Marrying NeRF with Feature Matching for One-step Pose Estimation

Ronghan Chen[1,2,3] Yang Cong[4]* Yu Ren[1,2,3]
[1]State Key Laboratory of Robotics, Shenyang Institute of Automation, Chinese Academy of Sciences
[2]Institutes for Robotics and Intelligent Manufacturing, Chinese Academy of Sciences
[3]University of Chinese Academy of Sciences
[4]College of Automation Science and Engineering, South China University of Technology
chenronghan@sia.cn, congyang81@gmail.com, renyu0414@gmail.com

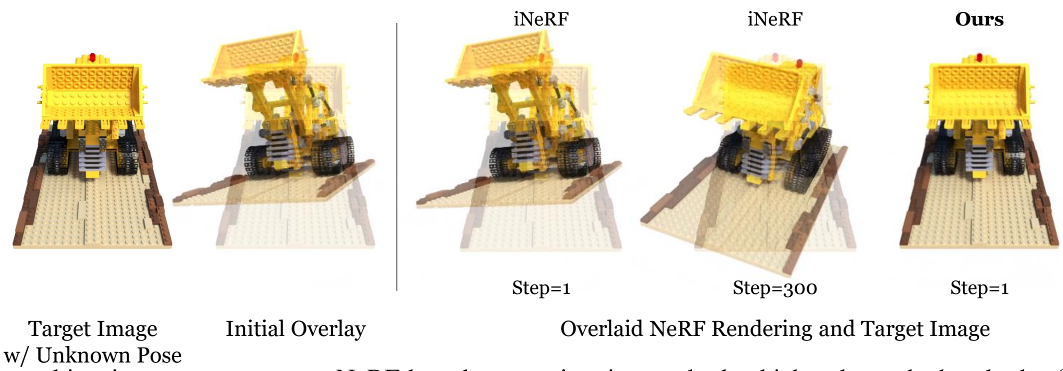

Fig. 1: Given an object image, we propose a NeRF-based pose estimation method, which reduces the hundreds of optimization steps in former NeRF-based method to *one* step, while avoiding being stuck in local minima, and obtaining more accurate poses. As a result, with only 5 minutes training of a fast NeRF [1], our method achieves CAD model-free real-time pose estimation on ***novel*** objects at 6FPS. **This is a shortened version of an accepted paper at ICRA 2024.**

Abstract— **Given the image collection of an object, we aim at building a real-time image-based pose estimation method, which requires neither its CAD model nor hours of object-specific training. Recent NeRF-based methods provide a promising solution by directly optimizing the pose from pixel loss between rendered and target images. However, during inference, they require long converging time, and suffer from local minima, making them impractical for real-time robot applications. We aim at solving this problem by marrying image matching with NeRF. With 2D matches and depth rendered by NeRF, we directly solve the pose in one step by building 2D-3D correspondences between target and initial view, thus allowing for real-time prediction. Moreover, to improve the accuracy of 2D-3D correspondences, we propose a 3D consistent point mining strategy, which effectively discards unfaithful points reconstruted by NeRF. Moreover, current NeRF-based methods naively optimizing pixel loss fail at occluded images. Thus, we further propose a 2D matches based sampling strategy to preclude the occluded area. Experimental results on representative datasets prove that our method outperforms state-of-the-art methods, and improves inference efficiency by 90×, achieving real-time prediction at 6 FPS.**

I. Introduction

Object pose estimation has wide applications in robot manipulation, augmented reality (AR) and mobile robotics [2]. Traditional methods typically require the CAD model of the object in advance, and searching for handcrafted features [3], [4] between the preregistered images or templates and the target image. However, obtaining such high-quality CAD model can be difficult and labor-intensive, or requires specialized high-end scanners. Recent methods have been applying deep neural network to regress the poses [5]–[9]. However, they can only estimate poses of known instances [5]–[7] or similar ones from the same category [8]–[11], and have to retrain on novel objects for hours. Moreover, they require large amount of training data, which is tedious to collect and annotate. Thus, it is difficult to apply such methods in real world due to unaffordable training time and human labor.

To further avoid tedious retraining for each novel object, recent methods [12], [13] learn from the traditional pipeline of SfM (Struction-from-Motion) to estimate object poses via feature matching. Given a small set of multi-view images, they first reconstruct sparse point cloud of the object via SfM, and then form 2D-3D correspondences to estimate the pose by solving the PnP [14] problem. Unfortunately, such methods rely on forming stably repeatable correspondences across all input frames, which usually cannot be guaranteed, thus leading to large pose error. On the other hand, recent advances in NeRF (Neural Radiance Fields [1], [15]–[17]) provide a mechanism for capturing complex 3D geometry in a few minutes. Following former render-and-compare methods for pose estimation [18]–[20], iNeRF [16] first trains a NeRF from image collection, and then during testing, it optimizes the pose by minimizing dense pixel error between the rendered and target image. Such dense supervision allows iNeRF to achieve more accurate alignment, but it also requires hundreds of iterations taking minutes. Moreover, its convergence relies on good initialization, and typically fails at large pose differences or occlusion.

In this work, we try to combine the best of both worlds by marrying image matching with NeRF to achieve *real-time* image-based pose estimation, without hundred steps of optimization. With 2D pixel matches and corresponding depth rendered by NeRF, we can build 2D-3D correspondences, and directly solve the pose with PnP [14]. This significantly reduces the iteration number and allows for real-time inference for NeRF based method. Moreover, comparing to former keypoint-based method [12], [13], this eases the difficulty of building 2D-3D correspondences in traditional SFM-based methods, which needs to find 2D matches between multiple input frames and the target image. With NeRF, our method only matches between two images once, and can convert arbitrary 2D matches to 2D-3D correspondences by backprojecting NeRF rendered depth into 3D space.

Moreover, owing to the implicit nature of NeRF, the rendered depth can be noisy and unfaithful [21]–[23]. To improve the quality of 2D-3D correspondences, we further propose a 3D consistent point mining strategy to discard unfaithful and noisy 3D points reconstructed by NeRF so that the PnP can obtain more accurate poses. Specifically, we render the 3D points from nearby viewpoints and regard the variation of them as the 3D consistency.

Our method also allows for further pose refinement from pixel error, like former render-and-compare methods [16], [18], [19]. However, this process is sensitive to occlusion, which bakpropagtes false gradient to the pose. We notice that the matching points indicate unoccluded area, and propose a matching point based sampling strategy for loss computation. We show that our proposed method improves the efficiency over former NeRF based methods by **90 times**, and can inference in real-time at 6FPS, while achieving higher pose accuracy and stronger robustness to occlusion.

Our contributions are three folds: 1) An efficient NeRF based pose estimation method is proposed by introducing image matching, which allows real-time image-based inference, and is free of CAD model or hours of pretraining. 2) We propose a 3D consistent point mining strategy to discard unfaithful 3D points to enable more accurate pose estimation. 3) In contrast to former render-and-compare based methods, our method can overcome the occlusion problem with a matching point based sampling strategy.

II. Background

NeRF. Given multi-view images with annotated camera parameters, NeRF [15] represents scenes via a 5D function:

$$\mathbf{c}, \sigma = \Phi(\mathbf{x}, \mathbf{d}), \tag{1}$$

which maps the query point location $\mathbf{x} \in \mathbb{R}^3$ to its density $\sigma \in \mathbb{R}^1$, and view-dependent color $\mathbf{c} \in \mathbb{R}^3$ at direction $\mathbf{d} \in \mathbb{R}^3$. To render an image from view P, the color $\hat{C}(\mathbf{p}, P)$ of a pixel $\mathbf{p} \in \mathbb{R}^2$ is obtained by accumulating the color along rays $\mathbf{r}$ that passes the pixel, following the volume rendering technique [46]:

$$\hat{C}(\mathbf{p}, P) = \sum_{i=i}^{N} \omega_i \mathbf{c}_i, \tag{2}$$

where $\omega_i = \sum_{i=i}^{N} T_i(1 - \exp(\sigma_i \delta_i))$ is the weight of each ray point, $T_i = \exp(-\sum_{j=1}^{i-1} \sigma_j \delta_j)$, and δ_i is the sample step along the ray. Similarly, we can also render an approximate depth at pixel $\mathbf{p}$ by

$$\hat{z}(\mathbf{p}, P) = \sum_{i=i}^{N} \omega_i t_i, \tag{3}$$

where $t_i \in \mathbb{R}^1$ is the depth at each ray point.

NeRF-based Pose Estimation. INeRF first proposes to estimate the pose of a novel object with NeRF. It first trains a NeRF model Φ with multi-view images of the object. Then, during inference, given a new target image I_t, iNeRF [16] recovers the camera pose $T \in SE(3)$ by optimizing:

$$\hat{T} = \underset{T \in \mathrm{SE}(3)}{\operatorname{argmin}} \|\Phi(T) - I_t\|_2, \tag{4}$$

where $\Phi(T)$ denotes NeRF rendered image from view T, and the function denotes an L2 loss between $\Phi(T)$ and the target image I_t. The NeRF weights are fixed in optimization.

III. Method

Our method aims at improving the convergence speed of NeRF-based pose estimation method. The key insight is to marry feature matching with NeRF to directly solve the pose from 2D-3D correspondences via PnP, which we introduce in III-A. Moreover, owing to the implicit nature of NeRF, 3D coordinates lifted from 2D pixels can be noisy and unfaithful. Thus, in Sec. III-B, we improve the 3D consistency by introducing a *3D consistent point mining* strategy before solving the pose. So far, without any refinement, our result is already more accurate than iNeRF [16] in most cases, which needs hundreds steps of refinement. Our method also allows further optimization to refine the initial pose. However, we notice that current pixel error (Eq. 4) cannot handle occluded images. For this, we propose a keypoint-guided occlusion robust refinement to tackle the occlusion problem, which is introduced in Sec. III-C.

A. One-step Pose Estimation via Feature Matching

Optimizing the pose from the photometric loss between rendered and target image following the formulation of iNeRF [16] (Eq. 4) can be extremely challenging, due to highly non-convex objective function. As a result, current methods are prone to being stuck in local minima. Here, we propose to estimate the pose by marrying image matching with NeRF. As shown in Fig. 2, the method has three main steps:

1) Matching: To estimate the pose of the target image I_t, we first render an image I_r from the initial guess of camera pose P with the trained NeRF model. Then, a pretrained off-the-shelf image matching model [45], [47], [48] is applied to form 2D-2D matches $[\mathbf{q}_i, \mathbf{p}_i]$ between the target image I_t and the rendered image I_r, with $\mathbf{q}_i \in I_t$ and $\mathbf{p}_i \in I_r$. We apply the recent proposed transformer-based image matching method LoFTR [47] in all our experiments.

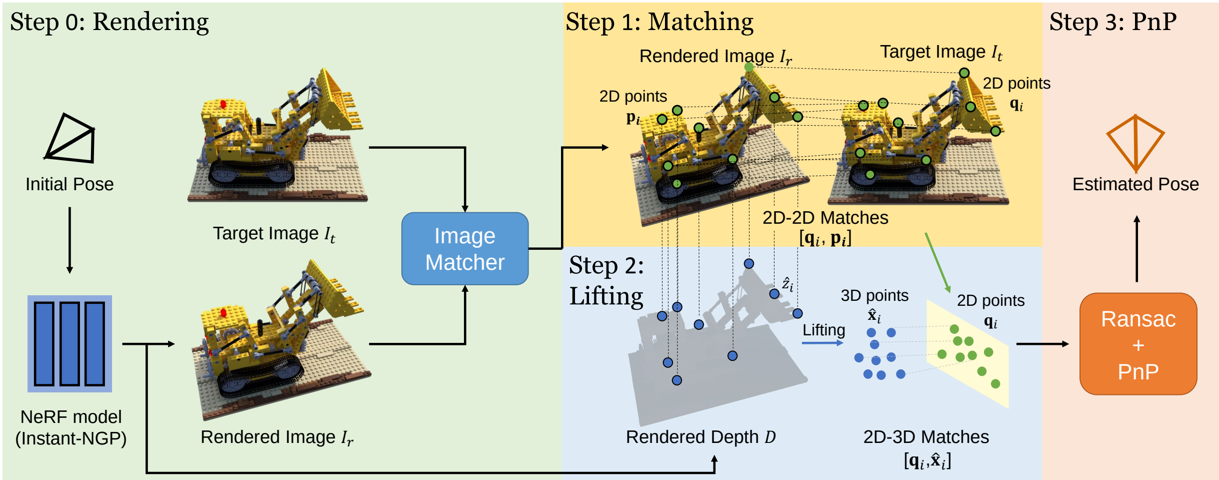

Fig. 2: Framework of the one-step pose estimation via feature matching strategy. Given the initial pose, we use NeRF [1] to render an RGB image I_r, and a depth image D. Then, an off-the-shelf image matcher [47] is applied to generate 2D-2D matches between the rendered and target image. Given location of matched 2D points and its depth rendered by NeRF, the 3D coordinates can be obtained, thus forming 2D-3D matches, from which the pose is finally solved via PnP+RANSAC.

2) Lifting: We then convert 2D-2D matches $[\mathbf{q}_i, \mathbf{p}_i]$ between target image I_t and rendered image I_r to 2D-3D correspondences $[\mathbf{q}_i, \mathbf{x}_i]$. We achieve this by lifting the matched 2D pixels $\mathbf{p} \in \mathbb{R}^2$ in NeRF rendered image I_r to 3D space. Specifically, we first obtain the depth $\hat{z}_i$ from the depth map D rendered by the trained NeRF model following Eq. 3. Then, the 3D coordinate of the corresponding point $\hat{\mathbf{x}}$ is obtained via backprojection, and transformed to world space via current camera pose P:

$$\hat{\mathbf{x}}_i = P \hat{z}_i K^{-1} \mathbf{p}_i \tag{5}$$

3) PnP: After obtaining the 2D-3D correspondences, the pose is computed via PnP [14] with RANSAC [49]. The above procedure already allows us to obtain good pose with only one rendering step, which is much faster than former NeRF based baselines [16], [17]. However, there may exists error due to inaccurate feature matches. In the following, we introduce a strategy to further improve the performance.

B. 3D Consistent Point Mining

In the above framework, one of the key factors that affect the pose accuracy is the precision of the 2D-3D matches, which are computed by a trained NeRF [15] model as stated in III-A.2. However, owing to the implicit nature of NeRF, the learned scene geometry can be unfaithful and noisy [21]–[23]. Moreover, the estimated 3D coordinates can be inconsistent when rendering from different views, resulting in large pose error. These problems become severer when the training images are limited, or the camera poses are noisy.

To counter the above problem, we propose to preclude the inconsistent 3D points by introducing a 3D consistent point mining strategy. Specifically, for each 3D keypoint $\mathbf{x}$ that is lifted from a matched 2D pixel $\mathbf{p}$, its consistency m is evaluated by re-estimating the 3D coordinates from nearby views, and computing how well these points are aligned with each other.

Specifically, given the current view P and estimated 3D keypoint $\mathbf{x}$, we first sample k nearby views $\mathcal{P} = \{P_i\}_{i=1}^k$. Then, we shoot rays $R = \{\mathbf{r}_i\}_{i=1}^k$ that pass the 3D keypoint $\mathbf{x}$ from each view in $\mathcal{P}$, and estimate the 3D coordinates $X = \{\mathbf{x}_i\}_{i=1}^k$ on these rays:

$$X = \hat{Z} \cdot \mathrm{norm}(\mathcal{P} P^{-1} \mathbf{x}), \tag{6}$$

where $\hat{Z} = \{\hat{z}_i\}_{i=1}^k$ is the depth value of rays R estimated by NeRF, and $\mathrm{norm}(\cdot)$ denotes vector normalization. We measure the point consistency with the location variance:

$$m = \frac{1}{k} ||X - \mathbf{x}||_2^2, \tag{7}$$

where larger m indicates lower consistency. Finally, we introduce a threshold γ to discard the points whose consistency $m > \gamma$, where γ is determined empirically.

C. Keypoint-guided Occlusion Robust Refinement

Current NeRF-based method cannot estimate the pose of occluded images. The reason is that the photometric loss computed from occluded area will backpropagate false gradients to the pose, which will aggravate the issue of being stuck in local minima.

Our image-matching based strategy provides a solution to this problem. Assuming the image matcher to be accurate enough, the matched keypoints naturally provide cues for unoccluded area, thus preventing the false gradients. We propose to compute the photometric loss with a new matched keypoint-guided sampling strategy. Specifically, after predicting matches, we apply 5×5 morphological dilation around the matched keypoint for n times to obtain the sample region.

IV. Experiments

We evaluate the pose estimation performance of our proposed method on NeRF synthetic dataset [15] and complex real-world scene from LLFF dataset [50].

TABLE I: 6-DoF pose estimation Results on the NeRF Synthetic and LLFF datasets, where RE / TE denote rotation / translation error, respectively. mRE / mTE denote mean rotation / translation error over all subjects.

Method	RE<5°(↑)	TE<0.05(↑)	mRE (↓)	mTE (↓)
NeRF Synthetic Dataset				
iNeRF [16]	0.585	0.56	10.33	0.559
pi-NeRF [17]	0.24	0.04	15.83	1.073
LoFTR [47]	0.785	-	6.15	-
Ours (1-step)	0.945	0.75	1.57	0.096
Ours	**0.95**	**0.88**	**1.25**	**0.077**
LLFF Dataset				
iNeRF [16]	0.50	0.55	16.46	0.0618
pi-NeRF [17]	0.00	0.00	133.37	3.999
LoFTR [47]	0.994	-	0.667	-
Ours (1-step)	**1.00**	**1.00**	0.325	**0.0027**
Ours	**1.00**	**1.00**	**0.135**	**0.0008**

A. Comparison Methods

We evaluate our method by comparing against state-of-the-art NeRF based pose estimation methods **iNeRF** [16], ***pi-NeRF*** [17], and image matching based method ***LoFTR [47]***.

To demonstrate the significance of the proposed feature matching strategy, we build Ours (1-step) baseline. It takes the PnP solved pose as final results, and does not apply further pose refinement.

B. Results on Synthetic Dataset

1) Setting: We choose Instant-ngp [1] as the NeRF model. For evaluation, we follow iNeRF [16] to sample test images, and add a rotation perturbation within $[10°, 40°]$, and a translation perturbation within 0.2.

2) Results: As shown in Tab. I, we report the pose correctness, *i.e.*, the rate of poses with rotation error $< 5°$, and translation error < 5 units, and mean rotation (mRE) and translation error (mTE).

On NeRF Synthetic dataset, *Ours (1-step)* already outperforms NeRF-based methods by 36% and 19% in terms of the rotation and translation accuracy. Moreover, pi-NeRF [17] achieves worse performance that iNeRF [16]. We assume the reason is that pi-NeRF fails to guess good initial pose under such severe pose perturbation, and abandoning the interest region based pixel loss used in iNeRF makes the convergence even harder. Our method is also superior than direct solving pose from LoFTR [47] 2D matches via epipolar geometry. With post refinement of 40 steps, our full method can further boost the correctness of rotation and translation from 94.5% / 75% to 95% / 88%.The qualitative results shown in Fig. 3 shows that our method achieves nearly perfect alignment under large initial pose differences.

C. Results on Real World Scene

On real-world scene, similar to NeRF Synthetic dataset, our method achieves the best results. This dataset is more challenging, because the scenes are captured with forward-facing images, which will result in larger image differences under the same rotation angle. As a result, it leads to performance degradation for the comparison methods. On

Fig. 3: Qualitative results of pose estimation on NeRF synthetic [15] and real-world LLFF dataset [50]. We visualize the results by overlying the target image and NeRF rendering image from the estimated pose.

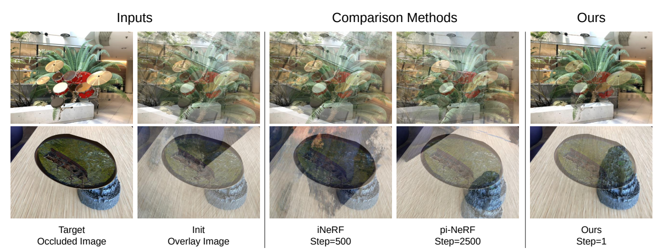

Fig. 4: Qualitative results of pose estimation on synthesized occluded data. The comparison methods fail to align the occluded images after hundreds of iterations, while our method aligns well in one step.

the contrary, our method even achieves better results (100%). This verifies the robustness of our method to large pose variations. Compared to synthetic dataset, the improved performance may be because the matcher [47] performs better on real-world data.

D. Results on Occluded Dataset

The occluded data is synthesized by composing the NeRF synthetic and LLFF dataset. The LLFF real-world images are used as background, and objects from synthetic dataset are overlayed as foreground. The qualitative results are shown in Fig. 4.

V. Conclusion

We have proposed a fast NeRF-based framework for imaged-based, CAD-free novel object pose estimation. By introducing keypoint matching, our method can directly solve the pose with one step. Moreover, we propose a 3D consistent point mining strategy to improve the quality of 2D-3D correspondences, and a matching keypoint based sampling strategy to improve the robustness to occluded images.

REFERENCES

[1] 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.

[2] Y. Cong, R. Chen, B. Ma, H. Liu, D. Hou, and C. Yang, "A comprehensive study of 3-d vision-based robot manipulation," *IEEE Trans. Cybern.*, vol. 53, no. 3, pp. 1682–1698, 2023.

[3] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proceedings of the International Conference on Computer Vision, Kerkyra, Corfu, Greece, September 20-25, 1999*. IEEE Computer Society, 1999, pp. 1150–1157.

[4] S. Hinterstoisser, C. Cagniart, S. Ilic, P. Sturm, N. Navab, P. Fua, and V. Lepetit, "Gradient response maps for real-time detection of textureless objects," *IEEE transactions on pattern analysis and machine intelligence*, vol. 34, no. 5, pp. 876–888, 2011.

[5] Y. Xiang, T. Schmidt, V. Narayanan, and D. Fox, "Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes," *Robotics: Science and Systems XIV*, 2018.

[6] C. Wang, D. Xu, Y. Zhu, R. Martín-Martín, C. Lu, L. Fei-Fei, and S. Savarese, "Densefusion: 6d object pose estimation by iterative dense fusion," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 3343–3352.

[7] S. Peng, Y. Liu, Q. Huang, X. Zhou, and H. Bao, "Pvnet: Pixel-wise voting network for 6dof pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4561–4570.

[8] Y. Di, R. Zhang, Z. Lou, F. Manhardt, X. Ji, N. Navab, and F. Tombari, "Gpv-pose: Category-level object pose estimation via geometry-guided point-wise voting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 6781–6791.

[9] X. Chen, Z. Dong, J. Song, A. Geiger, and O. Hilliges, "Category level object pose estimation via neural analysis-by-synthesis," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVI 16*. Springer, 2020, pp. 139–156.

[10] T. Lee, B.-U. Lee, M. Kim, and I. S. Kweon, "Category-level metric scale object shape and pose estimation," *IEEE Robotics and Automation Letters*, vol. 6, no. 4, pp. 8575–8582, 2021.

[11] K. Chen, S. James, C. Sui, Y.-H. Liu, P. Abbeel, and Q. Dou, "Stereopose: Category-level 6d transparent object pose estimation from stereo images via back-view nocs," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 2855–2861.

[12] J. Sun, Z. Wang, S. Zhang, X. He, H. Zhao, G. Zhang, and X. Zhou, "Onepose: One-shot object pose estimation without cad models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 6825–6834.

[13] X. He, J. Sun, Y. Wang, D. Huang, H. Bao, and X. Zhou, "Onepose++: Keypoint-free one-shot object pose estimation without CAD models," in *Advances in Neural Information Processing Systems*, 2022.

[14] V. Lepetit, F. Moreno-Noguer, and P. Fua, "Ep n p: An accurate o (n) solution to the p n p problem," *International journal of computer vision*, vol. 81, pp. 155–166, 2009.

[15] 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," in *ECCV*, 2020.

[16] 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, pp. 1323–1330.

[17] Y. Lin, T. Müller, J. Tremblay, B. Wen, S. Tyree, A. Evans, P. A. Vela, and S. Birchfield, "Parallel inversion of neural radiance fields for robust pose estimation," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 9377–9384.

[18] K. Park, A. Mousavian, Y. Xiang, and D. Fox, "Latentfusion: End-to-end differentiable reconstruction and rendering for unseen object pose estimation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 10 710–10 719.

[19] A. Palazzi, L. Bergamini, S. Calderara, and R. Cucchiara, "End-to-end 6-dof object pose estimation through differentiable rasterization," in *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, 2018, pp. 0–0.

[20] W. Chen, H. Ling, J. Gao, E. Smith, J. Lehtinen, A. Jacobson, and S. Fidler, "Learning to predict 3d objects with an interpolation-based differentiable renderer," *Advances in neural information processing systems*, vol. 32, 2019.

[21] L. Yariv, J. Gu, Y. Kasten, and Y. Lipman, "Volume rendering of neural implicit surfaces," *Advances in Neural Information Processing Systems*, vol. 34, pp. 4805–4815, 2021.

[22] L. Yariv, Y. Kasten, D. Moran, M. Galun, M. Atzmon, B. Ronen, and Y. Lipman, "Multiview neural surface reconstruction by disentangling geometry and appearance," *Advances in Neural Information Processing Systems*, vol. 33, pp. 2492–2502, 2020.

[23] P. Wang, L. Liu, Y. Liu, C. Theobalt, T. Komura, and W. Wang, "Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction," *arXiv preprint arXiv:2106.10689*, 2021.

[24] W. Kehl, F. Manhardt, F. Tombari, S. Ilic, and N. Navab, "Ssd-6d: Making rgb-based 3d detection and 6d pose estimation great again," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 1521–1529.

[25] M. Rad and V. Lepetit, "Bb8: A scalable, accurate, robust to partial occlusion method for predicting the 3d poses of challenging objects without using depth," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 3828–3836.

[26] B. Tekin, S. N. Sinha, and P. Fua, "Real-time seamless single shot 6d object pose prediction," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 292–301.

[27] C. Song, J. Song, and Q. Huang, "Hybridpose: 6d object pose estimation under hybrid representations," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 431–440.

[28] I. Shugurov, F. Li, B. Busam, and S. Ilic, "Osop: A multi-stage one shot object pose estimation framework," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 6835–6844.

[29] Y. Labbé, L. Manuelli, A. Mousavian, S. Tyree, S. Birchfield, J. Tremblay, J. Carpentier, M. Aubry, D. Fox, and J. Sivic, "Megapose: 6d pose estimation of novel objects via render & compare," in *Proceedings of the 6th Conference on Robot Learning (CoRL)*, 2022.

[30] G. Ponimatkin, Y. Labbé, B. Russell, M. Aubry, and J. Sivic, "Focal length and object pose estimation via render and compare," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 3825–3834.

[31] S. Liu, T. Li, W. Chen, and H. Li, "Soft rasterizer: A differentiable renderer for image-based 3d reasoning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 7708–7717.

[32] F. Petersen, B. Goldluecke, C. Borgelt, and O. Deussen, "GenDR: A Generalized Differentiable Renderer," in *IEEE/CVF International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[33] D. Maggio, M. Abate, J. Shi, C. Mario, and L. Carlone, "Loc-nerf: Monte carlo localization using neural radiance fields," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 4018–4025.

[34] 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, pp. 6229–6238.

[35] 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, pp. 12 786–12 796.

[36] F. Tosi, Y. Zhang, Z. Gong, E. Sandström, S. Mattoccia, M. R. Oswald, and M. Poggi, "How nerfs and 3d gaussian splatting are reshaping slam: a survey," *arXiv preprint arXiv:2402.13255*, 2024.

[37] J. Tang, S. Miller, A. Singh, and P. Abbeel, "A textured object recognition pipeline for color and depth image data," in *2012 IEEE International Conference on Robotics and Automation*. IEEE, 2012, pp. 3467–3474.

[38] M. Martinez, A. Collet, and S. S. Srinivasa, "Moped: A scalable and low latency object recognition and pose estimation system," in *2010 IEEE International Conference on Robotics and Automation*. IEEE, 2010, pp. 2043–2049.

[39] A. Collet, D. Berenson, S. S. Srinivasa, and D. Ferguson, "Object recognition and full pose registration from a single image for robotic manipulation," in *2009 IEEE International Conference on Robotics and Automation*. IEEE, 2009, pp. 48–55.

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

[41] E. Rosten and T. Drummond, "Machine learning for high-speed corner detection," in *Computer Vision–ECCV 2006: 9th European Conference on Computer Vision, Graz, Austria, May 7-13, 2006. Proceedings, Part I 9*. Springer, 2006, pp. 430–443.

[42] 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.

[43] K. M. Yi, E. Trulls, Y. Ono, V. Lepetit, M. Salzmann, and P. Fua, "Learning to find good correspondences," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 2666–2674.

[44] R. Ranftl and V. Koltun, "Deep fundamental matrix estimation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 284–299.

[45] P.-E. Sarlin, D. DeTone, T. Malisiewicz, and A. Rabinovich, "Superglue: Learning feature matching with graph neural networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 4938–4947.

[46] J. T. Kajiya and B. P. Von Herzen, "Ray tracing volume densities," *ACM SIGGRAPH computer graphics*, vol. 18, no. 3, pp. 165–174, 1984.

[47] J. Sun, Z. Shen, Y. Wang, H. Bao, and X. Zhou, "Loftr: Detector-free local feature matching with transformers," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 8922–8931.

[48] P.-E. Sarlin, A. Unagar, M. Larsson, H. Germain, C. Toft, V. Larsson, M. Pollefeys, V. Lepetit, L. Hammarstrand, F. Kahl *et al.*, "Back to the feature: Learning robust camera localization from pixels to pose," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 3247–3257.

[49] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[50] B. Mildenhall, P. P. Srinivasan, R. Ortiz-Cayon, N. K. Kalantari, R. Ramamoorthi, R. Ng, and A. Kar, "Local light field fusion: Practical view synthesis with prescriptive sampling guidelines," *ACM Transactions on Graphics (TOG)*, vol. 38, no. 4, pp. 1–14, 2019.

[51] V. Yugay, Y. Li, T. Gevers, and M. R. Oswald, "Gaussian-slam: Photo-realistic dense slam with gaussian splatting," *arXiv preprint arXiv:2312.10070*, 2023.

[52] H. Matsuki, R. Murai, P. H. Kelly, and A. J. Davison, "Gaussian splatting slam," *arXiv preprint arXiv:2312.06741*, 2023.

[53] C. Yan, D. Qu, D. Wang, D. Xu, Z. Wang, B. Zhao, and X. Li, "Gs-slam: Dense visual slam with 3d gaussian splatting," *arXiv preprint arXiv:2311.11700*, 2023.

[54] N. Keetha, J. Karhade, K. M. Jatavallabhula, G. Yang, S. Scherer, D. Ramanan, and J. Luiten, "Splatam: Splat, track & map 3d gaussians for dense rgb-d slam," *arXiv preprint arXiv:2312.02126*, 2023.