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

ONeK-SLAM: A Robust Object-level Dense SLAM Based on Joint Neural Radiance Fields and Keypoints

Yue Zhuge[1,2], Haiyong Luo[1], Runze Chen[3], Yushi Chen[3], Jiaquan Yan[3], Zhuqing Jiang[3]

***Abstract*— Neural implicit representation has recently achieved significant advancements, especially in the field of SLAM(Simultaneous Localization and Mapping). Previous NeRF-based SLAM methods have difficulties with object-level localization and reconstruction and struggle in dynamic and illumination-varied environments. We propose ONeK-SLAM, a robust object-level SLAM system that effectively combines feature points and neural radiance fields. ONeK-SLAM uses the joint information at the object level to improve localization accuracy and enhance reconstruction details. Moreover, our approach detects and eliminates dynamic objects based on the joint errors, while also harnessing the illumination invariance offered by feature points. Consequently, ONeK-SLAM achieves high-precision localization and detailed object-level mapping, even in dynamic and illumination-varying environments. Our evaluations, conducted on three public datasets that include both dynamic and variable lighting sequences, demonstrate that our method outperforms recent NeRF-based SLAM method in both localization and reconstruction.**

I. INTRODUCTION

In the field of computer vision, Dense Visual SLAM (Simultaneous Localization and Mapping) is a widely researched area, aimed at localization and mapping in unexplored environments. Our work aims to provide autonomous agents, like robots, with accurate localization and detailed object-level mapping.

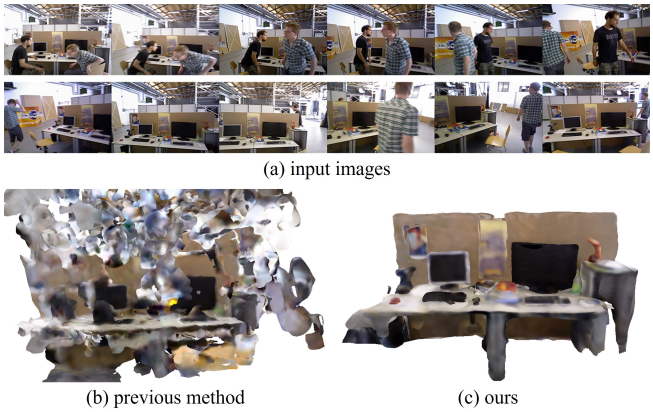
Traditional Dense Visual SLAM techniques [2]–[12] have primarily emphasized dense reconstruction but are constrained to areas that have been previously observed. Recent developments in Neural Radiance Fields (NeRF) [13]–[19] have demonstrated remarkable potential to finely render unobserved areas. Some works combine semantic information for object-level reconstruction [20]–[23], and others integrate NeRF into SLAM systems for improved localization and reconstruction [24]–[34].

This work was supported in part by the Strategic Priority Research Program of Chinese Academy of Sciences under Grant XDA28040500, the National Natural Science Foundation of China under Grant 62261042 and 62002026, the Key Research Projects of the Joint Research Fund for Beijing Natural Science Foundation and the Fengtai Rail Transit Frontier Research Joint Fund under Grant L221003, Beijing Natural Science Foundation under Grant 4232035 and 4222034, the Fundamental Research Funds for the Central Universities under Grant 2022RC13, Yibin City Introduction of High-Level Talent Project under Grant 2022YG03 and the Open Project of the Beijing Key Laboratory of Mobile Computing and Pervasive Device, Institute of Computing Technology, Chinese Academy of Sciences (Corresponding author: Haiyong Luo).

[1]Institute of Computing Technology, Chinese Academy of Sciences, China {zhugeyue21s, yhluo}@ict.ac.cn

[2]University of Chinese Academy of Sciences, Beijing, China

[3]Beijing University of Posts and Telecommunications, Beijing, China {chenrz925, chenyushi, YanJiaquan, jiangzhuqing}@bupt.edu.cn



(a) input images

(b) previous method (c) ours

Fig. 1. Reconstruction Results on the Dynamic Sequence of TUM RGB-D Dataset. (a) shows the input image, in which moving dynamic objects can be observed. (b) presents the reconstruction results obtained with NICE-SLAM [1]. (c) shows the reconstruction results achieved using our method. Our approach remains robust even in dynamic environments.

Most existing NeRF works do not distinguish between different objects, lacking fine-grained details. Some approaches, such as vMAP [35], achieve object-level scene reconstruction by separately modeling objects and decoupling scene representation. However, vMAP doesn't estimate pose on its own and relies on accurate initial pose input from another system. Moreover, the radiance field information learned by the model is not used to optimize pose estimation.

NICE-SLAM [1] builds upon iMAP [32], utilizing neural radiance fields for both pose estimation and reconstruction. It employs hierarchical grid-based neural implicit encoding to update local representations. However, NICE-SLAM's hierarchical grid-based neural radiance field model conflates various objects within a scene neural radiance field model, making it limiting in object-level details.

The methods previously mentioned exhibit limitations when applied to dynamic scenes, lacking the robustness needed to adapt to changing environmental conditions. Moreover, neural radiance field methods are sensitive to changes in lighting conditions, as it relies on photometric loss for training. In contrast, feature point algorithms like SIFT [36] are more stable under varying light conditions.

Adopting an object-level perspective, we propose ONeK-SLAM, a system that synergistically combines feature points and neural radiance fields for object-level SLAM. It provides high-precision localization and detailed object-level mapping without the need for external initial pose estimation. Remarkably, the ONeK-SLAM is robust to challenging com-

plex scenes with dynamic objects and imbalanced lighting. We segment the scene into individual objects. For each segmented object, we use both the re-projection errors of feature points and photometric and depth errors of neural radiance fields for object-level joint pose estimation. Moreover, we identify and exclude dynamic objects based on their joint errors, bolstering the system's robustness in dynamic environments. We leverage the lighting-invariant properties of feature points and employ an adaptive scene management strategy to ensure the robustness and reliability of our localization algorithm even in variable lighting conditions. The object-level joint pose estimation method enhances scene perception at a higher object-level, making our SLAM system less susceptible to the complex environment. Our contributions are as summarized follows:

- We propose ONeK-SLAM, a robust dense visual SLAM system based on object-level scene understanding and the joint information of objects, which provides precise localization and detailed mapping, robust to dynamic and illumination-varied environments.
- We design a novel joint object-level loss function that exploits feature points' lighting-invariance properties and the scene learning capabilities offered by NeRF. Additionally, we eliminate objects with joint errors that significantly exceed those of other objects and adapt the joint error for different environments.
- We have conducted experiments on various datasets: Replica [37], ScanNet [38], and TUM RGB-D [39], including both dynamic and variable lighting sequences. Through comparative evaluations against state-of-the-art NeRF SLAM methods, our system has demonstrated superior performance in both localization and mapping.

II. RELATED WORK

A. Dense Visual SLAM

The field of visual SLAM has evolved significantly since the foundational work of PTAM [40]. Traditional methods focus on generating point cloud maps, which vary in density from sparse [41]–[43] to dense [3], [4]. Recent advances have integrated deep learning into dense SLAM systems [7], [9], [44]–[46]. For instance, DeepV2D [8] employs an end-to-end model for motion and depth estimation, while CodeSLAM [47] optimizes camera pose and depth maps through auto-encoders. DROID-SLAM [48] employs a dense bundle adjustment layer to estimate optical flow field pose. Some methods also leverage voxel grids for mapping [5], [6], [10], [12]. Semantic SLAM methods [49]–[51] enrich maps by incorporating semantic information [52], [53]. NodeSLAM [54] and DSP-SLAM [55] focus on object-level pose and shape optimization for dense reconstruction. However, previous SLAM methods can only reconstruct observed parts and have limitations, such as incomplete reconstructions. They also face challenges in dynamic and variable illumination conditions. By combining illumination-invariant of keypoints with the shape and texture learning abilities of neural radiance fields on object level, we improve localization precision and dense reconstruction of the invisible regions.

B. Neural Field SLAM

Recent advances in neural implicit representations have significantly impacted 3D scene reconstruction [13]–[19], [56], [57]. Particularly, neural radiance fields have been used for camera pose estimation in SLAM system [27], [28], [34], [58]. Notable works like iMAP [32], NICE-SLAM [1] and NeRF-SLAM [31] have advanced SLAM pose estimation and dense visual SLAM, respectively. While existing methods offer solutions for large-scale reconstruction by partitioning the scene [26], [29], [30], they still face limitations, such as blurred object details within reconstructed scenes.

Semantic information has also been leveraged in 3D reconstruction. Unlike traditional methods that rely on prior knowledge [53], [59]–[61], object-level NeRF methods learn to directly reconstruct object information [20], [21]. vMAP [35] decouples different objects for individual modeling. However, these methods require a high-quality initial pose and do not optimize input poses using object-level NeRF data.

Our work is most related to NICE-SLAM [1] and vMAP [35]. NICE-SLAM's use of hierarchical grids partially mitigates the catastrophic forgetting problem but fails to achieve object-level granularity reconstruction. Our method uniquely compartmentalizes scenes at the object level and combines key point features for high-precision pose optimization and reconstruction. Unlike vMAP, which relies on external SLAM systems for initial poses, our approach simultaneously estimates poses by utilizing both keypoints and NeRF information of the objects. Moreover, we eliminate dynamic objects through object-level joint error and handle varying lighting conditions by using keypoints information, leading to enhanced robustness in complex environments.

III. METHOD

Our method adopts an object-centric perspective, leveraging the complementary strengths of both keypoints and neural radiance fields to achieve accurate localization and dense reconstruction. The overview of our method is shown in Fig 2. Initially, we segment the scene based on distinct objects, as shown in Section 3.A. Then we perform object-level pose estimation, as detailed in Section 3.B. We select keyframes for each object, as described in Section 3.C. For complex scenarios, we employ a strategy for eliminating dynamic objects and adaptive scene management, as discussed in Section 3.D.

A. Object-Level Scene Decomposition

Our approach first decomposes the current frame F^i into individual objects. Consistent instance segmentation masks across frames, sourced from dataset annotations or obtained from methods like [62], are used as input. Given the input image $C^i \in \mathbb{R}^{h\times w\times 3}$, depth map $D^i \in \mathbb{R}^{h\times w}$ and the instance segmentation map $M^i \in \mathbb{R}^{h\times w}$, we then define an object set O. When a new object appears, initialize the new object O_b and add it to the set O. Each O_b includes a image set $C_b = \{C_b^i\}$, a depth map set $D_b = \{D_b^i\}$, an object mask set $M_b = \{M_b^i\}$, 2D bounding box sets

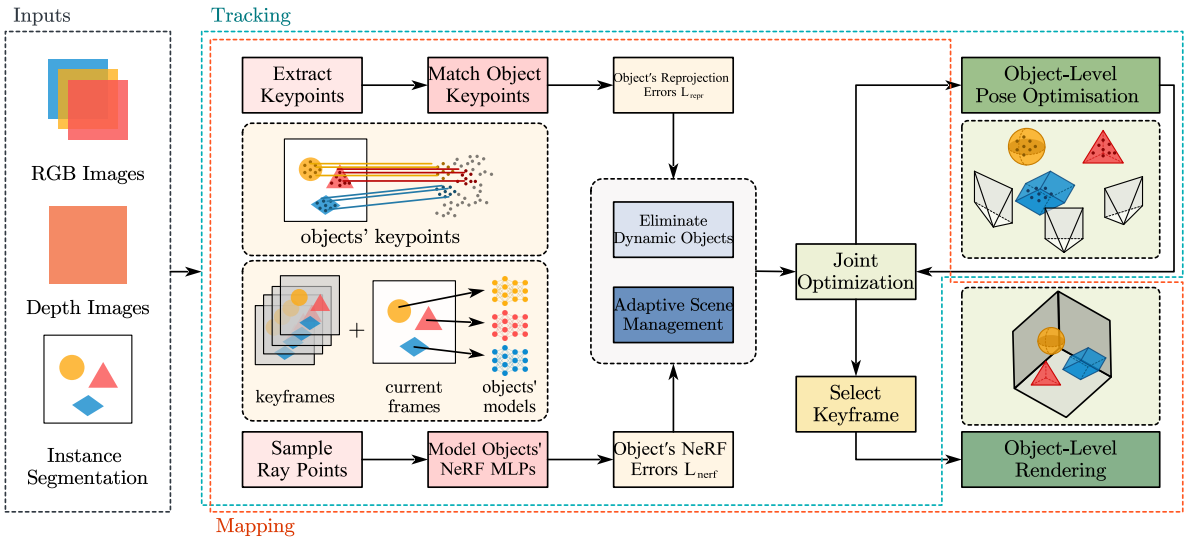


Fig. 2. **An overview of ONeK-SLAM.** Based on the input RGB-D images and segmentation results, we decompose the scene at the object level. Then, we jointly estimate the poses using both keypoints and neural radiance fields of objects. To handle complex scenes, we eliminate dynamic objects and adaptively adjust the joint loss of objects. Finally, we achieve object-level fine-grained scene reconstruction.

$B_b = \{bbox_b^i = \{x_{\min}, x_{\max}, y_{\min}, y_{\max}\}\}$, a map points sets $MP_b = \{{mp_1}^b, \ldots, {mp_n}^b\}$, and an object model F_b. To save computational power, only keyframe and current frame data are stored in each O_b. Here, b and i denote object instances and frame sequence numbers, respectively. Sets C_b^i, D_b^i, and M_b^i are obtained by cropping the current I, D, and M based on the object's bounding box $bbox_b^i$.

B. Object-Level Pose Estimation

1) Light Keypoints Pose Estimation: An initial pose $T^i = [R^i \mid t^i]$ for the current frame F^i is computed using a constant-velocity model. Keypoints $KP^i = \{kp_1^i, \ldots, kp_j^i, \ldots, kp_n^i\}$ are extracted from the current frame image C^i (SIFT keypoints are employed in this study, although other types can be used). We extract keypoints KP^i from the frame's image C^i, and utilize object mask set M_b to match keypoints to corresponding object map points MP_b, generating object-level matched point sets $Matches_b^i = \{kp_j^b, mp_j^b\}$. Each object's sum of reprojection errors is calculated as:

$$ {L_r}^b = \sum_{kp_j^b \in Matches_b^i} \| kp_j^i - K\left(\frac{1}{z_j^b}(R^i mp_j^b + t^i)\right) \|^2 \quad (1) $$

Here, K represents the camera's intrinsic matrix, and the map points are given as $mp_j^b = [x_j^b, y_j^b, z_j^b]^T$. The camera pose is optimized by minimizing the total reprojection error for all objects:

$$ \min_{T^i=[R^i|t^i]} L_r = \min_{T^i=[R^i|t^i]} \sum_{b, O_b \in O} {L_r}^b \quad (2) $$

2) Neural Radiance Field Optimization: After decomposing a scene into individual objects O_b, we utilize a neural radiance field to capture each object's shape and representation. For each object O_b, we select $N_f - 2$ keyframes $KF^i, 1 \le i \le N_f - 2$ along with the current frame F^i and the previous frame F^{i-1}, forming a pose set T_{N_f}. For each object O_b in its frame-specific 2D bounding box $bbox_b^i$, N_r pixel points $p^k = [u^k, v^k], 1 \le k \le N_r$ are chosen based on a normal distribution. This points sampling method, as opposed to using feature points, ensures adequate sampling even in texture-poor environments. The points' directions d^k in the camera coordinate system are calculated using the intrinsic matrix K and the object's 2D bounding box $bbox_b^i$. Using the pose set T_{N_f}, the corresponding rays r^k in the world coordinate system are then determined as:

$$ sp^{k,h} = r^k(sd^{k,h}) = t^i + R^i d^k sd^{k,h} \quad (3) $$

Where stratified sampling is performed along the ray r^k to sample N_s depths $sd^{k,h}$ for points $sp^{k,h}$. Uniform sampling is done around the depth value corresponding to pixel point p^k, capturing N_d depths $sd^{k,h}$ for points $sp^{k,h}$, where $1 \le h \le N_{sp} = N_s + N_d$. For every point $sp^{k,h}$, the model of object O_b, denoted as F_b, predicts the color $c^{k,h}$ and occupancy $o^{k,h}$ of that point. The termination probability of the ray at the point $sp^{k,h}$ is given by: $w^{k,h} = o^{k,h} \prod_{m<h} (1 - o^{k,m})$ [19]. The color and depth loss for object O_b is computed as:

$$ \hat{D}(p^k) = \sum_{h=1}^{N_{sp}} w^{k,h} sd^{k,h}, \hat{C}(p^k) = \sum_{h=1}^{N_{sp}} w^{k,h} c^{k,h} \quad (4) $$

$$ {L_c}^b = \sum_k |\hat{C}(p^k) - C_b(p^k)| \quad (5) $$

$$ {L_d}^b = \sum_k |\hat{D}(p^k) - D_b(p^k)| \quad (6) $$

To allow the neural radiance field to utilize the shape information of object O_b, occupancy is predicted in regions

where the object exists. The parameters of the neural radiance field are optimized by calculating the occupancy loss [35], as indicated in (7):

$$L_o{}^b = \sum_k |\hat{O}(p^k) - M_b(p^k)|, \hat{O}(p^k) = \sum_{h=1}^{N_{sp}} w^{k,h} \quad (7)$$

Where $M_b(p^k)$ is the value of pixel point p^k on its corresponding M_b^i. Similarly, the color $L_c{}^b$ and depth loss $L_d{}^b$ for object O_b is only predicted in the area where O_b exists.

For each object $O_b \in O$, the parameters θ_b of the model F_b are optimized by minimizing (7), (5), and (6), as described in (8):

$$\min_{\theta_b} L_n{}^b = \min_{\theta_b} \sum \alpha L_c{}^b + \beta L_d{}^b + \gamma L_o{}^b \quad (8)$$

where α, β, γ is the loss weighting factor.

3) Joint Object-Level Optimization: Building upon the training of object-level feature points and neural radiance fields, the optimization process combines both the geometric and global characteristics of objects to refine their pose. The overall objective function for the joint optimization is formulated as:

$$\min_{T^i=[R^i|t^i]} L = \min_{T^i=[R^i|t^i]} \sum_{b,O_b \in O} L_r{}^b + \lambda L_n{}^b \quad (9)$$

Here, λ serves as an adaptive parameter, regulating the optimization focus between feature points and neural radiance fields. Further details of λ will be discussed in Section 3.D.

C. Keyframe Selection

Our approach employs a keyframe management to enhance computational efficiency and system performance. Our framework maintains both a global keyframe list and object-specific keyframe lists. Keyframes are selected based on criteria such as co-visibility with other keyframes. When evaluating new keyframes for each object O_b, we choose the frame that has passed K_{step} frames since the last keyframe insertion for that object. Keyframes are also chosen based on information gain [1], [32] and the number of keypoints for tracking. The selected keyframe for the object O_b will also be inserted into the global list. Moreover, as new objects appear, the keyframe is selected. In contrast to traditional SLAM, which relies on a sufficient number of keypoints, our method benefits from object-level NeRF losses. This allows for effective pose optimization through photometric and depth losses, even when feature points are few.

D. Handling Complex Environments

1) Eliminating Dynamic Objects: To identify dynamic objects, we first calculate the reprojection error L_r^b for each object O_b and normalize it by the total number of its matched points N_{mp}^b, resulting in an average error aL_r^b. An object O_b is classified as dynamic if aL_r^b exceeds ten times the median average error of all other objects, as formulated in (10):

$$dyn^{O_b} = \begin{cases} 1 & aL_r{}^b \geq 10Me(aL_r{}^{bi}) \\ 0 & aL_r{}^b < 10Me(aL_r{}^{bi}) \end{cases}, \quad bi, O_{bi} \in O \, and \, O_{bi} \neq O_b \quad (10)$$

Here, dyn^{O_b} indicates whether object O_b is dynamic (1) or static (0). $Me()$ denotes the calculation of the median value. Similar thresholds are applied for average color and depth loss. If an object crosses this tenfold threshold, it is flagged as dynamic and excluded from further calculations and optimizations.

2) Adaptive Scene Management: We implement an adaptive parameter λ for pose optimization, with levels $\{\lambda_{low}, \lambda_o, \lambda_{high}\}$ and a default $\lambda = \lambda_o$. It adjusts dynamically to scene changes. For significant illumination changes increasing color loss, λ is lowered to λ_{low} for reduced NeRF weight in optimization. Conversely, in environments with sparse keypoint matches due to disturbances, λ is increased to λ_{high} to leverage NeRF data for object pose optimization.

IV. EXPERIMENTTS

Datasets. We have evaluated our approach on three benchmark datasets: Replica [37], ScanNet [38], and TUM RGB-D datasets [39]. Notably, ScanNet and TUM RGB-D datasets consist of data captured from real-world scenes.

Baselines. We compare our approach against several state-of-the-art NeRF-based SLAM algorithms, including iMAP [32], NICE-SLAM [1], and vMAP [35]. It should be noted that vMAP is dependent on pose estimations from other SLAM systems and does not carry out its own pose optimization. To ensure a fair comparison, we utilize pose estimations obtained from our method as input for vMAP. We also include the results of TSDF-Fusion [63] and DI-Fusion [64] as additional references.

Metrics. Consistent with the metrics used in NICE-SLAM [1], our evaluation focuses on both reconstruction and localization performance. For reconstruction, we examine 3D metrics, specifically Accuracy [cm], Completion [cm], and Completion Ratio [$<$ 5cm %]. For 2D metrics, we assess the Depth L1 [cm] loss, which is calculated between the reconstructed results and the ground truth across 1,000 randomly sampled depth maps. For localization, we employ ATE RMSE [39][cm] as the metric for camera pose estimation.

Implementation Details. All experiments were executed on a computer equipped with an Intel i9-13900 CPU and an NVIDIA GeForce RTX 3090 GPU. The code, written in Python, covers both tracking and mapping functionalities but excludes loop closure. Our implementation takes inspiration from both NICE-SLAM [1] and vMAP [35]. Each object-level NeRF model consists of a 4-layer MLP with a hidden size of 32 for each layer. We select a keyframe quantity of $N_f = 20$ for each optimization run, and the pixel points chosen are $N_r = 120$. We sample $N_s = 10$ and $N_d = 10$ points along the rays, with NeRF optimization objectives set to $\alpha = 1$, $\beta = 0.2$, $\gamma = 2$. The adjustable parameter λ values are $\{\lambda_{low} = 1, \lambda_o = 10, \lambda_{high} = 50\}$. The AdamW optimizer is used with a learning rate of 0.001.

A. Evaluation

Evaluation on Replica Dataset. We conducted experiments on eight RGB-D sequences extracted and rendered

TABLE I

RECONSTRUCTION AND LOCALIZATION RESULTS ON REPLICA. [CM]

	TSDF-Fusion	iMAP	DI-Fusion	NICE-SLAM	vMAP*	ours
Depth L1 ↓	7.57	7.64	23.33	3.53	3.36	**2.05**
Acc. ↓	**1.60**	6.95	19.40	2.85	4.33	3.00
Comp. ↓	3.49	5.33	10.19	3.00	3.53	**2.06**
Comp. Ratio ↑	86.08	66.60	72.96	89.33	90.00	**92.46**
ATE RMSE ↓	-	3.12	-	1.43	-	**0.46**

TABLE II

OBJECT RECONSTRUCTION RESULTS ON REPLICA. [CM]

	iMAP*	NICE-SLAM*	vMAP	ours
Object Acc. ↓	3.57	3.91	2.23	**1.85**
Object Comp. ↓	2.38	3.27	**1.44**	2.37
Object Comp. Ratio ↑	90.19	83.97	**94.55**	91.65

from the Replica dataset, as originally utilized by the authors of iMAP [32]. In Table I, results of TSDF-Fusion, iMAP, DI-Fusion, and NICE-SLAM are from [1]. The * on vMAP indicates that the pose estimates from our method are used as inputs. As depicted in Table I, our method outperforms other approaches in terms of both reconstruction and localization accuracy. Through the joint object-level pose estimation, our method is capable of precise localization and detailed object-level reconstruction. Fig 3 provides a more intuitive visualization, emphasizing the superior detail that our method can reconstruct. In Table II, object reconstruction results of iMAP*, NICE-SLAM*, and vMAP are from [35], with * indicating results trained using ground truth. Our algorithm demonstrates superior performance.

Evaluation on ScanNet Dataset. For the more extensive ScanNet dataset, we selected several scenes to validate the robustness of our approach, comparing it against other established methods. Because ScanNet does not provide comprehensive ground truth meshes, we conduct the evaluation on the RMSE of ATE. As shown in Table III, our method demonstrates higher localization accuracy. Moreover, qualitative analyses are presented through visualizations of the reconstructed scenes in Fig 4, demonstrating our method's capacity to distinctly reconstruct details even in larger scenes.

Evaluation on TUM RGB-D Datasets. Further evaluations were conducted on the TUM RGB-D datasets, focusing

TABLE III

LOCALIZATION RESULTS ON SCANNET. [CM]

Scene ID	0	59	106	169	207	Avg.
iMAP	55.95	32.06	17.5	70.51	11.91	36.67
DI-Fusion	62.99	128	18.5	75.8	100.19	78.89
NICE-SLAM	8.64	12.25	**8.09**	10.28	**5.59**	9.63
ours	**5.36**	**5.86**	8.82	**8.08**	6.76	**6.98**

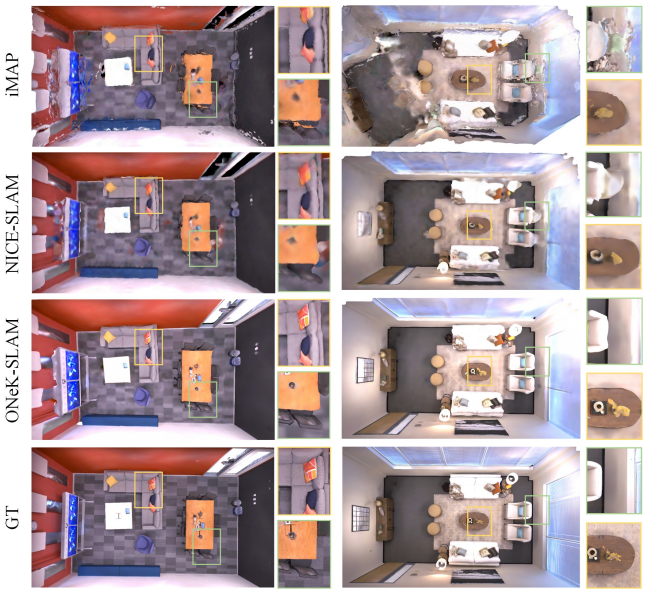


Fig. 3. Reconstruction Results on Replica.

Fig. 4. Reconstruction Results on ScanNet.

on camera localization performance. We selected sequences used by NICE-SLAM. To verify the robustness of our method in both dynamic and variable lighting environments, we also selected specific sequences: "fr3/w_h" for dynamic objects and "fr1/desk*" for lighting variations. Since this dataset also lacks ground truth meshes, we use the RMSE of ATE for evaluation. In Table IV, "fr3/w_h" represents the sequence "fr3/walking_halfsphere," which includes dynamic objects. "fr1/desk*" stands for the sequence that simulates changes in illumination. The results show that our method outperforms the currently best-performing NICE-SLAM in terms of localization capabilities. Fig 1 provides further insights by showcasing the reconstruction capabilities of our method in a dynamic environment, thereby illustrating its robustness.

TABLE IV
LOCALIZATION RESULTS ON TUM RGB-D. [CM]

	fr1/desk	fr2/xyz	fr3/office	fr3/w_h	fr1/desk*
iMAP	7.2	2.1	9.0	47.1	34.7
NICE-SLAM	2.7	1.8	3.0	38.2	3.2
ours	**1.5**	**0.3**	**1.1**	**15.1**	**1.5**

TABLE V
RUNTIME COMPARISON. [S]

	iMAP	NICE-SLAM	ours
Tracking	3.54	0.30	**0.24**
Mapping	24.08	2.51	**0.49**

TABLE VI
ABLATION STUDY ON REPLICA. [CM]

	w/o ONeK	w/o OK	w/o ONe	Full
Acc. ↓	2.60	3.42	**2.38**	2.78
Comp. ↓	2.30	2.45	2.26	**1.97**
Comp Ratio 1cm ↑	14.77	10.13	17.75	**19.05**
Comp. Ratio ↑	94.54	95.13	94.12	**96.93**
ATE RMSE ↓	3.11	1.93	0.97	**0.77**

B. *Performance Analysis*

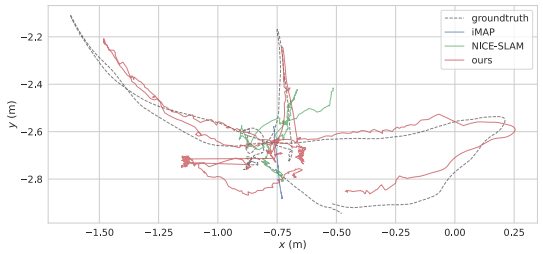


Fig. 5. Trajectories on the Dynamic Sequence of TUM RGB-D Dataset

Robustness in Dynamic Environments. We further extended our evaluations to the dynamic sequence from TUM RGB-D dataset that includes fast-moving dynamic objects to validate the robustness of our algorithm in such volatile environments. As shown in Table IV, under the column labeled "fr3/w_h" (which stands for the sequence 'fr3/walking_halfsphere'), our method maintains high localization accuracy even in the presence of dynamic objects. Fig 1 and Fig 5 validate that our reconstruction results are minimally affected by dynamic objects, while still providing high-precision localization. These results demonstrate that our ONeK-SLAM is robust in dynamic environments, both in terms of localization and reconstruction.

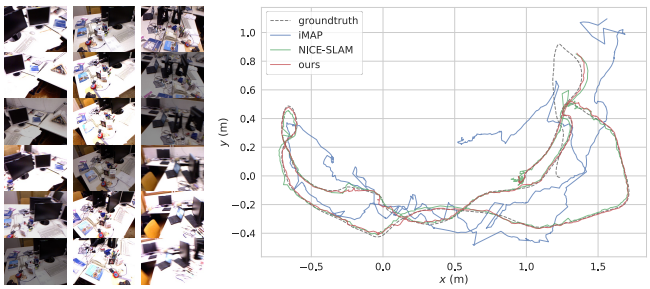


Fig. 6. Localization Results on the Illumination Variations Sequence of TUM RGB-D Dataset

Robustness in illumination-varying environments. We further conducted experiments on the "fr1_desk" sequence from the TUM RGB-D dataset to validate the robustness of our approach under varying lighting conditions. In this sequence, which comprises 610 images, we randomly selected 200 images and modified their light factors, ranging from 0.5 to 2, to simulate variable illumination conditions. As denoted in Table IV under the "fr1/desk*" column, and as shown in Fig 6, we observed that NICE-SLAM's localization accuracy suffers under varying lighting conditions. In contrast, our ONeK-SLAM, which relies on joint object-level pose estimation and is less susceptible to lighting changes, consistently provides accurate localization regardless of illumination variations.

Runtime Comparison. ONeK-SLAM outperforms iMAP and NICE-SLAM in runtime of each frame on Replica room0 sequence, as shown in Table V.

C. *Ablation Study*

We conducted an ablation study to validate the reliability of joint object-level pose estimation. Experiments were performed on the Room0 sequence of the Replica dataset, considering the following configurations: NeRF SLAM without object-level keypoints and neural radiance fields (w/o ONeK), object-level NeRF SLAM without keypoints(w/o OK), NeRF SLAM with keypoints but without object-level neural radiance fields (w/o ONe), and finally, NeRF SLAM with both object-level keypoints and neural radiance fields jointly optimized (Full). As delineated in Table VI, the full ONeK-SLAM exhibits superior reconstruction and localization accuracy.

V. CONCLUSION

We present ONeK-SLAM, an innovative Dense Visual system for object-level SLAM. Leveraging object-level scene understanding and joint information, ONeK-SLAM achieves precise localization and object-level mapping resilience in dynamic and illumination-varied environments. We introduce a novel joint object-level loss function that capitalizes on feature points' lighting-invariant properties and NeRF's scene learning capabilities while eliminating dynamic objects using the joint error. Our experiments on diverse datasets, including Replica, ScanNet, and TUM RGB-D, encompassing dynamic and lighting-variable sequences, reveal ONeK-SLAM's superior performance compared to state-of-the-art NeRF-based SLAM methods.

REFERENCES

[1] 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*, 2022, p. 12786–12796.

[2] R. A. Newcombe, S. J. Lovegrove, and A. J. Davison, "Dtam: Dense tracking and mapping in real-time," in *2011 international conference on computer vision*. IEEE, 2011, p. 2320–2327.

[3] F. Endres, J. Hess, J. Sturm, D. Cremers, and W. Burgard, "3-d mapping with an rgb-d camera," *IEEE transactions on robotics*, vol. 30, no. 1, p. 177–187, 2013.

[4] C. Kerl, J. Sturm, and D. Cremers, "Dense visual slam for rgb-d cameras," in *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2013, p. 2100–2106.

[5] O. Kähler, V. A. Prisacariu, and D. W. Murray, "Real-time large-scale dense 3d reconstruction with loop closure," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14*. Springer, 2016, p. 500–516.

[6] 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 Transactions on Graphics (ToG)*, vol. 36, no. 4, p. 1, 2017.

[7] C. Tang and P. Tan, "Ba-net: Dense bundle adjustment network," *arXiv preprint arXiv:1806.04807*, 2018.

[8] Z. Teed and J. Deng, "Deepv2d: Video to depth with differentiable structure from motion," *arXiv preprint arXiv:1812.04605*, 2018.

[9] S. Zhi, M. Bloesch, S. Leutenegger, and A. J. Davison, "Scenecode: Monocular dense semantic reconstruction using learned encoded scene representations," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, p. 11776–11785.

[10] T. Schops, T. Sattler, and M. Pollefeys, "Bad slam: Bundle adjusted direct rgb-d slam," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, p. 134–144.

[11] L. Koestler, N. Yang, N. Zeller, and D. Cremers, "Tandem: Tracking and dense mapping in real-time using deep multi-view stereo," 2021. [Online]. Available: https://www.semanticscholar.org/paper/69509fb1051cb7610ddd35e5e5f1a8b6dcf0886d

[12] A. Rosinol, J. J. Leonard, and L. Carlone, "Probabilistic volumetric fusion for dense monocular slam," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, p. 3097–3105.

[13] D. Azinovi'c, R. Martin-Brualla, D. B. Goldman, M. Nießner, and J. Thies, "Neural rgb-d surface reconstruction," *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. null, p. 6280–6291, 2021.

[14] A. Bozic, P. Palafox, J. Thies, A. Dai, and M. Nießner, "Transformerfusion: Monocular rgb scene reconstruction using transformers," *Advances in Neural Information Processing Systems*, vol. 34, p. 1403–1414, 2021.

[15] J. Choe, S. Im, F. Rameau, M. Kang, and I. S. Kweon, "Volumefusion: Deep depth fusion for 3d scene reconstruction," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, p. 16086–16095.

[16] C. Jiang, A. Sud, A. Makadia, J. Huang, M. Nießner, and T. Funkhouser, "Local implicit grid representations for 3d scenes," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, p. 6001–6010.

[17] S. Lionar, D. Emtsev, D. Svilarkovic, and S. Peng, "Dynamic plane convolutional occupancy networks," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, p. 1829–1838.

[18] 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, p. 99–106, 2021.

[19] M. Oechsle, S. Peng, and A. Geiger, "Unisurf: Unifying neural implicit surfaces and radiance fields for multi-view reconstruction," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, p. 5589–5599.

[20] W. Jang and L. Agapito, "Codenerf: Disentangled neural radiance fields for object categories," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, p. 12949–12958.

[21] B. Yang, Y. Zhang, Y. Xu, Y. Li, H. Zhou, H. Bao, G. Zhang, and Z. Cui, "Learning object-compositional neural radiance field for editable scene rendering," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, p. 13779–13788.

[22] S. Zhi, T. Laidlow, S. Leutenegger, and A. J. Davison, "In-place scene labelling and understanding with implicit scene representation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, p. 15838–15847.

[23] S. Zhi, E. Sucar, A. Mouton, I. Haughton, T. Laidlow, and A. J. Davison, "ilabel: Revealing objects in neural fields," *IEEE Robotics and Automation Letters*, vol. 8, no. 2, p. 832–839, 2022.

[24] C.-M. Chung, Y.-C. Tseng, Y. 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," *ArXiv*, vol. abs/2209.13274, p. null, 2022.

[25] M. M. Johari, C. Carta, and F. Fleuret, "Eslam: Efficient dense slam system based on hybrid representation of signed distance fields," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, p. 17408–17419.

[26] E. Kruzhkov, A. Savinykh, P. Karpyshev, M. Kurenkov, E. Yudin, A. Potapov, and D. Tsetserukou, "Meslam: Memory efficient slam based on neural fields," in *2022 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE, 2022, p. 430–435.

[27] C.-H. Lin, W.-C. Ma, A. Torralba, and S. Lucey, "Barf: Bundle-adjusting neural radiance fields," *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, vol. null, p. 5721–5731, 2021.

[28] S. Liu, Y. Zhang, S. Peng, B. Shi, M. Pollefeys, and Z. Cui, "Dist: Rendering deep implicit signed distance function with differentiable sphere tracing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, p. 2019–2028.

[29] D. Rebain, W. Jiang, S. Yazdani, K. Li, K. M. Yi, and A. Tagliasacchi, "Derf: Decomposed radiance fields," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, p. 14153–14161.

[30] C. Reiser, S. Peng, Y. Liao, and A. Geiger, "Kilonerf: Speeding up neural radiance fields with thousands of tiny mlps," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, p. 14335–14345.

[31] A. Rosinol, J. Leonard, and L. Carlone, "Nerf-slam: Real-time dense monocular slam with neural radiance fields," *ArXiv*, vol. abs/2210.13641, p. null, 2022.

[32] E. Sucar, S. Liu, J. Ortiz, and A. J. Davison, "imap: Implicit mapping and positioning in real-time," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, p. 6229–6238.

[33] Z. Wang, S. Wu, W. Xie, M. Chen, and V. Prisacariu, "Nerf-: Neural radiance fields without known camera parameters," *ArXiv*, vol. abs/2102.07064, p. null, 2021.

[34] 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 *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2021, p. 1323–1330.

[35] X. Kong, S. Liu, M. Taher, and A. J. Davison, "vmap: Vectorised object mapping for neural field slam," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, p. 952–961.

[36] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, p. 91–110, 2004.

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

[38] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner, "Scannet: Richly-annotated 3d reconstructions of indoor scenes," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, p. 5828–5839.

[39] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers, "A benchmark for the evaluation of rgb-d slam systems," in *2012 IEEE/RSJ international conference on intelligent robots and systems*. IEEE, 2012, p. 573–580.

[40] G. Klein and D. Murray, "Parallel tracking and mapping on a camera phone," in *2009 8th IEEE International Symposium on Mixed and Augmented Reality*. IEEE, 2009, p. 83–86.

[41] C. Campos, R. Elvira, J. J. G. Rodríguez, J. M. Montiel, and J. D. Tardós, "Orb-slam3: An accurate open-source library for visual, vi-

sual–inertial, and multimap slam," *IEEE Transactions on Robotics*, vol. 37, no. 6, p. 1874–1890, 2021.

[42] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardos, "Orb-slam: a versatile and accurate monocular slam system," *IEEE transactions on robotics*, vol. 31, no. 5, p. 1147–1163, 2015.

[43] R. Mur-Artal and J. D. Tardós, "Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cameras," *IEEE transactions on robotics*, vol. 33, no. 5, p. 1255–1262, 2017.

[44] J. Czarnowski, T. Laidlow, R. Clark, and A. J. Davison, "Deepfactors: Real-time probabilistic dense monocular slam," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, p. 721–728, 2020.

[45] R. Li, S. Wang, and D. Gu, "Deepslam: A robust monocular slam system with unsupervised deep learning," *IEEE Transactions on Industrial Electronics*, vol. 68, no. 4, p. 3577–3587, 2020.

[46] H. Zhou, B. Ummenhofer, and T. Brox, "Deeptam: Deep tracking and mapping," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, p. 822–838.

[47] M. Bloesch, J. Czarnowski, R. Clark, S. Leutenegger, and A. J. Davison, "Codeslam—learning a compact, optimisable representation for dense visual slam," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, p. 2560–2568.

[48] Z. Teed and J. Deng, "Droid-slam: Deep visual slam for monocular, stereo, and rgb-d cameras," 2021. [Online]. Available: https://www.semanticscholar.org/paper/67515d1f7df144683b059e684da7974e40aeaca1

[49] M. Hosseinzadeh, K. Li, Y. Latif, and I. Reid, "Real-time monocular object-model aware sparse slam," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, p. 7123–7129.

[50] Z. Liao, Y. Hu, J. Zhang, X. Qi, X. Zhang, and W. Wang, "So-slam: Semantic object slam with scale proportional and symmetrical texture constraints," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, p. 4008–4015, 2022.

[51] L. Nicholson, M. Milford, and N. Sünderhauf, "Quadricslam: Dual quadrics from object detections as landmarks in object-oriented slam," *IEEE Robotics and Automation Letters*, vol. 4, no. 1, p. 1–8, 2018.

[52] M. Runz, M. Buffier, and L. Agapito, "Maskfusion: Real-time recognition, tracking and reconstruction of multiple moving objects," in *2018 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE, 2018, p. 10–20.

[53] B. Xu, W. Li, D. Tzoumanikas, M. Bloesch, A. Davison, and S. Leutenegger, "Mid-fusion: Octree-based object-level multi-instance dynamic slam," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, p. 5231–5237.

[54] E. Sucar, K. Wada, and A. Davison, "Nodeslam: Neural object descriptors for multi-view shape reconstruction," in *2020 International Conference on 3D Vision (3DV)*. IEEE, 2020, p. 949–958.

[55] J. Wang, M. Rünz, and L. Agapito, "Dsp-slam: Object oriented slam with deep shape priors," in *2021 International Conference on 3D Vision (3DV)*. IEEE, 2021, p. 1362–1371.

[56] J. Sun, Y. Xie, L. Chen, X. Zhou, and H. Bao, "Neuralrecon: Real-time coherent 3d reconstruction from monocular video," *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. null, p. 15593–15602, 2021.

[57] Z. Yan, Y. Tian, X. Shi, P. Guo, P. Wang, and H. Zha, "Continual neural mapping: Learning an implicit scene representation from sequential observations," *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, vol. null, p. 15762–15772, 2021.

[58] 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.

[59] X. Kong, X. Yang, G. Zhai, X. Zhao, X. Zeng, M. Wang, Y. Liu, W. Li, and F. Wen, "Semantic graph based place recognition for 3d point clouds," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, p. 8216–8223.

[60] G. Li, Y. Li, Z. Ye, Q. Zhang, T. Kong, Z. Cui, and G. Zhang, "Generative category-level shape and pose estimation with semantic primitives," in *Conference on Robot Learning*. PMLR, 2023, p. 1390–1400.

[61] J. McCormac, R. Clark, M. Bloesch, A. Davison, and S. Leutenegger, "Fusion++: Volumetric object-level slam," in *2018 international conference on 3D vision (3DV)*. IEEE, 2018, p. 32–41.

[62] Y. Cheng, L. Li, Y. Xu, X. Li, Z. Yang, W. Wang, and Y. Yang, "Segment and track anything," *arXiv preprint arXiv:2305.06558*, 2023.

[63] B. Curless and M. Levoy, "A volumetric method for building complex models from range images," in *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, 1996, p. 303–312.

[64] J. Huang, S.-S. Huang, H. Song, and S. Hu, "Di-fusion: Online implicit 3d reconstruction with deep priors," *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. null, p. 8928–8937, 2020.