Affordance Transfer based on Self-Aligning Implicit Representations of Local Surfaces

Ahmet Tekden Marc Peter Deisenroth Yasemin Bekiroglu

***Abstract*—Objects we interact with and manipulate often share similar parts, e.g. handles, that allow us to transfer our actions flexibly due to their shared functionality. This corresponds to affordances, i.e. set of action possibilities offered by the environment [1]. In this work, we propose to learn affordances associated with implicit models of local shapes shared across object categories. Our approach takes an expert grasp demonstration on a given object, extracts the local geometry, and uses it as an anchor to align corresponding parts of objects from the same category. We show that the proposed implicit representation method can align objects within the same category under random pose perturbation. In addition, our general approach can align the local geometry to find grasp poses similar to the one demonstrated in the reference local shape. Finally, we show that we can identify the shared local geometry on novel objects from a different object category for affordance transfer.**

I. Introduction

Predicting grasp affordances is a topic widely studied [2], [3], [4], [5], [6], as grasping is one of the most widely used robotic skills. An approach to generate grasp poses is based on finding category-level dense point correspondences and using them for grasp generation [7]. Recently, Nerfs [8] and implicit representations [9], [10] have been used in robotic manipulation [11] for dense correspondence generation [12], [13]. However, as these approaches do not allow for transfer to new categories.

If we find correspondences explicitly with local surface similarities, we can transfer grasp poses across object categories that share parts with similar grasp affordances. A visualisation of this idea is shown in Figure 1 where, using our method, we can match the implicit representation learnt from the handle of a mug object to identify the grasp location on other objects that have similar parts. Previously, part prototypes were used to transfer grasps across novel object categories [14] using a shape similarity score between prototype parts and observed parts. However, shape similarity score alone may lead to poor generalization as this approach does not guarantee parts from same semantic category to match each other. Instead, in this work, we model shapes with implicit surface representations using position-based neural networks and perform shape inference by optimising a latent shape code [9]. For this optimization, such methods require all objects to be in canonical reference frames both in training and test time.

This work was supported by Chalmers AI Research Center (CHAIR) and Chalmers Gender Initiative for Excellence (Genie), the project AIMCoR - AI-enhanced Mobile Manipulation Robot for Core Industrial Applications and partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation.

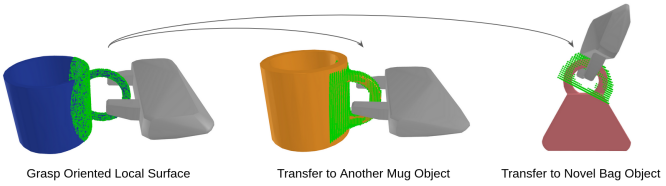

Fig. 1. We learn an implicit representation to model local surfaces aligned to generate grasp poses, which facilitates identifying similar local surfaces in novel objects from other categories to transfer grasps.

Aligning objects to the canonical poses is a hard problem, and often objects will still not completely align [15], [16]. Recently with Nerfs, it has been shown that camera poses can be estimated by optimising a pose embedding along with the Nerf reconstruction optimization [17], or by inverting the Nerf [18]. Similarly to the former one, in this work, we optimise a pose embedding that can be used to align objects to a canonical frame. The main contributions of this work can be summarized as follows:

- A novel pose and scale alignment approach that allows us to learn implicit representations even when there are pose variations between the objects up to an extent. This in turn, allow us to learn compositional implicit representations on explicitly provided locations even without complete surface match and alignment.
- We provide an approach for generalising expert grasp demonstration provided on a single object from an object category to other objects in the same category by simultaneously learning an implicit representation for the local surfaces of the objects and aligning them to have the grasp pose in the origin of the learnt local surface model.
- Finally, we show that the learned implicit representation can be used to identify similar local shapes on given a novel object to recover possible grasp locations.

II. Method

Our proposed approach is shown in Figure 2. In this section, we first describe how shape generation is done and then how the shape alignment process is integrated.

a) Preliminaries: Signed distance functions (SDFs) are implicit surface representations that are commonly used in computer vision to represent shapes of objects. An SDF evaluates to 0 for the points on the surface of objects, negative for the points inside the object and positive for the points outside of the object. For predicting SDF values, DeepSDF [9]

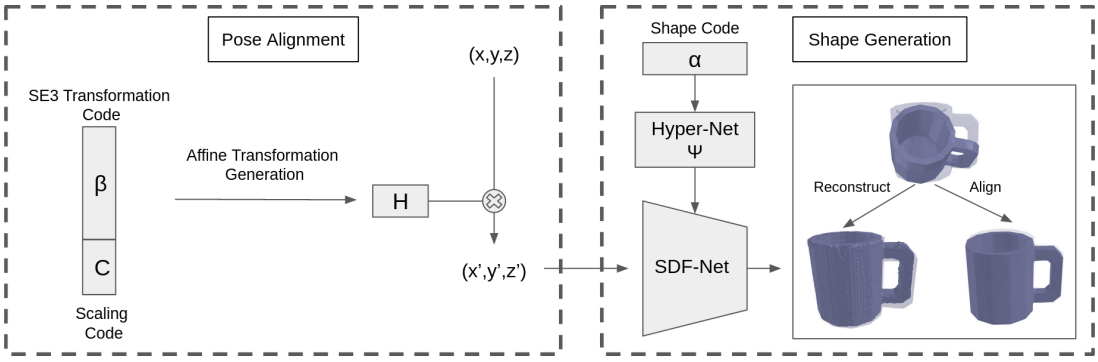

Fig. 2. The two part network architecture: First using an SE3 transformation and a scaling code, we generate an affine transformation matrix to transform the input positions. Using these positions, the shape generation module predicts SDF values for each given transformed input location.

introduced coordinate-based neural networks. In this network, a multilayer perceptron (MLP) predicts the SDF value for a given 3D $X = (x, y, z)$ location and a shape latent code (α). However, instead of an autoencoder training, DeepSDF used auto-decoders in which an object code is not produced by an encoder, but learned along with the MLP with gradient optimization. Later in Siren [19], the authors similarly used the autodecoder strategy, but instead of giving shape code to the network as input, the shape code is given to a hyper-network [20] to predict part of the MLP's weights. In our work, we similarly use hypernetworks along with position-based MLPs for shape generation.

One commonly used approach for representing high frequency shapes (i.e more complicated shapes) in coordinate-based MLPs is to use position encodings [21].Position encodings map 3D locations X to higher-dimensional inputs using the map $\gamma(X) = [X, \gamma_0(X), \gamma_1(X), ..., \gamma_{L-1}(X)] \in \mathbb{R}^{3+6L}$ where $\gamma_m(X) = [cos(2^m \pi X), sin(2^m \pi X)] \in \mathbb{R}^6$. In our work, we do not always use position encodings, as they require tuning of extra hyper-parameters.

b) Pose Alignment: SE3 transformations have 6 degree of freedoms (DOF), therefore ideally they should be represented with a 6 dimensional SE3 transformation code. Using the Lie algebra of SE(3) [22], we can map 6 dimensional embeddings to SE(3) transformations.

For finding the part scaling, we learn a 3-dimensional scaling embedding C that is then transformed to a $(4x4)$ transformation matrix by $diag([C, 1])$. Multiplying this matrix with the SE3 transformation matrix acquired before, we acquire an affine transformation matrix that we can use for pose and scale alignment. This affine transformation matrix is used to transform a given X location into X'. Optimising β and C together, which we call pose refinement code, while training the neural network allows the reconstructed shapes to be well aligned with each other.

c) Dynamically adapting sampling sphere: For training the neural network to only learn the local surface given on the object, we sample points only within the sphere with radius r and identity pose. The reason for this is to limit the network's capacity spent on the surface outside of the targeted one. However, one problem is that, since the alignment transformation is optimised at the same time with the network parameters, some of the points that should be in the sampling sphere are not inside, since the shapes are not aligned yet. For this reason, we use the current transformation matrix to sample points that will be in the unit sphere after the transformation.

d) Grasp Transfer: Since our algorithm can align objects within the same category, we can define a canonical pose with respect to the local surface of an object which can then be transferred to corresponding surfaces in the objects from the same category after the alignment. Ideally, this pose can be selected as the identity pose, so that our algorithm learns the local surface with respect to the pose.

This canonical pose corresponds to a grasp pose in our algorithm. We first get an expert grasp demonstration on one object, which we assign as the anchor object, and then we transform the shape with respect to this given pose. We apply the same transformation to all objects from the same category; however, usually the identity pose will not correspond to the ideal grasp pose for the other objects. For the anchor object, we fix the pose refinement codes so that other objects align around the anchor object. In the learnt implicit representation, the identity pose will correspond to the grasp pose.

For transferring to new objects, we can use the learnt implicit representation. For an object that has a similar local surface with the originally trained one, we can pick a several positions on close-by locations to this similar surface and then we can optimize a shape code along with pose refinement codes so that after the optimisation, if the reconstruction error is lower than a given threshold, we can use the pose refinement codes for identifying a grasp pose on the novel object.

III. Experiments

To evaluate our method, we use the category of cups and bags from the ShapeNet-Core V2 dataset [23] (note that all objects in ShapeNet-Core V2 are in canonical poses) and perform two experiments: Shape alignment and affordance transfer experiments.

A. Shape Alignment

To evaluate the performance of the shape alignment process, we create a dataset of mug objects whose poses are varied by applying random perturbations. For each object in the dataset, we also have the corresponding non-perturbed mug object. We

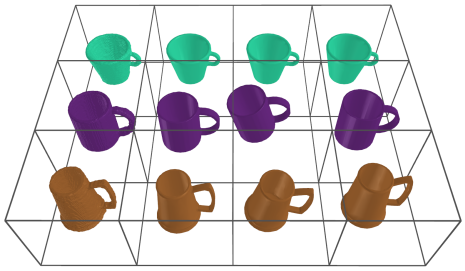

Fig. 3. Results from learned implicit representation and pose alignment. The first row: the anchor object, the first column: the reconstructed mesh, the second column: the mesh in canonical frame, the third column: the perturbed mesh, and the last column: re-aligned perturbed mesh using the learned pose embedding.

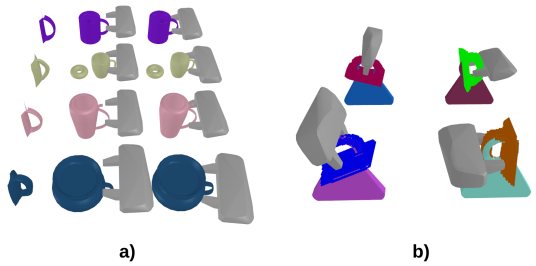

Fig. 4. a) Grasps that would have failed without the grasp alignment process. The initial grasps are either failing to contact the mug handle or colliding with it. After the local surface is learned and shapes are aligned, we acquire feasible grasp locations. b) Grasps transferred to a novel object. After optimizing the latent and pose refinement codes, we acquire grasps based on local surface fit.

train the self-aligning implicit representation network with one mug on its canonical reference frame as anchor, and 63 mugs with randomly perturbed position and orientations. For this experiment, positional encodings are used, because without positional encodings the network cannot model the hole in the handle of the mug. When learning implicit representations for these objects, we apply a smooth mask to input encodings in a similar way to BARF [17].In this experiment, we fix the scaling embeddings as mugs do not have a canonical size, and the learnt scaling cannot be evaluated quantitatively.

Visualization of object mesh reconstructions from the predicted SDF values, along with object meshes on ground truth location, pose perturbed location, and re-aligned perturbed pose location using the learn pose embedding, can be seen in Figure 3. The similarity between reconstructed shape, mesh in ground truth pose and re-aligned perturbed mesh shows that our network successfully aligns the implicit representation to the meshes on ground truth frames while maintaining the reconstruction quality.

We further evaluate how well our method aligns with the ground truth mesh by comparing Chamfer distances between points sampled from perturbed mesh and ground-truth mesh with reconstructed mesh and ground-truth mesh. To do this, we sample 10000 points from all meshes (except for the first one, which is in the canonical reference frame) and estimate the two-way Chamfer distance. As a result, we found that the Chamfer distance for the reconstructed meshes is about $0.001m^2$ and the pose perturbed meshes is about $0.019m^2$ This result show that our approach can align object meshes while learning an implicit representation for the objects.

B. Affordance Transfer

In this experiment we show that we can learn local geometry on corresponding parts of objects from the same category. We use handle of mugs to demonstrate and learn grasp affordances. We first provide a grasp pose reference on the anchor mug. This grasp pose is then naively transferred to other mugs using the relative position and orientation of the demonstrated grasp with respect to the centre of the object. However, this naive transfer provides a local surface dataset with perturbed locations that may cause grasps to fail. We then train our implicit representation network.

Visualisation of the expert grasp on the anchor mug and the mugs that originally had an unsuccessful grasp location if they had not been aligned can be seen in Figure 1, and Figure 4a. The first column corresponds to the surface reconstruction, the second column corresponds to the initial grasp location and the last column corresponds to the grasp location after alignment. As can be seen in Figure 4 alignment process moves the grasp location from infeasible to feasible locations.

Finally, we test whether we can use the learnt implicit representation to identify grasps on a novel bag object. For this, we generate reference frames with position close to the bag handle with random orientations. For each reference frame, we optimise a shape and alignment code. While our approach also succeeds with random position sampled from whole surface of the bag, it requires generation of a lot of reference frames on the object, which is not feasible overall. Visualisation of four different grasps generated by our approach with low reconstruction errorcan be seen in Figure 4b. Our approach enables generation of different grasp locations on novel objects based on surface reconstruction.

IV. Conclusion

In this work, we present an approach to learn self-aligning local implicit representations which can be used to transfer grasp affordances to new objects with similar local geometry. However, there are certain limitations to our approach. It requires training objects to be in poses close to canonical frames, and the scaling process has a limited performance in generalizing to significantly bigger or smaller object parts. In addition, in the affordance transfer process, our approach requires initial reference frames that are in close location to the shape to be aligned. In future work, we plan to use equivariant object representations [24], [25], possibly improve the scaling process by using approaches similar to DIF-NET [26], and to reduce the search space, we will use keypoint detection methods to identify similar local geometry in novel objects. Finally, we evaluate our approach using synthetic data with ground truth objects, point clouds and SDF values. We will further evaluate the transfer process on real data which can be partially observable with noise.

REFERENCES

[1] J. J. Gibson, *The ecological approach to visual perception: classic edition*. Psychology Press, 2014.

[2] P. Zech, S. Haller, S. R. Lakani, B. Ridge, E. Ugur, and J. Piater, "Computational models of affordance in robotics: a taxonomy and systematic classification," *Adaptive Behavior*, vol. 25, no. 5, pp. 235–271, 2017.

[3] C. Eppner, A. Mousavian, and D. Fox, "Acronym: A large-scale grasp dataset based on simulation," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 6222–6227.

[4] 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.

[5] A. Mousavian, C. Eppner, and D. Fox, "6-dof graspnet: Variational grasp generation for object manipulation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 2901–2910.

[6] Z. Jiang, Y. Zhu, M. Svetlik, K. Fang, and Y. Zhu, "Synergies between affordance and geometry: 6-dof grasp detection via implicit representations," *arXiv preprint arXiv:2104.01542*, 2021.

[7] L. Manuelli, W. Gao, P. Florence, and R. Tedrake, "kpam: Keypoint affordances for category-level robotic manipulation," *arXiv preprint arXiv:1903.06684*, 2019.

[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. 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*, 2019, pp. 165–174.

[10] L. Mescheder, M. Oechsle, M. Niemeyer, S. Nowozin, and A. Geiger, "Occupancy networks: Learning 3d reconstruction in function space," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4460–4470.

[11] D. Driess, J.-S. Ha, M. Toussaint, and R. Tedrake, "Learning models as functionals of signed-distance fields for manipulation planning," in *Conference on Robot Learning*. PMLR, 2022, pp. 245–255.

[12] 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," *arXiv preprint arXiv:2203.01913*, 2022.

[13] 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," *arXiv preprint arXiv:2112.05124*, 2021.

[14] R. Detry, C. H. Ek, M. Madry, and D. Kragic, "Learning a dictionary of prototypical grasp-predicting parts from grasping experience," in *2013 IEEE International Conference on Robotics and Automation*. IEEE, 2013, pp. 601–608.

[15] H. Wang, S. Sridhar, J. Huang, J. Valentin, S. Song, and L. J. Guibas, "Normalized object coordinate space for category-level 6d object pose and size estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2642–2651.

[16] C. Wang, R. Martín-Martín, D. Xu, J. Lv, C. Lu, L. Fei-Fei, S. Savarese, and Y. Zhu, "6-pack: Category-level 6d pose tracker with anchor-based keypoints," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 10 059–10 066.

[17] C.-H. Lin, W.-C. Ma, A. Torralba, and S. Lucey, "Barf: Bundle-adjusting neural radiance fields," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 5741–5751.

[18] 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 *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2021, pp. 1323–1330.

[19] V. Sitzmann, J. Martel, A. Bergman, D. Lindell, and G. Wetzstein, "Implicit neural representations with periodic activation functions," *Advances in Neural Information Processing Systems*, vol. 33, pp. 7462–7473, 2020.

[20] D. Ha, A. Dai, and Q. V. Le, "Hypernetworks," *arXiv preprint arXiv:1609.09106*, 2016.

[21] M. Tancik, P. Srinivasan, B. Mildenhall, S. Fridovich-Keil, N. Raghavan, U. Singhal, R. Ramamoorthi, J. Barron, and R. Ng, "Fourier features let networks learn high frequency functions in low dimensional domains," *Advances in Neural Information Processing Systems*, vol. 33, pp. 7537–7547, 2020.

[22] J.-L. Blanco, "A tutorial on se (3) transformation parameterizations and on-manifold optimization," *University of Malaga, Tech. Rep*, vol. 3, p. 6, 2010.

[23] A. X. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su, J. Xiao, L. Yi, and F. Yu, "ShapeNet: An Information-Rich 3D Model Repository," Stanford University — Princeton University — Toyota Technological Institute at Chicago, Tech. Rep. arXiv:1512.03012 [cs.GR], 2015.

[24] X. Li, Y. Weng, L. Yi, L. Guibas, A. L. Abbott, S. Song, and H. Wang, "Leveraging se (3) equivariance for self-supervised category-level object pose estimation," *arXiv preprint arXiv:2111.00190*, 2021.

[25] R. Sajnani, A. Poulenard, J. Jain, R. Dua, L. J. Guibas, and S. Sridhar, "Condor: Self-supervised canonicalization of 3d pose for partial shapes," *arXiv preprint arXiv:2201.07788*, 2022.

[26] Y. Deng, J. Yang, and X. Tong, "Deformed implicit field: Modeling 3d shapes with learned dense correspondence," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 10 286–10 296.