

NARF24: Estimating Articulated Object Structure for Implicit Rendering

Stanley Lewis¹, Tom Gao¹ and Odest Chadwicke Jenkins¹

Abstract—Articulated objects and their representations pose a difficult problem for robots. These objects require not only representations of geometry and texture, but also of the various connections and joint parameters that make up each articulation. We propose a method that learns a common Neural Radiance Field (NeRF) representation across a small number of collected scenes. This representation is combined with a parts-based image segmentation to produce an implicit-space part localization, from which the connectivity and joint parameters of the articulated object can be estimated, thus enabling configuration-conditioned rendering.

I. INTRODUCTION

Articulated objects pose significant challenges for robots due to their complex degrees of freedom compared to rigid-body objects, complicating tasks like pose estimation and grasp synthesis. The scarcity of suitable datasets exacerbates these challenges. This work introduces NARF24, a parts-based approach leveraging Neural Radiance Fields (NeRFs) to estimate prismatic and revolute joint parameters for non-loopy articulated objects using minimal observed configurations. Our method processes posed RGB images alongside image-space part segmentations. We validate our approach on a real-world robot-collected dataset, and another real-world dataset with sparse segmentation supervision. We show an ablation case for a pipeline component and present results from a serial chain manipulator in a simulated environment.

II. RELATED WORK

Articulated objects remain a difficult class of objects for robots to work with. Explicit methods using a parts-based approach combining 3d mesh models and URDF (Unified Robotics Description Format) files have seen success [1]. However, creating these models is laborious and difficult. Neural Radiance Fields (NeRF) [2] have also shown success in breaking the reliance on mesh models [3]. Previous works such as NARF22 have continued the parts-based approach to produce NeRF representations of articulated objects. However, they still require a-priori knowledge of the articulated object’s structure [3]. We utilize a shared-scene representation for part extraction and localization from segmentation masks. This enables traditional joint parameter estimation methods to produce a URDF model of the object’s structure [4].

¹S. Lewis, Tom Gao, and O.C. Jenkins are with the Robotics Department, University of Michigan, Ann Arbor, MI 48109 {stanlew, zimingga, ocj}@umich.edu

This work is supported in part by Ford Motor Company, in part by J.P. Morgan AI Research, and in part by Amazon.

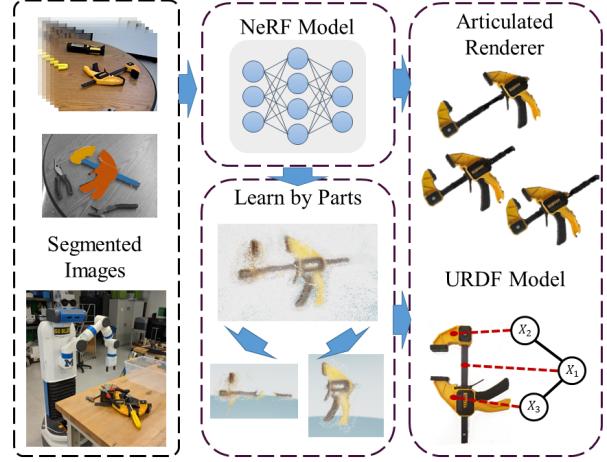


Fig. 1: NARF24 is a pipeline which takes in part-segmented images of an articulated object at a small number of configurations, then utilizes a scene-conditioned neural radiance field to estimate part poses and joint parameters. These create a URDF model and a subsequent articulation enabled NeRF for configurable rendering.

Other works have utilized deeply learned approaches to infer object articulations. URDFFormer [5] utilized simulation assets combined with known URDF models to learn an image-to-URDF transformer model. Weng et al. additionally inferred joint locations and angles from posed RGB input images via part segmentations [6].

III. METHOD

We start by collecting data on an articulated object at different articulation states. Each configuration example is referred to as a ‘scene’. We utilize Nerfstudio [7] to train a NeRF model that additionally contains a per-scene embedding. This embedding allows for scene-conditioned rendering and additionally makes the implicit space distribution more consistent between each scene.

We utilize the segmentation masks to create per-part point clouds within each scene. These clouds are registered to each other using Teaser++ [8] initialized with the results of a point-to-point iterative closest point (ICP) registration.

We utilize the approach in Sturm et al. [4] to estimate the joint type and parameters between each pair of parts using the registration coordinate frames. We estimate joint connectivity and classification using a chamfer distance computed between the part point clouds, and the expected point



Fig. 2: NARF24’s output when trained on the clamp in the ProgressTools dataset. **Left:** the original dataset image. **Middle:** The NARF24 output at the original pose and configuration values. **Right:** The NARF24 output at a counterfactual configuration (fully closed).



Fig. 3: NARF24’s output when trained on the clamp with only 5 percent segmentation labeling. **Left:** The ground truth image. **Center, Right** Two counterfactual renderings of the clamp at different configurations, overlaid on the greyscale original image for context.

clouds based on the predicted joint transforms. These joint predictions can then be used within the NARF22 framework [3] to perform configurable re-rendering of the object.

IV. RESULTS

A. Real World Datasets

1) *Progress-Tools Dataset*: For an initial experiment, we adopt the Progress-Tools Dataset from Pavlasek et al. [1], which contains robot-collected data on a handful of articulated tools, along with the ground truth poses and part segmentations. NARF24 is successfully able to extract the articulation estimates for the clamp, as shown in figure 2, which shows the learned articulated rendering at both the ground truth configuration, as well as at a counterfactual configuration as compared to the ground truth.

2) *Sparse Segments Dataset*: The most manually intensive, least scalable portion of NARF24’s inputs to obtain are the part segmentations. To test how well our approach performs when only a small number of these segmentations are provided, a dataset for the clamp was collected, but only 5% of the images were labelled with segmentation masks. Images without masks were used to train the shared-scene representation, but were not utilized in the parts separation or subsequent registration steps. Even with the lower segmentation coverage, figure 3 shows that acceptable articulation estimates were obtained.

3) *Registration Ablation*: The part registration step is the most sensitive part of the NARF24 pipeline with respect to estimating object structure. Poor per-part localization leads to inaccurate joint estimations, which hinders any downstream



Fig. 4: Qualitative ablation study on training a single-part NeRF subsequent to the part registration step. Top is the ground truth part, and bottom is the NeRF rendering.

Left (ICP only): Performs adequately.

Center (Teaser++ only): fails for single-part NeRF training.

Right (ICP & Teaser++): Produces the best output.

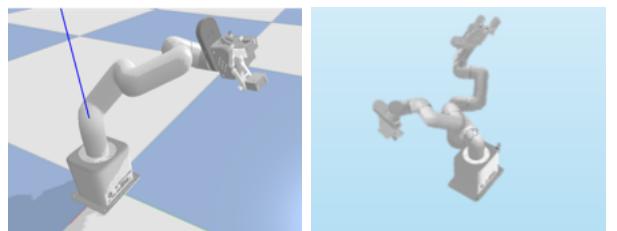


Fig. 5: **Left:** Example output from the simulator environment. **Right:** Renderings of the arm at two different configurations of every joint, after training on the generated sim data (overlaid at base part)

parts-based training process. We thus performed an ablation study to show that our ICP initialized Teaser++ method is effective. Figure 4 shows the resulting point clouds after training a single-part NeRF subsequent to estimating cross-scene part registration with ICP only, Teaser++ only, and Teaser++ with ICP initialization.

B. Simulated Articulated Arm

In order to demonstrate the best-case capabilities of the NARF24 system, a simulated dataset was created in PyBullet of a MyCobot 6 Degree of Freedom robot arm. Due to its simulated nature, this dataset has perfect camera poses, part segmentations, and part poses. Figure 5 shows an example rendering from the sim environment, along with a pair of renderings overlaid on top of each other, showing each of the arm’s joints being changed.

V. CONCLUSION

NARF24 adopts a parts-based approach to enable more scalable learning and rendering of NeRF based implicit models for articulated objects. Results show that real-world data can be used to generate configuration-conditioned renderers even with small amounts of segmentation labels, and a simulated data example shows the effectiveness on a complex robot arm.

REFERENCES

- [1] J. Pavlasek, S. Lewis, K. Desingh, and O. C. Jenkins, “Parts-based articulated object localization in clutter using belief propagation,” in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 10595–10602, IEEE, 2020.
- [2] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, “Nerf: Representing scenes as neural radiance fields for view synthesis,” *Communications of the ACM*, vol. 65, no. 1, pp. 99–106, 2021.
- [3] S. Lewis, J. Pavlasek, and O. C. Jenkins, “Narf22: Neural articulated radiance fields for configuration-aware rendering,” in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 770–777, IEEE, 2022.
- [4] J. Sturm, C. Stachniss, and W. Burgard, “A probabilistic framework for learning kinematic models of articulated objects,” *Journal of Artificial Intelligence Research*, vol. 41, pp. 477–526, 2011.
- [5] Q. Chen, M. Memmel, A. Fang, A. Walsman, D. Fox, and A. Gupta, “Urdformer: Constructing interactive realistic scenes from real images via simulation and generative modeling,” in *Towards Generalist Robots: Learning Paradigms for Scalable Skill Acquisition@ CoRL2023*, 2023.
- [6] Y. Weng, B. Wen, J. Tremblay, V. Blukis, D. Fox, L. Guibas, and S. Birchfield, “Neural implicit representation for building digital twins of unknown articulated objects,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3141–3150, 2024.
- [7] M. Tancik, E. Weber, E. Ng, R. Li, B. Yi, J. Kerr, T. Wang, A. Kristoffersen, J. Austin, K. Salahi, *et al.*, “Nerfstudio: A modular framework for neural radiance field development,” *arXiv preprint arXiv:2302.04264*, 2023.
- [8] H. Yang, J. Shi, and L. Carlone, “Teaser: Fast and certifiable point cloud registration,” *IEEE Transactions on Robotics*, vol. 37, no. 2, pp. 314–333, 2020.