is unsatisfying. Recently, deep learning has gained more and more attention, Droid-SLAM [19] estimates motion fields between frames, which is highly computationally expensive and requires a large memory footprint. TANDEM [20] uses a pre-trained MVSNet [21]-like neural network on monocular depth estimation. Unlike these methods, our work uses neural implicit fields to estimate camera poses and reconstruct the scene simultaneously, achieving better scene completeness and higher rendering quality for less observed regions.

B. Neural Implicit Field SLAM

Neural implicit fields have become a significant area of research in computer vision and robotics, offering a novel paradigm for scene representation and reconstruction. The Implicit Mapping and Planning (iMAP) framework [3] pioneers the use of deep implicit functions to represent 3D environments, providing a foundation for subsequent studies. The Neural Radiance Fields (NeRF) based LOAM (NeRF-LOAM) [22] extends this work by incorporating LiDAR odometry, enabling more accurate and efficient mapping. NICE-SLAM [4] and Co-SLAM [5] further expand on this by proposing improvements in efficiency and scalability, respectively. DIM-SLAM [23] and E-SLAM [6] demonstrate the versatility of neural implicit fields, showing how they can be used for dynamic scene reconstruction and event-based vision, respectively. Our work stands on the shoulders of these successful approaches, enhancing the robustness of existing neural implicit field SLAM methods.

C. Pose Optimization within NeRF

Another strand of research that bears similarity to our work pertains to the pose optimization within NeRF. However, the problem configuration diverges slightly from ours, as these methodologies do not necessitate the use of temporal data. The body of literature is growing, with key works including [24], which presents a novel approach to jointly optimize camera poses and scene representations. This has been further developed by the Bundle-Adjustable Radiance Field (BARF) [25], which integrates the bundle adjustment approach for more accurate and flexible 3D reconstructions. [26] makes a significant contribution by proposing a self-calibration mechanism for pose optimization, enhancing the accuracy of generated views. NoPe-NeRF [27] presents a novel approach for pose estimation using the neural implicit fields, which has important implications for pose optimization in NeRF. LocalRF [28] introduces a progressive optimization strategy to improve the robustness of view synthesis. It is worth noting that these methods primarily aim at reconstructing large-scale scenes. Consequently, the optimization process usually takes place offline and is associated with significant time expenditure.

III. HERO-SLAM

A. Overview

An overview pipeline of HERO-SLAM is shown in Fig. 2. The architecture of our SLAM system is similar to traditional dense SLAM systems, which has a tracking module to recover the pose of each frame, and a mapping module to reconstruct dense scenes from the tracked frames. We utilize a multi-resolution grid as the representation of the spatial feature, which can approximate an implicit function that

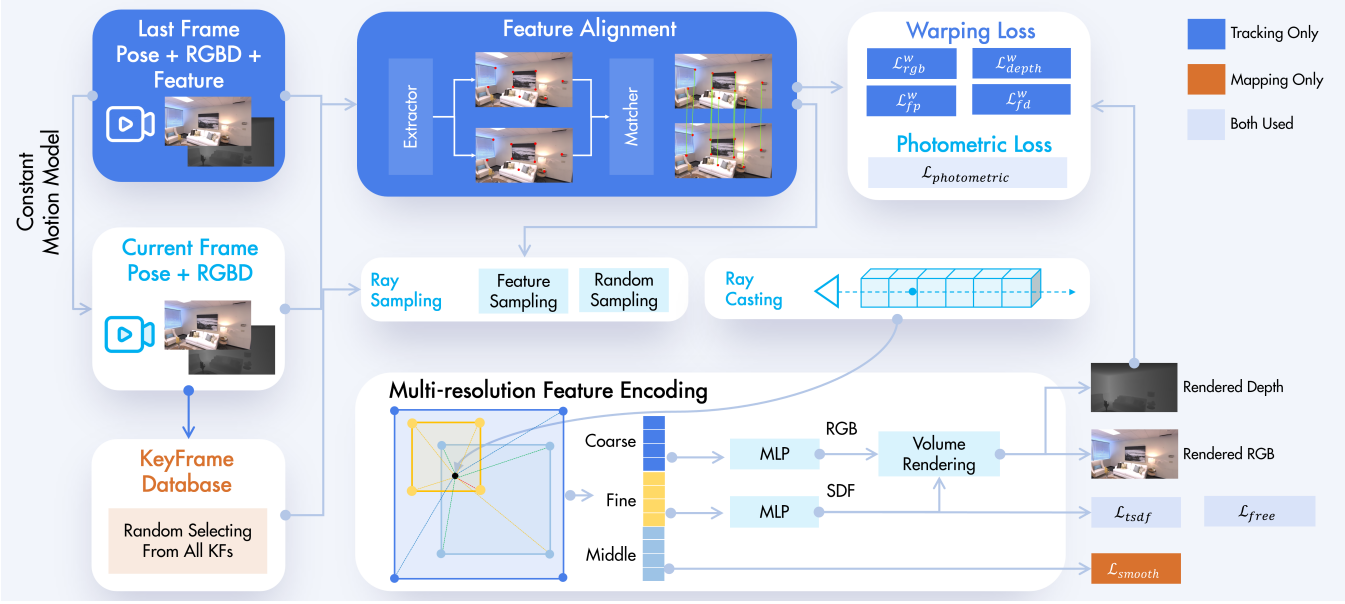

Fig. 2. The overview of HERO-SLAM. We use hybrid optimization to enhance the robustness of neural SLAM. Every newly captured frame would be aligned with the last frame for the camera pose estimation using feature-metric warping losses. The robustness and accuracy of tracking get improved, which in turn, facilitates the enhancement of mapping quality by optimizing the neural implicit field of multi-resolution feature encoding. The mapping module optimizes all keyframes from the keyframe database based on photometric reconstruction and depth supervision, following the volumetric rendering paradigm.

encodes the geometry and visual appearance of the scene. Through the process of sampling features from the volumetric grid along the viewing rays and subsequently querying these sampled features with the Multi-Layer Perceptron (MLP) decoder, we can use learning-based optimizers to optimize the rendering of each pixel's color and depth based on inferred camera parameters in a differentiable manner.

Our proposed system takes in a sequence of RGB-D frames over time, with varying data intervals and motion between frames. This sometimes results in challenging scenes for existing neural implicit field SLAM approaches but commonly happens in real-world applications. Our work advances the robustness of neural SLAM by proposing a hybrid enhanced robust optimization scheme, enabling us to leverage the neural SLAM in a variety of environments, achieving high-quality pose recovery and dense mapping.

B. Neural Implicit Field

1) Multi-resolution Neural Representation: A multi-resolution grid for implicit functions provides flexible and scalable means to encode complex geometrical and topological information. The grid's varying resolution allows for a greater level of detail where required, effectively capturing intricate aspects of the implicit function, while conserving computational resources in less detailed regions. The grid is used to encode the implicit functions representing the geometry of the 3D scene. The implicit function $f_\theta(\mathbf{x})$ at a location $\mathbf{x}$ in 3D space is represented as an MLP with parameters θ, which is trained to predict SDF values and appearance colors.

2) Color, Depth and Truncated Signed Distance: Following [3], [4], the representation of a scene can be effectively undertaken by employing multi-resolution representation with MLPs. A function in three dimensions, which accepts a spatial location $\mathbf{x} = (x, y, z)$ as an input, can be utilized to represent the scene:

$$\sigma, \mathbf{c} = f_\theta(\mathbf{x}). \tag{1}$$

when given a set of images I_i and the estimated poses P_i, we can sample particles to describe the intensity of the light that is either blocked or emitted along the ray. The color $\widehat{\mathcal{C}}(\mathbf{r})$ and depth $\widehat{\mathcal{D}}(\mathbf{r})$ of a ray $\mathbf{r}$ can be approximated by integrating the sampled particles along the ray as follows:

$$\widehat{\mathcal{C}}(\mathbf{r}) = \sum_{i=1}^{N} \exp(-\sum_{j=1}^{i-1} \sigma_i \delta_i)(1 - \exp(-\sigma_i \delta_i))\mathbf{c}_i, \tag{2}$$

$$\widehat{\mathcal{D}}(\mathbf{r}) = \sum_{i=1}^{N} \exp(-\sum_{j=1}^{i-1} \sigma_i \delta_i)(1 - \exp(-\sigma_i \delta_i)) \sum_{j=1}^{i} \delta_j, \tag{3}$$

where $\exp(-\sum_{j=1}^{i-1} \sigma_i \delta_i)$ represents the accumulated transmittance along the ray from the first sample to the i-th sample. The term $(1 - \exp(-\sigma_i \delta_i))$ denotes the alpha value of the current sample contributing to the rendered color and depth, while σ_i is the density of sample i, c_i is the predicted color of sample i, and δ_i is the distance from sample i to its next sample $i+1$. To supervise the training of f_θ, an L_2 photometric reconstruction loss is used:

$$\mathcal{L}_{photometric} = \sum_i \mathop{\mathbb{E}}_{\mathbf{r} \in I_i} ||\widehat{\mathcal{C}}(\mathbf{r}) - \mathcal{C}_i^{gt}(\mathbf{r})||_2^2, \tag{4}$$

where $\mathcal{C}_i^{gt}(\mathbf{r})$ is the ground truth color of $\mathbf{r}$ from image I_i. We follow [5] to make a conversion from the density to

the Truncated Signed Distance Field (TSDF) s_i by $\sigma_i = 1/((1+\exp^{s_i/\epsilon}) \cdot (1+\exp^{-s_i/\epsilon}))$, where ϵ is the parameter to truncate the distance field. Likewise, the depth supervision is applied to the TSDF as follows.

$$\mathcal{L}_{tsdf} = \sum_i \mathop{\mathbb{E}}_{\mathbf{r}\in I_i, |\mathcal{D}_i^{gt}(\mathbf{r})|<\epsilon} ||\widehat{\mathcal{D}}(\mathbf{r}) - \mathcal{D}_i^{gt}(\mathbf{r})||_2^2. \quad (5)$$

Besides, we adopt the same free space and smooth supervision from [5] to formulate $\mathcal{L}_{free}$ and $\mathcal{L}_{smooth}$.

C. Hybrid Enhanced Robust Optimization

The optimization scheme in Sec. III-B is computed in pixel-wise, while the underlying relationship in spatial-wise is not explicitly supervised by any loss function. We argue that this optimization scheme easily fails when relative motion between two consecutive frames is large, as the fact that the learning-based optimization uses the gradient descent method, which heavily relies on the initial guess and is easily stuck at local minima. Drawing the inspiration from [12], [29], [30], we additionally impose explicit supervision among the pairs of frames to facilitate tracking and mapping, by homography warping. Our hybrid enhanced robust optimization extends the neural implicit field SLAM from the perspective of feature metric matching, and all concluded into a set of warping losses to strengthen the supervision among different frames.

We first extend the pixel-wise photometric and depth supervision (Eq. 4 and 5) to multi-frame under the assumption of reprojection transformation obtained from the tracking module. To conquer the accuracy issue of warping, we adopt Structural Similarity Index (SSIM) [31] with 3×3 patches to compute $\mathcal{L}_{rgb}^w$ and $\mathcal{L}_{depth}^w$. Considering two frames I_i and I_j, the relative transformation in between is denoted as $\mathbf{R}_i^j$ and $\mathbf{t}_i^j$, we denote $\mathbf{H}_{ji}$ be the homography between them and $\mathbf{K}$ is the intrinsic parameter of the used camera.

$$\mathbf{H}_{ji} = \mathbf{K}(\mathbf{R}_i^j + \frac{\mathbf{t}_i^j \mathbf{n}_i^T \mathbf{R}_i^T}{\mathbf{n}_i^T(\mathbf{q}_i + \mathbf{R}_i^T \mathbf{t}_i)}), \quad (6)$$

where $\mathbf{n}_i$ is the normal. We denote a patch $\mathbf{P}_{\mathbf{q}_i}$ as the 3×3 patch centered at $\mathbf{q}_i$ in I_i, then its warped patch $\mathbf{P}_{\mathbf{q}_j}$ can be obtained by $\mathbf{H}_{ji}\mathbf{P}_{\mathbf{q}_i}$. We introduce the formulation of visibility mask $M_i(\mathbf{P}_{\mathbf{q}_i})$ from [32] to avoid warp invisible patches. Then we can define the warping losses of color and depth as follows.

$$\mathcal{L}_{rgb}^w = \frac{\sum_i M_i \cdot \text{SSIM}(I_i(\mathbf{P}_i), I_j(\mathbf{H}_{ji}\mathbf{P}_i))}{\sum_i M_i} \quad (7)$$

$$\mathcal{L}_{depth}^w = \frac{\sum_i M_i \cdot \text{SSIM}(D_i(\mathbf{P}_i), D_j(\mathbf{H}_{ji}\mathbf{P}_i))}{\sum_i M_i} \quad (8)$$

To further optimize the robustness of the proposed system, we opt to use the feature metric descriptor in image space to provide additional supervision to indicate the neural network optimized toward the guided direction. We use SuperPoint [33], which is a deep learning-based method designed for joint detection and description of interest points in an image, to extract the feature metric descriptor F from each frame. Then we apply LightGlue [34] to match the feature maps F_i and F_j, which augments the visual descriptors with context based on self- and cross-attention units with positional encoding. This helps introspect the feature maps and predicts a set of correspondences $\mathbf{S}_{ij}$ between the two frames, based on their pairwise similarity and unary matchability. Then, the projected pixel $\Pi(\mathbf{q}_i)$ in frame I_j of a pixel $\mathbf{q}_i$ lifted from frame I_i with homogeneous representation can be obtained by $\Pi(\mathbf{q}_i) = \mathbf{K}\left(D_{\mathbf{q}_i}\mathbf{R}_i^j\mathbf{K}^{-1}\mathbf{q}_i + \mathbf{t}_i^j\right)$, where $D_{\mathbf{q}_i}$ is the depth of $\mathbf{q}_i$.

To utilize the feature maps and the correspondences, we propose a hybrid enhanced robust optimization scheme, which optimizes the following feature points and feature maps pixel-wise loss functions during the tracking process,

$$\mathcal{L}_{fp}^w = \frac{\sum_{(\mathbf{q}_i,\mathbf{q}_j)\in S_{ij}} M_j ||\mathbf{q}_j - \Pi(\mathbf{q}_i))||_2}{\sum_{ij} M_j} \quad (9)$$

$$\mathcal{L}_{fd}^w = \frac{\sum_{(\mathbf{q}_i,\mathbf{q}_j)\in S_{ij}} M_j \cdot |F_j(\mathbf{q}_j) - F_j(\Pi(\mathbf{q}_i))|}{\sum_{ij} M_j} \quad (10)$$

The overall loss function $\mathcal{L}$ in optimizing the neural SLAM is defined as the sum of the above loss functions.

IV. Experiments

A. Experimental Setup

Datasets. We use three datasets for evaluation. Synthetic Replica [1] dataset is used to verify the quality of our reconstruction. Real-world ScanNet [36] and TUM RGB-D [37] datasets are used to evaluate pose estimation. Each dataset provides ground truth pose data. We follow [4] to pre-processing all testing data.

Metrics. To evaluate the quality of reconstruction, we utilize both 2D and 3D metrics. In the 2D metrics, we measure the Depth L1 (cm) by comparing the estimated and actual meshes at randomly selected points. On the other hand, in 3D, we evaluate the accuracy (cm), completion (cm), and completion ratio (%) with a threshold of 5cm. To achieve this, we employ the mesh culling strategy in Co-SLAM [5], which eliminates unobserved regions and noisy points outside the camera frustum and target scene. In terms of pose estimation evaluation, we use ATE RMSE (cm).

Baselines. We select several recent neural SLAM systems for comparison, iMAP [3], NICE-SLAM [4], DI-FUSION [38], Go-SLAM [39], Vox-Fusion [35] and Co-SLAM [5]. To examine the accuracy and quality at different image frequencies, results are presented using the notation 'Method[i=n]', where n represents the interval between images. No suffix means using all images for testing.

Implementation Details. The experiments are conducted on a desktop PC with a 3.60GHz Intel Core i9-9900K CPU and an NVIDIA RTX 2080ti GPU. Tracking and mapping were performed using 100 iterations with an interval of 5 for all methods, and 200 iterations with an interval of 10. For tracking, 1,024 pixels are sampled, while 2,048 pixels are sampled for global optimization in all keyframes.

TABLE I. Quantitative results of all eight scenes on the Replica dataset [1]. Our method achieves better reconstruction quality and has the best average performance in all metrics, even with low-frequency image sequences.

Metrics	Method	Replica								Avg.
		Office0	Office1	Office2	Office3	Office4	Room0	Room1	Room2	
Depth L1[cm]↓	iMAP [3]	3.79	3.76	3.97	5.61	5.71	5.08	3.44	5.78	4.64
	NICE-SLAM [4]	1.43	1.58	2.70	2.10	2.06	1.79	1.33	2.20	1.90
	Vox-Fusion [35]	3.44	1.77	3.52	1.82	4.84	1.76	2.52	3.58	2.91
	Co-SLAM [5]	1.24	1.48	1.86	1.66	1.54	1.05	**0.85**	2.37	1.51
	Co-SLAM[i=10]	1.13	2.57	64.69	1.66	41.32	75.14	2.99	2.31	✘
	Ours[i=10]	**1.12**	1.47	1.85	**1.52**	1.54	0.93	0.99	**2.18**	1.46
	Ours[i=5]	1.17	**1.41**	**1.72**	1.56	**1.46**	**0.91**	**0.85**	**2.18**	**1.41**
Acc.[cm]↓	iMAP	3.34	2.10	4.06	4.20	4.34	4.01	3.04	3.84	3.62
	NICE-SLAM	1.85	1.56	3.28	3.01	2.54	2.44	2.10	2.17	2.37
	Vox-Fusion	1.63	1.44	3.03	**2.33**	**2.02**	**1.77**	**1.51**	2.33	2.01
	Co-SLAM	1.57	1.31	2.84	3.06	2.23	2.11	1.68	1.99	2.10
	Co-SLAM[i=10]	1.68	1.35	46.72	2.73	11.09	17.58	3.22	1.95	✘
	Ours[i=10]	1.52	1.28	2.65	2.80	2.28	2.01	1.63	1.90	2.01
	Ours[i=5]	**1.51**	**1.26**	**2.55**	2.58	2.23	1.97	1.53	**1.87**	**1.94**
Comp.[cm]↓	iMAP	3.62	3.62	4.73	5.49	6.65	5.84	4.40	5.07	4.93
	NICE-SLAM	1.84	1.82	3.11	3.16	3.61	2.60	2.19	2.73	2.63
	Vox-Fusion	1.87	**1.44**	3.03	2.81	3.51	2.69	2.31	2.58	2.53
	Co-SLAM	1.56	1.59	2.43	2.72	**2.52**	**2.02**	1.81	1.96	**2.08**
	Co-SLAM[i=10]	1.61	1.77	11.19	2.74	15.47	17.10	3.13	2.08	✘
	Ours[i=10]	1.53	1.70	2.38	**2.68**	2.67	2.16	1.85	1.94	2.11
	Ours[i=5]	**1.50**	1.62	**2.34**	2.70	2.60	2.20	**1.78**	**1.93**	**2.08**
Comp. Ratio%↑	iMAP	83.59	88.45	79.73	73.90	74.77	78.34	85.85	79.40	80.50
	NICE-SLAM	94.93	94.11	88.27	87.68	87.23	91.81	93.56	91.48	91.13
	Vox-Fusion	93.86	94.40	88.94	89.10	86.53	92.03	92.47	90.13	90.93
	Co-SLAM	96.09	**94.65**	91.63	90.72	**90.44**	**95.26**	95.19	93.58	93.44
	Co-SLAM[i=10]	95.06	93.56	42.45	90.45	56.57	43.56	79.92	92.51	74.26
	Ours[i=10]	96.20	94.02	92.08	**91.13**	89.56	94.57	**95.54**	93.41	93.31
	Ours[i=5]	**96.44**	94.58	**92.44**	91.00	**90.44**	94.45	95.24	**93.86**	**93.53**

Fig. 3. Qualitative visualization of results among different approaches. Our reconstructions are smoother, more complete, and have fewer artifacts compared to other advanced methods on the ScanNet dataset [36].

B. Evaluation of Tracking and Mapping

Evaluation on Replica [1]. Detailed comparison results of all eight scenes are shown in Tab. I. Despite using lower-frequency images, our proposed method outperforms the baselines in both 2D and 3D metrics. In contrast, methods like NICE-SLAM [4] and Co-SLAM [5] rely solely on the uniform motion model, which can easily cause tracking drift and ultimately lead to reconstruction failure. Our method improves the robustness of the neural SLAM system by building feature correspondences between the current and former frames. Texture and feature metric warping constraints are used to optimize the camera pose. Furthermore, even as the image frequency decreases (from $i = 5$ to 10), our method still achieves good results with only a slight decrease, and a 100% success rate. On the contrary, Co-SLAM [5] fails to reconstruct several scenes at i=10, the average success rate is 62%. We present the evaluation of the reconstruction quality using the culling strategy from NICE-SLAM [4] in Tab. II. Even with low image frequency, our method exhibits the best overall performance.

Evaluation on TUM RGB-D [37]. We compare the accuracy of pose estimation on the TUM dataset with NeRF-based RGB-D SLAM [5]. However, Co-SLAM [5] requires continuous images, the poses of some scenes in the TUM dataset may not be continuous. We only test using the first continuous segment of these scenes. According to Table

TABLE II. Reconstruction results on Replica dataset [1] using NICE-SLAM [4] culling strategy (unit: cm).

Methods	Depth L1↓	Acc.↓	Comp.↓	Comp. Ratio↑
iMAP [3]	7.64	6.95	5.33	66.60
DI-FUSION [38]	23.33	19.40	10.19	72.96
NICE-SLAM [4]	3.53	2.85	3.00	89.33
Go-SLAM [39]	3.38	2.50	3.74	88.09
Co-SLAM [5]	1.58	**2.15**	2.21	92.99
Ours[i=5]	**1.41**	2.62	**2.15**	**93.22**
Ours[i=10]	1.46	2.73	2.14	93.13

TABLE III. Camera tracking results on TUM RGB-D dataset [37]. Our method achieves the best performance and is robust to large view changes. Non-continuous scenes are marked with an asterisk. Trajectories with errors larger than 30 cm are denoted as FAILED across the paper.

ATE RMSE (*cm*)	Co-SLAM [5]		Ours	
Interval	$i = 1$	$i = 5$	$i = 1$	$i = 5$
Tracking Iters	iter = 20	iter = 100	iter = 200	iter = 100
fr1/desk	**2.43**	FAILED		2.44
fr1/floor*	13.33	15.76	9.15	**5.15**
fr2/desk*	FAILED	FAILED	27.82	**3.40**
fr2/dwp	FAILED	FAILED	7.17	**7.14**
fr2/dishes	FAILED	24.02	8.75	**6.24**
fr2/pslam*	FAILED	FAILED	FAILED	**11.12**
fr3/office	2.40	2.40		**2.30**
fr3/ntnwl	FAILED	4.81	2.86	**2.13**
fr3/teddy	FAILED	FAILED	FAILED	**9.95**

III, our method achieves the highest and most reliable tracking performance. The TUM dataset contains many hand-held shooting scenes with significant viewpoint changes during movement. Our algorithm can effectively handle such changes through feature-metric optimization. Although increasing tracking iterations can lead to better results, our method still outperforms Co-SLAM [5] by a large margin. Fig. 4 illustrates how our method mitigates cumulative error and pose drift facing large viewpoint changes in challenging scenarios as the number of images increases.

Evaluation on ScanNet [36]. We evaluate the tracking results on real-world sequences from ScanNet, where the ground-truth trajectories are generated using BundleFusion [40]. In Fig. 3, a qualitative analysis of $scene0000$ and $scene0059$ is presented. Our method achieves better pose accuracy compared to NICE-SLAM [4] and Co-SLAM [5]. Moreover, Fig. 3 shows that our method exhibits better reconstruction quality with smoother surfaces, consistent geometries, and fewer artifacts.

C. Run-time and Performance Analysis

Table IV presents a comparison of run-time and performance for Replica [1] Office2 and TUM-RGBD [37] fr1/desk

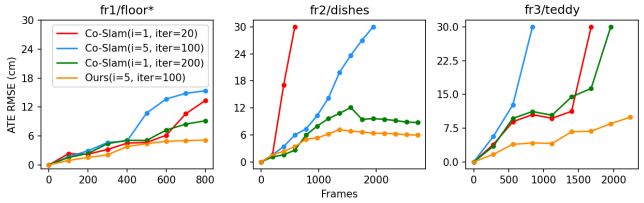

Fig. 4. A comparison of ATE RMSE trends as the number of images increases across different scenes.

TABLE IV. Run-time, frame rate comparison, and pose estimation performance with different iterations when tracking.

Method	Track (ms)↓	Map (ms)↓	FPS↑	ATE RMSE (*cm*)
NICE-SLAM	12.3×50	125.3×60	0.68	-
Co-SLAM	7.8×20	20.2×10	6.4	-
Ours[i=10]	11.3×200	11.7×200	4.43	1.75
Ours[i=10]	11.3×100	11.7×100	8.85	1.88
Ours[i=5]	11.3×200	11.7×200	2.22	2.37
Ours[i=5]	11.3×100	11.7×100	4.43	2.44

TABLE V. Camera tracking performance evaluated using different combinations of warping losses, without trajectory alignment.

Warp Losses				ATE RMSE (*cm*)	
RGB	Depth	KeyPoint	Feature	Office2	fr1/desk
✘	✘	✘	✘	FAILED	FAILED
✔	✘	✘	✘	4.28	FAILED
✔	✔	✘	✘	3.89	10.28
✔	✔	✔	✘	3.81	8.58
✔	✔	✔	✔	3.78	5.45

at different image frequencies. The run-time is measured in ms/iter × #iter. Our method's tracking performance remains largely unaffected even with fewer tracking iterations. This showcases the resilience of our method to significant changes in perspective and its ability to operate in real-time.

D. Ablation Study

We evaluate our feature-metric optimization by testing different warping loss combinations on tracking results, i.e., RGB and depth patch-wise warping, feature point, and feature map pixel-wise warping. Tab. V reports the results tested on Replica [1] and TUM-RGBD [37]. In this experiment, we evaluate the absolute trajectory errors (ATE) without estimating the rigid transformation to align the estimated trajectory with the ground truth, as commonly done in traditional SLAM. As shown in Tab. V, by using more warp losses, the trajectory tracking accuracy becomes better and better. Incorporating feature map pixel-wise warping losses into the TUM dataset has resulted in significant improvements in pose estimation. This is because detection and matching processes often encounter repeatability errors in feature points, which cannot be eliminated during pose optimization in tracking. Incorporating feature-metric supervision to maintain the semantic information's consistency around feature points is crucial for reducing errors.

V. Conclusion

This paper presents HERO-SLAM, a hybrid optimization solution for neural SLAM that stands for Hybrid Enhanced Robust Optimization. By fusing the prowess of both neural implicit field and feature-metric optimization, our hybrid method optimizes a multi-resolution implicit field and enhances robustness in challenging environments with sudden viewpoint changes or sparse data collection. The experimental results validate the effectiveness of our approach compared to existing methods, particularly in challenging scenarios.

REFERENCES

[1] J. Straub, T. Whelan, L. Ma, Y. Chen, E. Wijmans, S. Green, J. J. Engel, R. Mur-Artal, C. Ren, S. Verma *et al.*, "The replica dataset: A digital replica of indoor spaces," *arXiv preprint arXiv:1906.05797*, 2019.

[2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Proc. of the Conference on Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017.

[3] E. Sucar, S. Liu, J. Ortiz, and A. J. Davison, "imap: Implicit mapping and positioning in real-time," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2021, pp. 6229–6238.

[4] 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 *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 12 786–12 796.

[5] H. Wang, J. Wang, and L. Agapito, "Co-slam: Joint coordinate and sparse parametric encodings for neural real-time slam," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2023, pp. 13 293–13 302.

[6] M. M. Johari, C. Carta, and F. Fleuret, "Eslam: Efficient dense slam system based on hybrid representation of signed distance fields," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2023, pp. 17 408–17 419.

[7] 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," *Communications of the ACM*, vol. 65, no. 1, pp. 99–106, 2021.

[8] T. Müller, A. Evans, C. Schied, and A. Keller, "Instant neural graphics primitives with a multiresolution hash encoding," *ACM Trans. Graph.*, vol. 41, no. 4, pp. 102:1–102:15, Jul. 2022.

[9] T. Takikawa, A. Evans, J. Tremblay, T. Müller, M. McGuire, A. Jacobson, and S. Fidler, "Variable bitrate neural fields," in *ACM SIGGRAPH 2022 Conference Proceedings*, 2022, pp. 1–9.

[10] E. R. Chan, C. Z. Lin, M. A. Chan, K. Nagano, B. Pan, S. De Mello, O. Gallo, L. J. Guibas, J. Tremblay, S. Khamis *et al.*, "Efficient geometry-aware 3d generative adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 16 123–16 133.

[11] C. Sun, M. Sun, and H.-T. Chen, "Direct voxel grid optimization: Super-fast convergence for radiance fields reconstruction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 5459–5469.

[12] A. J. Davison, I. D. Reid, N. D. Molton, and O. Stasse, "Monoslam: Real-time single camera slam," *IEEE Trans. on Pattern Analalysis and Machine Intelligence (TPAMI)*, vol. 29, no. 6, pp. 1052–1067, 2007.

[13] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardos, "ORB-SLAM: a versatile and accurate monocular slam system," *IEEE Trans. on Robotics (TRO)*, vol. 31, no. 5, pp. 1147–1163, 2015.

[14] T. Qin, P. Li, and S. Shen, "Vins-mono: A robust and versatile monocular visual-inertial state estimator," *IEEE Trans. on Robotics (TRO)*, vol. 34, no. 4, pp. 1004–1020, 2018.

[15] J. Engel, T. Schöps, and D. Cremers, "Lsd-slam: Large-scale direct monocular slam," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Springer, 2014, pp. 834–849.

[16] R. Wang, M. Schworer, and D. Cremers, "Stereo DSO: Large-scale direct sparse visual odometry with stereo cameras," in *Proc. of the IEEE/CVF Intl. Conf. on Computer Vision (ICCV)*, 2017, pp. 3903–3911.

[17] R. A. Newcombe, S. Izadi, O. Hilliges, D. Molyneaux, D. Kim, A. J. Davison, P. Kohi, J. Shotton, S. Hodges, and A. Fitzgibbon, "Kinectfusion: Real-time dense surface mapping and tracking," in *Proc. of the Intl. Symposium on Mixed and Augmented Reality (ISMAR)*, 2011, pp. 127–136.

[18] T. Whelan, S. Leutenegger, R. Salas-Moreno, B. Glocker, and A. Davison, "Elasticfusion: Dense slam without a pose graph," in *Proc. of Robotics: Science and Systems (RSS)*, 2015.

[19] Z. Teed and J. Deng, "Droid-slam: Deep visual slam for monocular, stereo, and rgb-d cameras," *Proc. of the Conference on Neural Information Processing Systems (NeurIPS)*, vol. 34, pp. 16 558–16 569, 2021.

[20] L. Koestler, N. Yang, N. Zeller, and D. Cremers, "Tandem: Tracking and dense mapping in real-time using deep multi-view stereo," in *Conference on Robot Learning*. PMLR, 2022, pp. 34–45.

[21] Y. Yao, Z. Luo, S. Li, T. Fang, and L. Quan, "Mvsnet: Depth inference for unstructured multi-view stereo," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 767–783.

[22] J. Deng, X. Chen, S. Xia, Z. Sun, G. Liu, W. Yu, and L. Pei, "Nerf-loam: Neural implicit representation for large-scale incremental lidar odometry and mapping," in *Proc. of the IEEE/CVF Intl. Conf. on Computer Vision (ICCV)*, 2023.

[23] H. Li, X. Gu, W. Yuan, L. Yang, Z. Dong, and P. Tan, "Dense rgb slam with neural implicit maps," in *Proc. of the Int. Conf. on Learning Representations (ICLR)*, 2023.

[24] Z. Wang, S. Wu, W. Xie, M. Chen, and V. A. Prisacariu, "NeRF−−: Neural radiance fields without known camera parameters," *arXiv preprint arXiv:2102.07064*, 2021.

[25] C.-H. Lin, W.-C. Ma, A. Torralba, and S. Lucey, "Barf: Bundle-adjusting neural radiance fields," in *Proc. of the IEEE/CVF Intl. Conf. on Computer Vision (ICCV)*, 2021, pp. 5741–5751.

[26] Y. Jeong, S. Ahn, C. Choy, A. Anandkumar, M. Cho, and J. Park, "Self-calibrating neural radiance fields," in *Proc. of the IEEE/CVF Intl. Conf. on Computer Vision (ICCV)*, 2021, pp. 5846–5854.

[27] W. Bian, Z. Wang, K. Li, J.-W. Bian, and V. A. Prisacariu, "Nope-nerf: Optimising neural radiance field with no pose prior," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2023, pp. 4160–4169.

[28] A. Meuleman, Y.-L. Liu, C. Gao, J.-B. Huang, C. Kim, M. H. Kim, and J. Kopf, "Progressively optimized local radiance fields for robust view synthesis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2023, pp. 16 539–16 548.

[29] E. Zheng, E. Dunn, V. Jojic, and J.-M. Frahm, "Patchmatch based joint view selection and depthmap estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2014, pp. 1510–1517.

[30] C. Wu, J. Sun, Z. Shen, and L. Zhang, "MapNeRF: Incorporating map priors into neural radiance fields for driving view simulation," in *Proc. of the IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, 2023.

[31] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Trans. on Image Processing (TIP)*, vol. 13, no. 4, pp. 600–612, 2004.

[32] F. Darmon, B. Bascle, J.-C. Devaux, P. Monasse, and M. Aubry, "Improving neural implicit surfaces geometry with patch warping," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 6260–6269.

[33] D. DeTone, T. Malisiewicz, and A. Rabinovich, "Superpoint: Self-supervised interest point detection and description," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 224–236.

[34] P. Lindenberger, P.-E. Sarlin, and M. Pollefeys, "Lightglue: Local feature matching at light speed," in *Proc. of the IEEE/CVF Intl. Conf. on Computer Vision (ICCV)*, 2023.

[35] X. Yang, H. Li, H. Zhai, Y. Ming, Y. Liu, and G. Zhang, "Vox-fusion: Dense tracking and mapping with voxel-based neural implicit representation," in *Proc. of the Intl. Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE, 2022, pp. 499–507.

[36] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner, "Scannet: Richly-annotated 3d reconstructions of indoor scenes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 5828–5839.

[37] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers, "A benchmark for the evaluation of rgb-d slam systems," in *Proc. of the IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*. IEEE, 2012, pp. 573–580.

[38] J. Huang, S.-S. Huang, H. Song, and S.-M. Hu, "Di-fusion: Online implicit 3d reconstruction with deep priors," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2021, pp. 8932–8941.

[39] Y. Zhang, F. Tosi, S. Mattoccia, and M. Poggi, "Go-slam: Global optimization for consistent 3d instant reconstruction," *Proc. of the IEEE/CVF Intl. Conf. on Computer Vision (ICCV)*, 2023.

[40] A. Dai, M. Nießner, M. Zollhöfer, S. Izadi, and C. Theobalt, "Bundlefusion: Real-time globally consistent 3d reconstruction using on-the-fly surface reintegration," *ACM Trans. on Graphics (TOG)*, vol. 36, no. 4, p. 1, 2017.