

Bio-Inspired 3D Affordance Understanding from Single Image with Neural Radiance Field for Enhanced Embodied Intelligence

Zirui Guo, Xieyuanli Chen , Zhiqiang Zheng, Huimin Lu  and Ruibin Guo *

College of Intelligence Science and Technology, National University of Defense Technology, Changsha 410073, China; guozirui@nudt.edu.cn (Z.G.); xieyuanli.chen@nudt.edu.cn (X.C.); lhmnew@nudt.edu.cn (H.L.)

* Correspondence: guoruibin08@nudt.edu.cn

Abstract: Affordance understanding means identifying possible operable parts of objects, which is crucial in achieving accurate robotic manipulation. Although homogeneous objects for grasping have various shapes, they always share a similar affordance distribution. Based on this fact, we propose AFF-NeRF to address the problem of affordance generation for homogeneous objects inspired by human cognitive processes. Our method employs deep residual networks to extract the shape and appearance features of various objects, enabling it to adapt to various homogeneous objects. These features are then integrated into our extended neural radiance fields, named AFF-NeRF, to generate 3D affordance models for unseen objects using a single image. Our experimental results demonstrate that our approach outperforms baseline methods in the affordance generation of unseen views on novel objects without additional training. Additionally, more stable grasps can be obtained by employing 3D affordance models generated by our method in the grasp generation algorithm.

Keywords: neural radiance fields; 3D affordance models; robotic manipulation



Academic Editor: Xiaojie Wang

Received: 29 May 2025

Revised: 17 June 2025

Accepted: 18 June 2025

Published: 19 June 2025

Citation: Guo, Z.; Chen, X.; Zheng, Z.; Lu, H.; Guo, R. Bio-Inspired 3D Affordance Understanding from Single Image with Neural Radiance Field for Enhanced Embodied Intelligence. *Biomimetics* **2025**, *10*, 410. <https://doi.org/10.3390/biomimetics10060410>

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Affordance was first proposed by Gibson [1], referring to the ability of the environment to contain information relevant to interactions with the organisms. Since then, the concept of affordance has been widely used in the field of robotics [2,3]. Embodied intelligence is an important research topic in robotics, with extensive studies focusing on achieving stable and accurate grasping and manipulation. For specific objects like screwdrivers, drills, and other tools, considering their structure is crucial for performing appropriate manipulation beyond simple grasping. The different structures of the object are represented by various semantics in the image. Such specific semantics relevant to manipulation can be treated as affordance.

Although machine learning methods [4] have proven valuable in semantics prediction, which is relevant to affordance generation, they may suffer when handling the different structures of the object. Moreover, the algorithm cannot provide a complete 3D model of the object along with the prediction. Generating appropriate 3D affordance understanding for robotic manipulation is still challenging.

To tackle this, neural radiance fields (NeRF) [5] have been introduced as implicit representations that have achieved success for 3D scenario representations of geometry and texture. These methods are trained with posed images and can generate a 3D model of objects. Existing research focuses on enhancing rendering speed and enabling control-

lable synthesis to improve implicit representation. However, the exploration of unified affordance interpretation of homogeneous objects in neural rendering remains limited.

Unlike semantic understanding at the object class level [6], affordance often resides in specific object structures and can vary depending on object types. This requires the neural implicit representation to generate diverse syntheses controllably in response to various object categories. It is extremely difficult to control the synthesis directly since the neural network has extensive parameters. Bio-inspired by the neural mechanisms underlying human affordance understanding, this paper addresses the problem for specific category objects. Cognitive neuropsychology studies reveal that the left inferior parietal lobule plays a key role in processing seen-category objects, where such recognition critically depends on prior knowledge of shape and appearance [7,8]. Motivated by this functional segregation, we design AFF-NeRF with a dedicated shape and appearance feature extraction module, mimicking the specialized processing of the inferior parietal lobule. This module specifically addresses the 3D affordance understanding of objects within distinct categories.

Our main contributions are threefold:

- We design AFF-NeRF with a shape and appearance feature encoder to obtain the 3D affordance models. Our method can produce precise affordance with a single view of novel objects without any retraining or fine-tuning.
- We created a new dataset for affordance rendering validation, comprising different objects with diverse shapes in various poses and affordance annotations.
- The affordance models obtained by our approach can effectively improve the performance of downstream robotic manipulation tasks. We evaluate the performance of various grasp generation algorithms, and the results show that applying the affordance models can generate more stable grasps.

2. Related Work

Controllable synthesis rendering with NeRF: Implicit neural representation-based controllable synthesis generation can be broadly grouped into two aspects. One is to leverage the generative model framework to output synthesis by controlling the latent code [9–14], and the other is working in a supervised learning manner using various input images [15–17]. GRAF [12] is one of the early works to combine generative models with implicit representations, which has become the baseline for the subsequent work. Based on GRAF, GIRAFFE [9] decouples different objects and the background to arrange various latent codes. Ultimately, this method makes it possible to generate various objects with different poses. To be more user-friendly, semantic-driven generation with NeRF has been proposed [14,18]. Synthesis generation based on the generative model is more applicable to artistic creation, data provision, etc., and is not appropriate for generating specific objects. Liu et al. [16] uses a supervised learning method to generate various homogeneous objects. However, this method can only work on the training set and, therefore, does not have the ability to generalize to new objects. Jang et al. [15] propose an auto-decoder architecture with decoupled shape and appearance of latent spaces. For new objects, the method requires additional retraining.

Our approach differs from existing methods, and the comparison is shown in Table 1. We present a novel controllable affordance generation pipeline. Our method expands the representation to include operable affordance and demonstrates its ability to generalize to unseen objects with a single image without any retraining or fine-tuning.

Table 1. A comparison with prior works. Our approach can accomplish affordance understanding in a single view of the new object without additional training, unlike other methods.

	GIRAFFE [9]	EditNeRF [16]	CodeNeRF [15]	AFF-NeRF
Learns scene prior	✓	✓	✓	✓
Affordance outputs				✓
Allows zero-shot	✓		✓	✓
Without retraining	✓			✓

Affordance-based robot manipulation: To achieve autonomous manipulation, researchers have conducted extensive studies in perception-based methods [19–21]. These studies use original point clouds and RGB images as input, focusing on analyzing the geometry of the object, while using original sensor data may struggle with structurally complex objects [22]. To address this issue, Appius et al. [22] segment the structures of objects in the image and input the result of affordance understanding into the grasp generation algorithm to achieve a more stable grasp. With advancements in computer vision technology, obtaining 3D models of objects has become easier, enabling the generation of higher-quality manipulation strategies based on object models [23,24].

3. Methodology

We aim to design a controllable affordance generation pipeline that can handle both known and unseen objects. Figure 1 contains an overview of our method. We exploit deep residual networks to extract the shape and appearance codes and feed them into our AFF-NeRF, which is then trained with affordance labels. Meanwhile, a dataset containing various categories of objects is built to support our training and testing (Section 3.2).

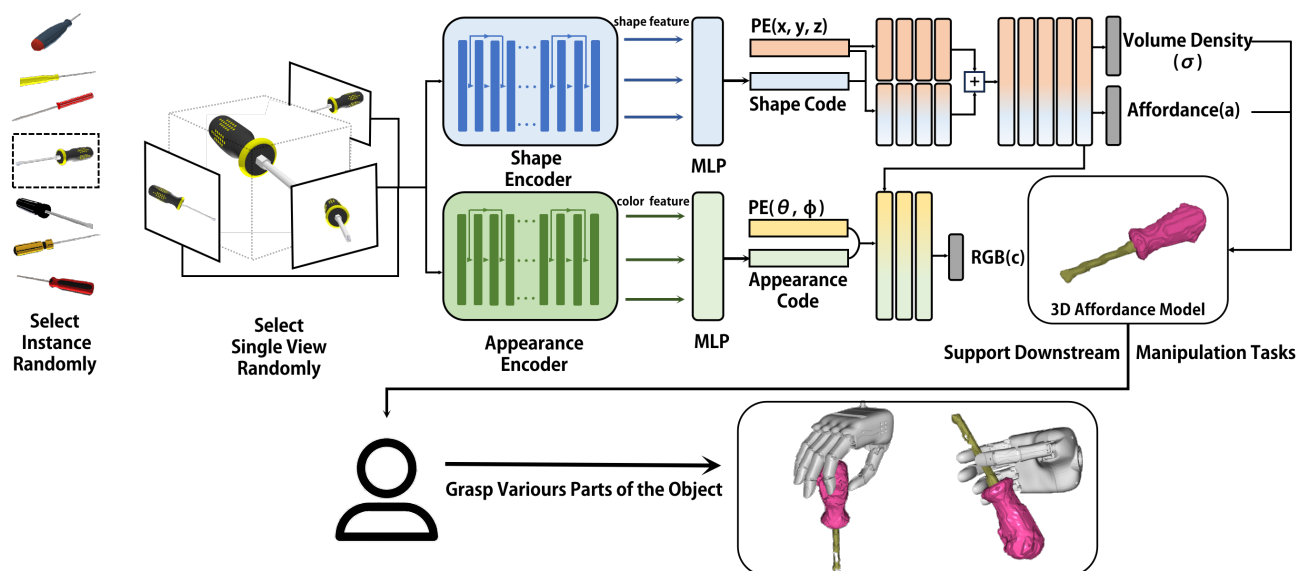


Figure 1. The pipeline of the proposed method. During training, our method first randomly selects a single image of the objects in the training set. Then, the images are fed into shape and appearance feature encoders. The 3D position (x, y, z) and viewing direction (θ, σ) are involved in the AFF-NeRF after position encoding (PE). The shape and appearance code are combined with position and viewing direction separately. Eventually, AFF-NeRF outputs volume density σ , affordance a , and color c . The 3D affordance models obtained from our method can be used in grasp generation algorithms to support downstream manipulation tasks.

3.1. Network Architecture and Training

Given a single image I from n images ($I_1 \sim I_n, n \geq 1$), we extract the shape and appearance feature using two deep residual networks, ResNet18 [25], for each image. The shape and appearance features of various images are then passed to the linear layer for integration to output two vectors, which we denote as shape code C_s and appearance code C_a :

$$C_s = SE(I), C_a = AE(I), \quad (1)$$

where SE and AE are the shape encoder and appearance encoder, as depicted in Figure 2.

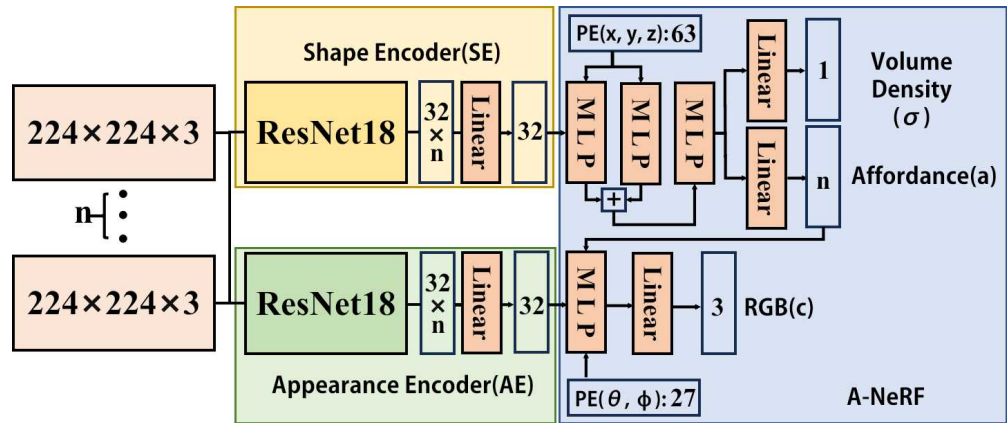


Figure 2. Network architecture. This figure shows a more detailed structure of the shape encoder and appearance encoder.

Our goal is to get implicit representations of homogeneous objects with affordance. Since most current NeRF methods cannot output affordance, we design AFF-NeRF with affordance output for homogeneous objects. Inspired by [6], we let the network output the affordance in combination with the volume density. On the front end of AFF-NeRF, we add a multi-layer perceptron (MLP) that integrates the shape and appearance codes, as shown in Figure 2.

More specifically, the shape and appearance codes are fed into the AFF-NeRF to adapt to various objects. As introduced in previous work [5], the NeRF network F has two stages of output F_{xyz} and $F_{\theta\phi}$: the former is view-independent, which is related to the geometry of the scene, and the latter is view-dependent, which is related to the texture of the scene.

$$F = F_{xyz} \odot F_{\theta\phi}. \quad (2)$$

The shape code C_s serves as the key to control the geometry. Thus, we combine it with (x, y, z) after position encoding to output volume density σ and affordance a , which are strongly correlated with the geometry.

$$F_{xyz} : (PE(x, y, z), C_s) \rightarrow (\sigma, a, q), \quad (3)$$

where q is the intermediation. The dimension of a is determined by the object structure. Since the appearance code C_a is responsible for the texture, we combine it with (θ, ϕ) after position encoding to generate color c .

$$F_{\theta\phi} : (PE(\theta, \phi), C_a, q) \rightarrow (c), \quad (4)$$

The whole procedure can be formulated as:

$$F : (PE(x, y, z), PE(\theta, \phi), C_s, C_a) \rightarrow (\sigma, c, a). \quad (5)$$

We follow the volume rendering method [6] to render the image with color and affordance pixel-by-pixel. Given the pixel and the focal point of the camera, we can obtain a ray: $r(t) = o + td$. Sample K random points $\{t_k\}_{k=1}^K$ on this ray between near and far bounds (t_n and t_f). The estimated color $C(r)$ and affordance $A(r)$ of the pixel can be expressed as:

$$\hat{C}(r) = \sum_{k=1}^K \hat{T}(t_k) \alpha(\sigma(t_k) \delta_k) c(t_k), \quad (6)$$

$$\hat{A}(r) = \sum_{k=1}^K \hat{T}(t_k) \alpha(\sigma(t_k) \delta_k) A(t_k), \quad (7)$$

where $\hat{T}(t_k) = \exp(-\sum_{k'=1}^{k-1} \sigma(t_{k'}) \delta_{k'})$, $\alpha(x) = 1 - \exp(-x)$, and $\delta_k = t_{k+1} - t_k$ is the distance between two adjacent sample points.

We train the network in a coarse-to-fine way, with photometric loss L_p and affordance loss L_a :

$$L_p = \sum_{r \in R} [\|\hat{C}_f(r) - C(r)\|_2^2 + \|\hat{C}_c(r) - C(r)\|_2^2], \quad (8)$$

$$L_a = - \sum_{r \in R} [\sum_{l=1}^n p^l(r) \log \hat{p}_c^l + \sum_{l=1}^n p^l(r) \log \hat{p}_f^l], \quad (9)$$

where R is the rays in the training batch. $C(r)$ and $p^l(r)$ are the ground truth for RGB color and affordance probability at class l of the sampled ray r . $\hat{C}_c(r)$ and $\hat{p}_c^l$ are the predicted color and affordance probability at class l of the coarse network, while $\hat{C}_f(r)$ and $\hat{p}_f^l$ are the output of the fine network. L_p and L_a are chosen separately as mean square error loss and multi-class cross-entropy loss.

The final loss L is a weighted sum of the above two losses:

$$L = L_p + \lambda L_a. \quad (10)$$

During the training process, a single image is utilized to extract shape and appearance features in each epoch. Moreover, the single image used for encoding is selected randomly in the train set, and the image corresponding to the sample ray r is not chosen. This makes the shape and appearance encoding less relevant to the view direction and passively enhances the feature extraction capability of the encoder. This image sampling strategy enables our network to function with limited views.

3.2. Dataset Construction

Existing affordance datasets provide affordance annotations, but the camera poses are missing, and there is a lack of images with multiple perspectives on a specified object [3]. For the above reasons, we have created a new dataset specifically for assessing our proposed AFF-NeRF from two sources. First, we obtain diverse models from open-source resources and then import the models into CoppeliaSim [26]. Color and depth images with poses are randomly captured on the hemisphere by controlling the camera in CoppeliaSim, as shown in Figure 3. After that, the images are annotated with affordance. For other objects, we use a similar method to obtain the same configured dataset from the PartNet [27]. The dataset includes a total of six types of objects, among which are screwdrivers, eyeglasses, buckets, bottles, cabinets, and teapots, and 120 images were captured from random viewpoints of each object. Figure 4 shows some of the images in the dataset.



Figure 3. Color and depth images with poses are randomly captured on the hemisphere by controlling the camera in the simulation environment.

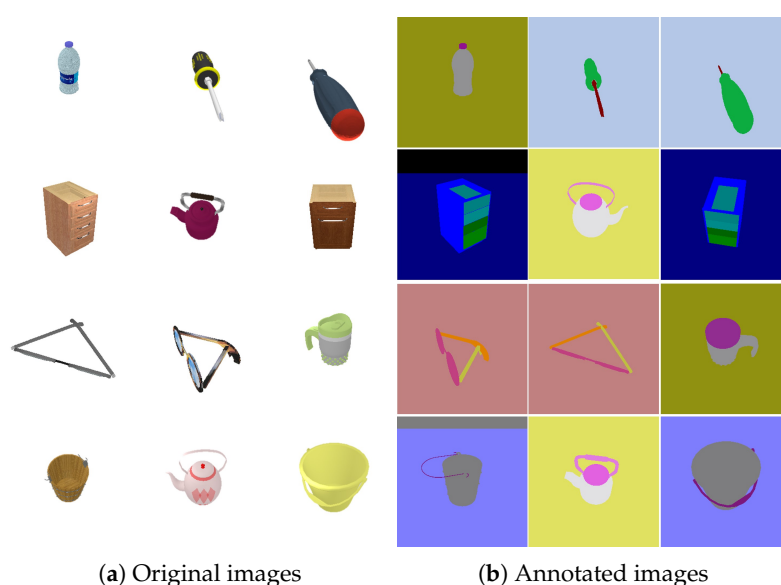


Figure 4. We have established a dataset containing objects of different categories with affordance annotation.

4. Experimental Evaluation

The main focus of this work is an affordance generation pipeline, working with a single input view. We take various category objects with different shapes to evaluate the proposed method. At the same time, we integrated the acquired 3D affordance model into the grasp generation algorithm to investigate whether affordance understanding improves grasp performance.

4.1. Experimental Setup

Implementation Details For the shape and appearance encoder SE and AE , we use the ResNet18 [25] backbone without pre-training. Both the shape and appearance code are 32-dimensional vectors. Following the original NeRF [5], the lengths of position encoding for 3D position and view direction are 10 and 4. For depth information in the dataset, we set the near and far sample boundaries to 0.1 m and 1.5 m. All the images were resized to 400×300 for all experiments. The model was implemented in PyTorch 1.12.1 [28] and trained on a single RTX3090 produced by Nvidia with 24GB memory. The system has an i9-10900K processor produced by Intel and 64 GB of RAM. The λ in the total loss was set to 4×10^{-4} . The batch size of rays was set to 1024, while the learning rate for the whole network was set to 5×10^{-4} .

Comparison Methods We compare our method with GSNeRF [29], Semantic-Ray [30], and modified PixelNeRF [31] on seen categories. The modified PixelNeRF renders new views with affordance using inputs from three views, while GSNeRF and Semantic-Ray are consistent with the original implementation. For unseen categories, we compare with Semantic-NeRF [6], GSNeRF, and Semantic-Ray. By limiting the objects in the training set, Semantic-Ray and GSNeRF can also work on the unseen categories.

Meanwhile, we choose to [32] evaluate the affordance understanding on grasp generation.

Evaluation Metrics Our approach focuses on affordance generation, and affordance refers to semantics in this paper. Therefore, we use the metrics for semantic segmentation to evaluate affordance. We use mean intersection over union (mIoU), mean pixel accuracy (MPA), and pixel accuracy (PA) to evaluate the quality of the rendered images with affordance. The detailed calculation of all metrics can be found in [4].

For grasp generation, we use the ϵ -metric [33] to evaluate, which is widely used to measure the grasp quality.

4.2. Experimental Results

4.2.1. Results of Affordance Understanding

The first experiment confirms that our approach consistently produces accurate affordance for various category objects that are not in training sets with only a single view, without requiring retraining, fine-tuning, or any labels.

Table 2 shows the quantitative results of affordance generation on various category objects. Our approach achieves the highest average mIoU, pixel accuracy, and mean pixel accuracy compared to the baseline methods on the categories of scissors, pens, glasses, and screwdrivers. Our approach with single view input averages 20.6% improvement in mIoU, 1.0% improvement in pixel accuracy, and 25.1% improvement in mean pixel accuracy. The qualitative results are shown in Figures 5 and 6. The results indicate a significant enhancement in mean pixel accuracy and mIoU. These results suggest that the proposed method exhibits a distinct advantage in multi-category affordance understanding. Our method is able to respond to small structures of the cross-category objects, such as pen caps, lenses, etc. In contrast, the baseline methods demonstrate affordance confusion when addressing small structures. We also conduct experiments on the object component in different positions. Figure 7 shows the generated affordance model. The result shows that our method can respond to the object in various states.

Table 2. Performance on seen categories of affordance understanding.

		Scissors	Pens	Glasses	Drivers	Average
mIoU	PixelNeRF_M	0.723	0.688	0.518	0.570	0.625
	SemRay	0.537	0.627	0.563	0.659	0.597
	GSNeRF	0.670	0.576	0.584	0.678	0.627
	Ours	0.784	0.892	0.586	0.709	0.743
PA	PixelNeRF_M	0.965	0.971	0.962	0.932	0.958
	SemRay	0.979	0.986	0.965	0.978	0.977
	GSNeRF	0.982	0.989	0.962	0.982	0.979
	Ours	0.989	0.992	0.967	0.977	0.981
MPA	PixelNeRF_M	0.729	0.403	0.547	0.613	0.573
	SemRay	0.645	0.656	0.682	0.690	0.668
	GSNeRF	0.770	0.759	0.687	0.684	0.725
	Ours	0.940	0.799	0.733	0.777	0.812

Bold denotes the best. PixelNeRF_M denotes the modified PixelNeRF. SemRay denotes Semantic-Ray.

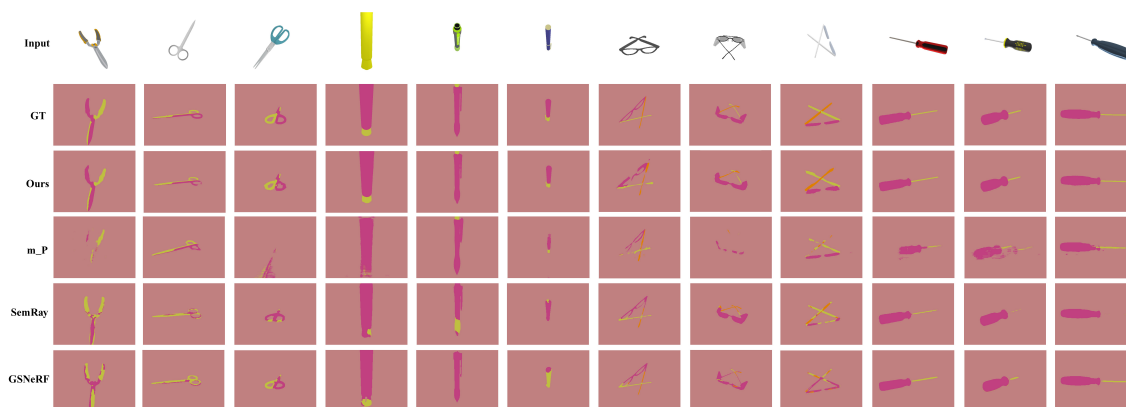


Figure 5. Results on the seen-category objects with single-view input. Our method is capable of generating affordance from novel perspectives of various categories of objects without retraining or fine-tuning. GT denotes the ground truth, m_P denotes the modified PixelNeRF, and SemRay denotes Semantic-Ray. With single image input, our method generates accurate affordance of novel perspectives, whereas the baseline yields affordance confusion.

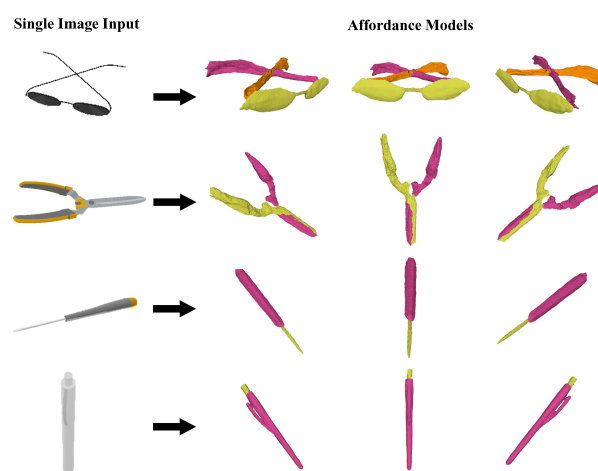


Figure 6. Affordance models generated by AFF-NeRF. Using single image inputs, AFF-NeRF affordance models of various seen-category objects are generated.

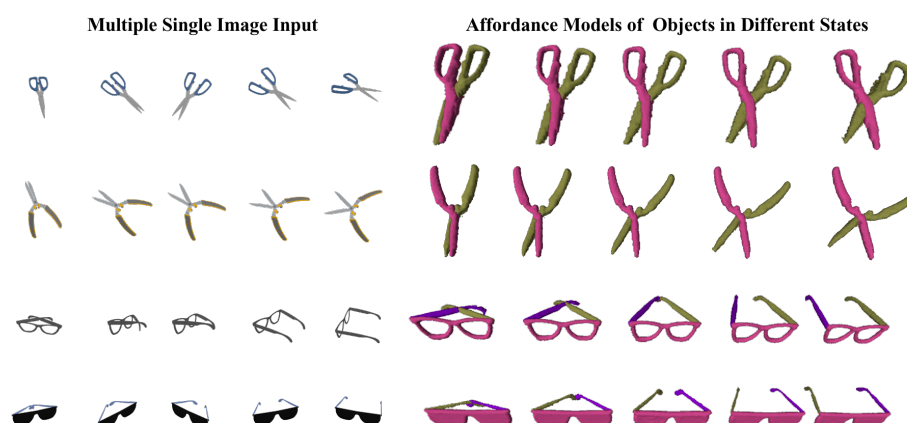


Figure 7. For articulated objects in the dataset, AFF-NeRF can generate affordance models of objects in various states.

4.2.2. Grasp Application

In addition to rendering new images, our method can also generate 3D models with affordance from a single image without extra manual processing. Affordance is strongly

correlated with robot manipulation tasks, wherein the generation of various strategies for different structures is often required. To verify the quality of the 3D model with affordance generated by our method, we input it into the grasp proposal algorithm [32], scoring the grasps using the metrics introduced in [33]. To compare, we also used the same grasp proposal method on the original object model without affordance understanding and compared the generated grasps. Table 3 shows the quantitative results of grasp scores on screwdrivers and eyeglasses. The qualitative results are shown in Figure 8. Our approach enables the grasp proposal algorithm to generate higher-quality grasps.



Figure 8. Visualization results of grasp generation algorithms with affordance model inputs.

Table 3. Grasp results with affordance model and original model.

	Screwdriver	Pen	Scissor	Eyeglass
Original Model (no affordance)	0.241	0.197	0.242	0.202
Affordance Model	0.272	0.198	0.306	0.242

5. Conclusions

This paper presented a novel NeRF method, AFF-NeRF, to address the problem of 3D affordance understanding for homogeneous objects. Our AFF-NeRF model decouples the shape and appearance of the object and extracts the corresponding code from only a single image. This allows our approach to adapt quickly to the various category objects in manipulation tasks. These codes are then integrated into the extended NeRF to render novel view images with affordance. To evaluate our approach, we also built a dataset containing different objects of various shapes with affordance annotation. We conducted experiments to validate the effectiveness of our AFF-NeRF model. The first experiment verified that our method achieves affordance rendering with novel objects without retraining or fine-tuning with single-view input. The experimental results demonstrate that our approach achieves significant improvements over the baseline and exhibits strong generalization to unseen objects. Meanwhile, our method has limitations when dealing with complex structures, partial occlusions, or extreme lighting. We will attempt to address these limitations in

future work. Furthermore, integrating our method with a grasp generation algorithm reveals that the affordance model enhances the generation of high-quality grasps.

Author Contributions: Conceptualization, Z.G. and X.C.; funding acquisition, X.C., Z.Z. and H.L.; investigation, Z.G.; project administration, Z.Z. and H.L.; software, Z.G.; validation, Z.G.; writing—original draft, Z.G. and X.C.; writing—review and editing, X.C., H.L. and R.G. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the National Science Foundation of China under Grant 62203460, U22A2059, and 62403478, Young Elite Scientists Sponsorship Program by CAST (No. 2023QNRC001), as well as the Innovation Science Foundation of National University of Defense Technology under Grant 24-ZZCX-GZZ-11.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data supporting this study's findings are available from the corresponding author or the first author upon reasonable request.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Gibson, J.J. The Theory of Affordances. In *The People, Place, and Space Reader*; Routledge: London, UK, 1977; Volume 1; pp. 67–82.
2. Horton, T.E.; Chakraborty, A.; Amant, R.S. Affordances for Robots: A Brief Survey. *AVANT. Pismo Awangardy-Filoz.-Nauk.* **2012**, *2*, 70–84. [[CrossRef](#)]
3. Hassanin, M.; Khan, S.; Tahtali, M. Visual Affordance and Function Understanding: A Survey. *ACM Comput. Surv. (CSUR)* **2021**, *54*, 1–35. [[CrossRef](#)]
4. Minaee, S.; Boykov, Y.; Porikli, F.; Plaza, A.; Kehtarnavaz, N.; Terzopoulos, D. Image Segmentation Using Deep Learning: A Survey. *IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI)* **2022**, *44*, 3523–3542. [[CrossRef](#)] [[PubMed](#)]
5. Mildenhall, B.; Srinivasan, P.P.; Tancik, M.; Barron, J.T.; Ramamoorthi, R.; Ng, R. Nerf: Representing Scenes as Neural Radiance Fields for View Synthesis. *Commun. ACM* **2021**, *65*, 99–106. [[CrossRef](#)]
6. Zhi, S.; Laidlow, T.; Leutenegger, S.; Davison, A.J. In-Place Scene Labelling and Understanding with Implicit Scene representation. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 11–17 October 2021; pp. 15838–15847.
7. Ishibashi, R.; Pobric, G.; Saito, S.; Lambon Ralph, M.A. The neural network for tool-related cognition: An activation likelihood estimation meta-analysis of 70 neuroimaging contrasts. *Cogn. Neuropsychol.* **2016**, *33*, 241–256. [[CrossRef](#)] [[PubMed](#)]
8. Ranganath, C.; Rainer, G. Neural mechanisms for detecting and remembering novel events. *Nat. Rev. Neurosci.* **2003**, *4*, 193–202. [[CrossRef](#)] [[PubMed](#)]
9. Niemeyer, M.; Geiger, A. Giraffe: Representing Scenes as Compositional Generative Neural Feature Fields. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 11453–11464.
10. Xu, Y.; Peng, S.; Yang, C.; Shen, Y.; Zhou, B. 3D-Aware Image Synthesis via Learning Structural and Textural Representations. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 18430–18439.
11. Xue, Y.; Li, Y.; Singh, K.K.; Lee, Y.J. Giraffe HD: A High-Resolution 3D-Aware Generative Model. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 18440–18449.
12. Schwarz, K.; Liao, Y.; Niemeyer, M.; Geiger, A. GRAF: Generative Radiance Fields for 3D-Aware Image Synthesis. In Proceedings of the Advances in Neural Information Processing Systems (NIPS), Virtual, 6–12 December 2020; Volume 33, pp. 20154–20166.
13. Chan, E.R.; Monteiro, M.; Kellnhofer, P.; Wu, J.; Wetzstein, G. Pi-GAN: Periodic Implicit Generative Adversarial Networks for 3D-Aware Image Synthesis. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 5799–5809.
14. Wang, C.; Chai, M.; He, M.; Chen, D.; Liao, J. Clip-NeRF: Text-and-Image Driven Manipulation of Neural Radiance Fields. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 3835–3844.
15. Jang, W.; Agapito, L. Codenerf: Disentangled Neural Radiance Fields for Object Categories. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 12949–12958.

16. Liu, S.; Zhang, X.; Zhang, Z.; Zhang, R.; Zhu, J.Y.; Russell, B. Editing Conditional Radiance Fields. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 5773–5783.
17. Yuan, Y.J.; Sun, Y.T.; Lai, Y.K.; Ma, Y.; Jia, R.; Gao, L. NeRF-Editing: Geometry Editing of Neural Radiance Fields. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 18353–18364.
18. Bao, C.; Zhang, Y.; Yang, B.; Fan, T.; Yang, Z.; Bao, H.; Zhang, G.; Cui, Z. Sine: Semantic-Driven Image-based NeRF Editing with Prior-Guided Editing Field. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023; pp. 20919–20929.
19. Wang, C.; Fang, H.S.; Gou, M.; Fang, H.; Gao, J.; Lu, C. Graspness Discovery in Clutters for Fast and Accurate Grasp Detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 15964–15973.
20. Gou, M.; Fang, H.S.; Zhu, Z.; Xu, S.; Wang, C.; Lu, C. RGB Matters: Learning 7-DOF Grasp Poses on Monocular RGBD Images. In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), Xi'an, China, 30 May–5 June 2021; pp. 13459–13466.
21. Sun, J.; Wang, Y.; Feng, M.; Wang, D.; Zhao, J.; Stachniss, C.; Chen, X. ICK-Track: A Category-Level 6-DoF Pose Tracker Using Inter-Frame Consistent Keypoints for Aerial Manipulation. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Kyoto, Japan, 23–27 October 2022; pp. 1556–1563.
22. Appius, A.X.; Garrabe, E.; Helenon, F.; Khoramshahi, M.; Chetouani, M.; Doncieux, S. Task-Aware Robotic Grasping by evaluating Quality Diversity Solutions through Foundation Models. *arXiv* **2024**, arXiv:2411.14917.
23. Czajewski, W.; Kołomyjec, K. 3D object detection and recognition for robotic grasping based on RGB-D images and global features. *Found. Comput. Decis. Sci.* **2017**, *42*, 219–237. [[CrossRef](#)]
24. Sahbani, A.; El-Khoury, S.; Bidaud, P. An overview of 3D object grasp synthesis algorithms. *Robot. Auton. Syst.* **2012**, *60*, 326–336. [[CrossRef](#)]
25. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
26. Rohmer, E.; Singh, S.P.; Freese, M. V-REP: A Versatile and Scalable Robot Simulation Framework. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Tokyo, Japan, 3–7 November 2013; pp. 1321–1326.
27. Mo, K.; Zhu, S.; Chang, A.X.; Yi, L.; Tripathi, S.; Guibas, L.J.; Su, H. PartNet: A Large-Scale Benchmark for Fine-Grained and Hierarchical Part-Level 3D Object Understanding. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 909–918.
28. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. Pytorch: An Imperative Style, High-Performance Deep Learning Library. In Proceedings of the Advances in Neural Information Processing Systems (NIPS), Vancouver, BC, Canada, 8–14 December 2019; Volume 32.
29. Chou, Z.T.; Huang, S.Y.; Liu, I.; Wang, Y.C.F. GSNeRF: Generalizable Semantic Neural Radiance Fields with Enhanced 3D Scene Understanding. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 16–22 June 2024; pp. 20806–20815.
30. Liu, F.; Zhang, C.; Zheng, Y.; Duan, Y. Semantic ray: Learning a generalizable semantic field with cross-reprojection attention. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023; pp. 17386–17396.
31. Yu, A.; Ye, V.; Tancik, M.; Kanazawa, A. PixelNeRF: Neural Radiance Fields from One or Few Images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 4578–4587.
32. Xu, W.; Zhang, J.; Tang, T.; Yu, Z.; Li, Y.; Lu, C. DiPGrasp: Parallel local searching for efficient differentiable grasp planning. *IEEE Robot. Autom. Lett.* **2024**, *9*, 8314–8321. [[CrossRef](#)]
33. Ferrari, C.; Canny, J. Planning optimal grasps. In Proceedings of the 1992 IEEE International Conference on Robotics and Automation, Nice, France, 12–14 May 1992; IEEE: Piscataway, NJ, USA, 1992; Volume 3, pp. 2290–2295.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.