CenterArt: Joint Shape Reconstruction and 6-DoF Grasp Estimation of Articulated Objects

Sassan Mokhtar, Eugenio Chisari, Nick Heppert, Abhinav Valada

***Abstract*— Precisely grasping and reconstructing articulated objects is key to enabling general robotic manipulation. In this paper, we propose CenterArt, a novel approach for simultaneous 3D shape reconstruction and 6-DoF grasp estimation of articulated objects. CenterArt takes RGB-D images of the scene as input and first predicts the shape and joint codes through an encoder. The decoder then leverages these codes to reconstruct 3D shapes and estimate 6-DoF grasp poses of the objects. We further develop a mechanism for generating a dataset of 6-DoF grasp ground truth poses for articulated objects. CenterArt is trained on realistic scenes containing multiple articulated objects with randomized designs, textures, lighting conditions, and realistic depths. We perform extensive experiments demonstrating that CenterArt outperforms existing methods in accuracy and robustness.**

I. Introduction

Manipulating articulated objects is crucial for many robotic applications such as household robots [1]–[3]. However, before a robot can manipulate an object, it needs to acquire a grasp on a moveable part. Prior research addresses the 6-DoF grasp pose generation (and articulation parameter estimation [4], [5]) problem for articulated objects through policy learning approaches utilizing reinforcement learning (RL) [6]–[8]. These approaches involve training a reinforcement learning agent to predict valid 6-DoF grasp poses, which are then used to generate trajectories for object manipulation. However, RL-based methods demand significant amounts of data and training time. Furthermore, while they perform well under controlled conditions in simulations, they lack generalization to applications characterized by diverse scenes, illumination conditions, and noisy sensor observations.

Inspired by recent advances, we adopt a vision-based approach and propose CenterArt for simultaneous 3D shape reconstruction and 6-DoF grasp poses estimation of articulated objects. CenterArt is an extension of CenterGrasp [9], a single-shot holistic grasp prediction approach for rigid-body objects. To train CenterArt, we set up two data generation procedures. First, we generate valid 6-DoF grasp poses for articulated objects in an object-centric manner. Second, we use the Sapien simulator [10] to design and render realistic kitchen scenes including multiple articulated objects, leveraging previously generated grasps.

Our primary contributions can be summarized as follows:

- The first approach for simultaneous 3D shape reconstruction and 6-DoF grasp poses estimation of articulated objects.
- A dataset containing valid 6-DoF ground truth grasp poses of articulated objects.
- Photo-realistic kitchen scenes consisting of several articulated objects.

Department of Computer Science, University of Freiburg, Germany
This work was funded by Carl Zeiss Foundation with the ReScaLe project.

II. Related Work

Areas related to this work include center-based object detection, neural implicit representations for articulated objects, and grasp distance functions.

Center-based Object Detection: Inspired by single-stage object detectors such as YOLO [11], Zhou *et al.* introduce CenterNet [12] which represents objects by a single point at their bounding box center, transforming object detection into a key point estimation problem. This approach improves accuracy and enhances predictions of object properties such as 6D pose estimation. CenterSnap [13] employs a center-based object detection method in a holistic manner to predict 6D poses and reconstruct 3D shapes. CenterArt follows a similar approach to CenterSnap, using a point-based representation to detect and represent the complete 3D information (6D pose, 3D shape, and joint state) of articulated objects in the scene.

Neural Implicit Representations for Articulated Objects: Compared to rigid objects, articulated objects have more complex structures, making their tracking [14] and reconstruction challenging. A-SDF (Articulated-SDF) [15] is one of the earliest works addressing this task with neural implicit representation. It represents articulated objects by disentangling codes for encoding shape and joint angle. CARTO [16] follows a holistic approach to detect, localize, and reconstruct articulated objects. Its decoder consists of two sub-decoders: a geometry decoder and a joint decoder. The shape reconstruction part of CenterArt is closely related to CARTO, where an MLP is trained to concatenate the shape and joint code of the objects with sampled points to output the SDF value.

Grasp Distance Functions: Inspired by advances in neural implicit fields, Weng *et al.* [17] introduced Neural Grasp Distance Fields (NGDF), extending the concept of neural implicit distance functions to the domain of grasping tasks. NGDF predicts scalar distance metrics representing valid grasp poses for objects. The distance-based representation offered by NGDF can be interpreted as a cost function, which can be minimized through an optimization process. CenterGrasp [9] proposes the Shape and Grasp Distance Function (SGDF), which is category-independent and handles multiple objects in the scene. We utilize SGDF from CenterGrasp and build upon it to develop a holistic approach for 3D shape reconstruction and 6-DoF grasp pose estimation of articulated objects.

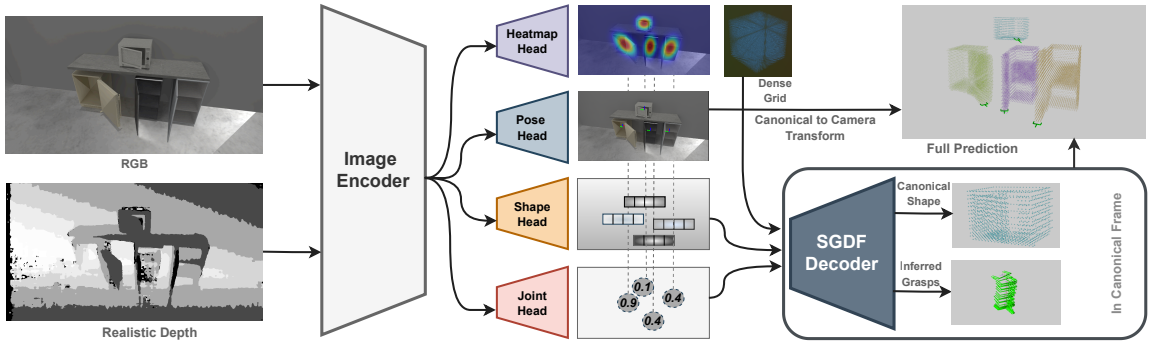

Fig. 1: Overview of CenterArt. First, input RGB-D images are encoded to predict object heatmaps, poses, shape codes, and joint codes in a per-pixel manner. Next, the peaks of heatmaps are used to detect the objects. The SGDF decoder then utilizes the predicted shape code and joint code to output the shape and grasp of detected objects. Finally, the estimated poses are used to transform the predicted 3D shapes and 6-DoF grasps from the canonical frame to the camera frame.

III. Technical Approach

Given an RGB-D image of a scene with multiple articulated objects, the goal is to reconstruct the objects and predict valid grasp poses. CenterArt consists of an image encoder that yields embedding vectors and joint states of each object in the scene, followed by a decoder that reconstructs the 3D shapes and determines valid 6-DoF grasp poses (see Fig. 1).

A. Image Encoder

Network Architecture: Inspired by the architecture of CenterSnap [13], we first pass the RGB-D image separately through a ResNet50 [18] to generate a low-resolution feature representation. Then, we concatenate the feature representations of RGB and depth images and feed them to a ResNet18-FPN backbone [19] to obtain a feature pyramid. Following a similar approach to SimNet [20], we feed the resulting pyramid of features to specialized heads. We utilize the same structure as CenterGrasp [9] for the heatmap, pose, and shape heads, adding a joint head as the fourth head to the encoder. It predicts the joint state of the articulated object for each pixel of the downsampled map. To ensure a consistent representation of joint states, we determine the global maximum joint state and consider normalized joint states as ground truth labels.

Losses: The total loss of the image encoder is given by

$$\mathcal{L}_{\text{encoder}} = w_{\text{heat}}\mathcal{L}_{\text{heat}} + w_{\text{pose}}\mathcal{L}_{\text{pose}} + w_{\text{shape}}\mathcal{L}_{\text{shape}} + w_{\text{joint}}\mathcal{L}_{\text{joint}}.$$

Each loss is calculated using the mean squared error.

Training: The image encoder network is trained for 105 epochs using the ADAM optimizer with a learning rate of 1e$-$3. Additionally, color jitter augmentation is applied to the RGB images.

B. Shape and Grasp Decoder

The decoder aims to reconstruct 3D shapes and predict valid 6D grasp poses. Inspired by CenterGrasp [9], we utilize the shape and grasp distance function (SGDF) decoder to map shape code, joint state, and 3D coordinate to shape and grasp distances.

Network Architecture: Inspired by DeepSDF [21], a deep feedforward multi-layer fully connected network is used for the decoder. The inputs of the network are shape code $\mathbf{z}^s \in \mathbb{R}^{32}$, joint code $z^j \in \mathbb{R}^1$, and a 3D point $\mathbf{x} \in \mathbb{R}^3$. Utilizing an 8-layer perceptron with 512 neurons at each layer, in the first layer shape code and 3D points are fed to the decoder. The joint code appends to the second layer. Moreover, the shape code, joint code, and the 3D point append to the output of the fourth layer. The activation function of hidden layers is ReLU, while the activation function of the last layer is the hyperbolic tangent.

Losses: For the SDF values, the clamp function is used, which constrains its input value. The loss is then defined by the $L1$ loss of the clamped SDF values. To have a uniform loss for the translation and rotation components of the grasp pose loss, we follow [22] to process the target and predicted grasp poses to represent the 6D grasp pose with five 3D points $\mathbf{o}^{gp}, \hat{\mathbf{o}}^{gp} \in \mathbb{R}^{3\times 5}$. Then, the grasp pose loss is simply the $L1$ distance between the target and predicted points. The third component of the loss is designed to regularize the shape codes. The total loss of the decoder is then given by

$$\mathcal{L}_{decoder} = w_{SDF}\mathcal{L}_{SDF} + w_{grasp}\mathcal{L}_{grasp} + w_{code}\mathcal{L}_{code}.$$

Training: Each articulated object is paired with 7 to 10 different joint states. For every object, we sample one joint state and include the corresponding object joint state pair in a validation set. Then, for each object joint state pair, we sample 100,000 points with corresponding SDF values. The grasp distance label is computed for each point by finding the closest ground-truth grasp. The SGDF network is trained for 600 epochs, with ADAM optimizer and step-based decay learning rate scheduler between 1e$-$3 to 25e$-$5. Additionally, dropout with probability 0.2 and weight normalization are applied for regularization.

C. Full CenterArt Inference

Given an RGB-D input, the image encoder produces per-pixel predictions for object heatmap, 6D pose, shape code, and joint code. Each peak in the heatmap is assumed to be the object's center, which is treated as the representative of the object. We then extract the 6D pose, shape code, and joint code corresponding to each object center to input into the SGDF decoder for per-object prediction. For each object,

we concatenate the shape code and joint code to create a latent code specific to the *object-joint state* pair. This latent code, along with sampled 3D coordinates in a dense grid, is fed into the decoder to predict an SDF value and a 6-DoF grasp for each sampled point. We identify object surface points and a set of valid grasp poses by considering iso-surfaces $SGDF(.) = 0$. Finally, the 6D pose of the object is utilized to transform the remaining points and grasps from their canonical frame to the camera frame.

Fig. 2: Generated kitchen scenes

D. Dataset Generation

Object-Centric 6-DoF Grasp Generation: Initially, 82 articulated objects were collected from the PartNet-Mobility dataset [23], covering five different categories: Microwave, Oven, Refrigerator, Dishwasher, and Storage Furniture. After performing a preprocessing step on the collected data, we consider ten different joint states for each object, which is obtained by evenly splitting the distance between the minimum and maximum joint state. Subsequently, for each *object-joint state* pair, a point cloud of the articulated link is generated. Then, the positions of ground-truth grasps are a subset of the generated point cloud.

To reduce the full rotation manifold and thus, speed-up data generation, we utilize the articulation axis and the joint state. We calculate the set of all valid orientations corresponding to three edges and the possible handle of the articulated link of the object. Then, for each point in the point cloud, we sample one orientation among a set of all valid orientations and form a candidate grasp pose. We then evaluate the candidate grasp in PyBullet simulator [24]. If the flying gripper can move the articulated link successfully, then the candidate grasp is regarded as a valid grasp pose.

We generate and store between 100 to 500 grasp poses for each *object-joint state* pair. Finally, all generated grasps were manually verified to exclude *object-joint state* pairs with insufficient grasp labels or where the labels did not cover all areas of interest in the link. Overall, $375,266$ grasp labels were generated for 766 *object-joint state* pairs.

Full Scene Generation: To train the encoder and evaluate the full pipeline, we set up a generation process for realistic kitchen scenes with single or multiple articulated objects. For training, we create about 25,000 random scenes. Each scene is rendered from four random camera poses, resulting in approximately 100,000 RGB-D images and labels. Rendering is done using the Sapien raytracing-based renderer [10], with realistic depth images [25]. Object heatmaps are generated by fitting a Gaussian to the ground truth masks, with the peaks indicating object locations in the image. (see Fig. 2)

IV. Experimental Results

We conducted experiments to evaluate the performance of CenterArt and compare against UMPNet [8], a state-of-the-art baseline for grasp estimation of articulated objects. UMPNet is an RL-based approach that estimates the 6-DoF grasp of articulated objects and predicts manipulation trajectories. Since CenterArt only estimates grasp poses, we used the corresponding part of UMPNet for comparison. UMPNet uses ground truth depth for training. To ensure a fair comparison, noise-free depths were provided to UMPNet. Additionally, UMPNet is trained on simple scenes with a single object and a floor. Thus, we alter our data generation process to exclude walls and multiple objects. For evaluating CenterArt, two scene variations are considered: scenes with a single object in a room, similar to UMPNet scenes, and more complex kitchen scenes with multiple objects.

Scene	Method	GT-depth		Noisy-depth	
		SR	RSR	SR	RSR
Single Obj	UMPNet	0.24	0.53	0.00	0.01
Single Obj	CenterArt	**0.52**	**0.95**	**0.53**	**0.97**
Single Obj	CenterArt + ICP	0.52	0.75	0.51	0.75
Multiple Objs	CenterArt	0.26	**0.72**	**0.51**	**0.94**
Multiple Objs	CenterArt + ICP	**0.33**	0.66	0.29	0.70

TABLE I: Evaluation of 6-DoF grasp pose estimation for unseen objects. SR = Success Rate (↑), RSR = Relaxed Success Rate (↑)

We utilize two metrics for evaluation. The Success Rate (SR) is defined as the successful movement of the articulated joint for at least 10 degrees. If the difference between the maximum joint state and the current joint state is less than 45 degrees, the goal is to close the joint; otherwise, the goal is to open the joint. Another metric is the Relaxed Success Rate (RSR), which determines if the predicted grasp pose is close enough to any ground truth grasps. In this metric, the prediction is regarded as successful if the minimum Euclidean distance between the predicted grasp position and any grasp label is less than 10% of the initial distance. The experiments are performed in the Sapien simulator [10] with a flying gripper.

We present the results in Tab. I. On the single object scenes, CenterArt doubles the success rate over the baseline UMPNet (52% compared to 24%). Additionally, CenterArt consistently performs well on both ground truth and noisy depth images, whereas UMPNet fails with noisy depth images. Even in complex kitchen scenes containing multiple objects, CenterArt still performs better than UMPNet in simple scenes with a single object. Finally, it is worth noting that while refining the predicted poses with ICP contributes to better results in CenterGrasp [9], as shown in Tab. I, it does not result in consistent improvement of CenterArt.

V. Conclusion

In this work, we introduced CenterArt, a vision-based approach that simultaneously performs shape reconstruction and 6-DoF grasp estimation of articulated objects. Additionally, we generated a dataset of valid 6-DoF grasp poses and realistic kitchen scenes with multiple articulated objects. Our experiments demonstrate that CenterArt improves the success rate of state-of-the-art baseline by 28%. Moreover, CenterArt achieves a consistent performance across various scenarios, including those with noisy depth and realistic kitchen scenes, highlighting its robustness in practical settings.

REFERENCES

[1] F. Schmalstieg, D. Honerkamp, T. Welschehold, and A. Valada, "Learning hierarchical interactive multi-object search for mobile manipulation," *IEEE Robotics and Automation Letters*, 2023.

[2] D. Honerkamp, T. Welschehold, and A. Valada, "N2m2: Learning navigation for arbitrary mobile manipulation motions in unseen and dynamic environments," *IEEE Transactions on Robotics*, 2023.

[3] D. Honerkamp, M. Buchner, F. Despinoy, T. Welschehold, and A. Valada, "Language-grounded dynamic scene graphs for interactive object search with mobile manipulation," *arXiv preprint arXiv:2403.08605*, 2024.

[4] A. Röfer, G. Bartels, W. Burgard, A. Valada, and M. Beetz, "Kineverse: A symbolic articulation model framework for model-agnostic mobile manipulation," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 3372–3379, 2022.

[5] R. Buchanan, A. Röfer, J. Moura, A. Valada, and S. Vijayakumar, "Online estimation of articulated objects with factor graphs using vision and proprioceptive sensing," *arXiv preprint arXiv:2309.16343*, 2023.

[6] K. Mo, L. J. Guibas, M. Mukadam, A. Gupta, and S. Tulsiani, "Where2act: From pixels to actions for articulated 3d objects," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6813–6823, 2021.

[7] H. Zhang, B. Eisner, and D. Held, "Flowbot++: Learning generalized articulated objects manipulation via articulation projection," *arXiv preprint arXiv:2306.12893*, 2023.

[8] Z. Xu, Z. He, and S. Song, "Universal manipulation policy network for articulated objects," *IEEE robotics and automation letters*, vol. 7, no. 2, pp. 2447–2454, 2022.

[9] E. Chisari, N. Heppert, T. Welschehold, W. Burgard, and A. Valada, "Centergrasp: Object-aware implicit representation learning for simultaneous shape reconstruction and 6-dof grasp estimation," *arXiv preprint arXiv:2312.08240*, 2023.

[10] F. Xiang, Y. Qin, K. Mo, Y. Xia, H. Zhu, F. Liu, M. Liu, H. Jiang, Y. Yuan, H. Wang, *et al.*, "Sapien: A simulated part-based interactive environment," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11097–11107, 2020.

[11] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 779–788, 2016.

[12] X. Zhou, D. Wang, and P. Krähenbühl, "Objects as points," *arXiv preprint arXiv:1904.07850*, 2019.

[13] M. Z. Irshad, T. Kollar, M. Laskey, K. Stone, and Z. Kira, "Centersnap: Single-shot multi-object 3d shape reconstruction and categorical 6d pose and size estimation," in *2022 International Conference on Robotics and Automation (ICRA)*, pp. 10632–10640, IEEE, 2022.

[14] N. Heppert, T. Migimatsu, B. Yi, C. Chen, and J. Bohg, "Category-independent articulated object tracking with factor graphs," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 3800–3807, IEEE, 2022.

[15] J. Mu, W. Qiu, A. Kortylewski, A. Yuille, N. Vasconcelos, and X. Wang, "A-sdf: Learning disentangled signed distance functions for articulated shape representation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13001–13011, 2021.

[16] N. Heppert, M. Z. Irshad, S. Zakharov, K. Liu, R. A. Ambrus, J. Bohg, A. Valada, and T. Kollar, "Carto: Category and joint agnostic reconstruction of articulated objects," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21201–21210, 2023.

[17] T. Weng, D. Held, F. Meier, and M. Mukadam, "Neural grasp distance fields for robot manipulation," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1814–1821, IEEE, 2023.

[18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[19] A. Kirillov, R. Girshick, K. He, and P. Dollár, "Panoptic feature pyramid networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6399–6408, 2019.

[20] T. Kollar, M. Laskey, K. Stone, B. Thananjeyan, and M. Tjersland, "Simnet: Enabling robust unknown object manipulation from pure synthetic data via stereo," in *Conference on Robot Learning*, pp. 938–948, PMLR, 2022.

[21] J. J. Park, P. Florence, J. Straub, R. Newcombe, and S. Lovegrove, "Deepsdf: Learning continuous signed distance functions for shape representation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 165–174, 2019.

[22] 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)*, pp. 13438–13444, IEEE, 2021.

[23] K. Mo, S. Zhu, A. X. Chang, L. Yi, S. Tripathi, L. J. Guibas, and H. Su, "PartNet: A large-scale benchmark for fine-grained and hierarchical part-level 3D object understanding," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[24] E. Coumans and Y. Bai, "Pybullet, a python module for physics simulation for games, robotics and machine learning," 2016.

[25] X. Zhang, R. Chen, A. Li, F. Xiang, Y. Qin, J. Gu, Z. Ling, M. Liu, P. Zeng, S. Han, *et al.*, "Close the optical sensing domain gap by physics-grounded active stereo sensor simulation," *IEEE Transactions on Robotics*, 2023.