DINA: Deformable INteraction Analogy

Zeyu Huang[a], Sisi Dai[b], Kai Xu[b], Hao Zhang[c], Hui Huang[a], Ruizhen Hu[a,*]

[a] College of Computer Science and Software Engineering, Shenzhen University, China
[b] School of Computer Science, National University of Defense Technology, China
[c] School of Computing Science, Simon Fraser University, Canada

Abstract

We introduce deformable interaction analogy (DINA) as a means to generate close interactions between two 3D objects. Given a single demo interaction between an anchor object (e.g. a hand) and a source object (e.g. a mug grasped by the hand), our goal is to generate many analogous 3D interactions between the same anchor object and various new target objects (e.g. a toy airplane), where the anchor object is allowed to be rigid or deformable. To this end, we optimize the pose or shape of the anchor object to adapt it to a new target object to mimic the demo. To facilitate the optimization, we advocate using interaction interface (ITF), defined by a set of points sampled on the anchor object, as a descriptive and robust interaction representation that is amenable to non-rigid deformation. We model similarity between interactions using ITF, while for interaction analogy, we transform the ITF, either rigidly or non-rigidly, to guide the feature matching to the reposing and deformation of the anchor object. Qualitative and quantitative experiments show that our ITF-guided deformable interaction analogy works surprisingly well even with simple distance features compared to variants of state-of-the-art methods that utilize more sophisticated interaction representations and feature learning from large datasets.

Keywords: Shape analysis, Interaction modeling, Imitation learning, Optimization

1. Introduction

Acquiring hand-object and object-object interaction data in 3D space can benefit several applications in AR/VR [1, 2, 3], functionality and affordance analysis [4, 5], computer animation and physical simulation [6, 7, 8], as well as imitation learning in robotics [9, 10]. Reconstructing such data from photographs or laser scans [11, 12] is challenging due to the ill-posedness of the problem, as well as object occlusions arising from close interactions. Data-driven approaches to generative modeling of 3D interaction data typically require large amounts of training data to begin with [13, 14], resulting in a "catch-22" situation.

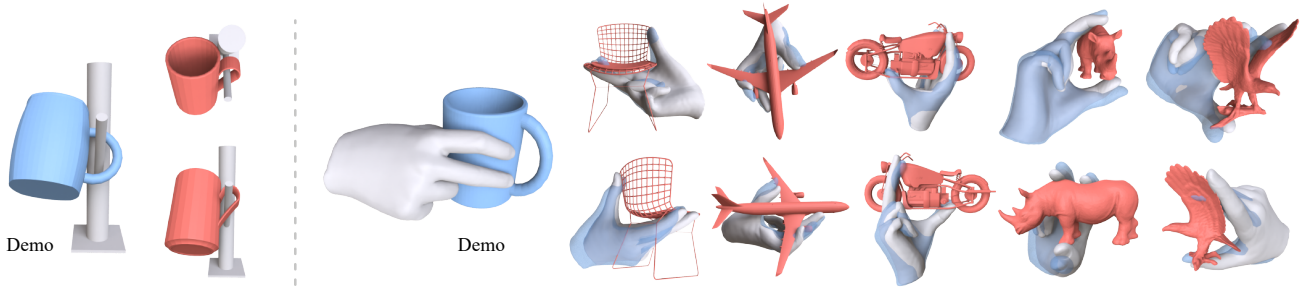
In this paper, we present *interaction analogy* as a means to generate close interactions between two 3D objects from a *single demo* interaction. Specifically, the demo interaction is between an *anchor object*, such as a rack, and a *source object*, such as a mug that can be placed on the rack; see Figure 1(a)-left. The goal of interaction analogy is to generate many *analogous* 3D interactions, each between the same anchor object and a new *target* 3D object such as a new mug or any other suitable object for the anchor. Importantly, we allow the anchor object, but not the source or target objects, to be non-rigid and hence *deformable*. In the latter case, our current work focuses on non-rigid articulated deformation of the hand, making it possible to generate novel hand grasping in 3D, as shown in Figure 1(b).

We coin our problem and solution as *deformable interaction analogy*, or DINA for short. To optimize relative positioning of the anchor and target objects, as well as deformation of the anchor, so as to mimic the demo interaction, we need to address fundamental questions about how to represent and optimize 3D interactions, and how to measure the similarity between the target and the demo interactions. Of the utmost importance is to find a carefully designed representation which should be: (1) sensitive to and characteristic of object-object interactions, rather than the individual 3D objects; (2) robust against shape variations of interacted objects; (3) amenable to non-rigid and adaptive deformations, e.g., of the high degrees-of-freedom human hand; (4) controllable to facilitate deformable modeling.

We represent close object-object interactions by an *interface* that is defined *on the anchor object* and *in close proximity* to the interaction. We call it the interaction interface or ITF for short. Specifically, the ITF consists of a set of points on the anchor object (see Figure 2) and they are determined by proximity to the intersection bisector surface (IBS) [15], which is defined by a partial set of points equidistant to the anchor and the source objects; see Figure 3. The IBS has been shown to be an informative spatial descriptor of object interactions, while robust against shape variations. Compared to the IBS and in the context of DINA, ITF is even more robust again variations of the *source object* since it resides on the anchor object which is fixed. In addition, ITF, like IBS, enjoys the other desirable properties listed above for interaction representation.

We model similarity, or analogy, between two object-object interactions based on *point-wise distances* from the ITF to the corresponding source object. Given a single demo interaction and a new target object with a random pose, we use ITF to

*Corresponding author. Email: ruizhen.hu@gmail.com



(a) Rigid (b) Deformable

Figure 1: With a *single demo* interaction, our method can generate an analogous interaction (shown in two views) for a new object (red) in both rigid (a) and *deformable* (b) settings. In (a), an interaction between the rigid rack and a new mug is generated. In (b), interactions with a variety of objects that vary significantly from the demo mug are generated by deforming the hand for grasping: demo hand pose in blue transparent color; deformed hand in grey.

guide the optimization of the pose, articulation, and/or shape of the anchor, adapting it to the target to produce a new interaction as similar to the demo as possible, while penalizing *inter-penetrations* between the interacting objects. Our DINA is carried out as a two-step optimization. First, we perform rigid interaction analogy (RINA) by computing a rigid posing of the anchor object with ITF guiding the distance feature matching. Then, from the RINA results, we locally deform the ITF, together with the anchor object, to bring the resulting point-wise distance features closer to those in the demo interaction.

As our main contribution is *deformable* interaction analogy, we perform extensive experiments on the hand grasping task to compare with variants of state-of-the-art hand pose optimization method ContactOpt [16] and other interaction representations based on the work of [17], both quantitatively and qualitatively. The results suggest that our ITF-guided interaction analogy works surprisingly well even with simple distance features, outperforming other alternatives without using contact maps or sophisticated feature learning from large-scale datasets. Moreover, experiments show that RINA guided by ITF can yield improvements over other options and it can also serve as a better initialization for non-rigid deformation in other frameworks. We summarize our contribution as follows:

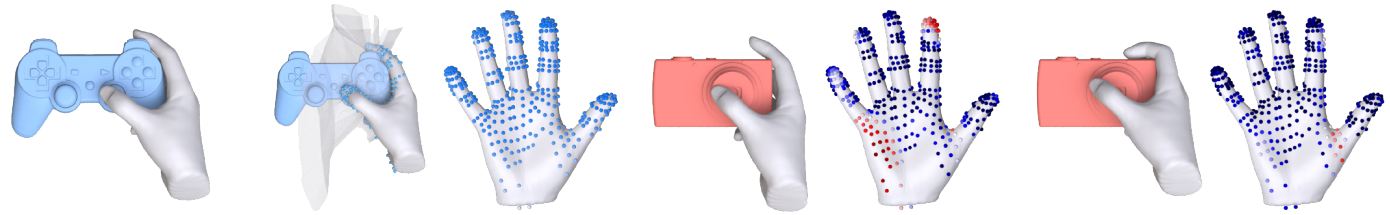
- We introduce deformable interaction analogy (DINA) as a means to generate interactions between two 3D objects;
- We propose interaction interface (ITF) which is a descriptive and robust interaction representation;
- We conduct extensive experiments to show the superiority of our approach on hand grasping generation task.

2. Related Work

Our problem setting bears resemblance to works on image [18] and shape analogies [19], where the input consists of three data entities, A, A', and B, while the goal is to generate a new analogous data entity B' that relates to B in "the same way" as A' relates to A. In DINA, all the data entities are 3D objects, and the relation to be modeled and imitated is an interaction between the objects, with one and only one demo interaction provided. Also relevant is motion retargeting [20, 21, 22, 23] of animatable characters to adapt to new interactions. These works focus on the human body as the anchor object, where there always exists a natural correspondence between the source and target objects, while we aim for a more general interaction analogy between objects. Thus, our work is related to geometric representations of interactions in 3D space, generative modeling of interactions, and one-shot imitation learning. In this section, we cover prior works most relevant to DINA.

Geometric representation of interaction. There is a large body of literature on representations of interactions between objects. Early practice utilizes relative vectors [24] from one object to another as the contact representation between interacting objects for scene synthesis, which is intuitive but not descriptive enough for representation of complex interaction in scenes. Zhao et al. [15] introduce a more sophisticated representation IBS that describes topological and geometric relationships between objects, which is effective for scene completion and generation [25, 26]. This idea is further developed to estimate and localize the functionality of 3D shapes [27, 28]. Pirk et al. [29] track particles on one of the interaction objects and build a spatial and temporal representation called interaction landscapes. All of these works only focus on analyzing the interaction between two objects and interaction retrieval. Our work, in the contrast, aims to facilitate the transition and deformation from one object to another for deformable interaction analogy

Generative modeling of interactions. General object-to-object interactions can be adopted to characterize functionalities of 3D objects and constitute interaction contexts to demonstrate how the object should be used [30] or synthesize interaction snapshots given high-level specifications such as language [31, 32, 33]. Modeling and analysis of more specific types of interactions like hand-object interaction have also been an active field of study, with a recent trend on generative modeling of human grasps. Karunratanakul et al. [34] propose an implicit representation of human-object interaction to model the joint distribution of the object and hand in the corresponding grasping pose. To improve the grasping quality, several works [35, 36, 16] learn an affordance model to predict the contact regions on the object surface or hand surface, providing guidance for grasp



(a) Demo interaction (b) Interaction representation (c) Rigid interaction analogy (d) Deformable interaction analogy

Figure 2: An illustrative overview of our deformable interaction analogy method. Given a demo interaction (a) of a hand (the anchor object) grasping a joystick (the source object), we first represent the interaction by an interaction interface (ITF) on the hand, with associated point-wise distances to the joystick. The ITF consists of a set of points on the hand and they are determined by the intersection bisector surface (IBS) between the hand and the joystick. For a new target object, e.g., a camera shown in (c), which is given in a random pose, to generate an analogous interaction (i.e., a grasping) by the hand, we first perform rigid interaction analogy (c) by optimizing the global transformation of the ITF together with the hand. Finally, the hand is deformed so that the ITF can better fit the geometry of the camera with a similar distance distribution to that in the demo, leading to a better hand grasping resembling the demo. The colormaps on ITF points shown in (c) and (d) indicate the distance errors against those extracted from the demo interaction.

generation. The generated grasps can be further used as demonstrations to train a policy that generalizes to unseen objects for dexterous hand manipulation by imitation learning [37]. This line of research learn and predict the grasping pose for a given object based on a large-scale grasp dataset. Our method, on the other hand, performs an interaction analogy with only a single grasp demonstration and allows the hand to be locally deformed, making it more adaptive to the geometry of the new target object.

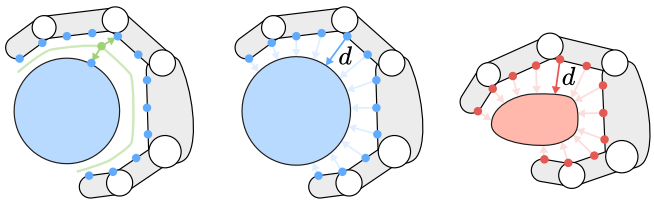
One-shot imitation learning. Teaching a robot to perform manipulation tasks using few-shot demonstrations has been a long-standing problem in robotics [38, 39, 40]. Yu et al. [10] proposes one-shot learning from a video of a human conducting tasks to build prior knowledge via meta learning. Combining the prior knowledge and only a single video demo, the robot can perform the task as the human did. The interaction analogy problem studied in our work is related to one-shot imitation learning since both require only a single demonstration as input. The main difference however, is that one-shot imitation learning in robotics usually involves learning policy for more complex behavior while we focus on a unit action such as an interaction snapshot.

Neural descriptor field. Recent works on neural descriptor fields (NDF) [17, 41] aim to solve a similar problem to DINA but makes the assumption that all the objects involved are *rigid*. These works characterize object interactions by sampling a fixed set of query points around the anchor object. Such a set of query points can be referred as a Basis Point Set (BPS) [42], which has been shown to provide an efficient and compact means to encode features of the *anchor object alone*, but not the interaction between the anchor and other objects. Also, neither NDF nor BPS has been considered for deformation modeling.

3. Method

3.1. Problem and representation

Given a demo interaction (O_a, O_s), where O_a is the anchor object that can be deformable and O_s is the source object that is to be replaced with a target object O_t given in a random pose,



(a) IBS & ITF (b) ITF distance feature (c) DINA

Figure 3: Determination of ITF based on IBS (a), ITF distance features (b), and deformation of ITF to guide deformable interaction analogy (c).

our goal is to optimize the global pose T as well as local deformation D of the anchor object O_a w.r.t O_t such that their interaction analogizes the demo interaction. Thus, the optimization is formulated as:

$$\{\bar{T}, \bar{D}\} = \arg\min_{\{T,D\}} dist\left(f(O_a, O_s), f(T \cdot D(O_a), O_t)\right). \quad (1)$$

where $f(O_1, O_2)$ is the interaction representation between object O_1 and O_2, and $dist(f_1, f_2)$ is the corresponding distance metric defined on two interactions.

Interaction representation. As our goal is to perform an interaction analogy, we would like the interaction representation to capture the geometric features of the most important region related to the interaction and guide the deformation of O_a towards O_t. Moreover, to deal with the target object given in arbitrary poses, the features should be SE(3)-invariant.

To better characterize the demo interaction (O_a, O_s), we utilize the interaction bisector surface (IBS) [15], as it has been shown to be robust to the local geometric details of interacted objects in indexing their spatial relations. To extract the IBS, we first sample a set of points on the surfaces of the two interacting objects uniformly and compute the Voronoi diagram for all those samples. The IBS is a surface mesh consisting of a subset of the ridges of the computed Voronoi diagram, which lies between the two objects. However, the computation complexity of both IBS itself and the corresponding distance measure makes it hard to use IBS directly for DINA.

To localize areas where interactions actually take place, we further select the set of points on O_a that determine this IBS to

form the interaction interface (ITF), denoted as $\mathcal{X}_a$. Compared to the IBS, which is separate from the interacting objects, ITF is located on the surface of the anchor object O_a automatically derived from IBS. Therefore, the deformation of the anchor object can be naturally driven by ITF. For each point $p \in \mathcal{X}_a$, we record its shortest distance to the source object O_s as $d(p, O_s) = \min_{q \in O_s} \|p - q\|_2$.

Figure 3 shows a 2D illustration of the computation and deformation of ITF, where the IBS is shown with a green line and the ITFs before and after deformation are shown with blue and red points, respectively.

3.2. Interaction analogy

Objective function. With the interaction between O_s and O_a represented by ITF points $\mathcal{X}_a$ with point-wise distances $\{d(p, O_s)\}$, the goal becomes to finding an optimal pose and valid deformation of $\mathcal{X}_a$ such that the point-wise distance to the target object $\{d(T \cdot D(p), O_t)\}$ is as close to $\{d(p, O_s)\}$ as possible, as shown in Figure 3 (c). Note that as the ITF points $\mathcal{X}_a$ are defined on the anchor object O_a, we constrain the deformation space of $\mathcal{X}_a$ to align with that of O_a. Thus, the objective function of Eq. (1) becomes

$$L_{\text{ITF}}(T, D) = \sum_{p \in \mathcal{X}_a} |d(p, O_s) - d(T \cdot D(p), O_t)|. \tag{2}$$

Moreover, to discourage heavy intersection between two objects in the final interaction, we further add an explicit penetration term that penalizes penetrations as in the work of [16]:

$$\begin{aligned} L_{\text{PEN}}(T, D) &= \sum_{x_i \in O_a'} \max((x_i - x_t) \cdot n_t, 0) \\ &+ \lambda \sum_{x_j \in O_t} \max((x_j - x_a) \cdot n_a, 0), \end{aligned} \tag{3}$$

where $O_a' = T \cdot D(O_a)$ is transformed and deformed anchor object, x_t is the nearest point of x_i on O_t with normal n_t, and x_a the nearest point of x_j on O_a' with normal n_a. We set $\lambda = 2$ as the normal of points on the anchor object is more reliable during the analogy. The final objective of our interaction analogy is defined as:

$$\{\bar{T}, \bar{D}\} = \arg\min_{\{T, D\}} L_{\text{ITF}}(T, D) + L_{\text{PEN}}(T, D). \tag{4}$$

Rigid-to-deformable optimization. Since the size and the geometry of the target object O_t can be quite different from those of the source object O_s, the original pose of O_a interacting with O_s might not fit well to O_t. We therefore need to not only transform but also deform (if deformable) the anchor object O_a to form an interaction better fitting the target object O_t.

To achieve this goal, a straightforward method is to optimize the global transformation T and local deformation D simultaneously. However, this method is prone to local minima since the search space is high-dimensional. It is common to add an additional L1 regularization term on D to restrict the deformation of the anchor object in a reasonable pose and shape. However, we find it difficult to balance the weights between the regularization term and our objective function to obtain a general solution. As shown in Figure 4 (a)-(b), optimization with weak regularization usually results in an unnatural hand pose, while strong regularization constrains the hand pose too much such that it cannot form a reasonable grasping for the target object.

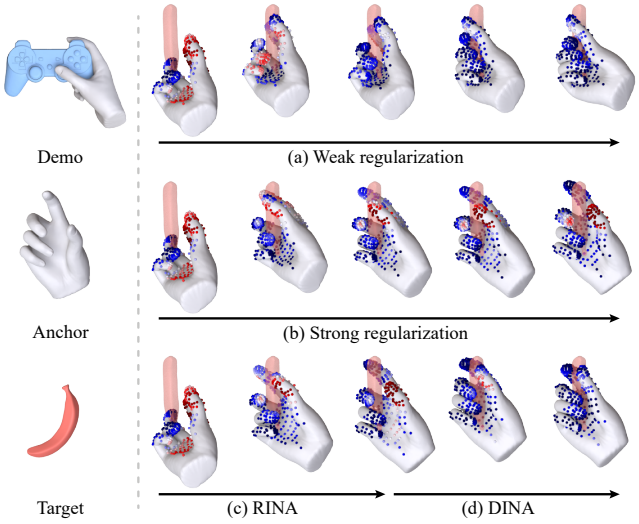


Figure 4: One-step vs. two-step optimization for interaction analogy.

To make the optimization more stable, we propose to use a two-step optimization. The first step confines the optimization on rigid transformation to obtain a relatively good initial interaction for bootstrapping the local deformation in the second step. We refer to the result of the first step as *RINA* for rigid interaction analogy and the final result as *DINA* for deformable interaction analogy.

To optimize the global pose of $\mathcal{X}_a$ relative to O_t, we first move the centroid of O_t to the centroid of O_s and randomly sample rotation and translation of $\mathcal{X}_a$ in this local frame. Specifically, the rotation is uniformly sampled from SO(3) and each translation component is sampled from $\mathcal{N}(0, 0.01)$. We then solve for the optimal rotation R and translation t that minimize the objective function in Eq. (4). The optimization is repeated for ten times with different initialization, and the optimal pose with the lowest error is taken as the final result. It takes 0.2s to extract the ITF representation from the demo interaction and 5s to optimize one target object with a NVIDIA RTX 3090 GPU.

With the RINA result, we can further enable the local deformation of the ITF to improve the quality of interaction and similarity to the demo. Since ITF is defined on the anchor object, we use the deformable MANO model [43] as the anchor object for hand grasping tasks. This model is parameterized by the pose of all joints in a low-dimensional PCA space. We then optimize the hand pose with the guidance of the ITF to obtain the deformed anchor object together with the deformed ITF.

The last row of Figure 4 shows one example of our two-step optimization process. We can see that the hand (anchor object) gradually moves closer to the tiny banana (target object), guided by the movement of the ITF points, and forms a rigid interaction analogy in the first step. Starting from this interaction as

an initialization, the hand is further deformed to form a tighter grasping. Note how the distance errors of the ITF points are minimized by both the rigid and the non-rigid optimization.

4. Results and Evaluation

We first perform quantitative and qualitative comparisons to show that our method outperforms state-of-the-art methods and justify our method design. Then, we show more diverse results to demonstrate the effectiveness and generality of our approach.

4.1. Experiment setup

Dataset. Our experiments are mainly conducted on the hand grasping task, based on the ContactPose dataset [44], which consists of 2306 grasping poses of 25 objects. We include all 77 scanned objects from YCB dataset [45] as additional target objects, which consists of objects of daily life with different shapes and sizes. To evaluate interaction analogy, we construct a test dataset with 5100 demo-target pairs on top of these interactions and objects. In more detail, we sample 50 grasps from ContactPose as demo interactions for each target object of the 102 objects from ContactPose and YCB dataset.

Evaluation metrics. To evaluate the similarity between the demo interaction and analogy result, we first adopt several metrics from ContactOpt [16] to measure the penetration and contact region similarity, which includes:

• **Penetration Volume and Difference** (cm^3) , denoted as V and $|\Delta V|$. We voxelize the target object with an edge length of 5mm, and calculate the volume of the voxels inside the anchor object surface. Meanwhile, we found that penetrations may occur in the demo interaction. Therefore, we further compute the difference between the penetration volumes of the demo and analogous interaction. A smaller penetration volume indicates a more natural interaction, and a smaller penetration difference suggests a more similar analogy to the demo interaction.

• **Contact Precision/Recall/F1**, denoted as P/R/F1. The contact region is defined as the points on the anchor object within 2mm to the interacting object surface as [16]. The contact region in each demo is used as the ground truth to compute the precision, recall, and F1-score of that in the analogous interaction. Larger values of these metrics indicate a more conformable analogy to the demo interaction, especially for the F1-score.

To measure the geometric similarity between demo interaction and analogous interaction, we further compute their **IBS difference** as in the work of [15]. A smaller value of this metric indicates higher similarity between the interaction type in the analogy and the demo interaction.

4.2. Comparison to variants of ContactOpt

To the best of our knowledge, no previous work has studied the exact deformable interaction analogy problem, thus we compare to several variants of the highly relevant work ContactOpt [16]. Given an initialized rough hand-object interaction, ContactOpt first uses a network called DeepContact pre-trained on the ContactPose dataset to predict both the object contact map and hand contact. Then, the hand pose is iteratively optimized to match the current contacts to the target contact maps predicted by DeepContact.

Table 1: Comparison to three variants of ConcactOpt [16].

Method	Penetration ↓		Contact ↑			Inter. ↓
	V	$\|\Delta V\|$	P	R	F1	IBS
ContactOpt	10.0	4.97	0.42	0.51	0.44	7.56
C_RINA_I	8.53	4.19	0.47	0.53	0.47	7.21
C_RINA_C	7.24	3.62	0.48	0.52	0.49	6.96
RINA	6.19	4.77	0.42	0.51	0.46	7.22
DINA	**5.43**	**3.34**	**0.52**	**0.57**	**0.53**	**6.83**

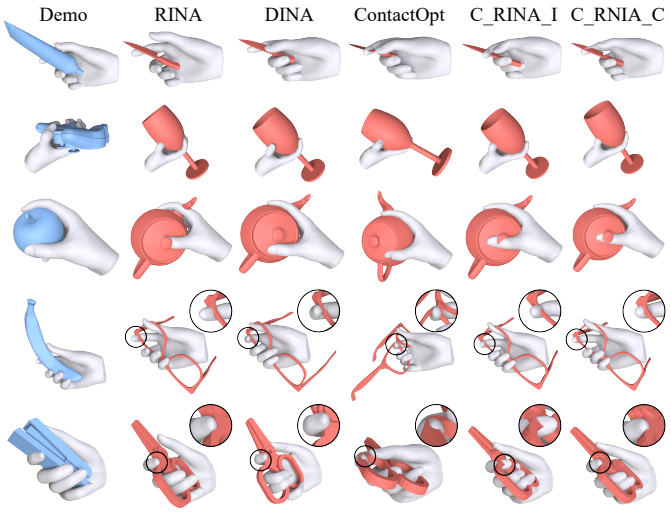


Figure 5: Visual comparison to variants of ContactOpt [16].

Variants of ContactOpt. As one demo interaction is given in the setting of interaction analogy, to let ContactOpt make full use of the demo, we extract the hand contact map from the demo interaction and always use it as the target contact map of the hand during the optimization, while for the target contact map of the target object, we use the one predicted by DeepContact.

Thanks to the two-step optimization, our method can get intermediate result RINA with rigid transformation, which can provide a better initialization for the following deformation for DINA. To test the effectiveness of RINA, we consider another two variants of ContactOpt: 1) C_RINA_I, which takes RINA as the initialized interaction for further hand pose optimization; 2) C_RINA_C, which further uses the target object contact map extracted from RINA instead of network prediction. Table 1 shows quantitative comparisons of our method (DINA) to all those three variants of ContactOpt.

Results. When comparing C_RINA_I to ContactOpt, we can see that based on the RINA initialization, the object contact map predicted by DeepContact has a larger contact area, which encourages the optimization to generate more contact between the hand and the target and leads to the highest recall among

all the methods while the precision is relatively low, as in original DeepContact. Moreover, closer interaction also leads to the highest penetration volume, which is the other reason for the recall increase. When further using the contact map extracted from RINA in C_RINA_C, the contact region is more accurate and consistent with that of the hand, leading to much higher precision and slightly lower recall, which results in an overall higher F1 score with less penetration.

It is interesting to see that the performance becomes better when using more and more information from RINA instead of the network prior, which indicates that information extracted by RINA from the demo is sufficient for deformable interaction analogy without the need to train on a large interaction dataset.

When comparing all the variants of ConcactOpt to DINA, we can see that DINA performs consistently better than the original ContactOpt concerning all the metrics by a large margin, and also beats the other two variants in most of the metrics, including penetration, F1, and IBS difference, thanks to our informative interaction representation and robust optimization.

Figure 5 shows some visual comparisons of results obtained using different methods. For the example shown in the first row, the target object is much smaller than the source object, so the contact region of RINA is relatively small when only rigid transformation is optimized. As a result, variant C_RINA_C that utilizes the most information of RINA cannot form a close interaction, while DINA that performs ITF-guided local deformation forms a tighter grasping. For the examples shown in the second and third row, RINA has clear penetrations due to significant geometry differences between the source and target shapes, which DINA resolves better than all ContactOpt variants. The last two rows show examples with complex topology, and we can see that DINA can better adapt the grasping pose to the topology change. Overall, DINA gets consistently better performance and visual results than all the ContactOpt variants.

4.3. Validation on interaction representation

To validate our interaction representation ITF associated with point-wise distance feature, we compare to several variants using different choices of interaction points, point-wise features. We also perform ablation study on each component of the objective of our interaction analogy.

Variants with different point selection. For the deformable interaction analogy, the interaction representation should be differentiable to the hand deformation, thus, ITF chooses to use a subset of points on the surface of the anchor object that directly participates in the interaction. Three other choices could be: 1) BPS [42], which was used in the work of [17] to guide the one-shot imitation learning of object manipulation; 2) Anchor, which uses all the surficial points of the anchor object regardless of whether each point plays a role in IBS determination; 3) Contact, which uses a pre-defined threshold to identify the contact region on the anchor object as explained in Section 4.1. Note that we take the same threshold (2mm) used for the evaluation, which means that this variant relies on more prior than our method.

Table 2: Quantitative comparison to variants of our method using different point selections and interaction representations.

Method	Penetration ↓		Contact ↑			Inter. ↓
	V	$\lvert\Delta V\rvert$	P	R	F1	IBS
BPS	2.80	4.89	0.34	0.18	0.21	8.98
Anchor	4.40	3.59	0.50	0.42	0.44	7.69
Contact	6.14	3.54	0.50	0.53	0.50	7.10
NDF-only	7.12	5.37	0.47	0.50	0.47	7.87
NDF+Dist	5.47	3.41	0.51	0.48	0.49	7.56
only L_{PEN}	**0.01**	6.85	0.01	0.00	0.00	40.1
only L_{ITF}	12.5	6.56	0.49	**0.59**	0.52	**6.72**
DINA	5.43	**3.34**	**0.52**	0.57	**0.53**	6.83

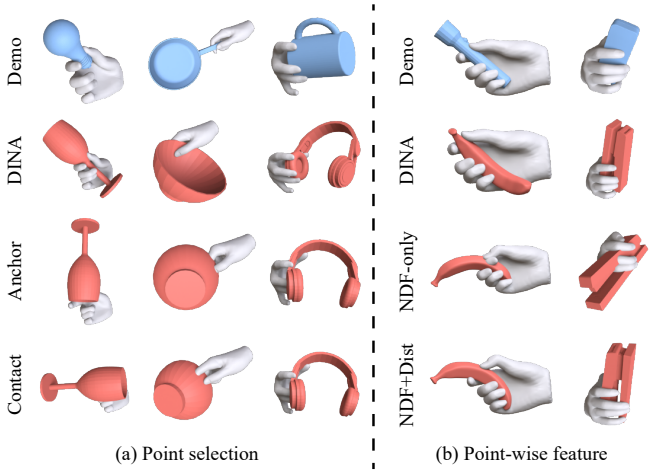


Figure 6: Visual comparison to variants of our method using different point selections (a) and different interaction representations (b).

The first three rows of Table 2 show the results of those three variants. Compared to our method's results shown in the last row, we can see that using BPS or the full anchor leads to lower absolute penetration and a more significant difference to the demo interaction, including lower recall, higher penetration difference, and higher IBS difference. This is because BPS or the whole anchor object over-constrains the optimization with too many unnecessary points, whose distance may become quite different for target objects with significant geometric differences. On the other hand, when using the contact region only, the optimization is under-constrained with too few points, which results in extensive penetration and lower precision, although the recall is relatively higher as it uses the contact region itself to guide the optimization. Some representative visual comparisons are shown in Figure 6(a), where we omit the result of BPS as it has a similar but worse result than Anchor.

Variants with different point-wise features. To deal with the target object given in arbitrary poses, point-wise features should be SE(3)-invariant, and our choice is the simple but effective distance feature. To further justify the use of distance features, we compare to two variants of our method using the SE(3)-invariant point-wise feature relative to the Neural Descriptor Fields (NDF) of the object proposed in the work of [17], which we denoted as NDF. We retrained the NDF network with our

Table 3: User study to score the generated results by geometric plausibility and interaction similarity to the demo interaction.

Method	Geometric plausibility ↑	Interaction similarity ↑
ContactOpt [16]	2.17	1.91
NDF [17]	1.63	1.78
Ours	**2.20**	**2.31**

test objects. So the two variants of our method using different point-wise features are: 1) NDF-only, which replaces the distance feature with NDF; 2) NDF + Dist, which uses both NDF and distance features to guide the optimization.

The fourth and fifth rows of Table 2 show the results of those two variants. When comparing to the variant using NDF only, our method with the distance term only yields consistently better results, with less penetration, while using both distance and NDF features did not show clear improvements. This comparison verifies that with our descriptive ITF points characterizing interactions, simple distance features already provide sufficient and even more accurate guidance for interaction analogy. Moreover, as NDF is a learning-based feature, we choose to use the non-learning-based distance features for robustness and better generality. Figure 6(b) shows some visual comparisons, where we can see that results obtained with the explicit distance term led to less penetration.

Ablation study on losses. The results of an ablation study on losses are shown in the sixth and seventh rows of Table 2. Specifically, these rows present the outcomes obtained when utilizing only L_{PEN} and only L_{ITF} in the interaction analogy optimization. It is evident that L_{ITF} plays a crucial role in generating interactions that closely resemble the demo, whereas L_{PEN} primarily contributes to reducing penetration between the hand and the object. By incorporating both losses in our method's objective function, we achieve a more balanced performance.

4.4. User study

We conducted a user study involving 30 participants who were asked to rank the generated results from our method, ContactOpt [16], and NDF [17] based on their geometric plausibility and similarity to the given demo interaction. Participants assigned scores ranging from 3 to 1, corresponding to the ranks of 1 to 3. The results are presented in Table 3, where our method obtained average scores of 2.20 and 2.31 for geometric plausibility and interaction similarity to the demo, respectively. These results suggest that our generated interactions were perceived to have higher quality according to human perception.

4.5. Qualitative results

Figure 7 shows examples of results where we fix the demo interaction and perform interaction analogy on different target objects. We see that the method can transfer the anchor from the same demo to different target shapes, adapting well to the significant geometry difference, for example, from the demo grasping a leg of the eyeglasses to a light bulb, mug and wineglass, and even to the scissors with totally different topology, as shown in the last row.

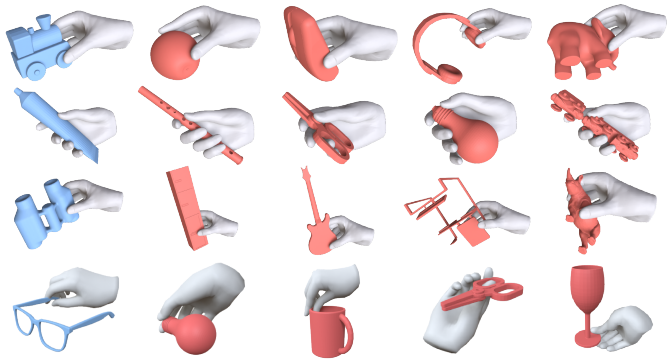


Figure 7: Examples of results where we fix the demo interaction and perform interaction analogy on different target objects.

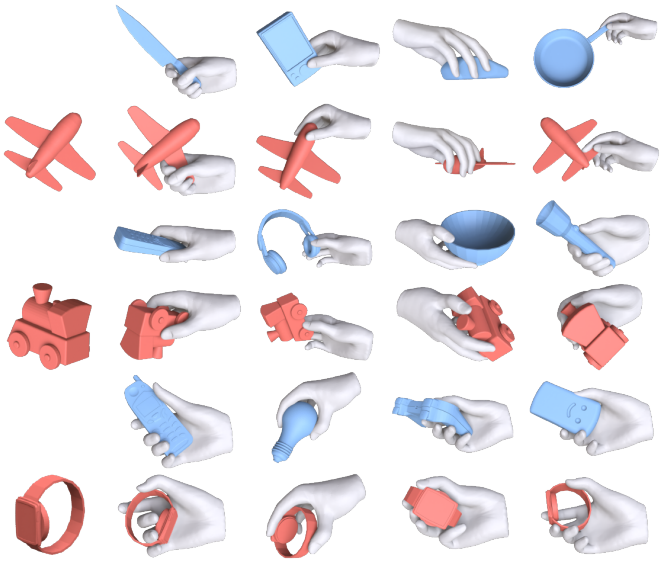


Figure 8: Examples of results where we fix an target shape and transfer the anchor object from different demo interactions.

Figure 8 shows the complementary scenario where we fix a target shape and transfer the anchor object from different demo interactions. We can see that different grasping poses can be generated for the same object when guided by various demos. All analogous grasping poses look pretty natural; see how different airplane model parts are grasped when following different demos.

Other than deformable interaction analogy with hand grasping as a running example, our method can also work on rigid anchor objects for applications like object manipulation and scene generation. Figure 9 shows some visual comparisons of RINA results guided by either ITF or BPS [42]. We can see that analogous interactions guided by ITF are generally more similar to the demo interactions than those guided by BPS, with a more accurate spatial relationship between two interacting objects. This shows that ITF leads to more accurate feature matching than BPS by focusing more on the points directly related to the interaction.

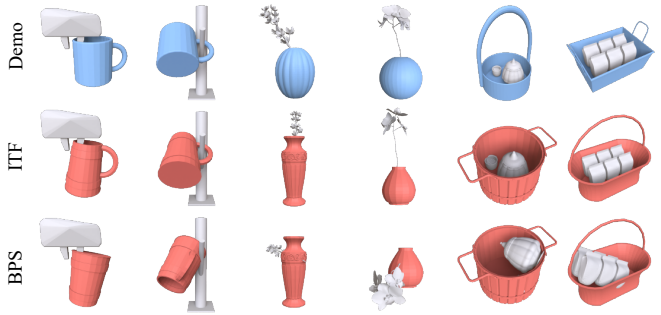


Figure 9: Visual comparison of RINA results guided by either ITF or BPS.

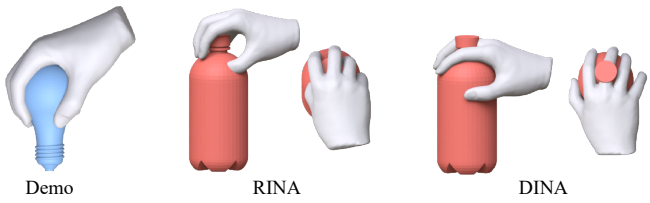


Figure 10: Failure case: hand deformation results in an overall more similar contact but introduces more penetration around the contact region for objects with quite different local geometry.

5. Discussion, Limitation, and Future Work

We introduce DINA, deformable interaction analogy, and a solution that is built on a descriptive interaction representation, i.e., interaction interface (ITF), which consists of a set of points on the anchor object identified by the computation of IBS. ITF together with the distance feature can be used to guide both global pose optimization and local deformation of the anchor object to form an analogous interaction that better resembles the demo interaction. Experiments show that this simple but informative representation performs surprisingly well compared to other options by replacing components with relevant state-of-the-art methods.

Our current method has the following limitations on which we plan to investigate in future works. First, we mainly experimented with DINA on hand grasping currently, which is clearly a prominent application. While the hand is expected to only perform the articulated motion, our ITF-driven optimization makes no assumption or guarantee of piece-wise rigidity on the anchor objects. In the future, it would be interesting to experiment with anchor objects that exhibit general soft-body deformations. Second, object intersections may still occur even with the penetration loss, especially when the demo has a large contact region while the target has a pretty different geometry around the contact region, as the failure example shown in Figure 10. A future improvement is to explore ways to find a better balance in this case. Further, our work focuses on interaction analogy with a single demonstration, and we found that different target objects may favor various demonstrations to form a better interaction. In this aspect, if multiple demonstrations are available, how to select the best one or even combine all of them to provide the optimal guidance for interaction analogy is worth exploring.

References

[1] S. Jörg, Y. Ye, F. Mueller, M. Neff, V. Zordan, Virtual hands in vr: motion capture, synthesis, and perception, in: SIGGRAPH Asia Course, 2020.

[2] M. Höll, M. Oberweger, C. Arth, V. Lepetit, Efficient physics-based implementation for realistic hand-object interaction in virtual reality, in: 2018 IEEE Conference on Virtual Reality and 3D User Interfaces (VR), IEEE, 2018, pp. 175–182.

[3] M.-Y. Wu, P.-W. Ting, Y.-H. Tang, E.-T. Chou, L.-C. Fu, Hand pose estimation in object-interaction based on deep learning for virtual reality applications, Journal of Visual Communication and Image Representation 70 (2020) 102802.

[4] R. Hu, M. Savva, O. van Kaick, Functionality representations and applications for shape analysis, in: Eurographics State-of-the-art Report, 2018.

[5] E. Corona, A. Pumarola, G. Alenya, F. Moreno-Noguer, G. Rogez, Ganhand: Predicting human grasp affordances in multi-object scenes, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 5031–5041.

[6] N. S. Pollard, V. Zordan, Physically based grasping control from example, in: ACM SIGGRAPH/Eurographics symposium on Computer animation, 2005, pp. 311–318.

[7] G. ElKoura, K. Singh, Handrix: animating the human hand, in: ACM SIGGRAPH/Eurographics symposium on Computer animation, 2003, pp. 110—119.

[8] C. K. Liu, Synthesis of interactive hand manipulation, in: ACM SIGGRAPH/Eurographics symposium on Computer animation, 2008, pp. 163—170.

[9] A. Hussein, M. M. Gaber, E. Elyan, C. Jayne, Imitation learning: A survey of learning methods, ACM Computing Survey 50 (2) (2018) 1–35.

[10] T. Yu, C. Finn, A. Xie, S. Dasari, T. Zhang, P. Abbeel, S. Levine, One-shot imitation from observing humans via domain-adaptive meta-learning, arXiv preprint arXiv:1802.01557 (2018).

[11] D. Tzionas, L. Ballan, A. Srikantha, P. Aponte, M. Pollefeys, J. Gall, Capturing hands in action using discriminative salient points and physics simulation, International Journal of Computer Vision 118 (2) (2016) 172–193.

[12] Y. Hasson, G. Varol, D. Tzionas, I. Kalevatykh, M. J. Black, I. Laptev, C. Schmid, Learning joint reconstruction of hands and manipulated objects, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 11807–11816.

[13] Z. Cao, I. Radosavovic, A. Kanazawa, J. Malik, Reconstructing hand-object interactions in the wild, in: ICCV, 2021.

[14] H. Zhang, Y. Zhou, Y. Tian, J.-H. Yong, F. Xu, Single depth view based real-time reconstruction of hand-object interactions, ACM Transactions on Graphics (TOG) 40 (3) (2021) 1–12.

[15] X. Zhao, H. Wang, T. Komura, Indexing 3d scenes using the interaction bisector surface, ACM Trans. Gr. 33 (3) (2014) 1–14.

[16] P. Grady, C. Tang, C. D. Twigg, M. Vo, S. Brahmbhatt, C. C. Kemp, Contactopt: Optimizing contact to improve grasps, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 1471–1481.

[17] A. Simeonov, Y. Du, A. Tagliasacchi, J. B. Tenenbaum, A. Rodriguez, P. Agrawal, V. Sitzmann, Neural descriptor fields: Se (3)-equivariant object representations for manipulation, in: 2022 International Conference on Robotics and Automation (ICRA), IEEE, 2022, pp. 6394–6400.

[18] A. Hertzmann, C. E. Jacobs, N. Oliver, B. Curless, D. H. Salesin, Image analogies, in: SIGGRAPH, 2001, pp. 327–340.

[19] C. Ma, H. Huang, A. Sheffer, E. Kalogerakis, R. Wang, Analogy-driven 3d style transfer, in: Eurographics, 2014.

[20] A. Bernardin, L. Hoyet, A. Mucherino, D. Gonçalves, F. Multon, Normalized euclidean distance matrices for human motion retargeting, in: Proceedings of the 10th International Conference on Motion in Games, 2017, pp. 1–6.

[21] T. Jin, M. Kim, S.-H. Lee, Aura mesh: Motion retargeting to preserve the spatial relationships between skinned characters, in: Computer Graphics Forum, Vol. 37, Wiley Online Library, 2018, pp. 311–320.

[22] J. Zhang, J. Weng, D. Kang, F. Zhao, S. Huang, X. Zhe, L. Bao, Y. Shan, J. Wang, Z. Tu, Skinned motion retargeting with residual perception of motion semantics & geometry, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 13864–13872.

[23] Z. Ye, J. Jia, J. Xing, Semantics2hands: Transferring hand motion semantics between avatars, in: Proceedings of the 31st ACM International Conference on Multimedia, 2023.
[24] R. A. Al-Asqhar, T. Komura, M. G. Choi, Relationship descriptors for interactive motion adaptation, in: Proceedings of the 12th ACM SIGGRAPH/Eurographics Symposium on Computer Animation, 2013, pp. 45–53.
[25] X. Zhao, R. Hu, P. Guerrero, N. Mitra, T. Komura, Relationship templates for creating scene variations, ACM Transactions on Graphics (TOG) 35 (6) (2016) 1–13.
[26] X. Zhao, R. Hu, H. Liu, T. Komura, X. Yang, Localization and completion for 3d object interactions, IEEE transactions on visualization and computer graphics 26 (8) (2019) 2634–2644.
[27] R. Hu, C. Zhu, O. van Kaick, L. Liu, A. Shamir, H. Zhang, Interaction context (icon) towards a geometric functionality descriptor, ACM Transactions on Graphics (TOG) 34 (4) (2015) 1–12.
[28] R. Hu, O. van Kaick, B. Wu, H. Huang, A. Shamir, H. Zhang, Learning how objects function via co-analysis of interactions, ACM Transactions on Graphics (TOG) 35 (4) (2016) 1–13.
[29] S. Pirk, V. Krs, K. Hu, S. D. Rajasekaran, H. Kang, Y. Yoshiyasu, B. Benes, L. J. Guibas, Understanding and exploiting object interaction landscapes, ACM Transactions on Graphics (TOG) 36 (3) (2017) 1–14.
[30] R. Hu, Z. Yan, J. Zhang, O. van Kaick, A. Shamir, H. Zhang, H. Huang, Predictive and generative neural networks for object functionality, ACM Transactions on Graphics 37 (4) (2018).
[31] M. Savva, A. X. Chang, P. Hanrahan, M. Fisher, M. Nießner, Pigraphs: learning interaction snapshots from observations, ACM Transactions on Graphics (TOG) 35 (4) (2016) 1–12.
[32] K. Zhao, S. Wang, Y. Zhang, T. Beeler, S. Tang, Compositional human-scene interaction synthesis with semantic control, in: European Conference on Computer Vision, Springer, 2022, pp. 311–327.
[33] Z. Su, Q. Fan, X. Chen, O. van Kaick, H. Huang, R. Hu, Scene-aware activity program generation with language guidance, ACM Transactions on Graphics (Proceedings of SIGGRAPH ASIA) 42 (6) (2023).
[34] K. Karunratanakul, J. Yang, Y. Zhang, M. J. Black, K. Muandet, S. Tang, Grasping field: Learning implicit representations for human grasps, in: 2020 International Conference on 3D Vision (3DV), IEEE, 2020, pp. 333–344.
[35] H. Jiang, S. Liu, J. Wang, X. Wang, Hand-object contact consistency reasoning for human grasps generation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 11107–11116.
[36] L. Yang, X. Zhan, K. Li, W. Xu, J. Li, C. Lu, Cpf: Learning a contact potential field to model the hand-object interaction, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 11097–11106.
[37] Y.-H. Wu, J. Wang, X. Wang, Learning generalizable dexterous manipulation from human grasp affordance, in: Conference on Robot Learning, PMLR, 2023, pp. 618–629.
[38] B. D. Argall, S. Chernova, M. Veloso, B. Browning, A survey of robot learning from demonstration, Robotics and autonomous systems 57 (5) (2009) 469–483.
[39] S. Song, A. Zeng, J. Lee, T. Funkhouser, Grasping in the wild: Learning 6dof closed-loop grasping from low-cost demonstrations, IEEE Robotics and Automation Letters 5 (3) (2020) 4978–4985.
[40] O. Biza, S. Thompson, K. R. Pagidi, A. Kumar, E. van der Pol, R. Walters, T. Kipf, J.-W. van de Meent, L. L. Wong, R. Platt, One-shot imitation learning via interaction warping, arXiv preprint arXiv:2306.12392 (2023).
[41] E. Chun, Y. Du, A. Simeonov, T. Lozano-Perez, L. Kaelbling, Local neural descriptor fields: Locally conditioned object representations for manipulation, in: 2023 IEEE International Conference on Robotics and Automation (ICRA), 2023, pp. 1830–1836. doi:10.1109/ICRA48891.2023.10160423.
[42] S. Prokudin, C. Lassner, J. Romero, Efficient learning on point clouds with basis point sets, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 4332–4341.
[43] J. Romero, D. Tzionas, M. J. Black, Embodied hands: modeling and capturing hands and bodies together, ACM Transactions on Graphics (TOG) 36 (6) (2017) 1–17.
[44] S. Brahmbhatt, C. Tang, C. D. Twigg, C. C. Kemp, J. Hays, Contactpose: A dataset of grasps with object contact and hand pose, in: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16, Springer, 2020, pp. 361–378.
[45] B. Calli, A. Walsman, A. Singh, S. Srinivasa, P. Abbeel, A. M. Dollar, Benchmarking in manipulation research: Using the yale-cmu-berkeley object and model set, IEEE Robotics & Automation Magazine 22 (3) (2015) 36–52.

Declaration of interests

☒The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: