

Fit-NGP: Fitting Object Models to Neural Graphics Primitives

Marwan Taher, Ignacio Alzugaray and Andrew J. Davison
Dyson Robotics Lab, Imperial College London
{m.taher, i.alzugaray, a.davison}@imperial.ac.uk

Abstract—Accurate 3D object pose estimation is key to enabling many robotic applications that involve challenging object interactions. In this work, we show that the density field created by a state-of-the-art efficient radiance field reconstruction method is suitable for highly accurate and robust pose estimation for objects with known 3D models, even when they are very small and with challenging reflective surfaces. We present a fully automatic object pose estimation system based on a robot arm with a single wrist-mounted camera, which can scan a scene from scratch, detect and estimate the 6-Degrees of Freedom (DoF) poses of multiple objects within a couple of minutes of operation. Small objects such as bolts and nuts are estimated with accuracy on order of 1mm.

I. INTRODUCTION

It remains a significant challenge to enable robots to manipulate objects around them with enough competence to unlock applications such as general domestic robotics, especially when these robots must rely only on their own on-board sensors such as cameras. While simple picking up and dropping can often be achieved via direct image-to-action control policies, more complex manipulation such as precise placing or insertion, can benefit from explicit reasoning about the 3D shape of objects.

While general object shape estimation is an interesting and important problem, in most application scenarios (e.g. office, factory, kitchen or household) a robot will usually be dealing with objects whose type is known in advance. Precise 3D models are often also available in the form of Computer-Aided Design (CAD) provided by the manufacturer, shared by other robots or estimated from the robot’s own past experiences. In this case, 3D scene understanding takes the form of *pose estimation* of known object models.

In this paper, we show for the first time that the recent real-time light field reconstruction method in the form of Instant-NGP [13] is ready to be straightforwardly used as an intermediate representation as part of a fully automatic system for highly accurate 3D object pose estimation in table-top settings for precise manipulation tasks. Our system comprises a robot arm with a single wrist-mounted RGB camera, which makes a rapid scan, reconstructs the scene, and fits object models all within two minutes of operation.

RGB light field reconstruction was recently revitalised by Neural Radiance Field (NeRF) [12], which uses a single neural network optimised via volume rendering to reconstruct scene density and illumination. Instant-NGP (Instant Neural Graphics Primitives) [13] is a development which uses a much more efficient hybrid grid/neural representation than NeRF to achieve efficient optimisation and rendering.

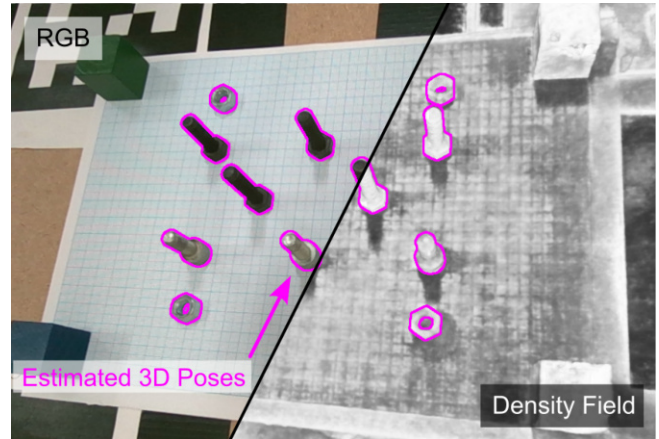


Fig. 1: A set of posed images (top-left) from a scene containing multiple objects is used to train a NeRF. The reconstructed density field (bottom-right) is employed to align model-to-scene poses for each object using multi-hypothesis optimisation, with the best pose silhouette reprojection overlaid (magenta).

Since it was designed primarily for visual fidelity rather than geometry estimation, the reconstructed scene density field is not necessarily accurate nor smooth. However, here we show that it contains details which are sufficiently suitable for object pose estimation, especially when heavily relying on object edges. We perform pose estimation via straightforward iterative optimisation of a cost function measuring the agreement between an object model and the Instant-NGP density field (see Fig. 1). Object poses are initialised automatically using off-the-shelf RGB object detection. Our full system includes automatic scene scanning, with camera pose estimates coming initially from robot arm kinematics and then refined using Instant-NGP’s camera pose optimisation function for increased accuracy. The use of kinematics means that the reconstruction is correctly scaled and object poses can be estimated with high metric precision.

This approach has many advantageous properties for object pose estimation in a robot manipulation setting. Firstly, it requires only a single RGB camera which the robot moves by itself. Alternatives such as depth cameras or stereo rigs typically have minimum and maximum range limits and limited reconstruction precision, as well as usually being more bulky and expensive. A single RGB camera can be moved arbitrarily close to the scene to gather great scene detail. We will show that this means that the pose of even tiny objects such as bolts and nuts can be estimated with an

accuracy of less than 1.5 mm. Further, light field estimation methods can cope with a wide range of lighting conditions and can build density maps which enable pose estimation even of objects which are usually difficult to deal with in computer vision due to issues such as shiny or metallic surfaces. Our method requires only geometric shape models of the objects, without any colour or texture information.

Fit-NGP is simple, accurate, and automatic, and is a method that other researchers will easily be able to copy and use in a wide range of manipulation settings. We present results which demonstrate the accuracy of the method in real robot experiments, and its ability to deal with different objects and lighting conditions.

II. RELATED WORK

6-DoF pose estimation of objects is a long-standing problem in computer vision, and much work has focused on methods which only require one input image, from classical pipelines using RANSAC+PnP [10] up to recent learning-based methods such as PoseCNN [22], DeepIM [11], or MegaPose [9].

Single-view methods are fundamentally sensitive to occlusions, poor lighting conditions or ambiguities. In manipulation-oriented setups where a robot controls the motion of a camera on its end effector, a robot can rapidly capture multiple frames while moving, and use all of the information to aid pose estimation. One option is to pool many single-frame pose estimates via multi-view constraints as in *CosyPose* [8]. Similar multi-view refinement has been taken to the scale of whole scenes in the object-based SLAM literature characterised by works such as *SLAM++* [17].

Alternatively, multiple frames can be used to build an intermediate 3D scene representation before attempting pose estimation directly against the reconstruction, which is the approach we follow in this paper. Related approaches designed for a manipulation setting include *MoreFusion* [21] which used a wrist-mounted depth camera to perform octree-based occupancy reconstruction [4] before fitting object CAD models and performing collision-aware pose refinement to deal with piles of objects. *Scan2CAD* [1] achieved something similar at a room scale using 3D CNNs for alignment.

During manipulation, cameras are close to the scene and the reconstruction accuracy and minimum range of depth cameras can often become a problem. New developments in light field estimation offer new possibilities to use RGB cameras to build a more accurate intermediate representation. NeRF [12] made a breakthrough by showing that a coordinate-based Multi-Layer Perceptron (MLP) can be trained through volume rendering to produce a photo-realistic scene representation from a set of posed RGB views and with no need for prior information, though at the cost of expensive off-line processing. *iMAP* [18] was the first real-time capable scene modelling system based on a NeRF-like MLP, but the requirement for a depth camera and a cut-down network size for speed meant that its reconstruction accuracy was not suitable for object pose estimation.

Recently, hybrid representations have emerged that train much faster than NeRF, and achieve higher view synthesis quality, especially *Instant-NGP*, which uses multi-resolution hash encoding of 3D voxel grids, indexing small MLPs, and converges in tens of seconds for many scenes.

NeRF and *Instant-NGP* were designed for high fidelity view synthesis, not accurate reconstruction, and the density fields they reconstruct are often noisy. It is possible to apply a regularisation prior [14] to improve surface smoothness. Instead, in our work we directly apply the strongest prior available — that the world is made of up objects for which we have known models — and directly fit these models against the raw density reconstruction. Although fuzzy in places, we have found that the reconstructions include accurate details on edges and high texture regions which allow extremely accurate object alignment, even for small objects with reflective surfaces which are very difficult to deal with in most RGB view-based methods. *Instant-NGP* can cope with these challenging issues and allows pose estimation to be purely based on the models and scene geometry.

NeRF is already beginning to be used in some robot systems such as in *Dex-NeRF* [5] and *Evo-NeRF* [6]. There is some work on aligning multiple NeRF reconstructions such as *nerf2nerf* [3], but we are not aware of other work attempting the alignment of object models against them.

III. METHODOLOGY

At its most general, *Fit-NGP* enables pose retrieval of multiple objects in an arbitrary scene given a set of RGB images with approximate camera poses, 3D models of the shape of the objects, and an off-the-shelf image segmentor capable of identifying the objects in one of these images, requiring no additional training data. We demonstrate a system designed for a potential indoor manipulation, where images are captured by an automatic scan from a single wrist-mounted RGB camera on a robot arm, with approximate camera poses coming from the arm’s known kinematics.

The core of our method is first to use the posed images to globally reconstruct the density and radiance fields of the scene. We then use segmentation in a single view to propose and initialise 3D object model pose hypotheses, which are refined by alignment with the density field reconstruction. In this paper we rely on *Instant-NGP* [13] for radiance field and density field reconstruction, as well as refinement of the original camera poses, which is crucial for accurate reconstruction and model fitting. In principle, any existing or future radiance and density reconstruction method could be used instead, though it would need to improve on *Instant-NGP*’s remarkable accuracy and efficiency for that to be worth it. An overview of the method is depicted in Fig. 2.

A. Object Model Representation

We represent object models \mathcal{M} flexibly as a set of surface points with normals $(\mathbf{x}_i, \mathbf{n}_i)$, $\mathbf{x}_i \in \mathcal{S} \subset \mathbb{R}^3$, $\mathbf{n}_i \in \mathcal{N} \subset \mathbb{R}^3$. This representation can be used for object models acquired or designed in different ways. For instance, in Sec. IV, we show the algorithm applied to both human-designed CAD

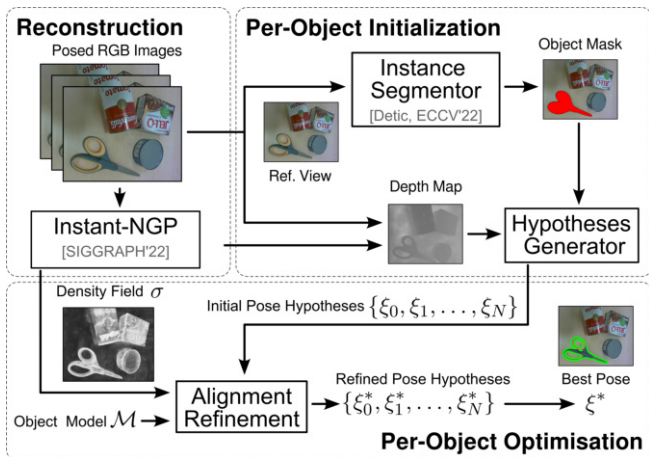


Fig. 2: Overview of the proposed framework: an Instant-NGP reconstruction is obtained from images captured from a robot’s wrist-mounted camera. Objects of interest are segmented from a reference view, and a depth map rendering from the same view is used to initialise a set of per-object pose hypotheses. Each hypothesis is optimised finding the best pose alignment using the Instant-NGP’s density field.

models and models built by 3D reconstruction. Our method tackles object pose alignment in purely geometrical terms, and our object models do not need any appearance information, granting the method robustness to varying lighting conditions. If necessary we pre-process models with high numbers of points and normals by uniformly sampling N_M samples across the surface for efficiency.

B. Density fields from Instant-NGP

We align the object models against a 3D density field $\sigma(\mathbf{x})$, which maps every 3D scene location $\mathbf{x} \in \mathbb{R}^3$ to its density or probability of being occupied. In our current system, this field comes from Instant-NGP, which optimises a density and radiance field against a set of captured images via volumetric rendering. We initialise the camera poses from robot arm kinematics, so no additional camera tracking system such as sparse visual SLAM/structure from motion is needed.

Instant-NGP does not explicitly aim for a high-fidelity density field and thus the quality of the retrieved 3D reconstruction is generally variable. However, we will demonstrate that it is sufficient to achieve millimetre-accuracy object pose estimation without any special post-processing. We only need to capture images and optimise Instant-NGP once per scene, and then multiple objects can be fitted to the same density reconstruction.

C. Multi-hypothesis Object-Pose Optimisation

Each object is allocated a fixed number of hypotheses of potential model-to-scene alignment poses $\mathcal{H} = \{\xi_0, \xi_1, \dots, \xi_{N_H}\}$, $\xi = \{R, \mathbf{p}\} \in \langle \mathbb{S}\mathbb{O}(3) \times \mathbb{R}^3 \rangle$, that map coordinates in the canonical space of the object onto the scene coordinate system as $\xi(\mathbf{x}) = R\mathbf{x} + \mathbf{p}$, $\forall \mathbf{x} \in \mathbb{R}^3$.

1) *Initialization*: To initialize these per-object hypotheses, we consider a specific view among all the posed images to be the reference view of the scene (e.g., the first captured

image, the most top-down view) and apply an off-the-shelf 2D instance segmentor [23] to identify all the relevant objects in the scene for which we have 3D models. We render a depth map from Instant-NGP for this reference view and extract a partial 3D point cloud from the 2D pixel mask of each identified object. We then generate multiple 3D pose hypotheses \mathcal{H} , using the centre of mass of the partial point cloud as the translation of every hypothesis and rotations are equally distributed in $\mathbb{S}\mathbb{O}(3)$. In Sec. IV we show that this simple initialization is sufficient to achieve good model fitting in real-world experiments.

2) *Alignment Refinement*: For each object model \mathcal{M} , we define a set of points \mathcal{X}^S closely distributed around the surface of the model and a set of points distributed outwards along the surface normals \mathcal{X}^N . Formally:

$$\mathcal{X}^S = \{\mathbf{x}^S \mid \mathbf{x}^S = \mathbf{x}_i + \mathbf{n}_i [-\delta^S, \delta^S]\} \quad (1)$$

$$\mathcal{X}^N = \{\mathbf{x}^N \mid \mathbf{x}^N = \mathbf{x}_i + \mathbf{n}_i (\delta^S, \delta^S + \delta^N)\} \quad (2)$$

$\forall \{\mathbf{x}_i, \mathbf{n}_i\} \in \{\mathcal{S}, \mathcal{N}\}$, with $\delta^S, \delta^N \in \mathbb{R}^+$ parametrising the uniform sampling in the defined intervals along the normals.

For a good model-to-scene fit, all the points near the surface of the model \mathcal{X}^S should be projected to high-density regions of the density field σ whereas the points projected outwards along the normals \mathcal{X}^N , away from the surface into free space, should be mapped to low-density regions. We retrieve an occupancy measurement from Instant-NGP’s density field as $s(\mathbf{x}) = 1 - \exp(-\exp(\sigma(\mathbf{x}))\beta)$, $\mathbf{x} \in \mathbb{R}^3$, which resembles the expression for the light transmittance used in differential rendering using a tuneable parameter $\beta \in \mathbb{R}^+$. As Instant-NGP’s density field is often quite irregular (see Fig. 3), δ^S and δ^N help in creating bands of points around the surface or out of it, i.e. \mathcal{X}^S or \mathcal{X}^N , which are largely expected to be occupied or empty, respectively, for the best model-to-scene fit. Formally, we define the model-to-scene fitness function:

$$f(\xi) = \frac{1}{|\mathcal{X}^S|} \sum_{\mathbf{x}^S \in \mathcal{X}^S} s(\xi(\mathbf{x}^S)) - \frac{1}{|\mathcal{X}^N|} \sum_{\mathbf{x}^N \in \mathcal{X}^N} s(\xi(\mathbf{x}^N)), \quad (3)$$

which can be used to convert all the initial pose hypotheses $\{\xi_0, \xi_1, \dots, \xi_{N_H}\}$ into refined ones $\{\xi_0^*, \xi_1^*, \dots, \xi_{N_H}^*\}$ via non-linear optimisation. Among all the refined pose hypotheses, the one that fits the best Eq. (3) is selected as the optimal pose of the object. Depending on the number of hypotheses per object N_H and the complexity of the geometry, it is not uncommon that multiple initial hypotheses are refined to the same final pose. In practice, each hypothesis is independent of the others and thus their refinement is executed in parallel as a batch.

IV. EXPERIMENTAL EVALUATION

We now showcase the capabilities of our method in a challenging robotic experimental setup. We explore a series of self-collected experiments with multiple objects loosely arranged in a table-top configuration (see Fig. 4). Many of the objects used are small, have complex geometries or would require tight tolerances in their manipulation, justifying the

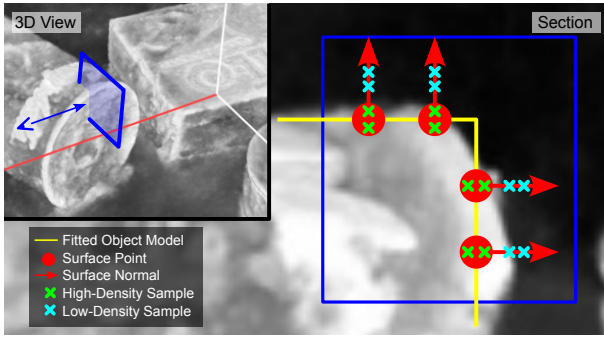


Fig. 3: Section of a reconstructed density field within object. Note that, while the RGB renders from a NeRF can achieve high-fidelity, the underlying density field can be as noisy as shown here, even after NeRF convergence. This noisy density field is used in our fitness function Eq. (3), promoting that points near the surface of the aligned model object \mathcal{X}^S and points along their normal \mathcal{X}^N fall within high-density or low-density region, respectively.

need for highly accurate pose estimation. Additionally, we focus on objects that are visually challenging for other techniques, with textureless or reflective surfaces.

A. Implementation Details and Experimental Setup

In our experiments, the scene is observed using a Franka Emika robot arm equipped with a wrist-mounted Real-sense Sense D455 1280×720 camera (only RGB data is used), interfaced via ROS [16] and hand-eye calibrated as per Tsai *et al.* [20]. For each experiment, we automatically execute a precomputed hemispherical trajectory, capturing up to 78 posed images pointing towards the centre of the scene which are used to train Instant-NGP. While the whole scanning process takes over a minute, it might be potentially too slow for some potential downstream applications. Therefore, in our experiments, we investigate how the performance of the system would be impacted in a less complete albeit faster scanning (see Sec. IV-D). Although poses for the captured views are initially retrieved via robot arm kinematics, we have found that potential inaccuracies (e.g., arm encoder noise, deviations in the hand-eye calibration) are handled by enabling the camera poses refinement in Instant-NGP, yielding higher fidelity density fields. Instant-NGP is trained once per scene and used in frozen form for pose estimation of all of the objects.

Once Instant-NGP has been trained, we employ Detic [23] to create 2D instance object masks for hypothesis initialization as described in Sec. III-C. Each object is allocated $N_{\mathcal{H}} = 216$ hypotheses, refined by optimising Eq. (3) via non-linear optimisation for 200 iterations using Adam [7] in Pytorch [15] and LieTorch [19], with different learning rates $l_{\text{Rot}} = 2.5e-2$ and $l_{\text{Trans}} = 1.0e-3$ for rotation and translation, respectively. In our experiments, we set $\beta = 0.01$. Note that while some experiments show objects arranged on a flat table-top, our system is not specifically tailored for this case, and always estimates full 6-DoF poses.

In our experiments, we consider a mixture of YCB [2] and other common objects such as bolts and washers, for

which 3D reconstructions and CAD designs are available, respectively. Each object model \mathcal{M} is uniformly subsampled to only consider a set of $N_S = 1280$ points along the surface and their normals. While the definition of \mathcal{X}^S and \mathcal{X}^N allows for the generation of many points along each normal according to the intervals defined by δ^S and δ^N , in our experience a single point on the object the surface ($\delta^S = 0$) and a single point in the direction of the normal at a fixed distance ($\delta^N = 5.0e-3$) is sufficient to retrieve accurate poses for simple object geometries, while greatly reducing the computational cost of querying the density field σ . This is further tested in the ablation study of Sec. IV-D.

The whole pipeline is run on a single machine with an NVIDIA Geforce RTX 3090 GPU, i7 Intel CPU and 64GB RAM. A single object is optimised in under 3 seconds; this time includes the concurrent optimisation of all 216 hypotheses. This duration can be significantly reduced by using better initial estimates which can allow us to run with a lower number of hypotheses.

B. Accuracy Evaluation

Quantitative Results. To assess the pose estimation accuracy, we collected our own dataset of 4 different scenes each containing a subset of standardised, industrial-grade low-tolerance objects with accurate and widely available CAD models. The dataset in aggregate is comprised of 43 object instances, containing: M8 nuts, M8x25 bolts, M8x30 hex bolts and M8x25 socket bolts. The small dimensions of these objects render external positioning systems largely inapplicable for highly accurate pose estimates. Hence, we manually aligned them using technical graph paper with a 1 mm grid division to establish our ground truth. We purposely placed the objects so that their relative 3D poses can be accurately retrieved from the graph paper up to 1 mm in translation and up to 4 degrees in rotation, given the geometry of the objects. We compare the estimated relative pose for all object-to-object combinations (considering object symmetries as in [8]) to their ground-truth counterparts. On this dataset, we report a median relative translation error of 1.6 mm and a median relative rotation error of 3.3 degrees. See Fig. 4 (top row) for illustrative examples.

Qualitative Results. We also showcase the capabilities of our system with a diverse set of objects, with scenes composed of YCB objects [2] and other common objects in randomly placed configurations. While these scenes offer a wider range of interesting configurations compared to the aforementioned dataset, millimetre-accurate ground-truth poses for such arrangements cannot be obtained without the use of specialised setups. Thus, we present here a qualitative evaluation instead, as depicted in 4. In these scenes, the proposed system is able to accurately retrieve the pose of each of the objects as evidenced by the object silhouettes reprojected onto the image space from their estimated poses. Note that these scenes include objects that are visually challenging to reconstruct due to, for instance, reflective surfaces (e.g. cans), or small dimensions (e.g. M8 washers), even for Instant-NGP. Additionally, our system is able to



Fig. 4: Self-collected real-world datasets where the poses of all the relevant objects are accurately estimated, as evidenced by the re-projection of the silhouettes of their models perfectly aligning in the image plane.

simultaneously operate on objects with very different scales (e.g. note the size difference of the M8 nuts with respect to the mustard bottle).

Fig. 5 shows a typical failure case of our algorithm. While Instant-NGP density field reconstruction is generally poor, in our experience, it is often sufficiently good for object pose estimation but only if the initial set of pose hypotheses are sufficiently close to the global optima. As our system only implements a simple initialization step based on rendered depth maps, where such a reconstruction is significantly erroneous (e.g. Instant-NGP cannot accurately disambiguate the local geometry), the initial set of poses will be too far from the optima to yield any meaningful pose estimation.

C. Performance over Varying Views

While the quality of the final trained Instant-NGP is dependent on many factors, one of the key elements is the number of views and how well they can characterise the scene (see Fig. 6). For applications targeting extremely high fidelity, a significant amount of effort can be dedicated to collecting many views of the scene with good coverage so that Instant-NGP performs the best. However, downstream tasks employing a robotic perception system are often time-bounded and thus only a limited number of views are expected available at any time. Here we explore how our system is impacted when a dense and lengthy scanning of the scene is not possible (see Fig. 7). Given that our method is limited in its capability to initialise pose hypothesis on degraded density fields (see Fig. 5), in the following results we initialise all the objects using the original set of densely captured views although only a limited subset of them is used in the training of Instant-NGP used later during the pose refinement optimisation. As per the quantitative experiment from Sec. IV-B we report the same median translational and rotational error in the scene for which we have ground-truth object poses.

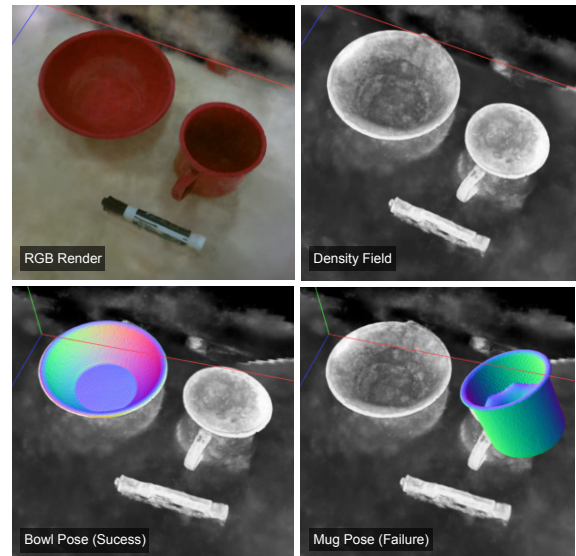


Fig. 5: Example of failure case of the proposed system. Instant-NGP is able to produce photorealistic RGB renderings (top left) even when the quality of the underlying density field is poor (top right). Despite this, object poses can still be retrieved provided that pose hypothesis initialisation is sufficiently close to the optima (bottom left), resulting in failure otherwise (bottom right).



Fig. 6: Density field with varying number of views.

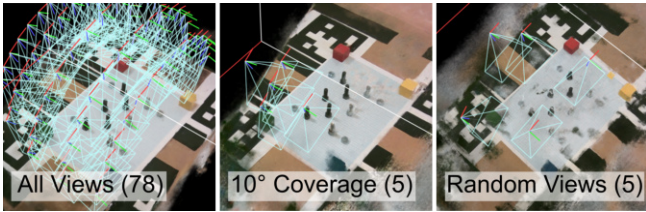


Fig. 7: The number of training views (in parentheses) imposes a trade-off between the quality of Instant-NGP (and our system’s performance) and the data acquisition time. Our experiments explore how the performance degrades when only a small set of close views with limited view coverage of the scene are considered (mid) or a sparse set of randomly sampled views (right).

Sparse Random Views. Here we randomly subsample the set of all the originally captured views that densely observe the the scene. For each of these subsets, we have a degraded yet generally good sparse coverage of the scene. Fig. 8 (top) illustrates how a minimal set of views is required to accurately reconstruct the scene to any meaningful degree so that our algorithm can be applied (around 20 views). Beyond the minimal set, only diminishing returns are achieved to a consistent translational and rotational error of less than 5 mm and 5 degrees, respectively.

Views with Limited Scene Coverage. It is not unusual that robots are not fully capable of observing the scene from an extensive number of viewpoints, for instance, due to limited reach or physical obstacles. Here we explore how the system is impacted by these situations by first randomly selecting a viewpoint looking at the center of the scene. Then only a subset of all the poses that are within a specified angle from the reference viewpoint are considered to train Instant-NGP. In this experiment, the scene coverage is poorer than using a random set of sparse views for the same number of training images, but the perceived parts of the scene are to be better reconstructed. However, this setup also imitates better the expected challenges encountered by a robotic perception system deployed in the wild. The presented results in Fig. 8 (bottom) indicate a strong dependency on this scene coverage factor, which is to be expected due to the fundamental limitations of multi-view geometry estimation.

D. Ablation Study

In this section, we provide an ablation over multiple design choices, reporting in Tab. I the resulting estimation error in the dataset with available ground truth.

We show that the camera pose obtained from the robot arm is high to a degree that renders the density field unusable for fitting objects using our method, producing estimates that more than the object diameters.

The optimisation function Eq. (3) requires both points on the objects’ model surface \mathcal{X}^S and along the normals \mathcal{X}^N to lie in high-density and low-density regions of the density field. While it would be expected that the points on the model’s surface would be sufficient to accurately estimate

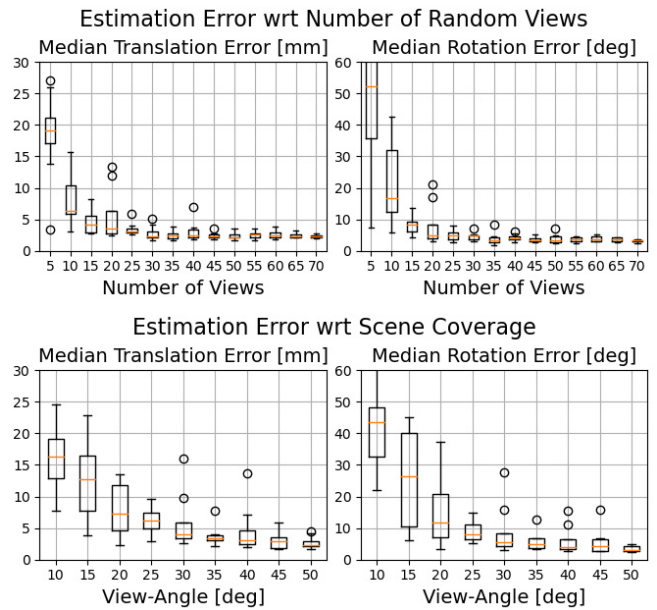


Fig. 8: Impact on pose estimation accuracy by randomly selecting a subset of the training views (top) and varying degrees of scene coverage (bottom). Results here aggregate the information of 8 randomly executed experiments under the same configuration.

| | Trans. Error [mm] | Rot. Error [deg] |
|------------------------------|-------------------|------------------|
| w/o camera pose optimisation | 17.3 | 49.4 |
| w/o samples along the normal | 43.2 | 70.1 |
| All | 1.6 | 3.3 |

TABLE I: Performance for different variants of the pipeline.

object pose, in practice, this is evidenced as insufficient. We believe that using both points on the model surface \mathcal{X}^S and along their normal \mathcal{X}^N alleviates the fact that the resulting Instant-NGP cannot accurately model the object’s geometry but can roughly estimate the transition between high-density and low-density regions.

V. CONCLUSIONS

We have presented a complete system with accurate and robust performance for model-based object pose estimation, suitable for operation in a high-precision manipulation setting where a robot arm is equipped with a single RGB camera. We make use of a state-of-the-art light field reconstruction method integrated with and calibrated against arm kinematics. Fit-NGP can estimate the full 3D pose of even small, metallic objects such as bolts and washers to within millimetre precision. Future work would expand on improving the performance of the system and, in particular, the initialisation of the pose hypotheses when considering only a small number of posed images for which we plan to explore active and data-driven approaches to scene scanning.

VI. ACKNOWLEDGEMENTS

Research presented here has been supported by EPSRC Prosperity Partnerships and Dyson Technology Ltd.

REFERENCES

- [1] Armen Avetisyan, Manuel Dahnert, Angela Dai, Manolis Savva, Angel X Chang, and Matthias Nießner. Scan2CAD: Learning cad model alignment in rgb-d scans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
- [2] Berk Calli, Arjun Singh, Aaron Walsman, Siddhartha Srinivasa, Pieter Abbeel, and Aaron M Dollar. The ycb object and model set: Towards common benchmarks for manipulation research. In *2015 international conference on advanced robotics (ICAR)*, pages 510–517. IEEE, 2015.
- [3] Lily Goli, Daniel Rebain, Sara Sabour, Animesh Garg, and Andrea Tagliasacchi. nerf2nerf: Pairwise registration of neural radiance fields. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9354–9361. IEEE, 2023.
- [4] Armin Hornung, Kai M. Wurm, Maren Bennewitz, Cyrill Stachniss, and Wolfram Burgard. OctoMap: An efficient probabilistic 3D mapping framework based on octrees. *Autonomous Robots*, 2013. Software available at <https://octomap.github.io>.
- [5] Jeffrey Ichnowski*, Yahav Avigal*, Justin Kerr, and Ken Goldberg. Dex-NeRF: Using a neural radiance field to grasp transparent objects. In *Conference on Robot Learning (CoRL)*, 2020.
- [6] Justin Kerr, Letian Fu, Huang Huang, Yahav Avigal, Matthew Tancik, Jeffrey Ichnowski, Angjoo Kanazawa, and Ken Goldberg. Evo-nerf: Evolving nerf for sequential robot grasping of transparent objects. In *6th Annual Conference on Robot Learning*, 2022.
- [7] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- [8] Yann Labbé, Justin Carpentier, Mathieu Aubry, and Josef Sivic. Cosypose: Consistent multi-view multi-object 6d pose estimation. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVII 16*, pages 574–591. Springer, 2020.
- [9] Yann Labbé, Lucas Manuelli, Arsalan Mousavian, Stephen Tyree, Stan Birchfield, Jonathan Tremblay, Justin Carpentier, Mathieu Aubry, Dieter Fox, and Josef Sivic. Megapose: 6d pose estimation of novel objects via render & compare. In *Proceedings of the 6th Conference on Robot Learning (CoRL)*, 2022.
- [10] Vincent Lepetit, Francesc Moreno-Noguer, and Pascal Fua. Ep n p: An accurate o (n) solution to the p n p problem. *International journal of computer vision*, 81:155–166, 2009.
- [11] 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)*, pages 683–698, 2018.
- [12] 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 *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.
- [13] 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.
- [14] M. Oechsle, S. Peng, and A. Geiger. Unisurf: Unifying neural implicit surfaces and radiance fields for multi-view reconstruction. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021.
- [15] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.
- [16] Morgan Quigley, Ken Conley, Brian Gerkey, Josh Faust, Tully Foote, Jeremy Leibs, Rob Wheeler, Andrew Y Ng, et al. Ros: an open-source robot operating system. In *ICRA workshop on open source software*, volume 3, page 5. Kobe, Japan, 2009.
- [17] R. F. Salas-Moreno, R. A. Newcombe, H. Strasdat, P. H. J. Kelly, and A. J. Davison. SLAM++: Simultaneous Localisation and Mapping at the Level of Objects. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.
- [18] E. Sucar, S. Liu, J. Ortiz, and A. J. Davison. iMAP: Implicit mapping and positioning in real-time. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021.
- [19] Zachary Teed and Jia Deng. Tangent space backpropagation for 3d transformation groups. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.
- [20] R.Y. Tsai and R.K. Lenz. A new technique for fully autonomous and efficient 3d robotics hand/eye calibration. *IEEE Transactions on Robotics and Automation*, 5(3):345–358, 1989.
- [21] Kentaro Wada, Edgar Sucar, Stephen James, Daniel Lenton, and Andrew J. Davison. MoreFusion: Multi-object reasoning for 6D pose estimation from volumetric fusion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.
- [22] Yu Xiang, Tanner Schmidt, Venkatraman Narayanan, and Dieter Fox. Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. *arXiv preprint arXiv:1711.00199*, 2017.
- [23] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In *ECCV*, 2022.