RealGrasper: Learning Human Hand Grasping from Multi-View Images

CHANDRADEEP POKHARIYA* and ISHAAN N SHAH*, IIIT Hyderabad, India
ANGELA XING, Brown University, United States
KEFAN CHEN, Brown University, United States
AVINASH SHARMA, IIIT Hyderabad, India
SRINATH SRIDHAR, Brown University, United States

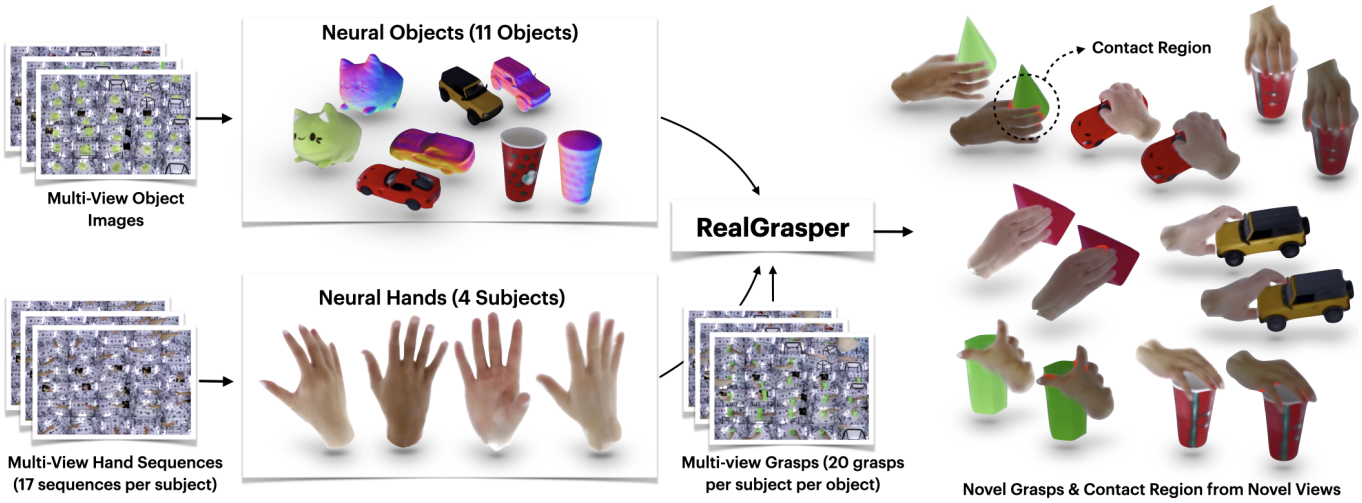

Fig. 1. We present a method to learn a model of human hand grasping directly from real-world multi-view images. First, we introduce RealGrasp, a large 53-view RGB dataset with over 362K frames spanning 12 different objects, 4 subjects, 17 free-hand sequences, and 20 grasps per object per subject. We introduce neural field representations of objects and hands that capture shape and appearance (middle). We show how our dataset and representation can be used to learn a generative grasp model RealGrasper that, when given an object and initial hand pose, estimates the final hand pose of the grasp. RealGrasper generalizes to previously-unseen objects and can visualize shape, appearance, and even contact regions (rightmost, denoted by red regions).

Understanding the way we grasp objects with our hands has important applications in problems ranging from activity recognition to building more dexterous robots. Yet, gaining this understanding has been hard due to challenges in capturing real-world data, the domain gap between simulation and the real world, and building representations that can model appearance, geometry, and contact. In this paper, we show that addressing the dataset and representation challenges can enable us to learn a model of human grasping directly from real-world multi-view images. First, we introduce RealGrasp, a 53-view RGB dataset with over **362K** frames spanning 11 different objects, 4 subjects, 17 free-hand sequences, and 20 grasps per object per subject. We next show how this dataset can help build high-fidelity template-free neural models of hands, objects, and grasps with minimal supervision. Rather than use mesh-like representations that may not faithfully capture appearance and contact properties, our neural models are neural fields that model shape, appearance, and even contact regions from arbitrary viewpoints. Finally, we introduce RealGrasper, a generative model consisting of a conditional variational autoencoder that, when given an initial 3D hand pose and object shape, estimates the final grasp pose of the hand. We show quantitative and qualitative results to evaluate our dataset and representation of grasping model.

CCS Concepts: • **Computing methodologies** → **Computer vision**; *Computer vision representations*; Reconstruction.

Additional Key Words and Phrases: grasping, neural fields, generative model

ACM Reference Format:
Chandradeep Pokhariya, Ishaan N Shah, Angela Xing, Kefan Chen, Avinash Sharma, and Srinath Sridhar. 2024. RealGrasper: Learning Human Hand Grasping from Multi-View Images. 1, 1 (November 2024), 11 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Each day, as we go about our daily lives, we effortlessly grasp more than a hundred different objects [Zuccotti 2015] thousands of times [Zheng et al. 2011]. Grasping, a task so ordinary for humans, remains tremendously difficult for machines as evidenced by its extensive study in robotics [Bicchi and Kumar 2000] and computer vision [Erol et al. 2007]. Understanding human grasping

*Both authors contributed equally to this research.

Authors' addresses: Chandradeep Pokhariya; Ishaan N Shah, IIIT Hyderabad, India; Angela Xing, Brown University, United States; Kefan Chen, Brown University, United States; Avinash Sharma, IIIT Hyderabad, India; Srinath Sridhar, Brown University, United States.

XXXX-XXXX/2024/11-ART $15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

has important applications for instance in robotics, mixed reality, and activity recognition. However, progress has been limited by challenges in capturing rich real-world grasping data and building suitable representations to capture hands, objects, and grasps.

Because of these challenges, previous work has resorted to simulation [Lundell et al. 2021a,b; Miller and Allen 2004; Turpin et al. 2022b; Ye and Liu 2012] as a way to model human grasps. In simulation, physical laws and heuristics are used to model the components of grasping including contact forces, friction, mass, and gravity. Inevitably, modeling every source of physical variation is difficult resulting in a domain gap between simulation and the real world [Bousmalis et al. 2018]. Some methods combine known physical constraints (e.g., contact) with real-world observations but: methods that use observations from markers can hinder free hand motion [DelPreto et al. 2022; Jiang et al. 2021; Taheri et al. 2020a; Zhang et al. 2021] while methods that operate on images use representations like parametric hand models [Cao et al. 2021; Hasson et al. 2019] that lack the expressive capability to easily model 2D hand boundaries, surface, and contacts. Furthermore, existing real-world grasping datasets [Brahmbhatt et al. 2019; Taheri et al. 2020b] have been limited to providing coarse 3D hand pose, use specially designed or instrumented objects, and do not enable modeling of both the **appearance and geometry** of grasps.

We show that addressing the dataset and representation challenges can enable us to learn a model of human hand grasping directly from real-world multi-view images. To this end, we present **RealGrasp**, a new 53-view RGB dataset with over **362K** image frames: multiple views of **11 different objects**, 17 multi-view videos of free hand articulation across **4 subjects**, and multiple views of **20 different grasps** on each of the objects and subjects – all captured without any special markers or sensors. We use this dataset to build high-fidelity neural hand and object models with minimal supervision (only 3D camera poses and hand poses obtained from off-the-shelf methods). Rather than use meshes as representations of the hand and object, we build on the latest advances in neural fields, specifically neural shape [Yariv et al. 2021] and articulating radiance fields [Li et al. 2022; Mildenhall et al. 2020]. Our **template-free** high-fidelity neural hand and object models learn appearance, geometry, and can model the **contact regions** of the grasp.

The neural hand and object models are then used to learn **RealGrasper**, a generative model consisting of a conditional variational autoencoder (CVAE) that, when given an initial 3D hand pose and an object model, estimates a final hand pose representing a plausible grasp of the object. During training, RealGrasper is only trained on multi-view RGB images and derived 3D hand poses of grasps without any other supervision. Our method produces novel natural grasps and can visualize the appearance and geometry of grasps from arbitrary viewpoints as shown in Figure 1.

We quantitatively evaluate our dataset and representation, and justify key design choices in Section 5. To our knowledge, ours is the first method to use neural fields to model grasping on real data making comparison with other methods challenging – but we provide some comparisons [Karunratanakul et al. 2020]. To sum up our contributions:

- **RealGrasp**, a large real-world 53-view RGB dataset (which we will release publicly) with over **362K** frames captured across 11 objects, 4 subjects, 17 free-hand sequences, and 20 grasps per object per subject.
- We model shape, appearance, and contact of hand–object grasping with **neural representations** faithfully to images.
- **RealGrasper**, a generative CVAE that learns to synthesize photorealistic grasping given an object and initial 3D hand pose from multi-view images.

2 RELATED WORK

In this review of related work, we focus on datasets, grasp simulation, perception of hand-object interactions, and representations.

Datasets: Datasets for human grasps are challenging to obtain because they need specialized hardware, extensive human annotation, and significant post-processing to make them useful. Some datasets use markers or special gloves to track the hand and object [Bernardin et al. 2005; DelPreto et al. 2022; Garcia-Hernando et al. 2018; Taheri et al. 2020a] but this hinders natural hand motion and introduces changes in image appearance. Therefore, work has focused on manual annotations [Ballan et al. 2012; Bullock et al. 2015; Rogez et al. 2015; Sridhar et al. 2016], optimization [Hampali et al. 2020], or automatic annotation [Simon et al. 2017] from RGB or depth. Many of these datasets are limited to only 3D hand poses and lack information about hand surface and contacts. Synthetic datasets [Hasson et al. 2019; Mueller et al. 2018, 2017] suffer from a domain gap that makes it challenging to generalize to real data. Other datasets like InterHand2.6M [Moon et al. 2020; Zimmermann et al. 2019] are limited to hand only without any objects, while others [Shan et al. 2020] focus on 2D understanding only.

ContactDB [Brahmbhatt et al. 2019] and ContactPose [Brahmbhatt et al. 2020] aim to address these limitations, and focus on scaling to many users and objects. While ContactDB is captured using thermal imaging, ContactPose uses multi-view RGB-D data. Both methods are limited to 3D hand poses only, objects are not real, and do not have sufficient views to support neural field representations. In this paper, we focus on providing a high-quality dataset with sufficient views to support neural field representations, enable capture of both appearance and geometry, and enable grasping models (including contact) to be trained from multi-view images.

Simulation for Grasping: Due to challenges in capturing real data, there has been extensive work on using simulation for modeling human grasps. GraspIt! [Miller and Allen 2004] is one of the most widely used methods and uses hand-designed heuristics and physics to obtain a final hand grasp pose given an object and initial hand pose. More recent multi-finger grasp simulation methods rely on analytic methods [Lundell et al. 2021a,b; Shao et al. 2020; Ye and Liu 2012] and can be used with a human hand model. Recently, D-Grasp [Christen et al. 2022] introduced a reinforcement learning method for dynamics grasp syenthesis, and Grasp'D [Turpin et al. 2022b] introduced differentiable simulation for grasping. ManipHand [Zhang et al. 2021] combines marker-based motion capture with a learning-based approach to synthesize manipulations of objects. All of the above methods suffer from a domain gap to real data [Bousmalis et al. 2018].

Perception for Grasping: Simultaneously in computer vision, significant work has studied capturing hands interacting with objects [Ballan et al. 2012; Hamer et al. 2010, 2009; Hasson et al. 2019; Karunratanakul et al. 2020; Romero et al. 2010; Tsoli and Argyros 2018; Tzionas and Gall 2015]. Several methods combine perception with physical constraints (proximity, contact, forces) during optimization or learning [Jiang et al. 2021; Oikonomidis et al. 2011; Pham et al. 2017; Sridhar et al. 2016; Tse et al. 2022]. To make hand shape and pose estimation easier, parametric hand models have been developed, notably MANO [Romero et al. 2017] and Total Capture [Joo et al. 2018], which are used by several methods for pose estimation [Cao et al. 2021]. However, parametric/template hand models cannot capture hand boundaries well resulting in a mismatch between shape and appearance. On the contrary, we propose to use template-free methods for obtaining hand and object models with better shape–appearance alignment.

Representations: Recent advancements in coordinate-based neural networks, or neural fields [Xie et al. 2022], have shown great success in encoding the geometry [Chen and Zhang 2019; Mescheder et al. 2019; Park et al. 2019] and appearance [Lombardi et al. 2019; Mildenhall et al. 2020; Sitzmann et al. 2019]. For example, neural radiance field (NeRF) [Mildenhall et al. 2020] uses an MLP to model the density and color and achieves photorealistic novel view synthesis. VolSDF [Yariv et al. 2021] and NeuS [Wang et al. 2021] improve the geometry representation and reconstruction of NeRF by deriving the density from a signed distance function (SDF) representing the distance to the closest surface of the scene geometry. Instant-NGP [Müller et al. 2022], Plenoxels [Yu et al. 2021], and TensoRF [Chen et al. 2022] greatly reduce the cost of building NeRF models. Many approaches also explore articulated neural fields to model dynamic human body [Li et al. 2022; Liu et al. 2021; Peng et al. 2021a,b; Weng et al. 2022]. LISA [Corona et al. 2022] proposes an implicit hand model, but code and datasets are not publicly available. TAVA [Li et al. 2022] proposes a template-free animatable neural representation for dynamic actors (e.g., human bodies), which is robust to unseen poses. We show how neural field representations, specifically TAVA [Li et al. 2022] and VolSDF [Yariv et al. 2021], can be used to build a neural representation of grasps from real data.

3 REALGRASP DATASET

We first describe our RealGrasp dataset, in particular, the hardware capture system, capture protocol, and annotation. The RealGrasp dataset was driven by three key considerations: (1) capture hands interacting with objects without any markers or special sensors like depth or thermal cameras, (2) capture both the appearance and geometry of grasps, and (3) support neural shape and radiance fields as representations for learning grasping. Achieving this goal from purely RGB videos requires a multi-view capture system with known camera poses. Many prior datasets (see Section 2) contain multi-view images or video of hand grasps [Hampali et al. 2020; Simon et al. 2017; Taheri et al. 2020a], but none have the large number of views needed to support neural field representations or are limited to hands only [Moon et al. 2020]. Thus, we built a custom system to capture a large 53-view real-world dataset of hand grasps.

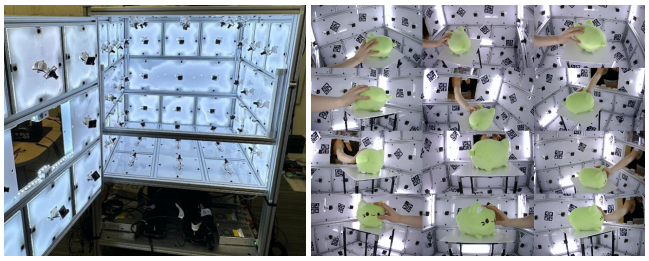

Fig. 2. (Left) Our data capture system where hands, objects, and grasps are captured by 53 cameras. (Right) Sample grasp frame from 12 out of the 53 views in our dataset.

Data Capture System: The data capture setup is shown in Figure 2 (left). It consists of 53 RGB cameras uniformly located inside a cubical capture volume with each cube face consisting of 9 cameras. The sides of the cube are illuminated evenly using LED lights with additional edge lights. Each RGB camera records at 120 FPS with a resolution of 1280×720. This system capture both static (for objects) and dynamic scenes (for hands and grasps). The cameras are software synchronized with a frame misalignment of no more than 3 ms. The multi-view system is calibrated for camera intrinsics and extrinsics using COLMAP [Schönberger and Frahm 2016; Schönberger et al. 2016] with fiducial markers on the walls.

RealGrasp Dataset: RealGrasp is a large real-world multi-view RGB dataset of hands grasping natural objects that we will publicly release. It contains **362K** image frames: multiple views of **11 different objects**, 17 multi-view video sequences of free hand articulation with **4 subjects**, and multiple views of **20 different grasps** on each of the 11 objects for all 4 subjects. Of the total frames, we use 360K to create neural hands, 636 frames for neural objects, and 1920 frames for grasp learning. Our goal is not to compete with existing datasets on quantity, but instead we focus on enabling the use of neural field representations for grasps. Figure 2 (right) shows some example data from our dataset.

Data Capture Protocol: Our capture protocol consists of four steps. First, we capture a sequence of an empty scene for camera calibration and background subtraction. Next, we collect multi-view videos of hands to build neural hand models (see Section 4.1), subjects reach their right hand into the center of the box and move their hand in different motions. Then, we collect static multi-view images of objects for neural object models (see Section 4.1). Finally, we record multi-view images of our subject's hand grasping the object in 20 different ways.

Automatic Annotation: The appearance and geometry of hand grasps are automatically extracted by our method as described in Section 4. Apart from this, we also provide 2D and 3D hand joint locations which we obtain from OpenPose [Simon et al. 2017] followed by 3D triangulation over the multi-view hand sequence. Then, we use inverse kinematics optimization [Sridhar et al. 2013] to obtain the joint angles and global orientation of the hand skeleton by optimizing them using gradient descent. We impose constraints to limit the degrees of freedom and joint angles for the rotation of the bones as described in Figure 2 in the supplement. To achieve temporal smoothness for the sequence, we apply the 1€ Filter [Casiez et al. 2012] on the estimated parameters.

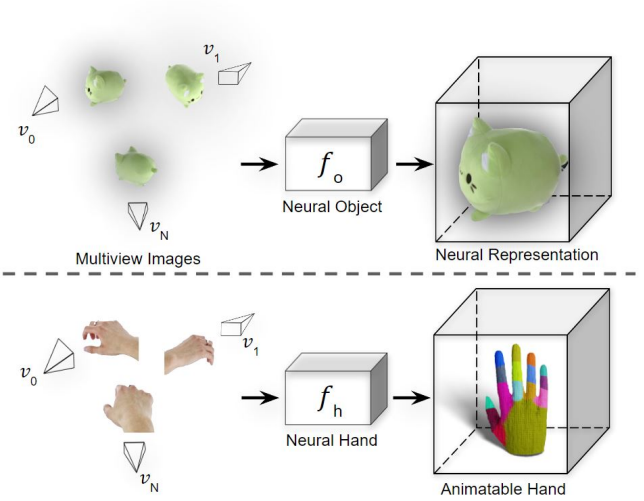

Fig. 3. To learn a neural representation of hand-object interaction, we train a neural radiance field of static objects using VolSDF [Yariv et al. 2021] and a dynamic neural hand model using TAVA [Li et al. 2022] from RealGrasp dataset captured by our multiview camera system.

To segment the hand and object from the background, we use a combination of both traditional and learning-based background subtraction methods [Cheng et al. 2020; Lin et al. 2020]. However, segmenting objects using the above methods fails in many cases due to complex object texture, so we use PhotoRoom [PhotoRoom 2023], a commercial application. To ensure segmentations are consistent across views, we first train InstantNGP [Müller et al. 2022] and extract an alpha mask.

4 METHOD

We aim to accurately capture hands, objects, and grasps from real-world observations with the ultimate goal of generating human grasps that are natural and realistic in terms of appearance, shape, and physical contact. Prior work in grasping has struggled to faithfully capture the appearance of hand-object interaction due to the limitations of commonly used mesh representations. We address this issue by leveraging neural shape and appearance fields to represent visual details of hands and objects as described in Section 4.1. Our neural representations are learned with only minimal supervision for 3D hand and camera poses obtained using off-the-shelf methods [Schönberger and Frahm 2016; Simon et al. 2017].

In Section 4.2, we introduce our generative model RealGrasper, a conditional variational autoencoder (CVAE) [Sohn et al. 2015] built on our neural representations to synthesize natural hand grasps given the encoded shape knowledge of the target object and initial hand pose. RealGrasper is trained solely on multi-view videos of hands grasping objects without any other supervision to generate high-quality photorealistic renderings of human grasps. In Section 4.3, we explain how the losses used to train RealGrasper.

4.1 Neural Grasp Representation

Prior works in hand-object interaction (see Section 2) heavily rely on mesh representations of object or parametric models such as MANO [Romero et al. 2017]. Due to the low dimensional nature of these template meshes, they can result in misalignment when fit to images, which adversely affects the estimation of hand-object contact and prevents the extensive study of real-world human grasping. Thus, it is important to use a representation that can reconstruct the appearance and geometry of hands and objects faithfully and avoid image misalignments. Inspired by the recent success of neural fields, we build an object representation upon VolSDF [Yariv et al. 2021] and a hand representation based on TAVA [Li et al. 2022] to learn a neural representation of hand grasps. These representations are derived from Neural Radiance Fields (NeRF) [Mildenhall et al. 2020] which encodes the geometry and view-dependent appearance of a scene as a continuous field of radiance $\mathbf{c}(\mathbf{x}, \mathbf{v})$ and volume density $d(\mathbf{x})$ using a multi-layer perceptron (MLP) $f : (\mathbf{x}, \mathbf{v}) \rightarrow (\mathbf{c}, d)$ where $\mathbf{x} \in \mathbb{R}^3$ is a 3D point and $\mathbf{v} \in \mathbb{R}^3$ is the corresponding viewing direction. The radiance field along each camera ray $\mathbf{r}$ is given as:

$$\mathbf{C}(\mathbf{r}) = \sum_{i=1}^{N} T_i(1 - exp(-d_i\delta_i))\mathbf{c}_i, \text{where } T_i = exp(-\sum_{j=1}^{i-1} d_i\delta_i), \quad (1)$$

δ_i is the distance between adjacent sample points on the camera ray and T_i denotes transmittance.

Neural Object Representation: To accurately represent the shape and appearance of objects, we train a VolSDF [Yariv et al. 2021] model from multi-view images. VolSDF is a neural field that models the volume density d as the Laplace's cumulative distribution function Ψ of a learnable signed distance function (SDF): $d(\mathbf{x}) = k\Psi(-SDF(\mathbf{x}))$. The zero-level set of the SDF defines the shape of object's surface. Such a formulation of density improves the geometry reconstruction compared with vanilla NeRF representation. To render the VolSDF model of the object, we follow Equation (1), but the radiance $\mathbf{c}(\mathbf{x}, \mathbf{v}, \mathbf{n})$ is also dependent on the level set's normal $\mathbf{n}(\mathbf{x}) = \nabla_{\mathbf{x}} SDF(\mathbf{x})$. This representation allows us to synthesize realistic novel views as well as reconstruct the object shape of the object with high fidelity and use it for grasp generation.

Neural Hand Representation: Unlike static objects, hands exhibit complex articulated poses and neural representation adopted for objects lacks the capacity to model dynamic, deformable scenes and articulated actors. To learn a neural representation of the hand model, we need the ability to animate the hand and generalize to out-of-distribution poses unseen during training. Hence, we adopt TAVA [Li et al. 2022], a template-free animatable neural radiance field that allows us to drive the hand model given novel poses at test time. The Neural Hand representation consists of a Lambertian neural radiance field that represents the shape and appearance of hands, and a neural blend skinning function to animate hands. The neural radiance field adopts Mip-NeRF [Barron et al. 2022]. The neural skinning function predicts skinning weights at each 3D points to blend all bone transformations using forward LBS-based deformation. We can render the deformed hand using Equation (1) after finding the radiance $\mathbf{c}(\mathbf{x}, \mathbf{v})$ and density $d(\mathbf{x})$ of the sampled points in their canonical space via inverse skinning. The Neural

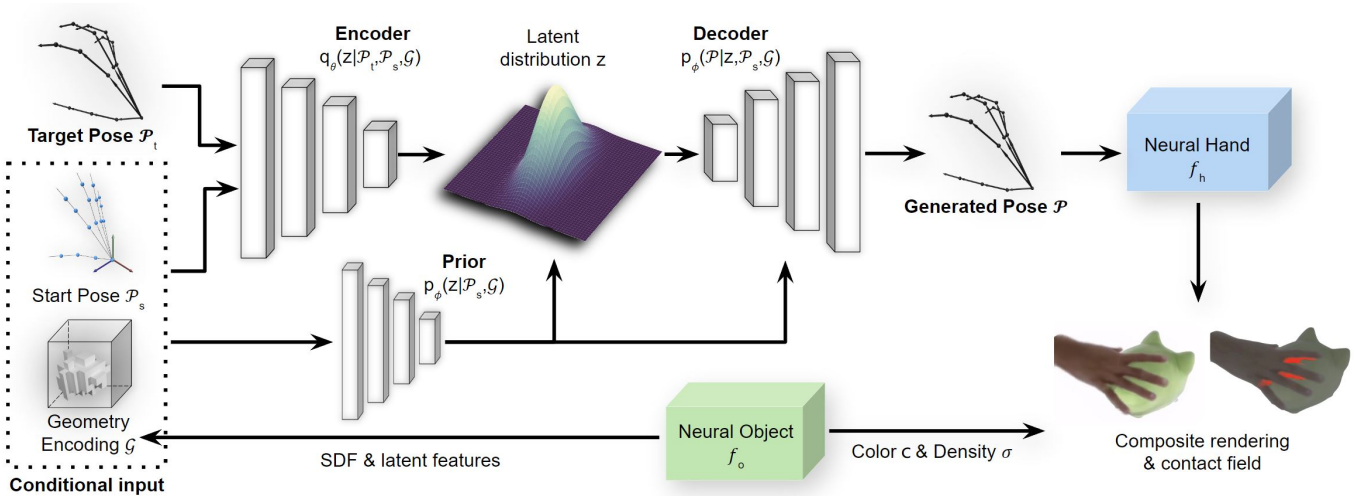

Fig. 4. **RealGrasper architecture.** We build on top of the CVAE model introduced in [Rempe et al. 2021]. Given the start pose $\mathcal{P}_s$, geometry encoding $\mathcal{G}$ of the target object from Neural Object model, and the ground truth target pose $\mathcal{P}_t$, the decoder reconstructs the hand pose by sampling from the estimated posterior from encoder during training. At inference, we can sample from the learned prior and generate novel grasps from decoder given the conditional input $\mathcal{P}_s$ and $\mathcal{G}$. The generated poses $\mathcal{P}$ can then be used to synthesize novel views of realistic hand grasps and contact regions (red).

Hand model are trained solely on multi-view images and 3D hand poses obtained from off-the-shelf methods [Simon et al. 2017].

Composite Grasp Representation: To model grasps, we combine the Neural Object and Neural Hand models in a photorealistic way. To combine the two neural fields, we use an additive composition of Equation (1) same as in [Wu et al. 2022]:

$$\mathbf{C}(\mathbf{r}) = \sum_{i=1}^{N} T_i((1 - exp(-d_i^o \delta_i))\mathbf{c}_i^o + (1 - exp(-d_i^h \delta_i))\mathbf{c}_i^h)$$
$$T_i = exp(-\sum_{j=1}^{i-1}(d_i^o \delta_i + d_i^h \delta_i)), \tag{2}$$

where $\mathbf{c}_i^o$, d_i^o denotes radiance and density of the object and $\mathbf{c}_i^h$, d_i^h denotes radiance and density of the hand. In the end, we can synthesize a photorealistic rendering of the hand grasping the object given any camera viewpoint.

Contact Region Reasoning: The neural field representation also allows us to extract the contact field between the object and the hand. Intuitively, contact is likely to occur in the vicinity of the object surface where the hand volume density is high. Thus, we query the volume density of Neural Hand at the sampled points close to the zero level set of the object SDF. If the density of a part of the hand is above a high threshold at regions in close proximity to the surface of the object, we set a positive mask for those parts of the hand and the contact field can be visualized using the mask which is visualized in red in Figure 5.

4.2 Neural Grasp Generation

The goal of RealGrasper is to synthesize novel grasps for target objects. By utilizing a generative model that learns to reconstruct a target hand pose from multi-view images of hands grasps given the target object and starting hand pose, RealGrasper is able to produce plausible grasps without any additional supervision. Built on Neural Grasp representation introduced in Section 4.1, our model can synthesize high-quality appearance and geometry of novel grasps from arbitrary viewpoints and model the contact between the hand and the object. Inspired by [Rempe et al. 2021], we adopt a conditional variational autoencoder (CVAE) [Sohn et al. 2015] architecture which formulates grasp reconstruction $p_\phi(\mathcal{P}_t|\mathcal{P}_s, \mathcal{G})$ as a latent variable model as shown in Figure 4.

Input Parameterization: We take the 21-joint hand skeleton in OpenPose [Simon et al. 2017] and parameterize the hand pose $\mathcal{P} = [\mathbf{t}\ \mathbf{r}\ \mathcal{J}\ \Phi]$ as the translation $\mathbf{t} \in \mathbb{R}^3$ and rotation $\mathbf{r} \in \mathbb{R}^3$ of the root joint together with joint positions $\mathcal{J} \in \mathbb{R}^{20\times3}$ and bone rotation $\Phi \in \mathbb{R}^{23}$ for the rest of the joints relative to the root joint. Since the hand skeleton is subject to a kinematic structure and a certain range of configurations, we can limit the degrees of freedom for the rotation of bones (see supplementary). The object geometry encoding $\mathcal{G} = [SDF(\mathcal{J})\ \mathcal{S}]$ consists of the SDF queried at the joint locations $SDF(\mathcal{J}) \in \mathbb{R}^{20}$, and shape features $\mathcal{S} \in \mathbb{R}^{32}$ on the object mesh extracted from a pretrained encoder [Peng et al. 2020].

Conditional Prior: We first learn a conditional prior from which the latent variables $\mathbf{z} \in \mathbb{R}^{24}$ represent the possible grasping motion transitions from the starting pose to the target pose on the object:

$$p_\phi(\mathbf{z}|\mathcal{P}_s, \mathcal{G}) = \mathcal{N}(\mathbf{z}; \mu_\phi, \sigma_\phi),$$

which parameterizes a Normal distribution with mean μ_ϕ and variance μ_ϕ. estimated by an MLP. Intuitively, the distribution of possible grasping motion could vary given different starting poses and objects. Thus, explicitly learning the prior helps the CVAE to generalize to diverse grasps and stabilize the training in our experiments.

Encoder and Decoder: The encoder learns the approximate posterior for training and parameterizes a Gaussian distribution

$$q_\theta(\mathbf{z}|\mathcal{P}_t, \mathcal{P}_s, \mathcal{G}) = \mathcal{N}(\mathbf{z}; \mu_\theta, \sigma_\theta).$$

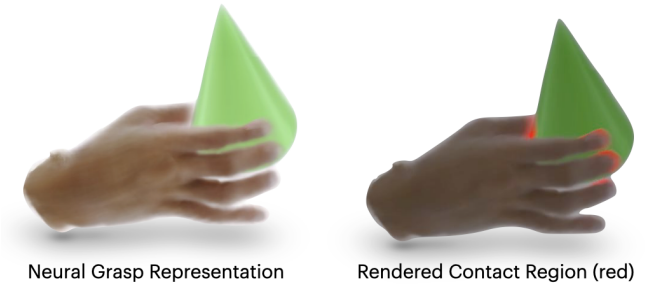

Fig. 5. **Contact Regions**. The red highlighted part of the fingers grasping the cone indicates close contact where the signed distance of the hand volume to the zero level set of the object is smaller than a threshold.

Fig. 6. **Contact Loss Ablation**. The left image shows the generated grasp when no contact loss is used. The right image shows a more plausible grasp generated when applying the contact loss demonstrating the effectiveness of the contact loss as a regularizer to discourage penetration.

We use KL divergence loss to regularizes the posterior to be near the prior (see Section 4.3). Conditioned on the latent variable $\mathbf{z}$ sampled from the encoder posterior, the starting hand pose and object geometry $\{\mathcal{P}_s, \mathcal{G}\}$, the decoder reconstructs the target pose $\mathcal{P}_t$ during training thereby learning the likelihood $p_\phi(\mathcal{P}_t|z, \mathcal{P}_s, \mathcal{G})$.

Generating Novel Grasps: At inference, the decoder takes the concatenation of the conditional input $\{\mathcal{P}_s, \mathcal{G}\}$ and sampled $\mathbf{z}$ from the learned prior $p_\phi(\mathbf{z}|\mathcal{P}_\mathbf{s}, \mathcal{G})$ to generate novel grasping poses $\mathcal{P}$. The generated grasping hand pose can be applied to synthesize photorealistic hand using our trained Neural Hand model. Since Neural Object and Neural Hand use radiance field representation, we can jointly render them using volumetric rendering in Equation (2) (see Section 4.1). During training, we can optimize the rendering loss between the synthesized composite hand-object image and the captured image, as explained in the next section.

4.3 Model Training and Loss

We first train Neural Object and Neural Hand following the original training setup in VolSDF [Yariv et al. 2021] and TAVA [Li et al. 2022] using multi-view images of objects and hands in our RealGrasp dataset. We then train RealGrasper with multi-view grasp images and hand poses to reconstruct grasp poses given initial hand pose and object encodings. The variational lower bound of the CVAE [Sohn et al. 2015] is optimized while Neural Object and Neural Hand are fixed:

$$\begin{aligned}\log p_\phi(\mathcal{P}_t|\mathcal{P}_s, \mathcal{G}) \geq \mathrm{E}[\log p_\phi(\mathcal{P}_t|z, \mathcal{P}_s, \mathcal{G})] \\ -KL(q_\theta(\mathbf{z}|\mathcal{P}_\mathbf{t}, \mathcal{P}_\mathbf{s}, \mathcal{G})\|p_\phi(\mathbf{z}|\mathcal{P}_\mathbf{s}, \mathcal{G})),\end{aligned} \tag{3}$$

where the first term measures the reconstruction error $L_\mathcal{J}$ of the decoder and the KL divergence $L_{\mathcal{KL}}$ regularize the posterior distribution approximated by the encoder to be close to the prior. In addition, we impose a rendering loss L_{rgb} on the composited image of the generated hand grasp, and a regularization loss to encourage physically plausible contact $L_{contact}$. In summary, our training loss consists of four terms:

$$L = L_\mathcal{J} + \alpha L_{\mathcal{KL}} + \beta L_{rgb} + \lambda L_{contact}, \tag{4}$$

where α, β, λ are hyperparameters to balance the loss terms. The primary objective of the model during training is to minimize the reconstruction error between the joint locations of the target hand $\mathcal{J}_t$ and the generated hand $\hat{\mathcal{J}}$:

$$L_\mathcal{J} = \|\mathcal{J}_t - \hat{\mathcal{J}}\|^2. \tag{5}$$

We encourage the posterior distribution to be close to the estimated prior distribution by minimizing the KL divergence:

$$L_{\mathcal{KL}} = KL(\mathcal{N}(\mathbf{z}; \mu_\theta, \sigma_\theta)\|\mathcal{N}(\mathbf{z}; \mu_\phi, \sigma_\phi)). \tag{6}$$

The volumetric rendering loss L_{rgb} between the final synthesized image and the ground truth image is similar to the photometric loss in NeRF [Mildenhall et al. 2020]:

$$L_{rgb} = \sum_\mathbf{r} \|\mathbf{C}(\mathbf{r}) - \hat{\mathbf{C}}(\mathbf{r})\|. \tag{7}$$

To avoid penetrating the object, we add a soft regularization to penalize negative distance to the object surface at hand joints:

$$L_{contact} = -\sum_{\mathbf{x}\in\mathcal{J}} (1 + \exp^{-k\cdot SDF(\mathbf{x})})^{-1}, \tag{8}$$

where k is a hyperparameter set to 20.

5 EXPERIMENTS & RESULTS

In this section, we show results and evaluate various components of our method including the quality of our dataset, representation, grasping, and ablate on design choices.

Implementation Details: Our model is implemented in PyTorch Lightning. The CVAE of RealGrasper is implemented using MLPs with 8 layers and Leaky ReLU as activation. We use ADAM optimizer with a learning rate of $5e^{-4}$ and batch size of 1 with gradient accumulation of 8 batches for training on a single RTX 2080Ti GPU. We train Neural Hand for each subject on four Tesla V100 for 72 hours and Neural Object on one V100 for 16 hours. We set the loss weights $\alpha = 1.0, \beta = 1.0, \lambda = 1.0$. To train the Neural Grasp, we consider 9 objects for training and 2 objects (prism and car1) for test split out of total 11 objects.

Dataset Quality: The hand pose estimation using OpenPose sometimes yields incorrect or missing keypoints for certain frames. We filter invalid frames by checking if the tracked hand skeleton is complete. After filtering, 95% of the frames in our dataset are reliable. Finally, we perform histogram equalization to improve the image contrast.

	Hands			Objects										
	subject1	subject2	subject3	cat	dog	couch	cup	table	car1	car2	car3	pyramid	prism	cone
PSNR↑	23.81	25.04	24.38	35.93	32.71	23.76	20.81	29.92	24.65	23.19	23.57	37.21	37.68	37.25
SSIM↑	0.87	0.81	0.78	0.95	0.94	0.79	0.82	0.91	0.84	0.87	0.87	0.97	0.98	0.97

Table 1. We report PSNR and SSIM as a measure of visual appearance quality of our representations Neural Hand and Neural Object on all subjects of hands and objects in our RealGrasp dataset. The higher the score the better the image quality.

	Training MPJPE↓			Test MPJPE↓	
	car2	cup	cone	car1	prism
w/ shape encodings	**2.23**	**1.98**	**0.55**	**4.53**	**3.64**
w/o shape encodings	3.20	4.85	2.56	4.95	4.52

Table 2. We show that the MPJPE errors of grasp pose reconstruction without shape encodings is consistently greater on both training and test objects. This indicates that shape encoding is critical to our model. We multiplied the error numbers by 1000 to adjust the scale.

5.1 Representation Evaluation

In this section, we evaluate the quality of our neural field representations of the hand, object and grasps.

Quantitative Evaluation: We measure the visual quality of our neural representation using PSNR and SSIM metrics (higher is better) on our RealGrasp dataset (see Table 1). For Neural Object and Neural Hand, we render five novel views and report the average metric value for all 3 hand subjects and 11 objects. Our PSNR quality is consistently over 25, with a few objects yielding lower PSNR/SSIM scores due to theur small size. To our knowledge, ours is the first neural field-based representation for grasping, thus there are no other methods we can compare against.

Qualitative Evaluation: We perform qualitative evaluation and visualize renderings of the novel grasp poses with varying novel views across multiple objects and associated contact field as well as rendering of same object grasps with different neural hand representations learnt on different individuals, as shown in Figure 7. Our proposed representation is able to produce realistic novel views of grasps for various combinations of hands and objects.

5.2 Comparison to Previous Work

To our knowledge, we are first method to model grasping using neural fields from real multi-view image data. However, this makes it challenging to directly compare with previous work. Works such as Grasp'D [Turpin et al. 2022a] and D-Grasp [Christen et al. 2022] are based upon physics simulation while others use parametric models [Cao et al. 2021; Jiang et al. 2021]. Furthermore, these methods work by taking object's geometry as their input and do not model appearance. On the other hand, our method can be trained and evaluated by considering only images as input. We therefore provide qualitative comparisons with GraspingFields [Karunratanakul et al. 2020] to show the difference in the quality of grasps and renderings.

Qualitative Comparisons: We show qualitative results with Grasping Field on two objects car and couch as shown in Figure 8. Both of the methods are able to grasp the object similarly, but our method produces better contact compared to Grasping Fields. Additionally, our method models appearance, can generate superior photorealistic composite rendering of the grasp from arbitrary viewpoints, and implicitly extracts contact regions.

5.3 Ablation Study

We ablate on different components of our method, in particular, contact loss and the need for shape encodings.

Effect of Contact Loss: We perform an ablation on the impact of contact loss on the performance of our results. Contact loss acts as a regularizer in the total loss term which penalizes the network to make predictions inside the object mesh. Specifically, contact loss acts as a soft regularizer if the SDF of the object at the joint bone locations is negative. We show in Figure 6 that the contact loss improves the physical contact between hand and object.

Effect of Shape Encodings: To condition our CVAE on the shape information about the object shape, we use [Peng et al. 2020] to encode the shape. This allows us to integrate local information and incorporate translational equivariance in the form of shape encodings. To test our hypothesis, we do two experiments one with shape encodings and another without shape encodings and report Mean Per Joint Position Error (MPJPE) on both training and test objects as shown in Table 2. The MPJPE measures Euclidean error averaged over all hand joints. Results show that shape encoding is essential to provide knowledge of the object shape to help training the CVAE.

6 CONCLUSION

In this paper, we addressed the dataset and representational challenges in understanding human grasping from multi-view video. We introduced RealGrasp, a large frame multi-view RGB dataset designed to support neural field representations to model grasping. It consists of over **362K** frames spanning 11 different objects, 4 subjects, 17 free-hand sequences, and 20 grasps per object per subject. We then showed how this data can support the representation of hands, objects, and grasps as neural fields. Finally, we use the neural representations to train RealGrasper, a grasping model that generates plausible grasps given an object and initial hand pose.

Limitations and Future Work: Our approach has several limitations. First, our dataset is currently limited in the number of objects/subjects and only static grasps – we plan to extend this work to dynamics grasps and in-hand manipulation. Our neural

models take a significant amount of time to train and generate composites which we hope to address by investigating faster neural fields [Müller et al. 2022; Yu et al. 2021]. Despite designed loss functions to avoid collision, our model could still fail at inference when the hand and object penetrate. We consider our method to be the first step towards building large-scale generative grasp and manipulation models from multi-view data.

REFERENCES

Luca Ballan, Aparna Taneja, Jürgen Gall, Luc Van Gool, and Marc Pollefeys. 2012. Motion capture of hands in action using discriminative salient points. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part VI 12*. Springer, 640–653.

Jonathan T. Barron, Ben Mildenhall, Dor Verbin, Pratul P. Srinivasan, and Peter Hedman. 2022. Mip-NeRF 360: Unbounded Anti-Aliased Neural Radiance Fields. *CVPR* (2022).

Keni Bernardin, Koichi Ogawara, Katsushi Ikeuchi, and Ruediger Dillmann. 2005. A sensor fusion approach for recognizing continuous human grasping sequences using hidden markov models. *IEEE Transactions on Robotics* 21, 1 (2005), 47–57.

Antonio Bicchi and Vijay Kumar. 2000. Robotic grasping and contact: A review. In *Proceedings 2000 ICRA. Millennium conference. IEEE international conference on robotics and automation. Symposia proceedings (Cat. No. 00CH37065)*, Vol. 1. IEEE, 348–353.

Konstantinos Bousmalis, Alex Irpan, Paul Wohlhart, Yunfei Bai, Matthew Kelcey, Mrinal Kalakrishnan, Laura Downs, Julian Ibarz, Peter Pastor, Kurt Konolige, et al. 2018. Using simulation and domain adaptation to improve efficiency of deep robotic grasping. In *2018 IEEE international conference on robotics and automation (ICRA)*. IEEE, 4243–4250.

Samarth Brahmbhatt, Cusuh Ham, Charles C Kemp, and James Hays. 2019. Contactdb: Analyzing and predicting grasp contact via thermal imaging. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 8709–8719.

Samarth Brahmbhatt, Chengcheng Tang, Christopher D Twigg, Charles C Kemp, and James Hays. 2020. ContactPose: A dataset of grasps with object contact and hand pose. In *European Conference on Computer Vision*. Springer, 361–378.

Ian M Bullock, Thomas Feix, and Aaron M Dollar. 2015. The Yale human grasping dataset: Grasp, object, and task data in household and machine shop environments. *The International Journal of Robotics Research* 34, 3 (2015), 251–255.

Zhe Cao, Ilija Radosavovic, Angjoo Kanazawa, and Jitendra Malik. 2021. Reconstructing Hand-Object Interactions in the Wild. In *ICCV*.

Géry Casiez, Nicolas Roussel, and Daniel Vogel. 2012. 1 € filter: a simple speed-based low-pass filter for noisy input in interactive systems. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (2012).

Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu, and Hao Su. 2022. TensoRF: Tensorial Radiance Fields. *arXiv preprint arXiv:2203.09517* (2022).

Zhiqin Chen and Hao Zhang. 2019. Learning Implicit Fields for Generative Shape Modeling. *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2019).

Ho Kei Cheng, Jihoon Chung, Yu-Wing Tai, and Chi-Keung Tang. 2020. CascadePSP: Toward Class-Agnostic and Very High-Resolution Segmentation via Global and Local Refinement. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2020), 8887–8896.

Sammy Christen, Muhammed Kocabas, Emre Aksan, Jemin Hwangbo, Jie Song, and Otmar Hilliges. 2022. D-Grasp: Physically Plausible Dynamic Grasp Synthesis for Hand-Object Interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Enric Corona, Tomas Hodan, Minh Vo, Francesc Moreno-Noguer, Chris Sweeney, Richard Newcombe, and Lingni Ma. 2022. LISA: Learning Implicit Shape and Appearance of Hands. In *CVPR*.

Joseph DelPreto, Chao Liu, Yiyue Luo, Michael Foshey, Yunzhu Li, Antonio Torralba, Wojciech Matusik, and Daniela Rus. 2022. ActionNet: A Multimodal Dataset for Human Activities Using Wearable Sensors in a Kitchen Environment. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Ali Erol, George Bebis, Mircea Nicolescu, Richard D Boyle, and Xander Twombly. 2007. Vision-based hand pose estimation: A review. *Computer Vision and Image Understanding* 108, 1-2 (2007), 52–73.

Guillermo Garcia-Hernando, Shanxin Yuan, Seungryul Baek, and Tae-Kyun Kim. 2018. First-person hand action benchmark with rgb-d videos and 3d hand pose annotations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 409–419.

Henning Hamer, Juergen Gall, Thibaut Weise, and Luc Van Gool. 2010. An object-dependent hand pose prior from sparse training data. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE, 671–678.

Henning Hamer, Konrad Schindler, Esther Koller-Meier, and Luc Van Gool. 2009. Tracking a hand manipulating an object. In *2009 IEEE 12th International Conference on Computer Vision*. IEEE, 1475–1482.

Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent Lepetit. 2020. Honnotate: A method for 3d annotation of hand and object poses. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 3196–3206.

Yana Hasson, Gül Varol, Dimitrios Tzionas, Igor Kalevatykh, Michael J Black, Ivan Laptev, and Cordelia Schmid. 2019. Learning joint reconstruction of hands and manipulated objects. In *CVPR 2019 - IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, Long Beach, United States, 11799–11808. https://doi.org/10.1109/CVPR.2019.01208

Hanwen Jiang, Shaowei Liu, Jiashun Wang, and Xiaolong Wang. 2021. Hand-Object Contact Consistency Reasoning for Human Grasps Generation. In *Proceedings of the International Conference on Computer Vision*.

Hanbyul Joo, Tomas Simon, and Yaser Sheikh. 2018. Total capture: A 3d deformation model for tracking faces, hands, and bodies. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 8320–8329.

Korrawe Karunratanakul, Jinlong Yang, Yan Zhang, Michael Black, Krikamol Muandet, and Siyu Tang. 2020. Grasping Field: Learning Implicit Representations for Human Grasps. In *2020 International Conference on 3D Vision (3DV)*.

Ruilong Li, Julian Tanke, Minh Vo, Michael Zollhofer, Jurgen Gall, Angjoo Kanazawa, and Christoph Lassner. 2022. TAVA: Template-free animatable volumetric actors. *European Conference on Computer Vision (ECCV)*.

Shanchuan Lin, Andrey Ryabtsev, Soumyadip Sengupta, Brian Curless, Steven M. Seitz, and Ira Kemelmacher-Shlizerman. 2020. Real-Time High-Resolution Background Matting. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2020), 8758–8767.

Lingjie Liu, Marc Habermann, Viktor Rudnev, Kripasindhu Sarkar, Jiatao Gu, and Christian Theobalt. 2021. Neural Actor: Neural Free-view Synthesis of Human Actors with Pose Control. *ACM Trans. Graph.(ACM SIGGRAPH Asia)* (2021).

Stephen Lombardi, Tomas Simon, Jason Saragih, Gabriel Schwartz, Andreas Lehrmann, and Yaser Sheikh. 2019. Neural Volumes: Learning Dynamic Renderable Volumes from Images. *ACM Trans. Graph.* 38, 4, Article 65 (July 2019), 14 pages.

Jens Lundell, Enric Corona, Tran Nguyen Le, Francesco Verdoja, Philippe Weinzaepfel, Grégory Rogez, Francesc Moreno-Noguer, and Ville Kyrki. 2021a. Multi-fingan: Generative coarse-to-fine sampling of multi-finger grasps. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 4495–4501.

Jens Lundell, Francesco Verdoja, and Ville Kyrki. 2021b. Ddgc: Generative deep dexterous grasping in clutter. *IEEE Robotics and Automation Letters* 6, 4 (2021), 6899–6906.

Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. 2019. Occupancy Networks: Learning 3D Reconstruction in Function Space. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*.

Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. 2020. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. In *ECCV*.

Andrew T Miller and Peter K Allen. 2004. Graspit! a versatile simulator for robotic grasping. *IEEE Robotics & Automation Magazine* 11, 4 (2004), 110–122.

Gyeongsik Moon, Shoou-I Yu, He Wen, Takaaki Shiratori, and Kyoung Mu Lee. 2020. Interhand2. 6m: A dataset and baseline for 3d interacting hand pose estimation from a single rgb image. In *European Conference on Computer Vision*. Springer, 548–564.

Franziska Mueller, Florian Bernard, Oleksandr Sotnychenko, Dushyant Mehta, Srinath Sridhar, Dan Casas, and Christian Theobalt. 2018. GANerated Hands for Real-Time 3D Hand Tracking from Monocular RGB. In *Proceedings of Computer Vision and Pattern Recognition (CVPR)*. 11 pages. https://handtracker.mpi-inf.mpg.de/projects/GANeratedHands/

Franziska Mueller, Dushyant Mehta, Oleksandr Sotnychenko, Srinath Sridhar, Dan Casas, and Christian Theobalt. 2017. Real-time Hand Tracking under Occlusion from an Egocentric RGB-D Sensor. In *Proceedings of International Conference on Computer Vision (ICCV)*. 10 pages. https://handtracker.mpi-inf.mpg.de/projects/OccludedHands/

Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. 2022. Instant Neural Graphics Primitives with a Multiresolution Hash Encoding. *ACM Trans. Graph.* 41, 4, Article 102 (July 2022), 15 pages. https://doi.org/10.1145/3528223.3530127

Iason Oikonomidis, Nikolaos Kyriazis, and Antonis A Argyros. 2011. Full dof tracking of a hand interacting with an object by modeling occlusions and physical constraints. In *2011 International Conference on Computer Vision*. IEEE, 2088–2095.

Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. 2019. DeepSDF: Learning Continuous Signed Distance Functions for Shape Representation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Sida Peng, Junting Dong, Qianqian Wang, Shangzhan Zhang, Qing Shuai, Xiaowei Zhou, and Hujun Bao. 2021a. Animatable Neural Radiance Fields for Modeling Dynamic Human Bodies. In *ICCV*.

Songyou Peng, Michael Niemeyer, Lars Mescheder, Marc Pollefeys, and Andreas Geiger. 2020. Convolutional Occupancy Networks. In *European Conference on Computer Vision (ECCV)*.

Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. 2021b. Neural Body: Implicit Neural Representations with Structured

Latent Codes for Novel View Synthesis of Dynamic Humans. In *CVPR*.

Tu-Hoa Pham, Nikolaos Kyriazis, Antonis A Argyros, and Abderrahmane Kheddar. 2017. Hand-object contact force estimation from markerless visual tracking. *IEEE transactions on pattern analysis and machine intelligence* 40, 12 (2017), 2883–2896.

PhotoRoom. 2023. PhotoRoom v5.0.9. https://www.photoroom.com/

Davis Rempe, Tolga Birdal, Aaron Hertzmann, Jimei Yang, Srinath Sridhar, and Leonidas J. Guibas. 2021. HuMoR: 3D Human Motion Model for Robust Pose Estimation. In *International Conference on Computer Vision (ICCV)*.

Grégory Rogez, James S Supancic, and Deva Ramanan. 2015. Understanding everyday hands in action from RGB-D images. In *Proceedings of the IEEE international conference on computer vision*. 3889–3897.

Javier Romero, Hedvig Kjellström, and Danica Kragic. 2010. Hands in action: real-time 3D reconstruction of hands in interaction with objects. In *2010 IEEE International Conference on Robotics and Automation*. IEEE, 458–463.

Javier Romero, Dimitrios Tzionas, and Michael J. Black. 2017. Embodied Hands: Modeling and Capturing Hands and Bodies Together. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)* 36, 6 (Nov. 2017).

Johannes Lutz Schönberger and Jan-Michael Frahm. 2016. Structure-from-Motion Revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. 2016. Pixelwise View Selection for Unstructured Multi-View Stereo. In *European Conference on Computer Vision (ECCV)*.

Dandan Shan, Jiaqi Geng, Michelle Shu, and David F Fouhey. 2020. Understanding human hands in contact at internet scale. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 9869–9878.

Lin Shao, Fabio Ferreira, Mikael Jorda, Varun Nambiar, Jianlan Luo, Eugen Solowjow, Juan Aparicio Ojea, Oussama Khatib, and Jeannette Bohg. 2020. Unigrasp: Learning a unified model to grasp with multifingered robotic hands. *IEEE Robotics and Automation Letters* 5, 2 (2020), 2286–2293.

Tomas Simon, Hanbyul Joo, Iain Matthews, and Yaser Sheikh. 2017. Hand Keypoint Detection in Single Images using Multiview Bootstrapping. In *CVPR*.

Vincent Sitzmann, Michael Zollhoefer, and Gordon Wetzstein. 2019. Scene Representation Networks: Continuous 3D-Structure-Aware Neural Scene Representations. In *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.), Vol. 32. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2019/file/b5dc4e5d9b495d0196f61d45b26ef33e-Paper.pdf

Kihyuk Sohn, Honglak Lee, and Xinchen Yan. 2015. Learning Structured Output Representation using Deep Conditional Generative Models. In *Advances in Neural Information Processing Systems*, C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett (Eds.), Vol. 28. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2015/file/8d55a249e6baa5c06772297520da2051-Paper.pdf

Srinath Sridhar, Franziska Mueller, Michael Zollhöfer, Dan Casas, Antti Oulasvirta, and Christian Theobalt. 2016. Real-time joint tracking of a hand manipulating an object from rgb-d input. In *European Conference on Computer Vision*. Springer, 294–310.

Srinath Sridhar, Antti Oulasvirta, and Christian Theobalt. 2013. Interactive Markerless Articulated Hand Motion Tracking using RGB and Depth Data. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 8 pages. http://handtracker.mpi-inf.mpg.de/projects/handtracker_iccv2013/

Omid Taheri, Nima Ghorbani, Michael J Black, and Dimitrios Tzionas. 2020a. GRAB: A dataset of whole-body human grasping of objects. In *European conference on computer vision*. Springer, 581–600.

Omid Taheri, Nima Ghorbani, Michael J. Black, and Dimitrios Tzionas. 2020b. GRAB: A Dataset of Whole-Body Human Grasping of Objects. In *European Conference on Computer Vision (ECCV)*. https://grab.is.tue.mpg.de

Tze Ho Elden Tse, Zhongqun Zhang, Kwang In Kim, Ales Leonardis, Feng Zheng, and Hyung Jin Chang. 2022. S^2Contact: Graph-based Network for 3D Hand-Object Contact Estimation with Semi-Supervised Learning. In *ECCV*.

Aggeliki Tsoli and Antonis A Argyros. 2018. Joint 3D tracking of a deformable object in interaction with a hand. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 484–500.

Dylan Turpin, Liquan Wang, Eric Heiden, Yun-Chun Chen, Miles Macklin, Stavros Tsogkas, Sven Dickinson, and Animesh Garg. 2022a. Grasp'D: Differentiable Contact-rich Grasp Synthesis for Multi-fingered Hands. In *ECCV*.

Dylan Turpin, Liquan Wang, Eric Heiden, Yun-Chun Chen, Miles Macklin, Stavros Tsogkas, Sven Dickinson, and Animesh Garg. 2022b. Grasp'd: Differentiable contact-rich grasp synthesis for multi-fingered hands. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VI*. Springer, 201–221.

Dimitrios Tzionas and Juergen Gall. 2015. 3d object reconstruction from hand-object interactions. In *Proceedings of the IEEE International Conference on Computer Vision*. 729–737.

Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. 2021. NeuS: Learning Neural Implicit Surfaces by Volume Rendering for Multi-view Reconstruction. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI)*. International Joint Conferences on Artificial Intelligence Organization. http://arxiv.org/abs/2106.10689v1

Chung-Yi Weng, Brian Curless, Pratul P. Srinivasan, Jonathan T. Barron, and Ira Kemelmacher-Shlizerman. 2022. HumanNeRF: Free-Viewpoint Rendering of Moving People From Monocular Video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 16210–16220.

Tianhao Walter Wu, Fangcheng Zhong, Andrea Tagliasacchi, Forrester Cole, and Cengiz Oztireli. 2022. D^2NeRF: Self-Supervised Decoupling of Dynamic and Static Objects from a Monocular Video. In *Advances in Neural Information Processing Systems*, Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (Eds.). https://openreview.net/forum?id=rG7HZZtIc-

Yiheng Xie, Towaki Takikawa, Shunsuke Saito, Or Litany, Shiqin Yan, Numair Khan, Federico Tombari, James Tompkin, Vincent Sitzmann, and Srinath Sridhar. 2022. Neural Fields in Visual Computing and Beyond. *Computer Graphics Forum* (2022). https://doi.org/10.1111/cgf.14505

Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. 2021. Volume rendering of neural implicit surfaces. In *Thirty-Fifth Conference on Neural Information Processing Systems*.

Yuting Ye and C Karen Liu. 2012. Synthesis of detailed hand manipulations using contact sampling. *ACM Transactions on Graphics (ToG)* 31, 4 (2012), 1–10.

Alex Yu, Sara Fridovich-Keil, Matthew Tancik, Qinhong Chen, Benjamin Recht, and Angjoo Kanazawa. 2021. Plenoxels: Radiance fields without neural networks. *arXiv preprint arXiv:2112.05131* (2021).

He Zhang, Yuting Ye, Takaaki Shiratori, and Taku Komura. 2021. ManipNet: Neural manipulation synthesis with a hand-object spatial representation. *ACM Transactions on Graphics (ToG)* 40, 4 (2021), 1–14.

Joshua Z Zheng, Sara De La Rosa, and Aaron M Dollar. 2011. An investigation of grasp type and frequency in daily household and machine shop tasks. In *2011 IEEE international conference on robotics and automation*. IEEE, 4169–4175.

Christian Zimmermann, Duygu Ceylan, Jimei Yang, Bryan Russell, Max Argus, and Thomas Brox. 2019. Freihand: A dataset for markerless capture of hand pose and shape from single rgb images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 813–822.

Paula Zuccotti. 2015. *Every Thing We Touch: A 24-hour Inventory of Our Lives*. Penguin UK.

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009

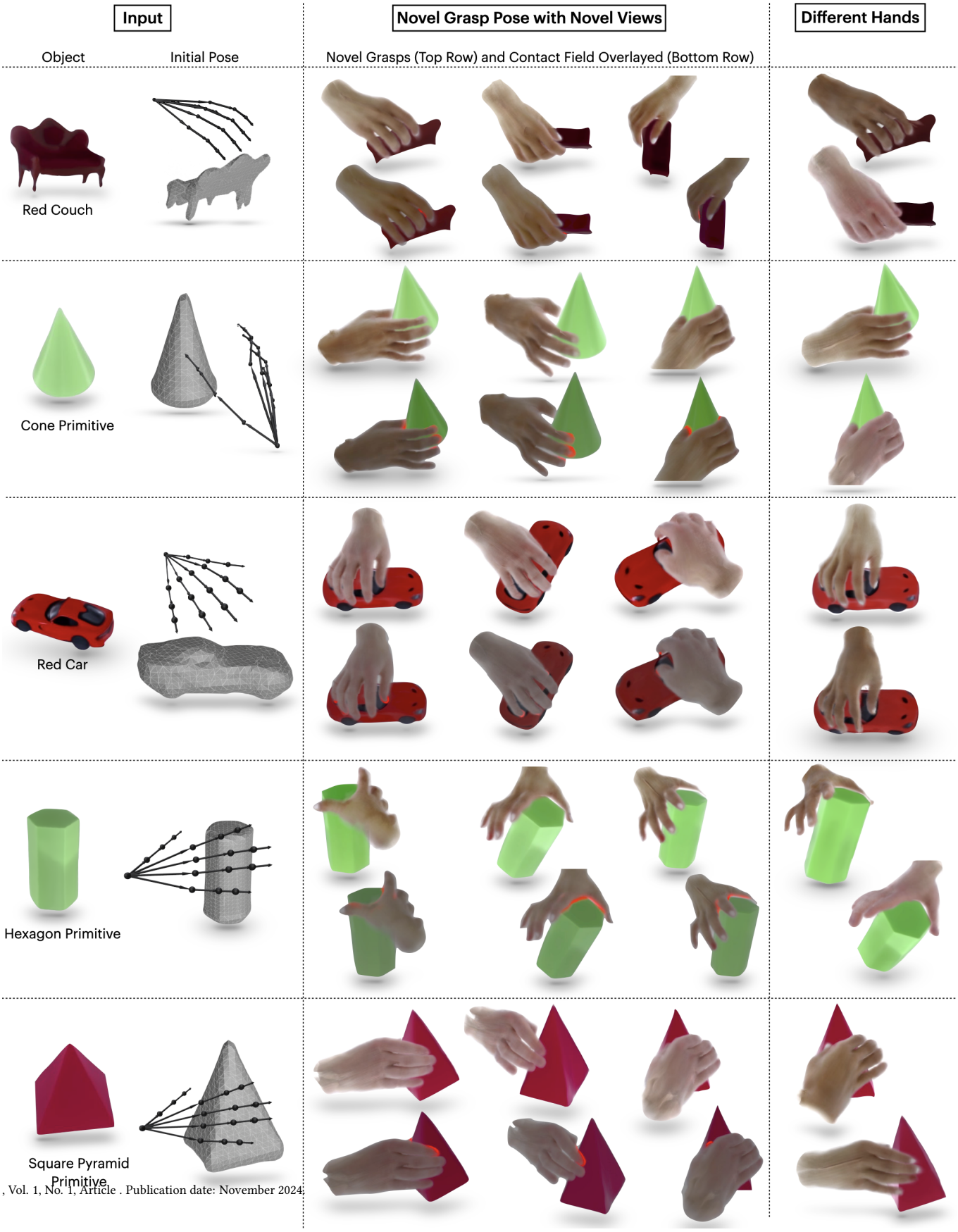

Fig. 7. Visualization of generated grasps and contact from RealGrasper given input object and initial hand pose shown in different views. The right column shows grasp synthesis on same input using Neural Hand model from different hand subject, which does not appear during training. The results demonstrate our grasp generation model can generalize to different hand subjects.

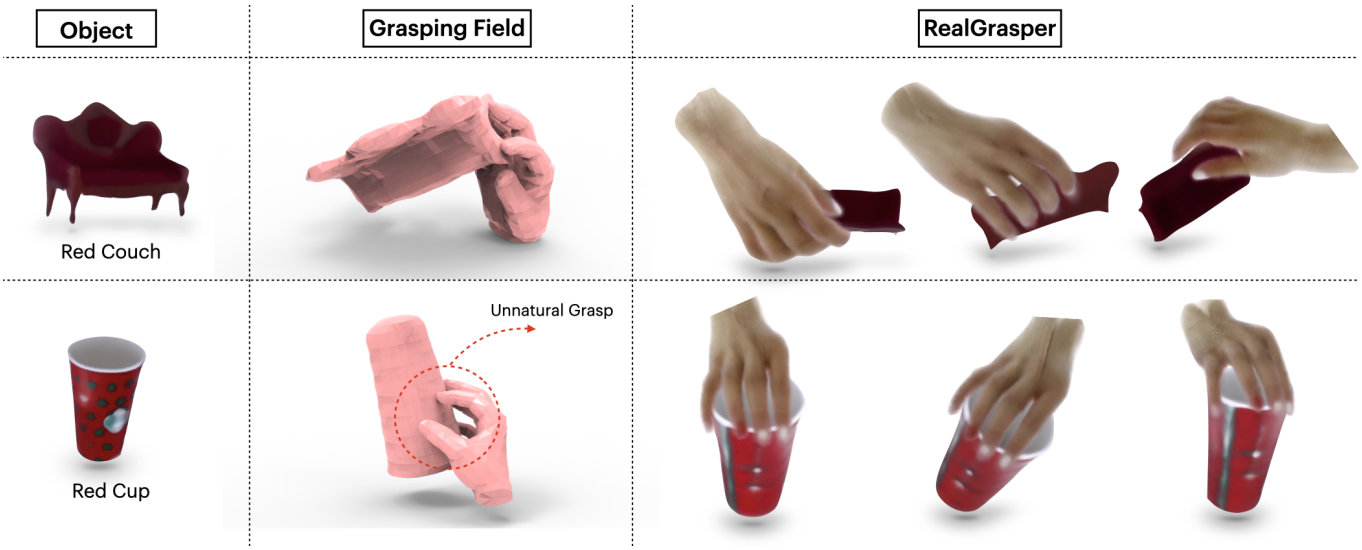

Fig. 8. Qualitative comparison with GraspingFields [Karunratanakul et al. 2020]. Our RealGrasper enables photo realistic rendering of hand-object grasping. We demonstrates more human-like natural grasps of the couch model and cup compared with Grasping Field and visually appealing rendering quality.

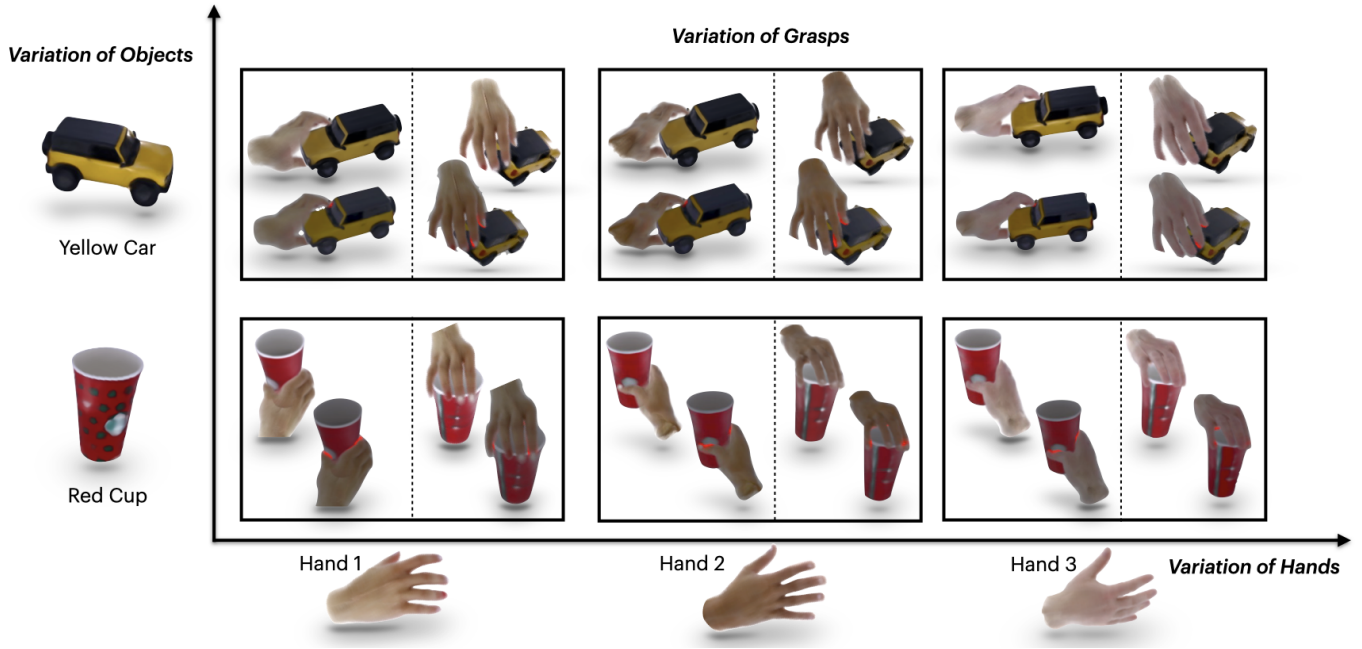

Fig. 9. **Variation of Hands, Objects and Grasps**: Our framework learns hand representations from different subjects incorporating subject's unique characteristics. We can swap hands and objects to generate various grasps in a plug and play manner. Hence, for a certain object, we can show variety of grasps with different hands.