

# Neural Fields for Robotic Object Manipulation from a Single Image

Valts Blukis<sup>1</sup>, Taeyeop Lee<sup>1,2</sup>, Jonathan Tremblay<sup>1</sup>, Bowen Wen<sup>1</sup>, In So Kweon<sup>2</sup>,  
 Kuk-Jin Yoon<sup>2</sup>, Dieter Fox<sup>1</sup>, Stan Birchfield<sup>1</sup>  
<sup>1</sup>NVIDIA, <sup>2</sup>KAIST

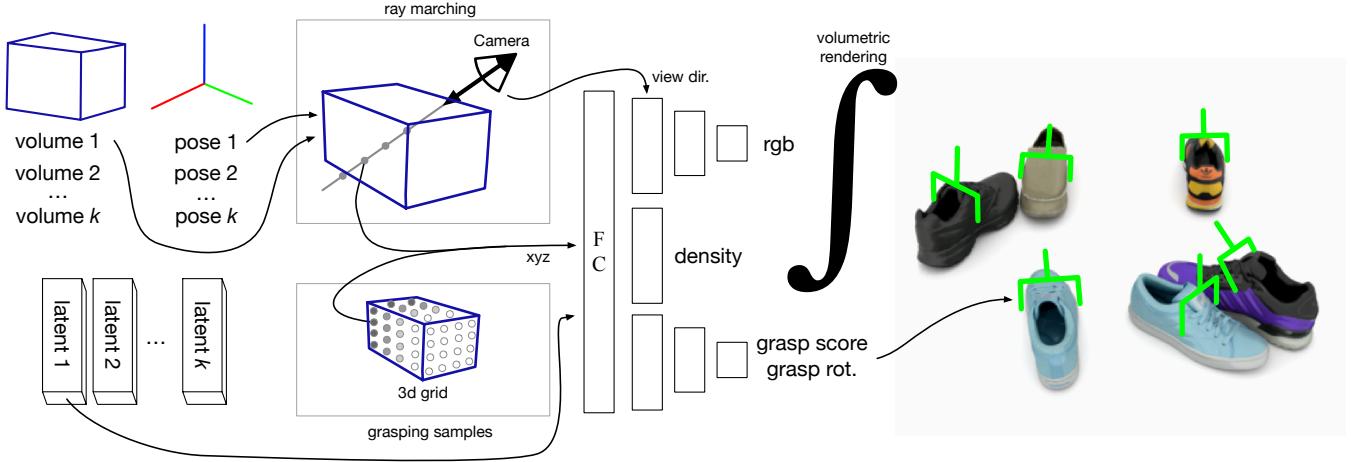


Fig. 1. We present a unified representation that can be used to re-render scenes and generate grasping poses. Given object poses (top left), volumes (top left), and learned latent codes (bottom left), the system renders the scene in novel configurations and annotates each object with a grasping pose (right image). The rendering ray-marches through a set of volumes (cuboids in which the objects reside). In order to generate grasps (shown in green on the right), we use a grid sampling approach that covers each object volume. Grasp orientation and score are predicted using a decoder which is partially shared with the volumetric rendering process. We select the grasp point that yields the highest confidence score. FC refers to a fully connected layer.

**Abstract**—We present a unified and compact representation for object rendering, 3D reconstruction, and grasp pose prediction that can be inferred from a single image within a few seconds. We achieve this by leveraging recent advances in the Neural Radiance Field (NeRF) literature that learn category-level priors and fine-tune on novel objects with minimal data and time. Our insight is that we can learn a compact shape representation and extract meaningful additional information from it, such as grasping poses. We believe this to be the first work to retrieve grasping poses directly from a NeRF-based representation using a single viewpoint (RGB-only), rather than going through a secondary network and/or representation. When compared to prior art, our method is two to three orders of magnitude smaller while achieving comparable performance at view reconstruction and grasping. Accompanying our method, we also propose a new dataset of rendered shoes for training a sim-2-real NeRF method with grasping poses for different widths of grippers.

## I. INTRODUCTION

Robotics perception systems are grounded in 3D understanding; in recent years we have seen state acquisition systems for robotics expand from RGB-only pose estimation methods for known objects [1], [2] to point-clouds used for grasping unknown objects [3], to tactile images [4], to unknown object pose estimation [2], [5]–[7]. These perception systems use various representations such as poses, volumes, depth images, point clouds, *etc.* These methods have different trade-offs in their representation of the world state. For example, a pose does not depict the size of an object, while a depth map does not capture any information in occluded regions. These trade-offs limit the ability to

combine multiple advanced applications in robotics within a single system, such as grasping unknown objects/categories, language grounding, scene editing, planning, robot motion generation, *etc.*

Neural Radiance Field (NeRF) [8] methods have swept through the computer vision and graphics community; they are also beginning to impact state representation in robotics, and are already used successfully for grasping [9]–[12], rendering tactile images [4], learning multi-object dynamics for manipulation [13], robot reinforcement learning [14], camera pose estimation [15], dynamic control [16], planning [17], scene reconstruction [18], and object reconstruction [19]. These methods provide a glimpse into the future of state acquisition for robotics applications but most come with some caveats: they can be slow to train, or need multiple image acquisitions; the final representation is likely to be inflexible (needing retraining if the scene arrangement changes), or the scene representation might be hard to reason about spatially, *e.g.*, hash tables [20].

In this work, we aim to present a Neural Fields system for Robotics Object Manipulation (NiFR). Since the original NeRF [8], researchers have extended the method to few-view novel image synthesis [21], which can be used to reconstruct an unseen scene from a single view, or few views. Methods have also been developed to leverage scene graphs to edit and modify NeRF scenes [22], [23]. We take inspiration from these works.

Figure 1 shows a high-level overview of our proposed system. When the system is deployed on a robot, it takes

as input a **single image** with basic object annotations, *e.g.*, object poses and volumes. Using that information, our system recovers object latent codes from which we can then re-edit the scene (render unseen views of the objects) and retrieve grasping poses for the objects.

We believe our representation is flexible enough to be applied to multiple robotics tasks. In this paper, we focus our effort on object reconstruction and regressing grasp poses from a single RGB observation.

Our model has fewer than  $\sim 70k$  parameters, making it lightweight compared to similar methods [21], [23], [24]. For comparison, AutoRF<sup>1</sup> and pixelNeRF have around 40 million and 28 million parameters, respectively. Combining this small model with an optimized CUDA implementation, we report fast rendering times of  $\sim 20$  ms at a resolution of  $128 \times 128$ , whereas pixelNeRF takes 1.3 seconds and AutoRF could take up to 12 seconds for the same resolution.

We also propose an original dataset of synthetic shoe objects to train our method for sim-2-real deployment on a real-world robotics system. We evaluate our method quantitatively on both grasp quality and reconstruction quality, and we compare it against strong baselines. Finally, we present qualitative results on a real robot experiment, showing that our system can be used to accomplish full grasping when integrated with modern computer vision systems.

## II. RELATED WORK

Recently neural radiance fields (NeRFs) [8] have emerged as a popular scene representation in computer vision and graphics. NeRF is a function parameterized by a neural network that maps a coordinate and direction vector in Euclidean space to two scalars: color and density. A NeRF can be queried on coordinates along rays to render an image of the scene from any viewpoint. These approaches are promising due to their ability to represent the geometry and appearance of a scene using a fixed parameter budget, although training can be time consuming and data hungry. Recent works have explored different approaches to accelerate NeRF training: instant-NGP uses a hash grid and CUDA acceleration [20], while Plenoxels uses a voxel based method to accelerate both training and rendering [25].

PixelNeRF is one of the first methods to address few-shot novel view synthesis. It trains a large category-level model on multiple scenes [21]. At test-time the method consumes a small number of images and camera poses and learns a grid representation that can be used to render the scene from new perspectives. EG3D [26] builds on this to use generative-adversarial training, and represents objects with latent codes that can be fitted from a single image by pivotal tuning inversion. 3D representations from few images have also been explored for producing textured meshes of zebras and birds [27].

We took inspiration from object focused NeRF methods, such as CodeNeRF [24], and AutoRF [23]. Both methods represent a radiance field with multiple per-object latent

codes. CodeNeRF requires tens of images at test-time, while AutoRF uses pre-trained models and can function from fewer test-time images. Like AutoRF, our proposed method also represents a complex scene by assigning each object a latent code. Whereas we find these codes through optimization, AutoRF focuses on an image encoder approach. Other impressive methods have explored editing of object-centric NeRFs, such as [22].

NeRF methods have also been applied to robotics, in reinforcement learning [14], for robot planning [13], [17], [28], [29], and to find depth maps from implicit functions for both thin objects [10] and transparent objects [12].

Robots have also been used by the graphics and computer vision community to generate interesting motion problems to solve [16], [30]. While impressive, many of these methods suffer from difficulty in generalising to new scenes, tedious capturing procedures, long training time, and/or long rendering time. We believe that our proposed method alleviates these concerns. Recently, Evo-NeRF [9] presented an exciting new method that can quickly edit an already-trained NeRF representation to take into consideration objects that have been grasped. They also presented a method that is trained on NeRF rendered depth images to produce grasp poses. In contrast, our method generates grasps directly from object latent codes, avoiding the need for separately trained downstream models.

Several other inspiring methods have also been proposed to retrieve grasping points. GraspNet [3] processes a point cloud using an auto-encoder to generate possible object grasps. Jiang *et al.* [31] used point clouds to generate implicit functions from which they optimize grasp points. Our proposed method for finding grasp poses through a grid search is similar to the concurrent work of Shridhar *et al.* [32], in which gripper poses are sampled on a voxel grid to find the correct next action. Category-level neural representations have been explored to allow robots to re-execute demonstrations in different configurations [33], [34].

## III. MODEL

### A. Scene Representation

We represent each of the  $M$  objects in the scene  $S$  with a tuple  $o = (P_o, V_o, f_o)$ , where  $P_o$  is a 6D object pose,  $V_o$  is a 3D object bounding volume, and  $f_o$  is a learned object latent code.  $f_o$  is intended to encode the shape, appearance, and grasp poses of the object.  $V_o$  is a simple volume such as a sphere or an oriented box to support fast ray-volume intersection checks. The objects form a flat scene graph with a single world node, similar to [23], [35]. This affords direct control over individual object poses to support visualizing or reasoning about the scene under object rearrangement. It also fits naturally with the computer vision systems which we leverage for the detection and tracking of objects.

The per-object latent code can be fed into neural network based decoders for downstream tasks. The two specific tasks that we study in this paper are volumetric rendering and predicting 6-DOF grasps for robot manipulation. For volumetric rendering, we use a neural radiance field decoder  $\Psi$ . To

<sup>1</sup>Thanks to Norman Müller for providing additional details.

predict grasps, we use a grasp decoder  $\Phi$ . The two decoders share the parameters of their first two layers.

### B. Volumetric Rendering with the Radiance Field Decoder

Our volumetric rendering approach consists of four steps: raytracing, raymarching, radiance field query, and ray integration. Raytracing uses the scene-graph, object volumes and poses to compute which rays may intersect with which objects. Raymarching samples coordinates along each ray where the radiance field decoder  $\Psi$  will be queried. These radiance field queries compute density and color at each marched coordinate along the ray. Finally the ray integral computes the ray color on each ray as imaged by the camera.

We make design choices in these components suitable for learning manipulation-friendly representations. Notably, the query coordinate  $\vec{p}$  and ray direction  $\vec{\omega}$  input to  $\Psi$  are expressed in the object-relative reference frame  $P_o$  for each object. This enables rearrangement of objects by changing their poses. We use a sinusoidal positional encoding [8] to encode  $\vec{p}$  and  $\vec{\omega}$ . The positional encoding of  $\vec{p}$  is concatenated with object latent code  $\mathbf{f}_o$ , before feeding into a fully-connected neural network. The network consists of a two fully-connected backbone layers, followed by separate decoder heads for the color and density  $\sigma$  outputs. The density  $\sigma$  does not depend on the viewing direction.

### C. Predicting Grasps with the Grasp Decoder

We use the grasp decoder  $\Phi$  to predict stable grasps on an object  $o = (P_o, V_o, \mathbf{f}_o)$  that we wish to manipulate. The grasp prediction consists of a grasp proposal stage and a filtering stage. The proposal is done by a neural network  $\Phi$  that directly consumes the scene representation, and the filtering relies on outputs of the radiance field described in the previous section. This highlights the flexibility of our representation.

1) *Grasp Decoder*: The grasp decoder  $\Phi$  takes as input a 3D coordinate  $\vec{p}$  and a latent code  $\mathbf{f}_o$  representing an object  $o$  in the scene. It predicts a grasp score  $S^g$  and a rotation matrix  $R^g$ . As for rendering, the input coordinate  $\vec{p}$  is in the object-relative reference frame  $P_o$ . The score  $S^g$  models the probability that there exists a stable grasp at coordinate  $\vec{p}$ .  $R^g$  is a 3x3 rotation matrix reflecting the grasp gripper orientation. Inspired by [36], to guarantee that  $R^g \in \mathbb{SO}(3)$  is an orthogonal matrix, or in other words, it models a pure rotation, the neural network outputs two 3-dimensional vectors  $\mathbf{a}$  and  $\mathbf{b}$ . We assemble the rotation matrix,  $R^g = [\mathbf{b} \quad \mathbf{a} \times \mathbf{b} \quad \mathbf{a}]$ , where  $\mathbf{b} = (\mathbf{a} \times \hat{\mathbf{b}}) \times \mathbf{a}$ ,  $\mathbf{a}$  is the gripper approach vector, and  $\mathbf{b}$  is a vector that points to gripper-left.

2) *Grasp Inference and Filtering*: To predict grasps on object  $o$ , we generate a set of coordinates  $\langle p_i^g \rangle_i$  on a dense 3D grid of fixed resolution  $res^3$  per axis within the volume  $V_o$ . We query  $\Phi$  to obtain  $res^3$  grasp proposals, each with a score  $S_i^g$ , rotation  $R_i^g$ , and the input position  $p_i^g$  of which we retain the top-K proposals with the highest score.

We perform a filtering stage to reject erroneous grasp proposals: We keep only grasps where the open gripper

mesh does not collide with the object, while the closed gripper mesh does. We compute object-gripper collisions by querying the radiance field decoder  $\Psi$ . We represent each gripper mesh as a pointcloud of 1000 randomly sampled points on the surface of the mesh. We transform the pointclouds according to the grasp pose  $(p_i^g, R_i^g)$ , and for each transformed point query the radiance field decoder for the density  $\sigma$ . We keep only grasps where the total density across the 1000 points is below a threshold  $T_{open}$  for the open gripper, and above a threshold  $T_{closed}$  for the closed gripper. This filtering process is made possible because our object representation captures the complete object shape, rather than a view-dependent surface pointcloud.

### D. Learning

1) *Pre-training*: We train our model on a dataset  $\mathcal{D}$  of  $N_{scenes}$  scenes. Each scene contains a ground truth scene graph (with object volumes  $V$  and poses  $P$ ),  $N_{views}$  ground truth images rendered from random camera poses, and a set of grasp annotations consisting of grasp position  $p^g$ , orientation  $R^g$ , and score  $\hat{S}^g$ . Each scene has  $N_{grasps}$  grasp annotations of high and low scores. We randomly initialize a separate latent  $\mathbf{f}$  for each object. The decoders  $\Phi$  and  $\Psi$  are shared across all scenes and objects. Unlike traditional NeRF, this results in learning a *universal* decoder capable of decoding any object within the training distribution / category.

We jointly optimize the NeRF reconstruction and grasp losses. The reconstruction loss  $\mathcal{L}_{rgb}$  is the pixel-wise mean squared error between the reconstructed and ground truth images. The grasp loss consists of two terms: grasp score loss  $\mathcal{L}_{gscore}$  and rotation loss  $\mathcal{L}_{grot}$ . Given the predicted grasp score and rotation  $S^g, R^g = \Phi(p^g, \mathbf{f}_o)$ , the grasp score loss is:

$$\mathcal{L}_{gscore} = \lambda \times (\text{ReLU}(S^g - \hat{S}^g))^2 + \text{ReLU}(\hat{S}^g - S^g)^2$$

where  $\lambda$  is a hyperparameter  $<< 1$ . There can exist multiple grasp annotations at position  $p^g$ , some with high scores and some with low. We want our network to predict high grasp scores if *any* of the grasps at  $p^g$  are stable, which we achieve using the weight  $\lambda$ . The grasp rotation loss is the mean squared error between the predicted and ground-truth rotation, weighed by the grasp score,  $\mathcal{L}_{grot} = \hat{S}^g(R^g - \hat{R}^g)^2$ . This loss encourages the predicted rotation at position  $p^g$  to be similar to high-scoring annotations, ignoring low-scoring ones. During pre-training, we optimize  $\mathcal{L} = \mathcal{L}_{rgb} + \mathcal{L}_{gscore} + \mathcal{L}_{grot}$  over the dataset  $\mathcal{D}$  using the RMSprop optimizer. We assume that a dataset of multiview images for a large scale number of objects are available, similar to [37], the Section IV-B describes in greater detail our proposed dataset for evaluating our method.

2) *Fine-tuning*: Given a universal decoder pre-trained as described above, when a previously unseen object is encountered, we can find a new latent code that represents this object, similar to GAN inversion [26]. We initialize a random latent code  $\mathbf{f}$ , and optimize it using gradient descent, keeping the decoders  $\Phi$  and  $\Psi$  frozen. Note that we do not

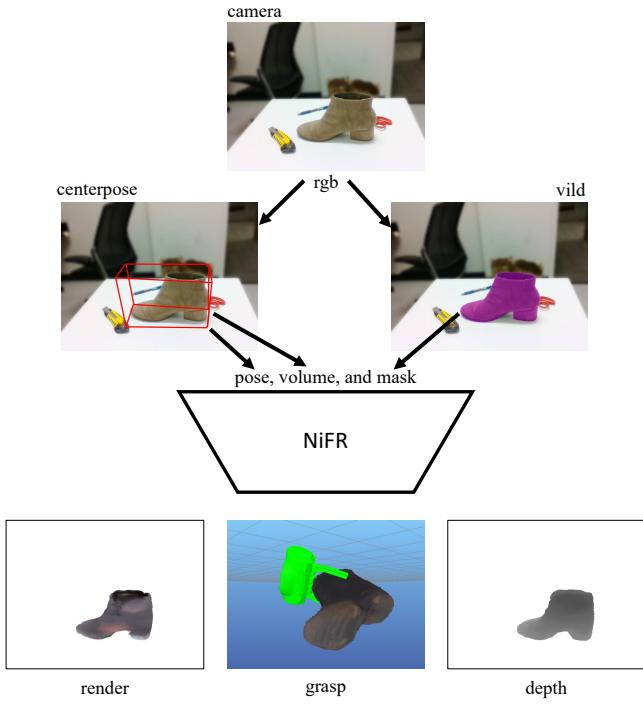


Fig. 2. Our system first consumes an image calibrated with respect to the robot. We retrieve the object pose and dimensions from the image using CenterPose [5]. In order to able to reconstruct the shoe properly, we leverage a masking algorithm (ViLD) [38]. Our system then takes as input the masked RGB image, the object pose, and volume. We then perform inverse rendering to optimize a suitable latent code for the object. The method can reconstruct the object at different poses, generate a valid 3D reconstruction, and find grasping poses (shown for Baxter's gripper).

assume any grasp annotations at test-time. The latent code  $\mathbf{f}$  is recovered only by optimizing the image reconstruction loss  $\mathcal{L}_{rgb}$  over a *single* image. Figure 2 gives an overview of a real-world robotics system using all the components in this section.

#### IV. EXPERIMENTAL RESULTS

##### A. Implementation Details

The original NeRF [8] can be slow to render and train. To mitigate these problems, we leverage the newly available WISP [39] framework. This lightweight library is built on top of PyTorch with appropriate CUDA acceleration. Although the performance is not yet comparable to NGP [20], it achieves greater flexibility, allowing (for example) easy implementation of the grasp related network outputs. We also implemented our own CUDA ray tracing acceleration with respect to multiple volumes to ensure that we can achieve interactive frame rates, *e.g.*, 24 frames per second. The model pre-training was done using eight NVIDIA v100 GPUs in  $\sim 12$  hours. The robotics experiments were done using an NVIDIA Titan Xp with a RealSense camera. The learned object latent codes are 32-dimensional, and the MLP Decoder has two hidden layers with dimension 128 in the shared backbone, and an additional 2 layers for each of the decoding heads.

##### B. Dataset - Shoes

In order to evaluate our proposed method in a robotics grasping application, we created a suitable training dataset. This dataset comprises images of a variety of scenes of multiple objects within a class - in our case, shoes - arranged without self intersections and viewed from many angles. The closest pre-existing dataset is composed of cars [37]. We selected 251 different 3D models of shoes from the google scanned dataset [40], from which we removed duplicates, and ensured they all shared the same reference orientation. We used the ray-traced renderer NVISII [41] to render  $800 \times 800$  pixel images with 2000 samples per pixel. For each shoe in the dataset, we rendered 400 random views sampled from a 360-view sphere at a fixed distance from the origin, see Figure 5 for examples. We refer to this dataset as **centered-shoe**. Accompanying these renders, we also generated 32 scenes where three to five shoes fell onto a table. For each of these scenes, we randomly selected 400 views from a hemisphere at a fixed distance from the origin of the scene, see Figure 4. We refer to this dataset as **falling-shoes**.

For each shoe within our dataset, we also include grasp poses for a 6 and 8 cm wide gripper, with quality scores. Inspired by [42], we sample grasp positions uniformly over the training object point cloud with approaches that are collision-free and achievable from the outside-in. In order to identify the grasp quality score  $S^g \in [0, 1]$  for each grasp  $g$ , we perform  $N$  trials ( $N = 50$  in our case) and count the number of successes in Pybullet [43]. Success is defined as an object staying in the gripper hand after being perturbed by random forces (under 5 mm, 5° magnitude) to its grasp pose  $\xi^g = (\hat{p}^g, \hat{R}^g)$  during physical simulation with the PyBullet library [43]. Therefore, the grasp score can be computed as:

$$S^g = \frac{\sum_{i=1}^N \Gamma(\Delta\xi_i \cdot \xi^g)}{N}$$

where  $\Gamma(\cdot)$  represents the physical simulation binary outcome about grasp success.  $\Delta\xi_i$  is a random grasp pose perturbation in each trial. Intuitively, a robust grasp pose should be consistently stable around its neighborhood in  $SE(3)$ . This has been shown to be effective for obtaining a continuous grasp quality score [42].

##### C. Novel View Synthesis

We focus our evaluation on few view synthesis. We pre-train a model on 235 shoes of the centered-shoe dataset, and we test specifically on 16 shoes, (see Figure 3). The scenarios we want to focus on are single-view and 3-view novel view synthesis. In greater detail, a pre-trained model is allowed to see either a single view or 3 views of the object, and then it is asked to re-render that object in different camera poses, see Figure 5 for a single view example. With respect to the falling-shoe dataset, we explore the same scenarios, made more challenging due to the partially occluded views of some shoes. We also propose a few-view scene editing task. Given a model trained on the centered-shoe dataset, the task is to render synthetic scenes with novel arrangements of specific shoes. The shoes chosen to be rendered are identified



Fig. 3. The 16 withheld shoes during training were selected for the interesting testing properties.



Fig. 4. Our method was trained on a single view (different than shown), with 100 epochs (less than 1 minute). The left column shows the ground truth, the middle is optimized from the scene data, and right is a scene editing using centered-shoe dataset to retrieve the object.

from single RGB images. First, we optimize a latent code for each chosen shoe from the provided single view image. Then, given a scene graph of poses, we render views of this new scene from any viewpoint. We quantitatively test our model, pre-trained on the centered-shoe dataset, using the scenes from the falling-shoe dataset.

For evaluation we use the standard NeRF image quality metrics [8]; Peak Signal to Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM) [44], and Learned Perceptual Image Patch Similarity (LPIPS) [45]. Regardless of the scenario, we evaluate the methods on 75 randomly selected views for each scene.

#### D. Results on synthetic data of shoes

In our final proposed method, we leverage latent code and decoder optimization to find the best fitting NeRF model for our scene. In these experiments, we are interested in understanding this design choice; as such, we design 3 different optimizations, 1) only the latent code is optimized, 2) only the decoder is optimized, whereas a random latent code is sampled, 3) both latent and decoder are optimized.

**Centered-shoe** - Table I shows different RGB-based reconstruction metrics after optimizing for 100 epochs (taking about 1 minute). We report the best optimization performance taken from 15 randomly selected viewpoints (which are shared across all experiments). We compare to PixelNeRF (pnerf) [21] as a baseline, using the same paradigm and

images for training. We make use of their released source code. Comparing the methods, pnerf has over 28 million parameters, and needs 1.25 seconds to render a 128x128 image. On the other hand, when doing novel view synthesis, our method only needs 40k parameters<sup>2</sup> and can render the same 128x128 frame in 20 milliseconds.

Even though our proposed method is multiple orders of magnitudes smaller than pnerf, we achieve similar results. Optimizing both the latent code and decoder allows for better reconstruction performance of the system than only optimizing the latent code. Figure 5 shows the selected single training view for testing a shoe (Reebok SH PRIME COURT MID) and its reconstruction counterpart, as well as a synthesised novel view. With the same experimental setting but with 3 training views and optimized for 100 epochs, our method achieves a PSNR of 19.89, 25.11, and 25.16, for the latent, decoder, and both, respectively.

**Falling-Shoe and Scene Editing** - We also evaluate our system in a complex environment where three to five test shoes are randomly placed on a tabletop. For this experiment we only looked at single image optimization. Figure 4 shows some qualitative results. Similar to the previous experiment, optimizing both the decoder and latent code shows greatest performance, with a PSNR of 19.84, 20.41, and 20.78 for the latent, decoder, and both, respectively. These results show that category-level scene reconstruction still needs to be explored as PSNR around 30 is more desirable. For the task of scene editing, Figure 4 shows quantitative results. We observe a PSNR of 20.94, which is better than doing single view optimization, although on the other metrics, it performed worse, e.g., 0.114 vs. 0.105 (lower is better) for LPIPS metric.

gt	p.nerf[21]	latent	decoder	both

Fig. 5. The first row shows the training view, while the second row is synthesized novel views. The second-row PSNR results are 29.18, 19.63, 27.58, and 27.64 (from left to right).

<sup>2</sup>This does not include the grasping head, which adds 30k parameters

TABLE I  
CENTER-SHOE DATASET RESULTS

	PSNR $\uparrow$				SSIM $\uparrow$				LPIPS $\downarrow$			
	pnerf[21]	latent	model	both	pnerf[21]	latent	model	both	pnerf[21]	latent	model	both
color	24.44	21.00	23.64	23.66	0.903	0.887	0.902	0.903	0.125	0.118	0.100	0.100
shape	21.35	17.88	20.50	20.43	0.883	0.840	0.875	0.875	0.140	0.147	0.121	0.122
shape and color	25.22	20.70	24.58	24.63	0.912	0.901	0.916	0.916	0.117	0.104	0.081	0.081
challenge	22.91	19.59	21.61	21.59	0.904	0.879	0.894	0.895	0.124	0.129	0.109	0.109
mean	<b>23.48</b>	19.79	22.58	22.58	<b>0.900</b>	0.877	0.897	0.897	0.126	0.125	<b>0.103</b>	<b>0.103</b>

### E. Grasping and Robotics Experiments

Following the same split for the synthetic shoe dataset, we report the grasping success rate. For each testing image we fine-tune our method and predict the 5 most confident grasps. As baseline we use GraspNet [3], [46] to generate grasps from the partial view point cloud. We evaluate grasp success using the PyBullet physics simulator [43]. We report a success rate for top-1 grasp prediction of 70.2% for GraspNet vs. 61.6% for our proposed method. While using top-5 decreases the performance of GraspNet to 69.7%, our method improves to 64.2%. While our performance is slightly lower than GraspNet, we achieve useful results with only 70k trainable parameters, compared to GraspNet’s 4 million.

We also tested our system on a real-world robotics system. The final deployed system on the robot is described in detail on Figure 2. We used CenterPose [5] to find the object poses  $P_o$  and volumes  $V_o$  from a single RGB image. We compute an instance segmentation mask of the object using the open-vocabulary object detector ViLD [38] to mask out rays that do not intersect with any object. We use the fine-tuning procedure from Section III-D.2 to optimize the latent code  $f_o$  for each object  $o$  and obtain the complete scene representation  $S$ . The whole process from capturing the RGB image to producing grasps takes less than 10 seconds, but has ample room for improvement using further CUDA-acceleration [20]. Using this system our Baxter robot was able to successfully grasp multiple diverse shoes from a single RGB image, including sneakers (shown here), Crocs, and heel boots with no more than two grasping attempts.



### V. CONCLUSION

In this paper we presented a unified NeRF-based representation that can be used to render objects and predict grasping poses. We demonstrated that our system works on a real robot and conducted quantitative comparisons to strong baseline methods. We leave as future work exploring other benefits of the presented compact scene representation, such as natural language driven tasks and motion planning.

### VI. ACKNOWLEDGEMENT

We would like to recognize the immense help of Alex Evans, Bala Sundaralingam, Yen-Chen Lin, Arsalan Mousavian, Stephen Tyree, and Towaki Takikawa.

### REFERENCES

- [1] J. Tremblay, T. To, B. Sundaralingam, Y. Xiang, D. Fox, and S. Birchfield, “Deep object pose estimation for semantic robotic grasping of household objects,” in *Conference on Robot Learning (CoRL)*, 2018, pp. 306–316.
- [2] Y. Labb  , 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 *6th Annual Conference on Robot Learning (CoRL)*, 2022. [Online]. Available: <https://openreview.net/forum?id=1zbWQxFIU->
- [3] A. Mousavian, C. Eppner, and D. Fox, “6-DOF GraspNet: Variational grasp generation for object manipulation,” in *Proceedings of the International Conference on Computer Vision (ICCV)*, 2019.
- [4] S. Zhong, A. Albini, O. P. Jones, P. Maiolino, and I. Posner, “Touching a neRF: Leveraging neural radiance fields for tactile sensory data generation,” in *6th Annual Conference on Robot Learning (CoRL)*, 2022. [Online]. Available: <https://openreview.net/forum?id=No3mbanRIZJ>
- [5] Y. Lin, J. Tremblay, S. Tyree, P. A. Vela, and S. Birchfield, “Single-stage keypoint-based category-level object pose estimation from an RGB image,” in *IEEE International Conference on Robotics and Automation (ICRA)*, 2022.
- [6] 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.
- [7] B. Wen and K. E. Bekris, “BundleTrack: 6D pose tracking for novel objects without instance or category-level 3D models,” in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2021.
- [8] 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 *European conference on computer vision*. Springer, 2020, pp. 405–421.
- [9] J. Kerr, L. Fu, H. Huang, J. Ichnowski, M. Tancik, Y. Avigal, A. Kanazawa, and K. Goldberg, “Evo-neRF: Evolving neRF for sequential robot grasping,” in *6th Annual Conference on Robot Learning (CoRL)*, 2022. [Online]. Available: <https://openreview.net/forum?id=Bxr45keYrf>
- [10] L. Yen-Chen, P. Florence, J. T. Barron, T.-Y. Lin, A. Rodriguez, and P. Isola, “NeRF-Supervision: Learning dense object descriptors from neural radiance fields,” in *IEEE Conference on Robotics and Automation (ICRA)*, 2022.
- [11] Y.-C. Lin, P. Florence, A. Zeng, J. T. Barron, Y. Du, W.-C. Ma, A. Simeonov, A. R. Garcia, and P. Isola, “MIRA: Mental imagery for robotic affordances,” in *6th Annual Conference on Robot Learning*, 2022. [Online]. Available: <https://openreview.net/forum?id=AmPeAFzU3a4>
- [12] J. Ichnowski\*, Y. Avigal\*, J. Kerr, and K. Goldberg, “Dex-NeRF: Using a neural radiance field to grasp transparent objects,” in *Conference on Robot Learning (CoRL)*, 2020.
- [13] D. Driess, Z. Huang, Y. Li, R. Tedrake, and M. Toussaint, “Learning multi-object dynamics with compositional neural radiance fields,” in *6th Annual Conference on Robot Learning*, 2022. [Online]. Available: <https://openreview.net/forum?id=qUvTmyGpnm7>
- [14] D. Driess, I. Schubert, P. Florence, Y. Li, and M. Toussaint, “Reinforcement learning with neural radiance fields,” *arXiv preprint arXiv:2206.01634*, 2022.
- [15] 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 *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2021.

- [16] Y. Li, S. Li, V. Sitzmann, P. Agrawal, and A. Torralba, “3D neural scene representations for visuomotor control,” in *CoRL*, 2021.
- [17] M. Adamkiewicz, T. Chen, A. Caccavale, R. Gardner, P. Culbertson, J. Bohg, and M. Schwager, “Vision-only robot navigation in a neural radiance world,” *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 4606–4613, 2022.
- [18] E. Sucar, S. Liu, J. Ortiz, and A. Davison, “iMAP: Implicit mapping and positioning in real-time,” in *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021.
- [19] J. Abou-Chakra, F. Dayoub, and N. Sünderhauf, “Implicit object mapping with noisy data,” *arXiv preprint arXiv:2204.10516*, 2022.
- [20] 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. [Online]. Available: <https://doi.org/10.1145/3528223.3530127>
- [21] A. Yu, V. Ye, M. Tancik, and A. Kanazawa, “pixelNeRF: Neural radiance fields from one or few images,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 4578–4587.
- [22] B. Yang, Y. Zhang, Y. Xu, Y. Li, H. Zhou, H. Bao, G. Zhang, and Z. Cui, “Learning object-compositional neural radiance field for editable scene rendering,” in *International Conference on Computer Vision (ICCV)*, October 2021.
- [23] N. Müller, A. Simonelli, L. Porzi, S. R. Buld, M. Nießner, and P. Kortschieder, “AutoRF: Learning 3D object radiance fields from single view observations,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022.
- [24] W. Jang and L. Agapito, “CodeNeRF: Disentangled neural radiance fields for object categories,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 12 949–12 958.
- [25] Fridovich-Keil and Yu, M. Tancik, Q. Chen, B. Recht, and A. Kanazawa, “Plenoxels: Radiance fields without neural networks,” in *CVPR*, 2022.
- [26] E. R. Chan, C. Z. Lin, M. A. Chan, K. Nagano, B. Pan, S. D. Mello, O. Gallo, L. Guibas, J. Tremblay, S. Khamis, T. Karras, and G. Wetzstein, “Efficient geometry-aware 3D generative adversarial networks,” in *CVPR*, 2022.
- [27] X. Li, S. Liu, K. Kim, S. D. Mello, V. Jampani, M.-H. Yang, and J. Kautz, “Self-supervised single-view 3D reconstruction via semantic consistency,” in *ECCV*, 2020.
- [28] M. Pantic, C. Cadena, R. Siegwart, and L. Ott, “Sampling-free obstacle gradients and reactive planning in neural radiance fields (NeRF),” *arXiv preprint arXiv:2205.01389*, 2022.
- [29] H. Turki, D. Ramaman, and M. Satyanarayanan, “Mega-NeRF: Scalable construction of large-scale NeRFs for virtual fly-throughs,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 12 922–12 931.
- [30] A. Noguchi, U. Iqbal, J. Tremblay, T. Harada, and O. Gallo, “Watch it move: Unsupervised discovery of 3D joints for re-posing of articulated objects,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 3677–3687.
- [31] Z. Jiang, Y. Zhu, M. Svetlik, K. Fang, and Y. Zhu, “Synergies between affordance and geometry: 6-DoF grasp detection via implicit representations,” in *Robotics Science and Systems (RSS)*, 2021.
- [32] M. Shridhar, L. Manuelli, and D. Fox, “Perceiver-actor: A multi-task transformer for robotic manipulation,” in *Proceedings of the 6th Conference on Robot Learning (CoRL)*, 2022.
- [33] B. Wen, W. Lian, K. Bekris, and S. Schaal, “You only demonstrate once: Category-level manipulation from single visual demonstration,” *RSS*, 2022.
- [34] A. Simeonov, Y. Du, A. Tagliasacchi, J. B. Tenenbaum, A. Rodriguez, P. Agrawal, and V. Sitzmann, “Neural descriptor fields: SE(3)-equivariant object representations for manipulation,” in *ICRA*, 2022.
- [35] J. Granskog, T. N. Schnabel, F. Rousselle, and J. Novák, “Neural scene graph rendering,” *ACM Transactions on Graphics (TOG)*, vol. 40, no. 4, pp. 1–11, 2021.
- [36] M. Sundermeyer, A. Mousavian, R. Triebel, and D. Fox, “Contact-graspnet: Efficient 6-dof grasp generation in cluttered scenes,” in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 13 438–13 444.
- [37] V. Sitzmann, M. Zollhöfer, and G. Wetzstein, “Scene representation networks: Continuous 3d-structure-aware neural scene representations,” *Advances in Neural Information Processing Systems*, vol. 32, 2019.
- [38] X. Gu, T.-Y. Lin, W. Kuo, and Y. Cui, “Open-vocabulary object detection via vision and language knowledge distillation,” in *International Conference on Learning Representations*, 2022. [Online]. Available: <https://openreview.net/forum?id=L3lnMbR4WU>
- [39] T. Takikawa, O. Perel, C. F. Tsang, C. Loop, J. Litalien, J. Tremblay, S. Fidler, and M. Shugrina, “Kaolin wisp: A pytorch library and engine for neural fields research,” <https://github.com/NVIDIAAGameWorks/kaolin-wisp>, 2022.
- [40] L. Downs, A. Francis, N. Koenig, B. Kinman, R. Hickman, K. Reyman, T. B. McHugh, and V. Vanhoucke, “Google scanned objects: A high-quality dataset of 3D scanned household items,” *arXiv preprint arXiv:2204.11918*, 2022.
- [41] N. Morrical, J. Tremblay, Y. Lin, S. Tyree, S. Birchfield, V. Pascucci, and I. Wald, “NVISII: A scriptable tool for photorealistic image generation,” 2021.
- [42] B. Wen, W. Lian, K. Bekris, and S. Schaal, “Catgrasp: Learning category-level task-relevant grasping in clutter from simulation,” in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 6401–6408.
- [43] E. Coumans and Y. Bai, “PyBullet, a python module for physics simulation for games, robotics and machine learning,” 2016.
- [44] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, “Image quality assessment: from error visibility to structural similarity,” *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.
- [45] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, “The unreasonable effectiveness of deep features as a perceptual metric,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 586–595.
- [46] J. Lundell, “6-dof graspnet pytorch,” 2020.