Robust and Efficient Edge-guided Pose Estimation with Resolution-conditioned NeRF

Liesbeth Claessens[1]
liesbeth.claessens@mavt.ethz.ch
Fabian Manhardt[2]
fabianmanhardt@google.com
Ricardo Martin-Brualla[2]
rmbrualla@google.com
Roland Siegwart[1]
rsiegwart@ethz.ch
Cesar Cadena[1]
cesarc@ethz.ch
Federico Tombari[2]
tombari@google.com

[1] Autonomous Systems Lab
ETH Zurich
Zurich, Switzerland
[2] Google, Inc.
Zurich, Switzerland

Abstract

Monocular 6D pose estimation attempts to obtain the 3D location and rotation of an object from a single input image. Deep Learning based methods have lead to large improvements in this area, but still require highly realistic 3D CAD models or manual data labeling. Neural Radiance Fields (NeRF) have made significant progress in solving the inverse problem of realistically rendering an object from a novel pose *without* these requirements. Nevertheless, current methods that try to harness them to extract the object's pose from a given image using an analysis-by-synthesis approach lack robustness and speed, and require hours of pretraining on the object of interest. We propose a novel pose estimation pipeline that makes NeRF-based analysis-by-synthesis reliable and fast. Our proposal improves the quality of the optimization by 1) changing the representation of the pose to a decoupled and continuous parameterization, 2) increasing the model's robustness to changes in scale by means of conditioning it on the used resolution, and 3) developing an edge-based sampling strategy that focuses on shooting rays near image regions with a strong learning signal. These improvements, along with our backbone choice, allow us to estimate the pose with more than 5% higher recall and more than 4 times faster than prior work, while reducing the pretraining time from hours to minutes.

1 Introduction

Monocular 6D object pose estimation aims at predicting the 3D rotation and translation of an object with respect to the camera from a single RGB image. This is a crucial problem

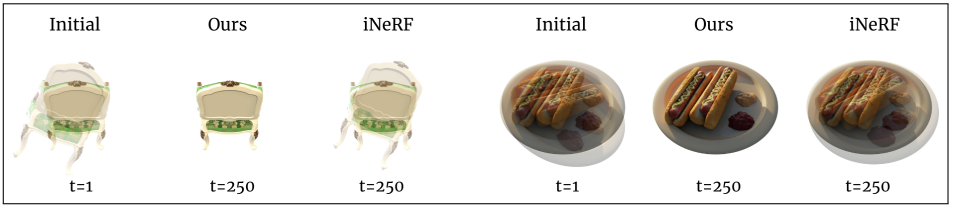

Figure 1: **Radiance Fields for Pose Estimation.** We show two exemplary pose optimization examples using the Chair and Hotdog objects from the Blender dataset. In contrast to iNeRF, our method is capable of converging to a visually optimal solution, thanks to our improved pose parameterization and smart sampling strategy.

in robotic perception as it represents a key challenge for autonomous agents to effectively interact with the surroundings, e.g. via robotic manipulation and grasping [39]. 6D object pose estimation is relevant also in other applied domains such as autonomous driving [25], 3D scene understanding [15], and virtual and augmented reality [4].

Despite large recent improvements, existing methods for pose estimation still exhibit several limitations [14]. Most state-of-the-art approaches, for instance, depend heavily on the availability of a high-quality 3D CAD model for each object of interest [10, 38]. Obtaining such CAD models can be a labor intensive process which requires a high level of expertise and tends to not scale well with the number of objects [36]. Moreover, it is well known that the accurate capturing and rendering of certain types of objects, *e.g* with non-Lambertian surfaces, is far from trivial. Once a CAD model is obtained, a dataset has to be created that captures the correspondence between points on the CAD model and image pixels. The creation of such a dataset requires a significant and error prone annotation effort [14]. When this requirement is bypassed through synthetic generation, the methods trained on the data tend to suffer from domain gap when being deployed in real settings [36, 38].

Neural Radiance Fields (NeRF) have recently emerged as a powerful tool for the reverse problem of synthesizing lifelike images from novel poses [27]. The scene representation learned by NeRF is highly accurate and captures complex photo-metric effects such as transparencies and reflection [35], whilst only requiring a small set of posed training images. Since the image synthesis process of NeRFs is fully differentiable, it is only natural to leverage it for pose estimation via analysis-by-synthesis [43]. Nevertheless, while this paradigm allows for model-free pose estimation of objects with very challenging surfaces, it is limited in its applicability since it requires hours of pre-training on the object of interest and almost a minute to run at inference time.

In this work, we propose a novel pose estimation method developed to make the NeRF-based analysis-by-synthesis paradigm more efficient and reliable. We adopt the fast-training Instant NGP backbone from [28], and make pose optimization with this backbone viable by altering several aspects of the optimization. First, we reparameterize the pose. We propose to decouple rotation from translation parameters [38] and further improve the rotation representation by using the novel formulation from [45], which is fully continuous in SO(3). Second, we make our model more suitable for pose estimation by increasing its rendering quality by conditioning it on the resolution. Please note that this conditioning does not increase the supervision requirements of our system. Finally, borrowing ideas from classical edge-based pose refinement [11], we develop a novel ray sampling mechanism that focuses

on image regions that have a strong signal to measure misalignment in the pose. To prove the efficacy of our method, we evaluate it on the NeRF Realistic Synthetic 360° Dataset [27] and show that we outperform the state-of-the-art [43] in terms of accuracy and speed.

2 Related work

Neural Radiance Fields NeRFs [27] are a recently proposed method for conducting high-fidelity novel view synthesis given a set of posed images. Due to NeRF's impressive results, several follow-up works have been introduced that improve upon different aspects and tackle different challenges in novel view synthesis. MipNeRF, for example, reduces aliasing by making NeRF robust to scale changes by aggregating the pixel color from a cone instead of a single ray [2]. Another line of work investigates the use of different ways to encode the 3D locations to speed up training and obtain more detailed renderings [7, 13, 28]. Some methods propose to optimize the camera pose together with the NeRF to circumvent errors in pose introduced by COLMAP [16, 22, 34, 40]. Another branch of NeRF literature focuses on explicitly modeling different aspects of the rendering process, such as e.g the Bidirectional Reflectance Distribution Function (BRDF), in order to improve the rendering quality and controllability [5, 35]. For a more complete overview of the recent literature on Radiance Fields, we kindly refer the reader to [42].

Monocular 6D pose estimation To retrieve the 6D pose, early methods use local or global features and search for key point correspondences on CAD models [3, 8, 9, 23]. Recently, deep learning methods dominate the field and solve the task using convolution neural networks (CNNs) supervised by annotated data to extract deep features. After the feature extraction, there are two main approaches to estimate the pose, correspondence-based and regression/classification-based approaches. Correspondence-based methods establish 2D-3D correspondences [29, 30, 31, 32, 42], prior to leveraging a variant of the RANSAC&PnP paradigm to solve for pose. Regression-based approaches, on the other hand, directly regress or classify the pose of the detected object. Early regression-based methods usually have lower performance due to the existence of ambiguities [24] such as pose symmetries [41]. Later methods like SSD-6D [17] discretize the rotation space to circumvent this issue. In combination with better continuous representations for rotation [45], these methods have started to demonstrate high effectiveness [10, 37]. Nevertheless, all the previously mentioned methods heavily depend on the availability of accurate 3D CAD models. Obtaining such 3D models is a difficult and time-consuming process. In addition, most state-of-the-art modeling pipelines are incapable of capturing non-Lambertian surfaces, such as glass. NeRF, in contrast, has the capability to incorporate such surfaces and, in addition, is fully differentiable. Thus, iNeRF has recently proposed to leverage neural radiance fields to enable model-free pose estimation via analysis-by-synthesis [43]. Nevertheless, while this works generally well, it still requires several hours/days of training. In contrast, our method trains and evaluates fast (in the order of seconds) and yields higher accuracy on similar hardware.

3 Background

Neural Radiance Fields NeRF is a novel view synthesis method that is trained on a set of pairs of posed images $\{(\mathbf{P}_1^{c2w}, \mathbf{I}_1), ..., (\mathbf{P}_N^{c2w}, \mathbf{I}_N)\}$ by minimizing the photo-metric loss

between ground truth pixels $\hat{\mathbf{I}}_{i,\mathbf{p}}$ and rendered pixels $\mathbf{I}_{i,\mathbf{p}}$:

$$\mathcal{L}_{photo} = ||\hat{\mathbf{I}}_{i,\mathrm{p}} - \mathbf{I}_{i,\mathrm{p}}||_2. \tag{1}$$

NeRF does not predict the pixel color directly, but is modeled as an MLP ϕ that takes a point in 3D space $\mathbf{x} \in \mathbb{R}^3$ and viewing direction $\mathbf{d} \in \{\mathbf{d} \in \mathbb{R}^3 : \|\mathbf{d}\| = 1\}$ as input to predict the emitted color $\mathbf{c} \in \mathbb{R}^3$ and density $\sigma \in \mathbb{R}^+$ with $\phi(\mathbf{x},\mathbf{d}) = (\mathbf{c},\sigma)$. To improve the fidelity of the rendering, positional encoding $\gamma : \mathbb{R}^3 \longrightarrow \mathbb{R}^{6L}$ is typically employed, where L is the number of frequency channels of the encoding [27]. Prepending the positional encoding turns ϕ into:

$$\phi(\gamma_x(\mathbf{x}), \gamma_d(d)) = (\mathbf{c},\sigma). \tag{2}$$

This formulation then allows to render a pixel $\mathbf{p} = (x, y, 1)$ with volumetric rendering. To this end, the ray $\mathbf{r} = (\mathbf{a},\mathbf{d})$ cast by the camera with pose $\mathbf{P}^{c2w} = (\mathbf{R}^{c2w}|\mathbf{t}^{c2w}) \in SE(3)$, with 3D rotation $\mathbf{R}^{c2w} \in SO(3)$ and translation $\mathbf{t}^{c2w} \in \mathbb{R}^3$, in world space is computed according to $\mathbf{r} = (\mathbf{t}^{c2w}, \mathbf{R}^{c2w}\mathbf{K}^{-1}\mathbf{p})$. Finally, to render the color $\hat{\mathbf{I}}_{i,\mathrm{p}} \in [0,1]^3$ at pixel $\mathbf{p}$, we sample M discrete depth values t_m along the ray within the near and far plane $\left[t_n, t_f\right]$, and query the radiance field ϕ at the underlying 3D points $\mathbf{x}_m = \mathbf{a} + t_m\mathbf{d}$ along the viewing direction $\mathbf{d}$. The final color at pixel $\mathbf{p}$ is then composited as

$$\hat{\mathbf{I}}_{i,\mathrm{p}} = \hat{I}(\mathbf{p};\theta,\mathbf{P}_i) = \sum_{m=1}^{M} \alpha_m \mathbf{c}_m\,, \tag{3}$$

$$\text{where} \quad (\mathbf{c}_m, \sigma_m) = \phi\left(\gamma_x(\mathbf{x}_m), R\gamma_d(\mathbf{d})\right), \tag{4}$$

$$\alpha_m = T_m\left(1 - \exp(-\sigma_m\delta_m)\right), \tag{5}$$

$$T_m = \exp\left(-\sum_{m'=1}^{m} \sigma_{m'}\delta_{m'}\right). \tag{6}$$

Thereby, T_m denotes the accumulated transmittance along the ray from t_n to T_m, and $\delta_m = t_{m+1} - t_m$ is the distance between adjacent samples.

NeRF for pose estimation via analysis-by-synthesis We propose to estimate the 6D object pose $\mathbf{P}^{w2c}$, that brings the object from world to camera coordinate system, via a render-and-compare approach similar to [43]. To this end, we first train a NeRF on a set of posed images of the object of interest. To speed up the training from hours to a mere couple of minutes we adopt the Instant NGP architecture and replace the positional encodings from NeRF with a grid of trainable hash encodings [28]. This increases the convergence speed of the training by allowing the system to not have to update all parameters for every input-output pair [20]. During inference, we make use of the obtained radiance field to solve for the 6D object pose $\mathbf{P}^{w2c}$. To this end, we take the input image together with the optimized and frozen NeRF and minimize the photometric loss $\mathcal{L}_{photo}$ with respect to the pose $\mathbf{P}^{w2c}$.

4 Method

In this section, we introduce the key elements of our proposed method for fast, robust pose optimization. In section 4.1, we explain how we reparameterize the pose. In section 4.2, we explain how we make the analysis-by synthesis process more effective by changing the model architecture and sampling strategy. For the architectural details of our method and optimization parameterization, we kindly refer the reader to the Supplementary Material.

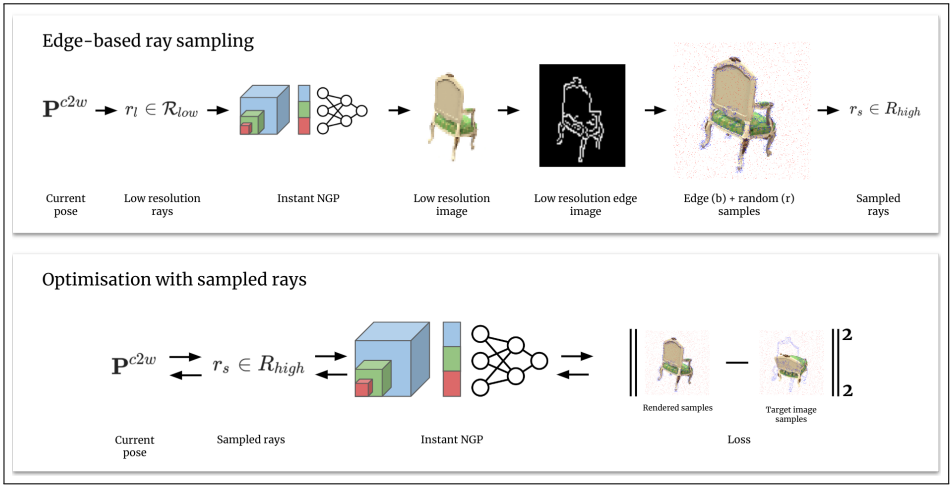

Figure 2: **Schematic Overview of Edge-Based Ray Sampling.** Given the current pose in our optimization, we render the respective view in low resolution. Subsequently, we extract edges from the low resolution image to obtain regions of interest for our optimization. We then upscale the edge image to a higher resolution and convert it to a probability distribution. After that, we sample both randomly and from the edge-based probability distribution to obtain the pixels for our optimization.

4.1 6D Pose Parameterization

When estimating the 6D object pose $\mathbf{P}^{w2c}$, several representations have been proposed in literature [17, 31, 33]. Utilizing the right representation can thereby have a huge impact on the quality of the output pose [38]. In this work, we leverage recent advances in pose representation to formulate it in a way that makes it more amenable to our optimization. The main features of our parameterization are that its rotation $\mathbf{R}^{w2c} \in SO(3)$ is continuous and decoupled from the translation parameters $\mathbf{t}^{w2c} \in \mathbb{R}^3$. This has been empirically shown to lead to better optimizations [21, 38, 45].

Interestingly, Li *et al.* have shown that when estimating the object translation, it is beneficial to work purely in image space [21]. In line with [21], we thus parameterize the translation $\mathbf{t}^{w2c}$ as the 2D center $\mathbf{c}$ in pixel space and object depth z. Using the camera intrinsics matrix $\mathbf{K} \in \mathbb{R}^{3\times3}$, one can then easily deduce the respective translation $\mathbf{t}^{w2c} = \mathbf{K}^{-1}\mathbf{c}z$. When estimating the 3D rotation, a mere lateral translation of the object can lead to the effect that visually different structures possess the same 3D rotation (see Figure 3). This is unfavorable as the network, thus, needs to have an understanding of the underlying camera parameters and 3D location in order to obtain the correct rotation. To circumvent this issue, we instead disentangle rotation and translation by means of the allocentric representation [18, 19, 26] $\mathbf{R}_a^{w2c}$. Intuitively, the allocentric parameterization automatically accounts for changes in the appearance of the object caused by its translation by compensating for the changes in viewpoint. Given the object's estimated allocentric rotation $\mathbf{R}_a^{w2c}$, the rotation $\mathbf{R}_{a2e}$ between the camera principal axis $\mathbf{l} = [0,0,1]^T$ and the ray through the object center projection $\mathbf{o} = \mathbf{K}^{-1}\mathbf{c}$

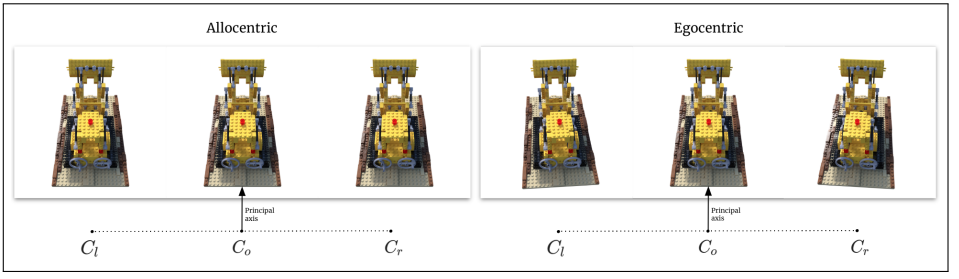

Figure 3: **Allocentric vs egocentric rotation.** In the above image [26] the object was translated along a line perpendicular to the principal axis of the camera with the rotation parameters kept constant. When using egocentric parameterization (right), despite constant rotation parameters the object appears rotated to the viewer. This is not the case when keeping the allocentric rotation constant (left), as this parameterization accounts for the viewpoint [19].

is computed. Essentially, $\mathbf{R}_{a2e}$ takes vector $\mathbf{l}$ to align it with $\mathbf{o}$ according to

$$\mathbf{R}_{a2e} = \mathbf{I}_3 + (sin\,\alpha)[\mathbf{a}]_X + (1 - cos\,\alpha)[\mathbf{a}]_X^2, \tag{7}$$

with $\mathbf{I}_3$ representing the identity matrix in $\mathbb{R}^{3\times3}$, $\mathbf{a} = \frac{\mathbf{l}\times\mathbf{o}}{||\mathbf{l}\times\mathbf{o}||}$ being the axis between the object ray $\mathbf{o}$ and the optical center ray $\mathbf{l}$, $\alpha = \arccos(\mathbf{l}\cdot\mathbf{o})$ describing the angle between them, and $[\cdot]_X$ being the skew-symmetric matrix. The final egocentric rotation can be derived as

$$\mathbf{R}^{w2c} = \mathbf{R}_{a2e}\mathbf{R}_a^{w2c}. \tag{8}$$

It is well known in literature that any representation for the 3D rotation with four or fewer parameters possesses discontinuities in the Euclidean space. When optimizing for the rotation, this can lead to larger errors close to the discontinuity boundaries [38]. To overcome this issue, Zhou *et. al.* have presented a novel 6-dimensional representation for $\mathbf{R}_a^{w2c}$, which is fully continuous [45]. In particular, the 6-dimensional representation $\mathbf{R}_{a6d}^{w2c}$ is defined as the first two columns of $\mathbf{R}_a^{w2c} = [\mathbf{r}_1, \cdots, \mathbf{r}_3]$:

$$\mathbf{R}_{a6d}^{w2c} = [\mathbf{r}_1 \mid \mathbf{r}_2]. \tag{9}$$

Given the current 6-dimensional estimate of $\mathbf{R}_{a6d}^{w2c} = [\mathbf{r}_1|\mathbf{r}_2]$, the rotation matrix $\mathbf{R}_a^{w2c} = [\mathbf{r}_{a\cdot1}^{w2c}|\mathbf{r}_{a\cdot2}^{w2c}|\mathbf{r}_{a\cdot3}^{w2c}]$ can be computed according to

$$\begin{cases} \mathbf{r}_{a\cdot1}^{w2c} = \zeta(\mathbf{r}_1) \\ \mathbf{r}_{a\cdot3}^{w2c} = \zeta(\mathbf{r}_{a\cdot1}^{w2c} \times \mathbf{r}_2) \\ \mathbf{r}_{a\cdot2}^{w2c} = \mathbf{r}_{a\cdot3}^{w2c} \times \mathbf{r}_{a\cdot1}^{w2c} \end{cases}, \tag{10}$$

where $\zeta(\bullet)$ denotes the vector normalization operation.

4.2 Efficient Render-and-Compare

Finally, using the current pose in camera space, we transform it to world space as $\mathbf{P}^{c2w} = (\mathbf{R}^{w2c^T} | -\mathbf{R}^{w2c^T}\mathbf{t}^{w2c})$, and query our trained NeRF to render the respective image $\hat{\mathbf{I}}$. We then

	Batch	Chair	Drums	Ficus	Hotdog	Lego	Materials	Mic	Ship	Mean
		Rotation error < 5 degrees								
iNeRF [43]	2048	0.88	0.81	0.9	0.97	0.97	0.82	**0.8**	**0.73**	0.86
Ours	8192	**0.93**	**0.95**	**0.92**	**1.0**	**1.0**	**0.85**	0.74	0.66	**0.88**
		Translation error < 0.02 units								
iNeRF [43]	2048	0.86	0.81	0.85	0.81	0.97	0.79	**0.77**	**0.7**	0.82
Ours	8192	**0.91**	**0.92**	**0.88**	**0.98**	**1.0**	**0.85**	0.71	0.63	**0.86**

Table 1: **Comparison with state-of-the-art.** Results on the NeRF Synthetic 360° dataset.

minimize the photo-metric error to solve for the camera pose. However, since rendering the full image $\hat{\mathbf{I}}$ is very expensive, we attempt to first understand what the most informative areas of our image to sample from are. To this end, we propose to condition the NeRF on the output resolution, in order to cheaply generate high quality images of low resolution, which can then be harnessed to infer the right locations for sampling.

Resolution-conditioned Rendering As aforementioned, we explicitly condition both the color and the density model on the resolution during training as well as inference by means of feeding the applied scaling factor as additional input. To this end, we randomly scale the original images to a set of discrete resolutions. The training batches are then composed by pixels from the different resolutions (uniformly mixed across all scales). We find that training on a multi-resolution dataset while explicitly conditioning on the resolution yields better results than just training on a multi-resolution dataset [2], while still not requiring any additional supervision signal.

Edge-based Ray Sampling To reduce the number of rays to sample during optimization, we place our samples in the pixel regions that are the most structurally informative to the *current pose*. The reason for this is that we observed that there is significant overlap between these structurally relevant regions and the regions with the strongest pose estimation gradients. To optimize for the pose or track the object in 3D space, classical computer vision commonly relied on edges since they possess the best signal to measure misalignment in 2D space [11, 17]. Inspired by this, we propose to also query the NeRF mostly around the edges of the object. Unfortunately, we do not have the information on where the edges are located. Nevertheless, we can leverage our resolution-conditioned model to efficiently render the current pose hypothesis at a very low resolution. In particular, we employ a resolution of 64×64, shooting in total 4096 rays. We then run the Canny edge detector [6] on this low resolution image to obtain the edges for this hypothesis. Finally, we use the retrieved edges to sample another 4096 rays in the original image resolution, which we combine with another 4096 randomly sampled rays. Subsequently, we compute the photo-metric loss and update our pose hypothesis. This process is then repeated until we hit convergence (commonly we require around 250 steps per image). We visualize the sampling process and subsequent optimization in Figure 2.

5 Experiments

Datasets Following the state of the art in NeRF-based pose optimization [43], we employ the NeRF Realistic Synthetic 360° Dataset [27] to evaluate our method. This dataset consists of

	Runtime			Convergence	
	Worse (s)	Best (s)	Average (s)	Rotation error $< 5^\circ$	Translation error < 0.02 Units
iNeRF	-	-	50	0.86	0.82
Ours (random sampling)	-	-	**4.4**	0.85	0.82
Ours (SLS)	10.1	4.4	5.3	0.88	0.87
Ours (edge sampling)	-	-	11	0.88	0.86
Ours (SLS+edge sampling)	23.2	11	12.9	**0.91**	**0.9**

Table 2: **Runtime and convergence overview.** Average refers to the average duration of a full pose estimation on a V100 GPU. Even the slowest version of our method is significantly faster than iNeRF, on the same hardware resources.

8 objects with different surface properties, such as completely opaque (e.g. Chair) and very reflective (e.g. Materials), rendered with Blender. To ensure a fair comparison, we also utilize the same perturbations as done in [43], which were provided to us by the authors.

Metrics To compare the pose estimation quality, we report the recall of successful optimization. As for rotation, we define an optimization to be successful when the error of the pose estimate is below a threshold of 5°. Similarly, for the translation, we consider an optimization to be successful if the translational error is less than 0.02 in COLMAP space [43].

5.1 Comparison with state of the art

We compare our method against iNeRF [43], which is the main representative of the NeRF-based analysis-by-synthesis paradigm. As one can see in TABLE 1, our method is able to outperform iNeRF on 6 out of 8 objects of the NeRF Realistic Synthetic 360° Dataset [27]. On average we exceed iNerf with a recall of 0.88 compared to 0.86 for rotation and 0.86 compared to 0.82 for translation. When examining results for individual scenes, we highlight how our method is particularly strong for well textured objects. For example, on Drums we can improve the numbers from iNeRF by more than 10% for both rotation and translation. This is quite expected, since textured objects possess strong edges even at low resolution, enabling us to sample very precisely from the full resolution image. In contrast, objects with barely any texture, such as Mic, cannot benefit equally well from our sampling strategy as little edges can be retrieved. Further, when comparing to the state of the art in terms of speed, we obtain similar results (see TABLE 2). In particular, while the fastest version of our method converges on average in less than 5 seconds on a single v100 GPU, a single optimization usually requires around 50 seconds for iNeRF using the same hardware. In TABLE 2, we show a comparison between different variations of our method in terms of run time and accuracy. Our method with random sampling runs 10 times faster than iNeRF, while yielding a similar recall. When we enhance our method with a simple Stochastic Local Search (SLS) strategy to escape local minima we surpass iNeRF in terms of recall, while still running almost 10 times faster. When combining SLS with edge sampling, our method yields the highest recall but runs significantly slower than its other variations.

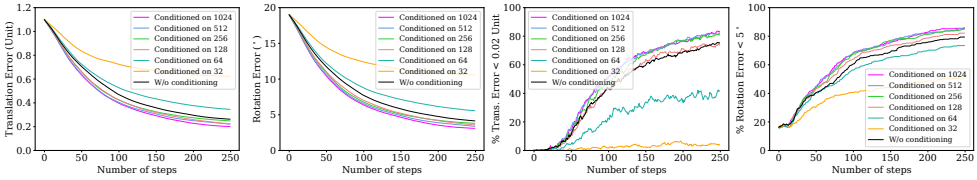

Figure 4: **Resolution-conditioning ablation study.** Conditioning on higher resolutions leads to higher accuracy.

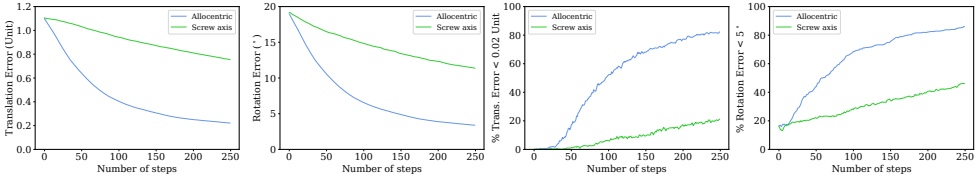

Figure 5: **Parameterization ablation study.** Our allocentric orthogonal 6D parameterization leads to higher accuracy.

5.2 Ablation studies

Ablation of the resolution conditioning. In this experiment, we illustrate that using a resolution-conditioned architecture significantly improves the quality of the pose estimation without requiring any additional supervision signal. This improvement can be observed even when the backbone is conditioned on a resolution lower than that of the original dataset (*i.e* 800×800). In particular, when conditioning the NeRF on a resolution of 512, we can report an average improvement of 6% in recall for rotation and translation compared to the model without resolution-conditioning. A more extensive comparison of how conditioning on specific resolutions affects the optimization is provided in Figure 5. The results show a pronounced but diminishing correlation between conditioning on higher resolutions and increased pose estimation quality.

Ablation on the pose parameterization. In Figure 5 we empirically show that our pose parameterization is crucial for a successful convergence. To this end, we compare our presented parameterization for pose with the more commonly employed screw axis parameterization, as also utilized in iNeRF. As it can be easily observed, using the continuous and allocentric parameterization allows our model to consistently converge to better optima.

6 Limitations

As an iterative optimization operating on a non-convex loss landscape, our method is susceptible to local minima. However, this can already be counteracted to some degree using a simple Stochastic local search (SLS) strategy [1, 12] to escape from local minima. A potential SLS mechanism could consist of: a) generating a set of allocentric rotation perturbations to the current pose in the optimization when it stagnates and b) restarting the optimization from the rotation perturbation that leads to the lowest photo-metric loss. As shown in TABLE 2, this strategy significantly boosts performance. Moreover, when combined with edge sampling, it leads to the best results (at the cost of running slower). Another significant

limitation of our method is that it requires a set of posed images for pre-training the NeRF.

7 Conclusion

In this work we present a fast and robust pose estimation method that follows the NeRF-based analysis-by-synthesis paradigm. We significantly improve performance and speed by means of reparameterizing the pose into a continuous and decoupled formulation, making our model more robust to changes in scale by explicitly conditioning it on different resolutions during training, and focusing our ray batches on regions of the image that are structurally relevant to the pose. Our proposed design choices allow our method to run faster than prior work on the same hardware, whilst being more accurate. Another important advantage of our method is that our choice of backbone allows us to reduce the total duration of the pipeline - which includes training the method on the scene prior to running the pose estimation - to minutes instead of hours. Nevertheless, there is still potential for improvement: our edge-guided sampling, for instance, is outperformed by random sampling on smooth objects such as Mic and Ship. One interesting future direction could therefore be to make the sampling procedure adaptive to the nature of the scene to further increase the accuracy.

References

[1] Eric Angel. *A Survey of Approximation Results for Local Search Algorithms*, page 30–73. Springer-Verlag, Berlin, Heidelberg, 2006. ISBN 3540322124.

[2] Jonathan T. Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P. Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *ICCV*, pages 5855–5864, October 2021.

[3] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. SURF: speeded up robust features. In Ales Leonardis, Horst Bischof, and Axel Pinz, editors, *Computer Vision - ECCV 2006, 9th European Conference on Computer Vision, Graz, Austria, May 7-13, 2006, Proceedings, Part I*, volume 3951 of *Lecture Notes in Computer Science*, pages 404–417. Springer, 2006. doi: 10.1007/11744023_32.

[4] Hayet Belghit, Abdelkader Bellarbi, Nadia Zenati, and Samir Otmane. Vision-based pose estimation for augmented reality: Comparison study. In *3rd IEEE International Conference on Pattern Analysis and Intelligent Systems (PAIS 2018)*. IEEE, 2018.

[5] Mark Boss, Raphael Braun, Varun Jampani, Jonathan T Barron, Ce Liu, and Hendrik Lensch. Nerd: Neural reflectance decomposition from image collections. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12684–12694, 2021.

[6] John Canny. A computational approach to edge detection. *IEEE Transactions on pattern analysis and machine intelligence*, (6):679–698, 1986.

[7] Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu, and Hao Su. Tensorf: Tensorial radiance fields. In *European Conference on Computer Vision (ECCV)*, 2022.

[8] Alvaro Collet and Siddhartha S. Srinivasa. Efficient multi-view object recognition and full pose estimation. In *IEEE International Conference on Robotics and Automation, ICRA 2010, Anchorage, Alaska, USA, 3-7 May 2010*, pages 2050–2055. IEEE, 2010. doi: 10.1109/ROBOT.2010.5509615.

[9] Alvaro Collet, Manuel Martínez, and Siddhartha S. Srinivasa. The MOPED framework: Object recognition and pose estimation for manipulation. *Int. J. Robotics Res.*, 30(10): 1284–1306, 2011. doi: 10.1177/0278364911401765.

[10] Yan Di, Fabian Manhardt, Gu Wang, Xiangyang Ji, Nassir Navab, and Federico Tombari. So-pose: Exploiting self-occlusion for direct 6d pose estimation. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 12376–12385. IEEE, 2021. doi: 10.1109/ICCV48922.2021.01217.

[11] Tom Drummond and Roberto Cipolla. Real-time visual tracking of complex structures. *IEEE Transactions on pattern analysis and machine intelligence*, 24(7):932–946, 2002.

[12] Thach-Thao Duong, Duc Pham, Abdul Sattar, and M A Hakim Newton. Weight-enhanced diversification in stochastic local search for satisfiability. pages 524–530, 08 2013.

[13] Peter Hedman, Pratul P. Srinivasan, Ben Mildenhall, Jonathan T. Barron, and Paul Debevec. Baking neural radiance fields for real-time view synthesis. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct 2021. doi: 10.1109/iccv48922.2021.00582.

[14] Tomas Hodan, Frank Michel, Eric Brachmann, Wadim Kehl, Anders GlentBuch, Dirk Kraft, Bertram Drost, Joel Vidal, Stephan Ihrke, Xenophon Zabulis, et al. Bop: Benchmark for 6d object pose estimation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 19–34, 2018.

[15] Siyuan Huang, Siyuan Qi, Yinxue Xiao, Yixin Zhu, Ying Nian Wu, and Song-Chun Zhu. Cooperative holistic scene understanding: Unifying 3d object, layout, and camera pose estimation. *CoRR*, abs/1810.13049, 2018.

[16] Yoonwoo Jeong, Seokjun Ahn, Christopher Choy, Anima Anandkumar, Minsu Cho, and Jaesik Park. Self-calibrating neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5846–5854, 2021.

[17] Wadim Kehl, Fabian Manhardt, Federico Tombari, Slobodan Ilic, and Nassir Navab. SSD-6D: making rgb-based 3d detection and 6d pose estimation great again. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 1530–1538. IEEE Computer Society, 2017. doi: 10.1109/ICCV.2017.169.

[18] Abhijit Kundu, Yin Li, and James M. Rehg. 3d-rcnn: Instance-level 3d object reconstruction via render-and-compare. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 3559–3568. Computer Vision Foundation / IEEE Computer Society, 2018. doi: 10.1109/CVPR.2018.00375.

[19] Abhijit Kundu, Yin Li, and James M Rehg. 3d-rcnn: Instance-level 3d object reconstruction via render-and-compare. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3559–3568. IEEE Computer Society, 2018.

[20] Jae Yong Lee, Yuqun Wu, Chuhang Zou, Shenlong Wang, and Derek Hoiem. Qff: Quantized fourier features for neural field representations. *arXiv e-prints*, pages arXiv–2212, 2022.

[21] Yi Li, Gu Wang, Xiangyang Ji, Yu Xiang, and Dieter Fox. Deepim: Deep iterative matching for 6d pose estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[22] Chen-Hsuan Lin, Wei-Chiu Ma, Antonio Torralba, and Simon Lucey. Barf: Bundle-adjusting neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5741–5751, October 2021.

[23] David 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*, pages 1150–1157. IEEE Computer Society, 1999. doi: 10.1109/ICCV.1999.790410.

[24] Fabian Manhardt, Diego Martin Arroyo, Christian Rupprecht, Benjamin Busam, Tolga Birdal, Nassir Navab, and Federico Tombari. Explaining the ambiguity of object detection and 6d pose from visual data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6841–6850, 2019.

[25] Fabian Manhardt, Wadim Kehl, and Adrien Gaidon. ROI-10D: monocular lifting of 2d detection to 6d pose and metric shape. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 2069–2078. Computer Vision Foundation / IEEE, 2019. doi: 10.1109/CVPR.2019.00217.

[26] Fabian Manhardt, Gu Wang, Benjamin Busam, Manuel Nickel, Sven Meier, Luca Minciullo, Xiangyang Ji, and Nassir Navab. Cps++: Improving class-level 6d pose and shape estimation from monocular images with self-supervised learning. *arXiv preprint arXiv:2003.05848*, 2020.

[27] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, pages 405–421. Springer, 2020. URL https://www.matthewtancik.com/nerf.

[28] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Trans. Graph.*, 41(4): 102:1–102:15, July 2022. doi: 10.1145/3528223.3530127.

[29] Kiru Park, Timothy Patten, and Markus Vincze. Pix2pose: Pixel-wise coordinate regression of objects for 6d pose estimation. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 7667–7676. IEEE, 2019. doi: 10.1109/ICCV.2019.00776.

[30] Sida Peng, Yuan Liu, Qixing Huang, Xiaowei Zhou, and Hujun Bao. Pvnet: Pixel-wise voting network for 6dof pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019*, pages 4561–4570. Computer Vision Foundation / IEEE, 2019. doi: 10.1109/CVPR.2019.00469.

[31] Mahdi Rad and Vincent Lepetit. BB8: A scalable, accurate, robust to partial occlusion method for predicting the 3d poses of challenging objects without using depth. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 3848–3856. IEEE Computer Society, 2017. doi: 10.1109/ICCV.2017.413.

[32] Ivan Shugurov, Sergey Zakharov, and Slobodan Ilic. Dpodv2: Dense correspondence-based 6 dof pose estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[33] Martin Sundermeyer, Zoltan-Csaba Marton, Maximilian Durner, Manuel Brucker, and Rudolph Triebel. Implicit 3d orientation learning for 6d object detection from rgb images. In *Proceedings of the european conference on computer vision (ECCV)*, pages 699–715, 2018.

[34] Prune Truong, Marie-Julie Rakotosaona, Fabian Manhardt, and Federico Tombari. Sparf: Neural radiance fields from sparse and noisy poses. *arXiv e-prints*, pages arXiv–2211, 2022.

[35] Dor Verbin, Peter Hedman, Ben Mildenhall, Todd Zickler, Jonathan T Barron, and Pratul P Srinivasan. Ref-nerf: Structured view-dependent appearance for neural radiance fields. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5481–5490. IEEE, 2022.

[36] Gu Wang, Fabian Manhardt, Jianzhun Shao, Xiangyang Ji, Nassir Navab, and Federico Tombari. Self6d: Self-supervised monocular 6d object pose estimation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 108–125. Springer, 2020.

[37] Gu Wang, Fabian Manhardt, Federico Tombari, and Xiangyang Ji. Gdr-net: Geometry-guided direct regression network for monocular 6d object pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 16611–16621. Computer Vision Foundation / IEEE, 2021.

[38] Gu Wang, Fabian Manhardt, Federico Tombari, and Xiangyang Ji. Gdr-net: Geometry-guided direct regression network for monocular 6d object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16611–16621, 2021.

[39] Pengyuan Wang, Fabian Manhardt, Luca Minciullo, Lorenzo Garattoni, Sven Meier, Nassir Navab, and Benjamin Busam. Demograsp: Few-shot learning for robotic grasping with human demonstration. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5733–5740. IEEE, 2021.

[40] Zirui Wang, Shangzhe Wu, Weidi Xie, Min Chen, and Victor Adrian Prisacariu. Nerf-: Neural radiance fields without known camera parameters. *CoRR*, abs/2102.07064, 2021. URL https://arxiv.org/abs/2102.07064.

[41] Yu Xiang, Tanner Schmidt, Venkatraman Narayanan, and Dieter Fox. Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. In Hadas Kress-Gazit, Siddhartha S. Srinivasa, Tom Howard, and Nikolay Atanasov, editors, *Robotics: Science and Systems XIV, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA, June 26-30, 2018*, 2018. doi: 10.15607/RSS.2018.XIV.019.

[42] Yiheng Xie, Towaki Takikawa, Shunsuke Saito, Or Litany, Shiqin Yan, Numair Khan, Federico Tombari, James Tompkin, Vincent Sitzmann, and Srinath Sridhar. Neural fields in visual computing and beyond. *Computer Graphics Forum*, 2022. ISSN 1467-8659. doi: 10.1111/cgf.14505.

[43] Lin Yen-Chen, Pete Florence, Jonathan T. Barron, Alberto Rodriguez, Phillip Isola, and Tsung-Yi Lin. inerf: Inverting neural radiance fields for pose estimation. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1323–1330, 2021. doi: 10.1109/IROS51168.2021.9636708.

[44] Sergey Zakharov, Ivan Shugurov, and Slobodan Ilic. DPOD: 6d pose object detector and refiner. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 1941–1950. IEEE, 2019. doi: 10.1109/ICCV.2019.00203.

[45] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5745–5753, 2019.