Learning to Perceive and Predict Everyday Hand-Object Interactions

Yufei Ye

CMU-RI-TR-24-59

August, 2024

The Robotics Institute
Carnegie Mellon University
Pittsburgh, Pennsylvania 15213

Thesis Committee:
Shubham Tulsiani, Co-Chair
Abhinav Gupta, Co-Chair
Deva Ramanan
Andreas Geiger, University of Tübingen
Angjoo Kanazawa, UC Berkeley

Thesis proposal submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Robotics

Abstract

This thesis aims to equip computers with the ability to understand everyday hand-object interactions in the physical world – both perceiving ongoing interactions in 3D space and predicting possible interactions. This ability is crucial for applications such as virtual reality, robotic manipulations, and augmented reality. The problem is inherently ill-posed due to the challenges of one-to-many inference and the intricate physical interactions between hands and objects. To address these challenges, we explore a learning approach that mines priors from everyday data to enhance computer perception of interactions. Our goal is to develop methods for building 3D representations that respect the physical world's inherent structure and can generalize to novel everyday scenes.

We first explore how to scale up 3D object priors for single-view object reconstruction in isolation, by introducing a learning technique for unsupervised, category-level 3D object reconstruction from unstructured image collections. Furthermore, we argue that interactions between hands and objects should not be marginalized as occlusion noise, but rather explicitly modeled to improve 3D reconstruction. To this end, we propose an approach to reconstruct hand-object interactions from a single image by leveraging hand pose information to better infer in-hand objects. Our research then extends the core idea to reconstruction from short video clips, where we combine multi-view cues with data-driven priors for accurate 3D inference. While perceiving ongoing interactions allows for predicting possible interactions, we also explore interaction synthesis – predicting spatial arrangements of human-object interactions. We propose a generative method that leverages a large-scale pretrained model to achieve realistic, controllable, and generalizable predictions of novel everyday objects. Finally, this thesis presents a unified generative prior for hand-object interactions, allowing for both reconstruction and prediction tasks. We also make efforts to scale up the training data by aggregating multiple existing real-world interaction datasets. We demonstrate that the resulting joint prior can facilitate interaction reconstruction and prediction, outperforming current task-specific methods.

Acknowledgements

The journey has been exciting not only because of the research I've devoted my passion to, but also, more importantly, because of the people I've had the privilege to meet. It is almost a doomed attempt to include everyone I am grateful to along this exciting journey over the years. I would not have come this far without each one of you.

First and foremost, I would like to thank my advisors, Shubham Tulsiani and Abhinav Gupta. Shubham is the smartest, kindest, warmest, wisest, and most supportive advisor one could ever ask for. After working on a problem in my background for almost a month, he can immediately propose a much cleaner solution in the first meeting about it. I am still amazed by how sharp he is after all these years. While extremely sharp, he always finds an angle that is constructive and kind. With endless patience, he hones my every research skill. Throughout all these years, I have never seen this great mind overtaken by his emotions, good or bad, even once. He has shown me the power of rationality, calmness, and consistent diligence, despite the ups and downs. I also want to thank him for his generous support, which allowed me to explore opportunities to my benefit. I am especially thankful to him for pushing me to renew my visa in Mexico within 20 hours, which essentially changed my career path. It is a great fortune to be advised by such a role model for six years.

Like ice and fire, Abhinav is another possibility of being a great advisor. He showed me that passion and energy are the ultimate driving forces for research. I have learned a lot from him about maintaining childlike curiosity and enthusiasm. He leads me to question the very fundamental basics and never lose sight of the big picture, often with his iconic pause followed by, ”Why do we even do this?” I am always grateful to him for his blunt honesty, guidance, and support throughout my PhD.

I also want to thank my committee members, Deva, Andreas, and Angjoo, for their insightful feedback, career advice, and flexibility in making the process happen. I appreciate your time and efforts. Deva's seminar was one of the most insightful classes I've ever had. Just from the seminar alone, Deva's passion for research has already inspired me. A special thank you to Angjoo—your physical presence at my proposal injected special energy into the event.

I had the fortune to meet Michael, who brought me to a workshop where I had no idea how I got in or what journey it would unravel. He kindly hosted me at MPI, and it's been one of the happiest times of my life and research. His excitement for research, articulate arguments, humility, and humanity are a special sight to see. I also want to thank my PostDoc advisor, Karen Liu. Both Michael and Karen appeared in my career at a very special time when I was confused and overwhelmed. They helped a young researcher reinforce the belief in high-quality work and curiosity-driven research.

I am fortunate to work with my amazing collaborators. Xiaolong Wang is my mentor who led me into serious research. While his daily schedule—noon to 3 a.m.—is bizarre to most people, his motto, "Every day is so enjoyable because there is something new," sums it up perfectly. I am lucky to work with Kris and learn to think from the perspective of leveraging data and resources. Maneesh generously offered weekly feedback during my weekly meetings when I was a master's student. I spent great time at NVIDIA. Sifei taught me to appreciate even the smallest progress. Shalini taught me important lessons about being a professional research scientist in the industry. Stan reminded me to simplify the problem and always keep alternative technical paths in mind. Xueting showed me resilience when she kept failing to meet her promise of "this is my last week to try this project." Jiaming has been the mentor I looked up to since undergrad. Sudeep and Homanga are the go-to persons for all of my dumb questions about robotics. Dhiraj Gandhi helped me pull off my first robotics paper, and we've had so much fun together since my summer internship. Omid works on very relevant research topic to mine and it's great to collaborate closely in my late PhD stage. I learned a lot from Yao, Tim, Weiyang and Haven about LLM/foundation model and being so passionate about research. Thank you, Poorvi and Ishaan, for your hard work and letting me participate in your early research careers.

Every greatness comes from a tiny seed, and it's my labmates and colleagues who witnessed the dirt and helped foster the seed to bloom. Thank you for being there for my immature ideas, bombed presentations, first drafts, and cheering for my little achievements. Special thanks to those ahead of me who lifted me up in my career, and I hope I can do the same for those who come after. People from AGI (Abhinav Gupta Intelligence) group: David, Xiaolong, Abhinav, Saurabh, Yin, Pedro, Ishan, Lerrel, Adithya, Jacob, Gunnar, Senthil, Victoria, Helen, Raunaq, Sudeep, Shikhar, Nadine, Pratyusha, Tian, Tao, Wenxuan, Nilesh, Gaurav, Jianren, Mohan, Ishan, *etc.* PPL group with Shubham: Jason, Homanga, Yehonathan, Himangi, Hanzhe, Yanbo, Qitao, Z, Bharath, Poorvi, Naveen, Mayank, Yen-Chi, Paritosh, Amy, and Lucas. And also my PS/MPI friends: Victoria, Yao, Omid, Haven, Shashank, Marilyn, Soubhik, Markos, Yuliang, Yandong, Sai, Zhen, Georgio, Hanz, Rick just to name a few. The CMU vision group is a collaborative environment. I had a great time in discussing with them, such as Minh, Rohit, Aayush, Mihir, Dinesh, Leo, Ziyan, Zen, Gaurav, Jinkun, Nupur, Neehar, Akash, among many others.

Beyond research, my life in Pittsburgh has been enriched by my dearest friends. We share our cats, homesickness, and happiness at hotpot parties with Wenxuan, Gengshan, Yisha, Xianyi, Shuyan, Anqi, Sheng-Yu, Ruihan, Emma, Kangle, Chaoyang, Mengtian, Yan, *etc.*. Thank you Helen, Jason, Sudeep, Shikhar, Alex, along with others, who (introduced American football to me and) made me feel welcome in the new country. I also want to thank my friends from badminton community, running club, and All-University Sympony Orchestra, who constitutes a supporting system outside of work.

Thank you Suzanne and Christine who act as shields for us against logistical burdens.

Finally, I want to thank my parents, Lixin Li and Changchun Ye, for their boundless love and unconditional support, and my grandparents, some of whom are in heaven. I am proud to be part of the Li and Ye family. Lastly, to Donglai Xiang, my best friend and my love, thank you, and I look forward to all the years ahead of us.

Contents

1 Introduction **1**

2 Background **5**
2.1 Human Hand Pose Reconstruction 5
2.2 Reconstructing Generic Objects in Isolation 6
2.3 Hand-Object Interaction Reconstruction 7
2.4 Hand-Object Interaction Prediction 8

3 Reconstructing Generic Objects in the Wild **9**
3.1 Method 12
3.2 Experiments 15
3.3 Discussion 23
3.4 Implementation Details and Additional Results 23

4 Reconstructing Generic Objects in Hands from Single Images **27**
4.1 Method 28
4.2 Experiments 33
4.3 Discussion 41
4.4 Implementation Details and Additional Results 41

5 Reconstructing Generic Objects in Hands from Monocular Videos **51**
5.1 Preliminary: Diffusion Models 52
5.2 Method 54
5.3 Experiments 59
5.4 Discussion 68
5.5 Implementation Details and Additional Results 68

6 Predicting Hand-Object Interactions via Image Synthesis **75**
6.1 Method 77
6.2 Experiments 82

6.3 Discussion 88
6.4 Appendix 89
6.5 Implementation Details and Additional Results 89

7 Generative Hand-Object Prior for Reconstruction and Prediction 105
7.1 Method 107
7.2 Experiments 112
7.3 Discussion 120
7.4 Implementation Details and Additional Results 120

8 Conclusions 134

List of Figures

3.1 Given a single image, we predict a mesh with textures (rendered from the predicted view and a novel view). The models can learn directly from collections of images with only foreground masks, without supervision of mesh templates, multi-view association, camera poses, semantic annotations, *etc.*. 10

3.2 **Volumetric Representation Prediction and Rendering.** Left: Given an input image, the encoder-decoder network infers a semi-implicit volumetric representation (S_o, S_f) and a camera pose v. The semi-implicit volume is then projected from the predicted camera pose to obtain foreground image and mask. The semi-implicit volume is also projected from a novel view v'. The projections are required to both match the 2D observation and appear realistic. Right: Projection process mimicking ray marching. 11

3.3 **Instance-level specialization.** We convert the inferred volumetric occupancy to an initial mesh. The mesh geometry and textures are then iteratively refined to better match the given input. 14

3.4 Visualization of categorical volumetric representation across different methods. 17

3.5 Ablation study: refining mesh initialized from the predicted volume (col 2/5) and another volumes (col 3/6). 18

3.6 Visualizing categorical volumetric representation across different methods on CUB-200-2011, Quadrupeds, Chairs in the wild. 18

3.7 Ablation study: comparing reconstructed volumes when the model disables different loss terms. 19

3.8 Visualizing our refined shapes from the predicted view (2nd column in each quadruplet) and a novel view (3rd+4th / quadruplet) on CUB, Quadrupeds, Chairs in the wild. We are able to capture both the shared shapes in categories and instance-specific differences. 20
3.9 Test on COCO: visualization of lifting detection results to meshes via the shelf-supervised models. 22
3.10 Visualizing our reconstructed meshes from the predicted view (2nd in each quadruplet) and a novel view (3rd and 4th in each quadruplet) trained on multiple categories on Open Images. 25
3.11 Left: qualitative results on CUB replacing annotation with prediction in training and/or test time. Middle: quantitative results on synthetic chairs by adding random noise on masks during both training and inference. Right: masks used to train the models in the paper. 26
3.12 Results on training models with different viewpoint priors. . . 26

4.1 Given an RGB image depicting a hand holding an object, we infer the 3D shape of the hand-held object (rendered in the image frame and from a novel view). 28
4.2 Given an image of a hand-held object, we first use an off-the-shelf system to estimate hand articulation θ_A and the camera pose π_w. With the predicted articulated hand along with the image, the object shape is reconstructed by an implicit network. For each query point $\mathbf{x}$ in canonical hand wrist frame, it is transformed to image space $\mathbf{x}_p$ to get visual feature $\phi = g(\mathbf{x}_p, I)$. In parallel, we also encode its articulation-aware representation $\psi = h(\mathbf{x}; \theta)$. Then we use an implicit decoder to predict signed distance value $s = f(\mathbf{x}, \phi, \psi)$. 29
4.3 Visualizing reconstruction from our method and two baselines [73, 96] on ObMan dataset from the image frame and a novel view. 33
4.4 Visualizing reconstruction of our method and two baselines [73,96] on HO3D dataset in the image frame and from a novel view. 35
4.5 Visualizing reconstruction of our method and two baselines [73,96] on MOW dataset in the image frame and from a novel view. 36

4.6 Visualizing reconstruction of hand-held object with or without explicitly considering hand pose. 38
4.7 Top: Object reconstruction given hand pose corrupted by Gaussian on Obman dataset. Bottom: Object reconstruction given hand pose corrupted by prediction error on HO3D dataset. . 39
4.8 Visualizing hand-object reconstruction before and after test-time refinement in the image frame and from two novel views. 39
4.9 Visualizing reconstruction from our method and two baselines [73, 96] on ObMan dataset from the image frame and a novel view. 44
4.10 Visualizing reconstruction from our method and two baselines [73, 96] on HO3D dataset from the image frame and a novel view. 45
4.11 Visualizing reconstruction from our method and two baselines [73, 96] on RHOI dataset from the image frame and a novel view. 46
4.12 Visualizing reconstruction from our method and two baselines [73, 96] on RHOI dataset from the image frame and a novel view. 47
4.13 **Cross-dataset generalization:** we show quantitative results on HO3D for models pretrained on ObMan and RHOI. 48
4.14 Visualizing reconstruction of hand-held object with or without explicitly considering hand pose on ObMan, HO3D and RHOI. 49
4.15 Visualizing hand-object reconstruction before and after test-time refinement in the image frame and from two novel views. 50

5.1 Given a video clip depicting a hand-object interaction, we infer the underlying 3D shape of both the hand and the object. **Top:** sampled input frames; **Middle:** reconstruction in the image frame; **Bottom:** reconstruction from a novel view. . . . 52
5.2 **Method Overview :** We model the HOI scene (middle) by a time-persistent implicit field ϕ for the object, hand meshes H^t parameterized by hand shape β, hand articulation θ^t_A, along a time-varying rigid transformation $T^t_{h\to o}$ for object pose. We register the cameras in the hand frame. We optimize a video-specific scene representation using reprojection loss from the original view and diffusion distillation loss from a novel view v'. 54

5.3 **Geometry-informed Diffusion Method:** Our diffusion model takes in a noisy geometry rendering of the object, the geometry rendering of the hand, and a text prompt, to output the denoised geometry rendering of objects. 56

5.4 **Generations from conditional diffusion model:** Given the geometry rendering of hand G_h (only showing surface normals) and a text prompt C, we visualize 4 different generations from the diffusion model. Middle row shows the generated surface normal of the objects and bottom row visualizes the generated object masks overlayed on the given hand masks. Note the left and middle column share the same text condition while middle and right column share the same hand condition. 58

5.5 **Qualitative evaluation on HOI4D:** We show reconstruction by our method (DiffHOI) along with two baselines [85, 232] in the image frame (left) and another novel view with (top right) or without (bottom right) hand. Please see project website for reconstruction videos. 60

5.6 Visualizing HOI reconstruction comparisons of our method with other variants of diffusion models that only incorporate category prior, hand prior, and no prior. (Top: image frame, bottom: novel view) 63

5.7 **Ablation Study:** Visualizing HOI reconstruction comparison of our method and variants that do not distill on depth, mask, and normals. (Top: image frame, bottom: novel view) 64

5.8 **Comparing with template-based method:** We show reconstruction in the image frame (top) and from a novel view (bottom) by our method and HOMAN [71] when provided with ground-truth templates, a random template, and the most dissimilar template in the training split. 67

5.9 Comparing reconstructions of our method and the iHOI baseline [232] on 8 in-the-wild video clips taken from the Internet (Left: image frame, top right: novel view HOI, and bottom right: novel view object-only). 68

5.10 **Generations from conditional diffusion model.** Given the geometry rendering of hand (i) (row 1-4 visualizing surface normal, depth, mask, and uv coordinate), as well as a text prompt with category information, we visualize 5 different generations (ii-vi) from the diffusion model. Row 1-3 in col ii-vi shows the generated geometry rendering of the object, and row 4 visualizes overlayed hand and object masks for a better view of the hand-object relations. All examples on the left use the ground truth paired hand and category information while each example to its right uses another random category but remain hand the same. 73

6.1 Given a single RGB image of an object (first column), we synthesize plausible images of hand-object interactions from which feasible 3D hand poses can be directly extracted (remaining columns). 76

6.2 The proposed method consists of two image-conditioned diffusion models: LayoutNet and ContentNet. Given an object image, we first use LayoutNet (left) to predict a HOI spatial arrangement $\boldsymbol{l}_0$. For every diffusion step, the LayoutNet splats the noisy layout parameter into image space, concatenates it with the object image and their blending, and predicts the denoised layout. We apply the diffusion loss in the splatted 2D space $\mathcal{L}_{\text{mask}}$. Then the ContentNet (right) takes in the predicted layout along with the object image to synthesize an HOI image. The two modules are connected by the articulation-agnostic hand proxy (middle top). 77

6.3 **Paired Data Generation:** Given an HOI image, we first segment out hand (b) and remove it by inpainting (c). Then we use SDEdit [136] to reduce inpainting artifact (d). As inpainting introduce discrepancy between mask and unmasked region (f) while SDEdit undesirably modifies the unmasked object region, we mix up *both* object image sets in training. . . . 81

6.4 Visualizing HOI synthesis from our method and three baselines [88,101,174] on HOI4D (left) and EPIC-KITCHEN dataset (right). 82

6.5 Visualizing 3D affordance prediction from our method, GAN-Hand [32] and diffusion model [174] that directly predicts 3D pose on HOI4D (left) and EPIC-KITCHEN dataset (right). . . 84

6.6 **Layout Editing**: Visualizing HOI synthesis when the conditioned layouts gradually change location and orientation. . . 87

6.7 **Heatmap-guided synthesis:** Given a heatmap, LayoutNet is guided to generate layout at the sampled location, from which HOI images are synthesized and 3D poses are extracted. . . . 88

6.8 **Scene-level Integration:** Given a cluttered scene, we detect each object and synthesize its interactions individually. Each object's layout scale is guided to appear in the same size when transferred back to the scene. 89

6.9 Visualizing more examples of the constructed paired training data. We train all the models with a mixture of inpainted and SDEdited object images. 95

6.10 Visualizing the proposed paired data construction applied to EPIC-KITCHEN. 96

6.11 Visualizing more comparisons of the generated HOI images from the proposed method and other image synthesis baselines [88, 101, 174] on the HOI4D dataset. 97

6.12 Visualizing more comparisons of the generated HOI images from the proposed method and other image synthesis baselines [88, 101, 174] on the EPIC-KITCHEN dataset. 98

6.13 Visualizing more comparisons of the extracted 3D hand pose from the proposed method and other 3D affordance baselines [32, 174] on the HOI4D dataset. 99

6.14 Visualizing more comparisons of the extracted 3D hand pose from the proposed method and other 3D affordance baselines on the EPIC-KITCHEN dataset. 100

6.15 Visualizing the ablation of ContentNet for its LDM-based and GLIDE-based implementations (Sec 6.5.2). 101

6.16 Visualizing more layout editing results. 102

6.17 Visualizing more results of heatmap-guided synthesis. 103

6.18 Visualizing more scene integration results with the individual prediction from crops. 104

7.1 G-HOP can generate plausible hand-object interactions across a wide variety of objects (top). The learned generative prior can also guide inference for tasks such as reconstructing everyday interaction clips and synthesizing human grasps given object meshes. 106

7.2 **Method Overview of Generative Hand-Object Prior:** Hand-object interactions are represented as interaction grids within the diffusion model. This interaction grid concatenates the (latent) signed distance field for object and skeletal distance field for the hand. Given a noisy interaction grid and a text prompt, our diffusion model predicts a denoised grid. To extract 3D shape of HOI from the interaction grid, we use decoder to decode object latent code and run gradient descent on hand field to extract hand pose parameters. 107

7.3 **Reconstructing Interaction Clips:** We parameterize HOI scene as object implicit field, hand pose, and their relative transformation (left). The scene parameters are optimized with respect to the SDS loss on extracted interaction grid and reprojection loss (right). 109

7.4 **Grasp Synthesis:** We parameterize human grasps via hand articulation parameters and the relative hand-object transformation (left). These are optimized with respect to SDS loss by converting grasp (and known shape) to interaction grid (right). 111

7.5 **Dataset Statistics:** number of training samples for each category when training our generative prior. Zoom in for better view. 111

7.6 **Generations from Generative Hand-Object Prior:** Given a text prompt (only showing class label), we visualize two generated interactions from G-HOP . Categories are sorted from most common to least common in training (left to right). Generations are diverse in terms of object shape such as teapots, hand articulation such as mouse, and use intent like hammer. 111

7.7 **Qualitative Evaluation on HOI4D:** We show reconstruction by G-HOP and two other video reconstruction baselines [85, 233] in the image frame (left) and from another view with (top right) or without (bottom right) reconstructed hand. Please see our project page for reconstruction videos from all methods. 113

7.8 **In-the-Wild Reconstruction:** reconstruction on interaction clips from novel datasets [35,68]. 114

7.9 **Visualizing Grasp Generations:** Given an object mesh (left) from HO3D or ObMan, we sample two grasps from each method. 114

7.10 **Contact Map on Hand:** We visualize contact probability on hand over all generated samples from G-HOP and GraspTTA [90] on the HO3D dataset. 116

7.11 **Comparison with Baselines:** Preference percentages from users for pairwise method comparison on HO3D and ObMan. 116

7.12 **Grasp Diversity:** 10 random grasps of a power drill. Although GraspTTA generates more diverse grasps, some of them are not plausible as they disregard object functions. 118

7.13 **Ranking Grasps:** We visualize grasps with two highest scores (top) and two lowest scores (bottom) among 100 generated grasps from G-HOP. 119

7.14 **Comparing Hand Representation in Generative Hand-Object Prior:** Top 2 rows show the diffusion model that represents hand shape as pose parameters; bottom 2 rows show the diffusion model (ours) that represents hand shape as skeletal distance field. The homogeneous grid space is easier for the network to reason about interaction. 122

7.15 **2D Joint Prior (DiffHOI-J):** reconstruction and grasp synthesis results guided by 2D joint prior. 125

7.16 GF assumes known object pose when evaluating. Randomizing object pose affects their performance. 127

7.17 **Effect of Mesh Refinement:** We visualize synthesized grasps before (G-HOP†) and after (G-HOP) refinement. 128

7.18 **User Study Interface:** We visualize user study interface including the user instruction page and the survey page. 129

List of Tables

3.1 Comparing ours to other image-based supervised works in terms of supervision and outputs. 11

3.2 Quantitative results (3D IoU / F-score) on synthetic data comparing different methods for shape reconstruction. 16

3.3 Quantitative results (3D IoU) on synthetic data to ablate the effect of each loss term. 19

4.1 Quantitative results for object reconstruction error using F-score $(5mm, 10mm)$, Chamfer distane (mm) and intersection volume (cm^3). We compare our method with prior works [73, 96] on Obman, HO3D, MOW datasets. 33

4.2 **Cross-dataset generalization:** we report quantitative results on HO3D for models pretrained on ObMan and MOW. 37

4.3 **Analysis of articulation-conditioning:** we report quantitative results of object error in F-score, Chamfer distane (CD), intersection volume on 3 datasets and compare ours with the ablation that only consider visual feature. 38

4.4 **Analysis of articulation-aware encoding:** We compare different ways to incorporate hand articulation: articulation-aware positional encoding and pose parameters. Star indicates reconstruct object shape given ground truth hand articulation. 40

4.5 Error analysis against hand pose noise. σ is the average prediction error. * marks our unablated method. 40

4.6 **Test-time refinement.** We report object error, intersection volume, simulation displacement and hand error before and after test-time refinement. 41

4.7 Our customized split on HO3D dataset. 43

5.1 **Comparison with baselines:** Quantitative results for object reconstruction error using F1@5mm and F1@10mm scores and Chamfer Distance (mm). We compare our method with prior works HHOR [85] and iHOI [232] on the HOI4D dataset. . . . 59

5.2 **Analysis of the effect of data-driven priors:** Quantitative results on HOI4D for object reconstruction error in the object-centric frame (F@5, F@10, CD) and for hand-object alignment in the hand frame (CD_h). We compare our method with ablations that does not use prior, or use other variants of diffusion models that only conditions on hand or category. . . . 62

5.3 **Ablation without surface normal, mask and depth:** Quantitative results on the HOI4D dataset for object reconstruction error using mean F1 scores (5mm, 10mm), CD in object frame and for hand-object alignment using CD in hand frame ($\mathrm{CD_h}$). We compare our method with other ablations that do not distill normals, masks, and depths respectively. 62

5.4 **Error analysis against hand pose noise:** * marks our unablated method. Numbers in parentheses are per-frame prediction errors before optimization. 65

5.5 **Comparison with template-based baseline:** Quantitative results on the HOI4D dataset for object reconstruction error in the object-centric frame (F@5, F@10, CD) and for hand-object alignment (CD_h). We compare our method with HOMAN [71] with the ground truth template (-GT), with random templates from the training split (and reporting the average), and with furthest template from the ground truth (-furthest). 66

5.6 **Full ablation results of object reconstruction:** Quantitative results for object reconstruction error using F1@5mm and F1@10mm scores and Chamfer Distance (mm). We compare our method with variants that do not optimize per-frame object poses (Sec.5.5.8), blend hand and object masks in a hard way (Sec.5.5.9), or do not distill certain geometry modality (Sec. 4.2, Tab. 4) 69

5.7 **Full ablation results of HOI alignment:** Quantitative results for hand-object alignment using Chamfer distance (mm) in hand frame ($\mathrm{CD_h}$). We compare our method with variants that do not optimize per-frame object poses (Sec.5.5.8), blend hand and object masks in a hard way (Sec.5.5.9), or do not distill certain geometry modality (Sec. 4.2, Tab. 4). 69

6.1 Quantitative results for HOI synthesis using contact recall, FID score, and a user study on the HOI4D and EPIC-KITCHEN datasets. We compare our method with prior works [88, 101, 174]. 82

6.2 **Analysis of data augmentation**: contact recall (CR%) and FID score on the real and the inpainted object image set of HOI4D and comparisons of ours with the ablations of excluding aggressive common data augmentation (CmnAug) or SDEdit [136]. 85

6.3 User study for 3D affordance prediction on HOI4D and EPIC-KITCHEN dataset. We compare our method with GANHand [32] and a diffusion model that directly predicts 3D poses. . . 85

6.4 **Few-shot Adaption:** Quantitative results using contact recall when finetuning the proposed HOI synthesis model and a pretrained inpainting model with 32 samples from new categories. 86

7.1 **Comparing HOI reconstruction:** object error (F@5mm, F@10mm, CD), hand-object alignment CD_h, and hand error (MPJPE, AUC) on HOI4D. We compare G-HOP with baselines and also ablate if reconstruction benefits from priors in the 3D space or from joint modeling hand and object. 115

7.2 **Comparison with Baselines:** We compare our synthesized human grasps against GraspTTA [90] and annotated grasps provided by datasets (GT) on HO3D and ObMan. We report table the intersection between meshes, displacement distance in simulation, and hand contact ratio and area. 117

7.3 **Ranking Grasps:** plausibility on HO3D over all grasps, along with the top and bottom 10% grasps ranked by G-HOP. . . . 119

7.4 **Network architecture for G-HOP.** 121

7.5 **Comparing Object Error of HOI Reconstruction on HOI4D.** 123

7.6 **Comparing Hand Error of HOI Reconstruction on HOI4D.** . 123

7.7 **Comparing Hand-Object Alignment (CD_h $\downarrow$) of HOI Reconstruction on HOI4D.** 123

7.8 **Additional Ablation Studies of HOI reconstruction:** We report object error (F@5mm, F@10mm, CD), hand-object alignment CD_h, and hand error (MPJPE, AUC) on HOI4D. We analyze the effect of other implementation details, including dynamic noise thresholding and choice of text prompt templates. 124

7.9 **Comparison with Baselines:** We compare human grasp synthesis along with prior work GF [96]. * denotes GF's evaluation setting with known object pose. 126
7.10 **Effect of Refinement:** We report human grasp synthesis before and after mesh refinement. G-HOP† denotes generated grasps before mesh refinement. 126
7.11 We provide list of class names and their attributes used in the text prompt. The class names are manually merged across different datasets while the attributes are automatically generated by large language model [151]. 130

Chapter 1

Introduction

We humans live in a physical 3D world. Not only do we live as passive observers to name objects in this world, but we also live to actively interact with this world every single day in order to survive and thrive. Among the daily activities, a significant amount of interactions is performed via human hands [91]. Humans have evolved a great ability to perceive and predict their own and others' hand interactions. From one look at another holding a cup, we understand hand gestures relative to the cup and can effortlessly replicate their hold by ourselves. Even children from their second year of life are able to use spoons with their hands [31]. Not only do we understand what *is* happening, but we also *anticipate* what could happen. A bowl would be picked up by humans in certain ways. A stack of bowls would tipple if the bottom one is knocked. We argue that an intelligent agent aiming to mimic human interactions or a virtual assistant striving to aid in them must understand such generic everyday interactions. In this thesis, we aim to equip computers with similar abilities: both perceiving the ongoing interactions in 3D space and reasoning about the interactions to predict possible futures.

The problem is inherently ill-posed. Multiple 3D shapes can result in identical 2D projections and it is often impossible to disambiguate one from another via pure geometry rules [192]. In addition, it adds another layer of complexity due to the physical interaction in terms of contact and dynamics between hand and object [8, 162]. Though ill-posed and extremely complicated from pure mathematical and physical views, humans respond instantaneously and understand the everyday interactions of common objects in one shot. One of the key reasons behind this amazing ability is the prior that humans adopt – we adapt to this existing physical world and the common senses in social norms. We simply prefer some explanations to the alternatives (*e.g.* a mug typically having a handle, not putting fingers inside a mug, holding knives by handles). These priors are exhibited implicitly in our daily lives and are applicable to perceive and predict general objects. In this thesis, we explore a learning approach to mine such priors from daily life data to better perceive interactions.

To achieve the goal, we first need to build a 3D representation of the scenes that respect the inherent structure of the physical world. This includes 3D inference of any *individual* objects (object prior) and 3D inference of *relations* between the human and objects (interaction prior).

A desired system should be able to infer those 3D structures from *any* 2D visual inputs. In Chapter 3, we first scale up category-level objects prior for 3D reconstruction in an unsupervised manner. To alleviate expensive 3D supervision, we explore leveraging abundant data sources, *i.e.* unstructured image collections, and use only automatic segmentation outputs from off-the-shelf recognition systems as supervisory signals. Our insight is that different instances within one category are geometrically related and regularize each other at a category level from which the shape of each instance can be specialized. We supervise the model by enforcing consistency between the projection of the predicted 3D and the observed images, while also appearing realistic from a novel view. We show that the method can generalize to in the wild data for 50 categories with variant topology and shapes, an order of magnitude more than existing work. This work is published as Ye, Yufei, *et al.* "Shelf-supervised mesh prediction in the Wild." CVPR 2021.

While building category-level prior for generic individual objects, we assume a clear view of the objects of interest or occlusions are considered as noise to be marginalized over. However, interacting with objects naturally introduces a special form of mutual occlusion. For example, pinched fingers indicate a thin structure within the hands. It can be explicitly leveraged rather than to be marginalized over.

In Chapter 4, we propose an approach to reconstruct hand-object interaction without any templates from *a single image*. We study to build interaction prior – hand articulation is highly predictive of the object shape. In particular, given an image depicting a hand-held object, we first use off-the-shelf systems to estimate the underlying hand pose and then infer the object shape in a normalized hand-centric coordinate frame. We parameterize the object by signed distance which is inferred by an implicit network that leverages the information from both visual feature and articulation-aware coordinates to process a query point. We perform experiments across three datasets and show that our method consistently outperforms baselines and is able to reconstruct a diverse set of objects. We analyze the benefits and robustness of explicit articulation conditioning and also show that this allows the hand pose estimation to further improve in test-time optimization. This work is published as Ye, Yufei, *et al.* "What's in your hands? 3D Reconstruction of Generic Objects in Hands." CVPR 2022.

After building object prior and interaction prior for reconstruction from single images, in Chapter 5, we extend the data-driven priors to reconstructing HOI from short video clips. The input video naturally provides more multi-view cues to guide 3D inference than single images as the input. However, they are insufficient on their own due to occlusions and limited viewpoint variations in everyday interaction clips. To obtain accurate 3D, we augment the multi-view signals with generic data-driven priors to guide reconstruction. Specifically, given an input video, our proposed approach casts 3D inference as a per-video optimization and recovers a neural 3D representation of the object shape, as well as the time-varying motion and hand articulation. We empirically evaluate the current approach on egocentric videos across 6 object categories, and observe significant improvements over prior single-view and multi-view methods. We also demonstrate our system's ability to reconstruct arbitrary clips from YouTube, showing both 1^{st} and 3^{rd} person interactions. This work is published as Ye, Yufei, *et al.* "Diffusion-Guided Reconstruction of Everyday Hand-Object Interaction Clips." ICCV 2023.

Perceiving the ongoing interactions allows for the prediction of the possible interactions. In Chapter 6, we study to predict different spatial arrangements of human-object interactions. Given an image of an object. In particular, given an RGB image of an object, we synthesize plausible images of a human hand interacting with it. We propose a two-step generative approach: a high-level sampling that samples an articulation-agnostic hand-object-interaction layout, and a low-level sampling that synthesizes images of a hand grasping the object given the predicted layout. Both are built on top of a large-scale pretrained diffusion model to make use of its latent representation. Compared to baselines, the proposed method is shown to generalize better to novel objects and perform surprisingly well on out-of-distribution in-the-wild scenes. The resulting system allows us to predict descriptive affordance information, such as hand articulation and approaching orientation. This work is published as Ye, Yufei, *et al.* "Affordance diffusion: synthesizing hand-object interactions." CVPR 2023.

Finally, we draw inspiration from previous works and explore a unified data-driven prior that allows for both reconstruction and prediction (Chapter 7). We propose G-HOP, a generative prior for hand-object interactions that allows modeling both the 3D object and a human hand, conditioned on the object category. To learn a 3D spatial model that can capture this joint distribution, we propose a suitable 3D representation that represents the human hand via a skeletal distance field to obtain a representation aligned with the (latent) signed distance field for the object. We show that this hand-object prior can then serve as generic guidance to facilitate other tasks like reconstruction from interaction clip and human grasp synthesis. Additionally, we also put efforts into scaling up the available data by seven diverse real-world interaction datasets spanning across 155 categories. It results in a first approach that allows jointly generating both hand and object. Our empirical evaluations demonstrate the benefit of this joint prior in video-based reconstruction and human grasp synthesis, outperforming current task-specific baselines.

Excluded Research In order to keep the thesis clean, I exclude my work that learns to transfer semantic knowledge across categories by knowledge graphs [216]. It is published as Xiaolong Wang*, Yufei Ye*, Abhinav Gupta "Zero-shot Recognition via Semantic Embeddings and Knowledge Graphs." CVPR2018. (* denotes equal contribution). I also exclude my work that predicts interaction in scenes via scene graph, for both video prediction [234] and robot manipulations [231]. They are published as Ye, Yufei, *et al.* "Compositional video prediction." ICCV 2019 and Ye, Yufei, *et al.* "Object-centric model predictive control." CoRL 2019.

Chapter 2

Background

2.1 Human Hand Pose Reconstruction

Approaches tackling hand pose estimation from RGB(-D) images can be broadly categorized as being model-free and model-based. Model-free methods [29, 87, 142, 144, 153, 172, 173, 246] typically detect 2D keypoints and lift them to 3D joints position or hand skeletons. Some works [29, 53, 160] then directly predict 3D meshes vertices from the 3D skeleton by coarse-to-fine generation. Model-based methods [10, 180, 197, 240, 243] leverage statistical models like MANO [177] whose low-dimensional pose and shape parameters can be directly regressed [10, 180] or optimized [197, 240, 243]. These model-based methods are generally robust to occlusion, domain gap *etc.*, and we build on these in our works.

In particular, throughout the thesis, we use off-the-shelf systems [180] to get initial estimation of hand pose in images/videos. Although better initial hand poses [159] are expected to lead to better interaction understandings, we keep the same off-the-shelf system for consistent comparision. The reconstruction method is model-based which directly regresses a 45-dimensional articulation parameter (θ_A) and a 6-dim global rotation and translation (θ_w) along with a weak perspective camera. We rig the parametric MANO model by the predicted articulation pose θ_A to obtain an articulated hand mesh in a canonical frame around the wrist. To relate a point in the wrist frame to the image space, we first transform the hand by the predicted global transformation and then project it by the camera matrix. As an implementation detail, we convert the predicted weak perspective camera to a full perspective one as it helps to account for large perspective effects. In summary, we relate a query point in the canonical wrist frame to the image by

$$\mathbf{x}_p = \pi_{\theta_w}(\mathbf{x}) = KT_{\theta_w}\mathbf{x}$$

where K is the camera intrinsic and T_{θ_w} is the global rigid transformation of the hand.

While the thesis focuses on interactions rather than improving hand pose estimators, we also show in Chapter 4, 5, and 7 that jointly reasoning about the geometric interaction between the predicted 3D object and the inferred hand pose can also help improve the hand pose estimate.

2.2 Reconstructing Generic Objects in Isolation

While model-based method is widely used for hand reconstruction to inject our prior knowledge of human hands, it is significantly more challenging to obtain a unified object model that can be used for general object reconstruction because of the significantly wide diversity of object geometry. Model-free methods [1, 55, 154, 193, 220] learn a manifold of shape by first mapping the input to a latent space from which 3D shape is generated. These methods typically suffer from losing finer details as the reconstruction only rely on less expressive latent code. While all methods above have presented impressive results, they crucially require 3D supervision. In contrast, our approach in Chapter 3 aims for a coarse-to-fine inference where neither 3D nor pose annotation is available.

With a similar motivation as ours in Chapter 3 to relax the need of supervision, several approaches study the reconstruction task with only multi-view or even single-view supervision. The key is to ensure reprojection consistency of the predicted 3D with available observations. While this relaxes the requirement for tedious 3D supervision, manual annotations are still required in different forms, such as semantic key-points [93,105], multi-view association [30,126,204,219], categorical template [58,93,107,108], or camera pose annotation [76,77,97,126,227]. Some recent works use self-supervised semantic co-part segmentation [114], foreground masks [51,78], or symmetry [221] to further relax the manual annotation. Our work in chapter 3 has similar setup while ours does not require semantic in training, and reconstructs textured full 3D meshes with various topology and shapes.

It is worth noting the remarkable progress in 3D/4D object reconstruction [84, 121, 244] after our work in chapter 3 is published. The progress is mainly due to large-scale data [39,170], foundational models [42], and better 3D representations [21, 137, 145, 154, 230]. However, the state-of-the-art methods typically assume isolated and unoccluded objects in images – and cannot be directly leveraged for reconstructing hand-held objects. Even approaches that are robust to occlusion consider it as noisy context to marginalize over, instead of a source of signal for the shape of the underlying object. In contrast, this thesis advocates that explicitly taking hand pose into account helps infer the 3D structure of objects more accurately.

2.3 Hand-Object Interaction Reconstruction

2D Understanding of Hand-Object-Interaction. In order to understand hand-object-interaction, efforts have been made to locate the active objects and hands in contact in 2D space, via either bounding boxes detection [5, 140, 187] or segmentation [49, 188]. Furthermore, temporal understanding of HOI videos [56, 66, 163, 165, 196] aims to locate the key frames of state changes and time of contact. In this thesis, we focus on the geometry aspect of HOI.

HOI Reconstruction from Images. Reconstructing hand-objects interactions is even more challenging than isolated object reconstruction due to heavy mutual occlusions. Most of the prior works make the simplifying assumption of knowing the instance-specific object template and then reduce this ill-posed problem to 6DoF pose estimation [17,65,171,176,198,206,236]. Chapter 4 as well as previous work [73,96] explore a template-free approach to reconstruct more general objects by learning data-driven priors of interaction from large-scale datasets. While it is able to generate reasonable per-frame predictions, it is not trivial to aggregate information from multiple views in one sequence and generate a time-consistent 3D shape.

HOI Reconstruction from Videos. There have been many efforts in capturing hand-object interactions with multiple cameras or monocular RGBD cameras. Known (scanned) templates of either rigid or articulated objects are fitted to multiple sequences and can achieve very accurate reconstructions to even serve as pseudo ground truth of datasets [12, 46, 68, 186, 201, 205]. Another line of works recover the 6D object pose from monocular RGB videos [70, 71, 158]. While all previous works assume the reconstructed object to be known, a few very recent works focus on template-free in-hand scanning from monocular videos [67,85]. They directly leverage recent neural radiance field and neural implicit fields [117, 138, 154, 155, 164, 225, 230], that have shown great potential in novel view synthesis and representing generic 3D/4D scenes. However, their scanning setup [67, 85] requires every region of the objects to be fully observed, which is often not true for everyday video clips. In contrast to all prior works, we tackle template-free 3D HOI reconstruction from everyday video clips in Chapter 5 and 7.

2.4 Hand-Object Interaction Prediction

Visual Affordance from Images. Affordance is defined as functions that environments could offer [54]. Although the idea of functional understanding is core to visual understanding, it is not obvious what is the proper representation for object affordances. Some approaches directly map images to categories, like holdable, pushable, liftable, *etc.* [14, 79, 112, 147]. Some other approaches ground these action labels to images by predicting heatmaps that indicate interaction possibilities [47, 86, 123, 146, 152]. While heatmaps only specify *where* to interact without telling *what* to do, recent approaches predict richer properties such as contact distance [96], action trajectory [123, 141], grasping categories [60, 133], *etc.*. Instead of predicting more sophisticated interaction states, in Chapter 6, we explore directly synthesizing HOI images for possible interactions because images demonstrate both *where* and *how* to interact comprehensively and in a straightforward manner.

3D Grasp Synthesis. Interaction represented in 2D can not be directly used to command a robot to grasp an object in 3D. There are extensive works in robotics that predict 3D robot grasp [2, 11, 116, 132] for different end-effectors. Meanwhile, human grasp as a special end-effector receives great attention [13, 45, 61, 90, 96, 98, 124]. Most relevant work including GF [96] and GraspTTA [90] model a conditional probability of human hand given an object mesh. Chapter 6 explores the possibility of mining human interaction prior from large-scale image synthesis models to predict 3D human grasps without knowing 3D object models. The coarse but generalizable 3D hand pose prediction is shown as useful human prior for dexterous manipulation [4, 38, 104, 133, 167, 222]. When the object geometry is known, we directly leverage a generic joint hand-object generative prior (Chapter 7) and show that this leads to more natural human grasps than task-specific methods.

Chapter 3

Reconstructing Generic Objects in the Wild

We live in a 3D world where 3D understanding plays a crucial role in our visual perception. Yet most computer vision systems in the wild still perform 2D semantic recognition (classification/detection). Why is that? We believe the key reason is the lack of 3D supervision in the wild. Most recent advances in 2D recognition have come from supervised learning but unlike 2D semantic tasks, obtaining supervision for 3D understanding is still not scalable.

While some recent approaches [57, 220] have attempted to build supervised 3D counterpart of 2D approaches, the concerns about scalability still remain. Instead, a more promising direction is to learn models of single image 3D reconstruction by minimizing the amount of manual supervision needed. Early approaches in this direction focused on using multi-view supervision [219, 227]. However, obtaining multiple views of the same objects/scene is still not easy for the data in the wild. Therefore, recent approaches [93, 108, 148] have attempted to learn single-image 3D reconstruction models from image collections. These approaches have targeted use of category templates, pose supervision and keypoints to provide supervision (See Table 3.1). However, such supervision still limits the scalability to hundreds of categories.

Our work is inspired by recent approaches that forgo supervision by exploiting meta-supervision from the category structure and geometric nature of the task. More specifically, the two common supervisions used are: (a) **rendering supervision** ([108, 114]): any given image of an instance in a category is merely a rendering of a 3D structure under a particular viewpoint. We can therefore enforce that the inferred 3D shape be consistent

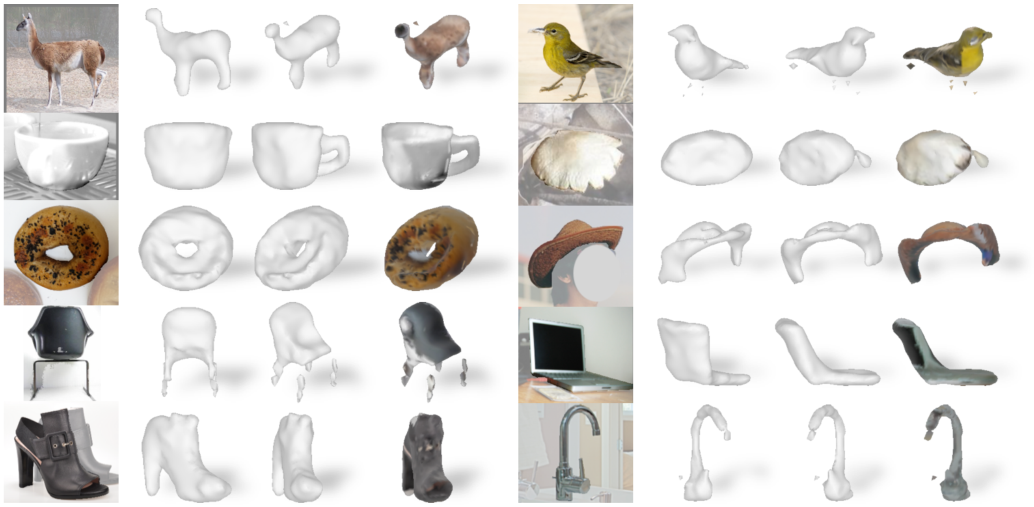


Figure 3.1: Given a single image, we predict a mesh with textures (rendered from the predicted view and a novel view). The models can learn directly from collections of images with only foreground masks, without supervision of mesh templates, multi-view association, camera poses, semantic annotations, *etc.*.

with the available image evidence when rendered; (b) **adversarial supervision** ([148]): in addition, the availability of an image collection also allows us to understand what renderings of 3D structures should look like in general. This enables us to derive supervisory signal not just from renderings of predictions in the input view, but also from novel views, by encouraging the novel-view renderings to look realistic. Prior work has exploited these supervisions but individually they pose several limitations for scaling 3D reconstruction models. For example, [58, 108] still requires template models. Similarly, [148] exploits the adversarial supervision and ignores the explicit geometric supervision. Therefore, such an approach only works on categories with strong structure and curated image collections. Specifically, table 3.1 summarizes the differences of our method with others in terms of supervision and outputs.

This chapter attempts to build upon the very recent successes in meta-supervision and provide an approach to scale learning of single image 3D reconstruction in the wild. We present a two-step approach: the first step relies on category-level understanding for coarse 3D inference (learned via meta-supervision). The second step specializes coarse models to match the details in the input image. Our approach can learn using only unannotated

Table 3.1: Comparing ours to other image-based supervised works in terms of supervision and outputs.

	[77]	[107]	[93]	[97]	[204]	[221]	[148]	[51]	[114]	ours
pose	✓		✓	✓						
template		✓	✓	✓						
semantic			✓						(✓)	
multi-view					✓					
mask	(✓)	✓	✓	✓	✓			✓	✓	✓
3D recon.		✓	✓	✓	✓	(✓)		✓	✓	✓
topology					✓			✓		✓
texture	✓		✓	✓	✓	✓	✓		✓	✓

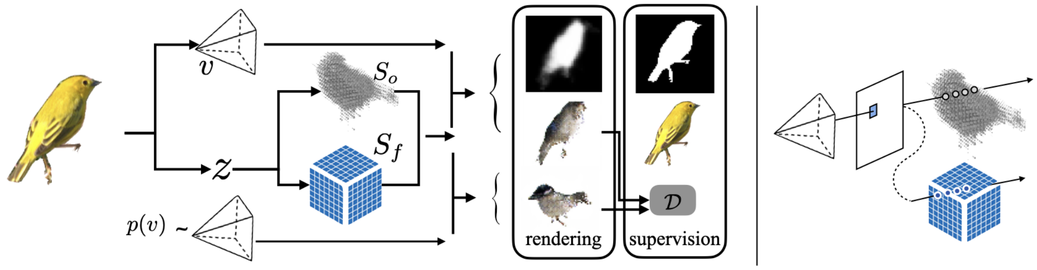


Figure 3.2: **Volumetric Representation Prediction and Rendering.** Left: Given an input image, the encoder-decoder network infers a semi-implicit volumetric representation (S_o, S_f) and a camera pose v. The semi-implicit volume is then projected from the predicted camera pose to obtain foreground image and mask. The semi-implicit volume is also projected from a novel view v'. The projections are required to both match the 2D observation and appear realistic. Right: Projection process mimicking ray marching.

image collections, without requiring any ground-truth 3D [9,57,214], multi-view [204,227], category templates [58,107], or pose supervision [97,219]. This not only allows our approach to infer accurate 3D, but also enables it to do so beyond the synthetic settings, using in-the-wild image collections in a 'shelf-supervised' manner: with only approximate instance segmentation masks obtained using off-the-shelf recognition systems as supervision. Yet our biggest contribution is the demonstration of scalability – we show results on order of magnitude more classes than existing papers.

3.1 Method

Our goal is to learn a model that, given an input image segmented with object mask, outputs a 3D shape in the form of a triangle mesh with texture and the corresponding camera pose. We use a two-step approach. First, we predict a canonical-frame volumetric representation and a camera pose to capture the coarse 3D structure which is consistent with categorical priors. We then convert this coarse volume to a memory-efficient mesh representation which is refined to better match the instance-level details.

We propose to learn the category level model from image collections using geometric and adversarial meta-supervision signals. More specifically, our key insight is that the projection of the predicted 3D should explain the observed images and masks, while also appearing realistic from a novel view.

3.1.1 Volumetric Reconstruction Model

First we define the volumetric reconstruction model which is learned separately for each category. Given an image this model predicts a volumetric representation in a canonical frame with corresponding camera pose. Note that unlike approaches that use a deformable category-level shape space, a volumetric representation allows us to capture larger shape and topology variations.

Concretely, we adopt a semi-implicit representation comprised of an explicit occupancy grid S_o, with an implicit 3D feature S_f, i.e. $S = (S_o, S_f)$. The latter can help capture appearance, texture, material, lighting, etc. This allows synthesizing both mask and appearance from a query view, and thereby lets us use both RGB images and foreground masks as supervision. The overall method is depicted in Figure 3.2.

Encoder-Decoder Architecture. We learn an encoder-decoder style network ϕ to predict this semi-implicit representation $(S, v) \equiv \phi(I)$ where v is the camera pose. The encoder maps the input image to a low-dimensional latent variable z and predict the camera pose, i.e. $(z, v) \equiv \phi_E(I)$. The latent variable z is then decoded to the volumetric representation, $S \equiv \phi_D(z)$. The key here is that the view-independent decoder learns to predict the shape in a canonical pose across all instances in the category. To further regularize the network, we leverage the observation that many objects exhibit reflection symmetry, and enforce a fixed symmetric plane $(x = 0)$ via averaging predicted features in symmetrically related locations.

Volumetric Rendering. Our goal is to supervise the volumetric model using only 2D observations. Therefore, what we need is a rendering function (π) which projects volumetric representation to obtain 2D images and masks from a query view i.e. $(I, M) \equiv \pi(S_f, S_o, v)$. Similar to other volumetric neural renderers [129,204], we use a geometrically informed projection process by mimicking ray marching.

For a given pixel $\mathbf{p}$, we use D samples along the ray to obtain a 'rendered' feature and mask value. Let us denote the coordinate of the d-th sample on the ray as $C_v + \lambda_d e_\mathbf{p}$ where $e_\mathbf{p}$ is the corresponding ray direction. We sample both implicit feature and occupancy at these locations, denoted as $S_f[C_v + \lambda_d e_\mathbf{p}]$ and $S_o[C_v + \lambda_d e_\mathbf{p}]$. We then composite these samples to obtain a per-pixel feature $s_f^\mathbf{p}$ and mask $s_m^\mathbf{p}$, by using the expected value with respect to ray stopping probability [203]:

$$\begin{aligned} s_f^\mathbf{p} = &\sum_{d=1}^{D} (S_o[C + \lambda_d e_\mathbf{p}] \prod_{h=1}^{d-1} (1 - S_o[C + \lambda_h e_\mathbf{p}])) \\ &\cdot S_f[C + \lambda_h e_\mathbf{p}] \end{aligned}$$

The pixelwise mask value $s_m^\mathbf{p}$ is similarly rendered by setting S_f to constant 1. While this process lets us directly compute the rendered mask M, we use a few upconvolutional layers to transform the rendered 2D feature image to the output color image.

Training. We supervise this network with only unannotated images and foreground masks. We use three different kinds of supervision (or terms in the loss function):

Pixel consistency loss. Our first term is the simplest one. Any predicted volumetric representation when rendered in the same camera view should explain the input image and mask. This is performed in color space, mask space [125], and perceptual space [238] .

$$\begin{aligned} \mathcal{L}_{rgb} &= \|\hat{I} - I\|_1 \\ \mathcal{L}_{mask} &= 1 - \frac{\|\hat{M} \otimes M\|_1}{\|\hat{M} \oplus M - \hat{M} \otimes M\|_1} \\ \mathcal{L}_{perc} &= \|h(\hat{I}) - h(I)\|_2^2 \end{aligned}$$

where $\hat{I}, \hat{M}$ are rendered image and mask; h is the feature extracted by a pretrained AlexNet [106] and $\oplus \otimes$ are element-wise summation and multiplication respectively.

View synthesis adversarial loss. A degenerate solution could arise such that the shape is only plausible from the predicted view. To avoid it, we require

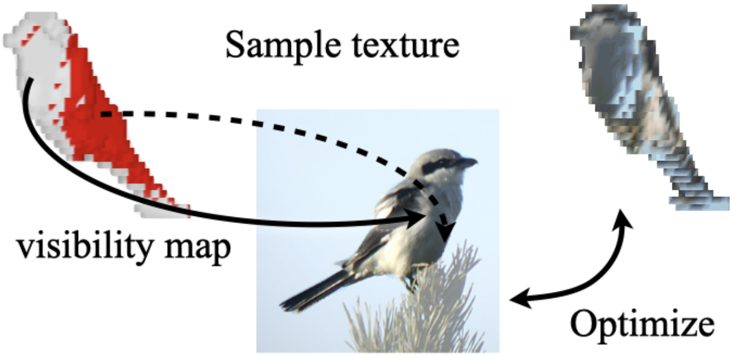


Figure 3.3: **Instance-level specialization.** We convert the inferred volumetric occupancy to an initial mesh. The mesh geometry and textures are then iteratively refined to better match the given input.

the projection of predicted shape to appear realistic from a random novel view. Specifically, we sample another camera pose from a fixed prior to render the novel view: $I' = \pi(\phi_D(z), v'), v' \sim p(v)$. We then present this generated image to an adversarial discriminator with an objective to fool it. We similarly encourage photo-realism when rendering from the predicted camera pose. The loss is minimized in a vanilla GAN [59] scheme.

$$L_{adv} = \log \mathcal{D}(I) + \log(1 - \mathcal{D}(\pi(S, v)) + \log(1 - \mathcal{D}(\pi(S, v'))$$

Content consistency loss. To further regularize the network we build on a insight that the encoder and decoder networks should be self-consistent. Given a synthesized image from the decoder, the encoder should predict the actual content (latent variable with camera pose) that generated that image. Formally,

$$\begin{aligned}\mathcal{L}_{content} = &\|\phi_E(\pi_S(S, v)) - (z, v)\|_2^2 \\ &+\|\phi_E(\pi_S(S, v')) - (z, v')\|_2^2\end{aligned}$$

Empirically, we found $\mathcal{L}_{content}$ important to stablize training.

Optimization. The neural renderer and decoder are trained to minimize all of the above objectives. But the encoder is not optimized with the adversarial loss, as in [111].

3.1.2 Instance-Level Specialization

As shown in Figure 3.3, the volumetric representation captures general category-level structure to hallucinate a full 3D shape. However, this shape is coarse as it is: a) limited by the volumetric resolution, and b) generated only from a low-dimensional latent variable. On the other hand, a mesh representation is more flexible and can allow capturing the finer shape details. We therefore go beyond this coarse volumetric prediction, and capture details specific to this instance by converting the volume to an initial mesh, which is then adjusted to better match the input image.

Volumetric to Mesh. We first obatin an initial mesh from the predicted volumetric occupancy. This is done similar to Mesh-RCNN [57] by binarizing the occupancy grid S_o and extracting its surfaces. Next, every vertex is projected to the image to obtain visibility and texture at the vertex. At this step we leverage the symmetry of the mesh to fuse the textures from its reflective symmetric vertex. The final associated texture for each vertex is an average of itself and its visible symmetric neighbors.

Mesh refinement. We optimize the geometry and refine the texture of the mesh iteratively. Given a posed textured input mesh, we first optimize the vertex location and the camera pose such that the projection of the mesh matches the observation. After every step of mesh geometry update, vertex textures are re-sampled from the image given the adjusted projected location. More specifically, we use a mesh-based differential renderer [125] to project and render. The rendered images and masks $(\hat{I}, \hat{M}) \equiv \pi_G(G, v)$ are encouraged to be consistent with the input image and foreground mask. We regularize the optimization by penalizing large vertex displacement $\|\delta X\|_2^2$ and encourage Laplacian smoothness $\|\Delta X\|_2^2$.

3.2 Experiments

Our goal is to highlight how our approach learns to predict 3D meshes from image collections in the wild. Specifically, we show 3D reconstruction for 50 object categories from OpenImages dataset [110]. Note that this diverse set of reconstructions is an order of magnitude larger than those of any existing approaches.

However, there is no ground truth for OpenImages. Also, most baseline approaches fail to work on uncurated image collections. In order to provide comparisons, we perform two additional experiments. First, we compare on data drawn from 3D Warehouse [217], using rendered images as image

Table 3.2: Quantitative results (3D IoU / F-score) on synthetic data comparing different methods for shape reconstruction.

	airplane	car	chair
HoloGAN [148]	0.28/ 0.31	0.43 / **0.44**	0.26 / 0.25
PrGAN [51]	0.29/ 0.18	0.48 / 0.37	0.28 / 0.28
Ours	**0.33** / 0.46	**0.55** / 0.43	**0.31** / 0.29
Ours (refined)	—- / **0.49**	—- / 0.42	—- / **0.31**

collection. Using synthetic data allows us to provide quantitative evaluation and perform ablative analysis. Second, we also compare with some of the other curated common datasets used in the literature (CUB, Chair-in-Wild, ImageNet Quadrapeds). This helps us to qualitatively compare with some baseline approaches.

3.2.1 Synthetic Data

We first evaluate our method on models from 3D Warehouse [217], using the subset recommended by Chang *et al.* [22]. We select three categories which are commonly used to evaluate single-view reconstruction: aeroplane, car, and chair. Note that within a category, the shapes across instances can have a large variation and even different topology, especially for chairs. Each 3D model is rendered from 20 views, with uniformly sampled azimuth $[0°, 360°]$ and elevation elevation $[-60°, 60°]$. However, the network is not provided with multi-view associations in training.

Evaluation metrics. We report 3D IoU with resolution 32^3 and F-score in the canonical frame for volumetric reconstruction and report F-score [200] for mesh refinement. The F-score can be interpreted as the percentage of correctly reconstructed surface. As our predictions (and those of baselines) can be in an arbitrary canonical frame that is different from the ground truth frame, we explicitly search for azimuth, elevation for each instance and binarizing threshold for each category to align the predicted canonical space with the ground-truth.

Baselines. We compare our approach to [51,148]. We adapt HoloGAN [148] by training their system on our data, and obtaining a 3D output by adding a read-off function from the learned volumetric feature to occupancy by enforcing the reprojection consistency with foreground masks. We implement PrGAN [51] using our encoder-decoder network. Our implementation provides a boost to original PrGAN.

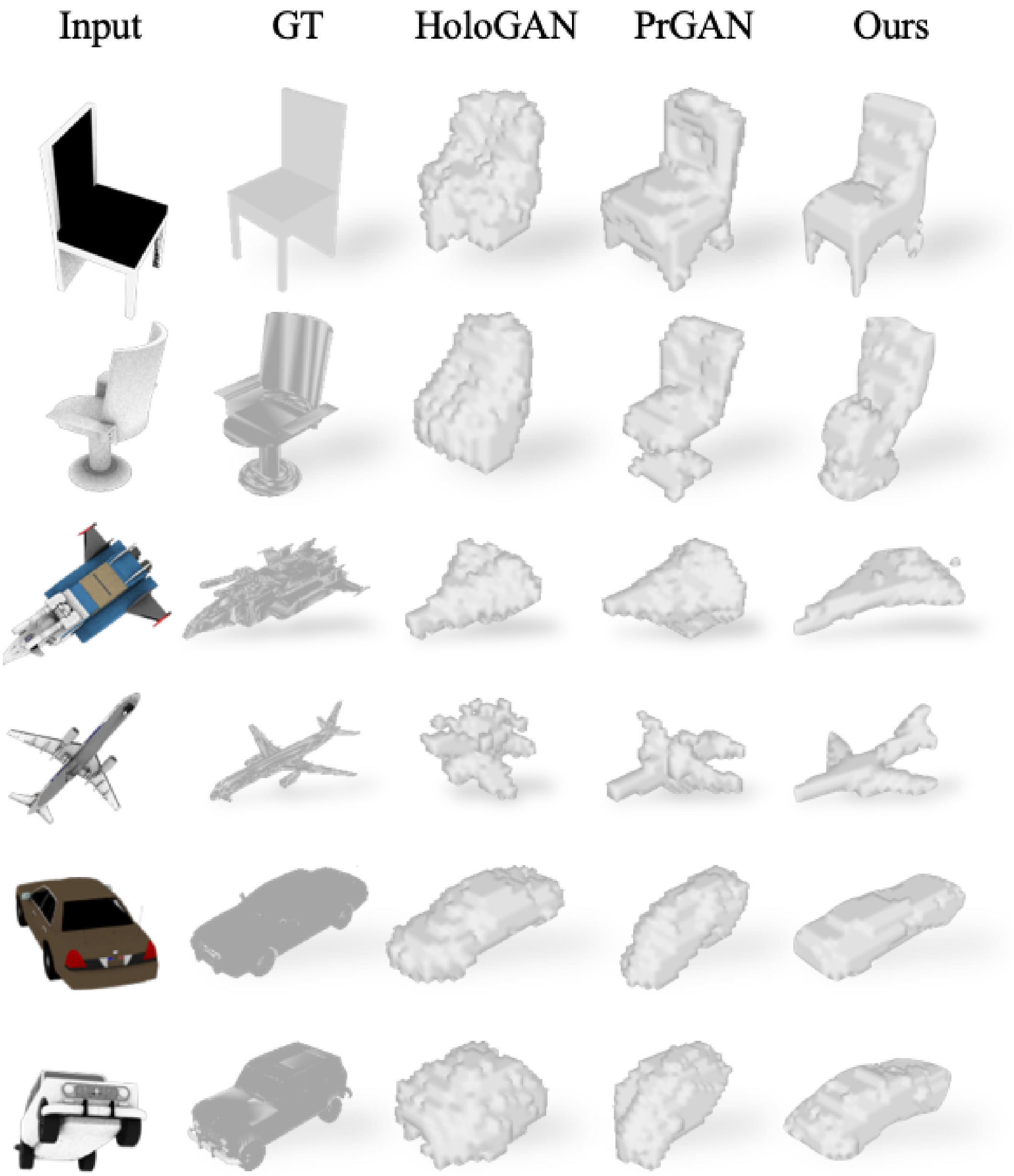


Figure 3.4: Visualization of categorical volumetric representation across different methods.

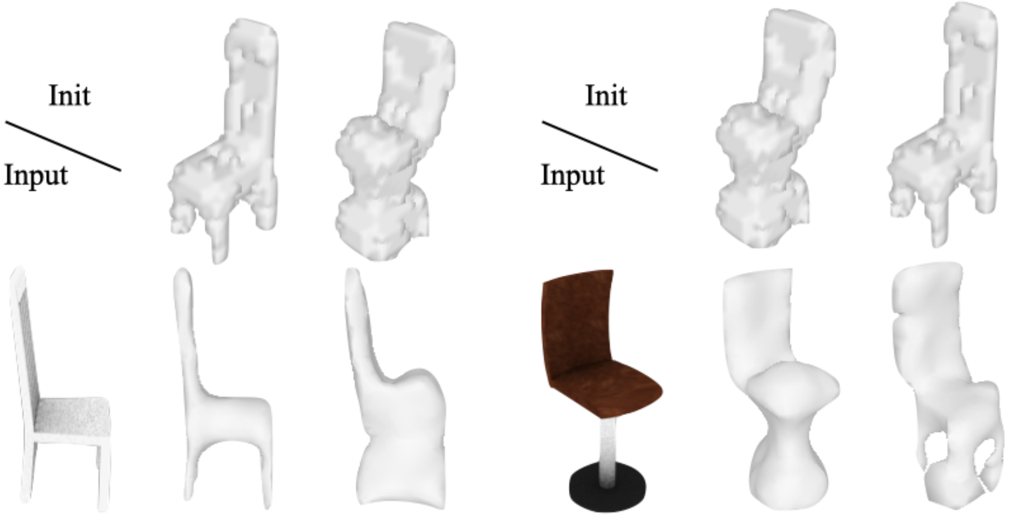


Figure 3.5: Ablation study: refining mesh initialized from the predicted volume (col 2/5) and another volumes (col 3/6).

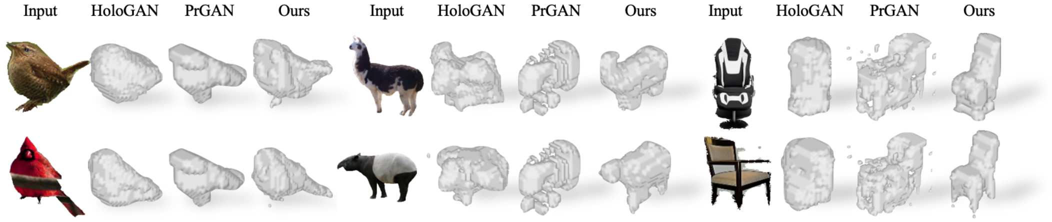


Figure 3.6: Visualizing categorical volumetric representation across different methods on CUB-200-2011, Quadrupeds, Chairs in the wild.

Figure 3.4 visualizes the reconstructions in a canonical frame on 3 categories. HoloGAN is able to reconstruct a blobby shape, but as it does not explicitly represent 3D occupancies, it struggles to generate a coherent shape. PrGAN is able to capture the coarse shape layout but it lacks some details like flat body of aeroplanes. In contrast, we reconstruct the shape more faithfully to the ground-truth as we leverage information from both appearance and foreground masks. Quantitatively, we report the 3D IoU on these categories in Table 3.2 and, consistent with the qualitative results, observe empirical gains across all categories.

Mesh Refinement. Table 3.2 also reports the evaluation of the mesh refinement stage. Compared with the initial meshes converted from volumetric representation, our specialized meshes match the true shape better.

Table 3.3: Quantitative results (3D IoU) on synthetic data to ablate the effect of each loss term.

	airplane	car	chair
Ours	**0.33**	**0.55**	**0.31**
Ours $-\mathcal{L}_{adv}$	0.25	0.44	0.22
Ours $-\mathcal{L}_{cont}$	0.24	0.54	0.23

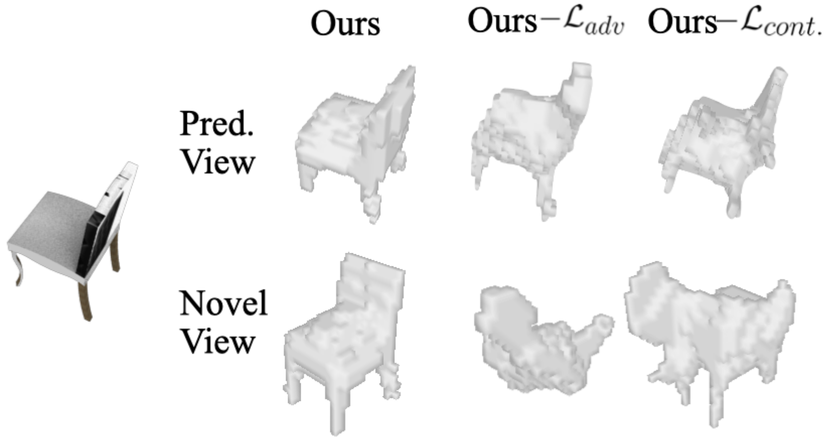


Figure 3.7: Ablation study: comparing reconstructed volumes when the model disables different loss terms.

In Figure 3.5, we visualize the refinement results with an interesting ablation to further highlight the importance of mesh initialization. Instead of initializing with our predicted volume, we initialize the mesh from another chair consisting of different numbers of chair legs. The refinement fails to specialize well. This indicates that the meshes for all instances cannot be adjusted from one single shape especially when shapes have a large variance, and that our volumetric prediction, though coarse, provides an important initialization for the instance-level refinement.

Loss ablation. We provide quantitative (Table 3.3) and qualitative (Figure 3.7) results to show each loss term is necessary. Without adversarial loss, the model collapses to generate shapes only looking similar to the input from the predicted view. It does not even look like a chair from another view, since this degenerate solution is not penalized by other losses. Without content loss, the performance also drops, especially on categories with larger shape variance like chairs. The consistency loss is not ablated because it is needed for the task of reconstruction.

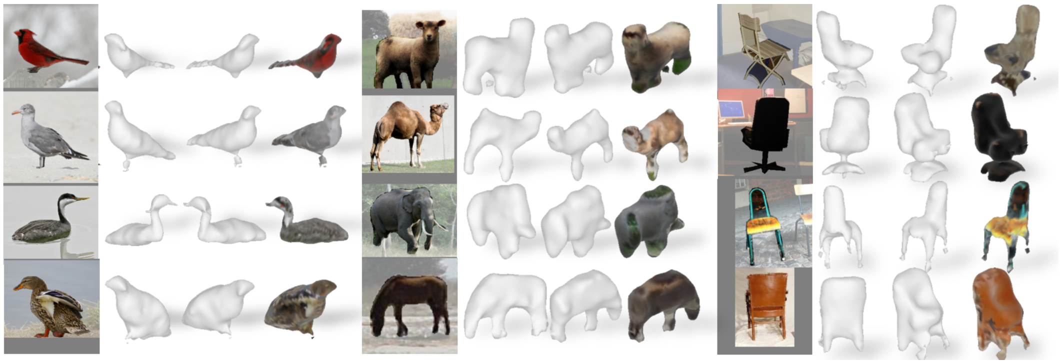


Figure 3.8: Visualizing our refined shapes from the predicted view (2nd column in each quadruplet) and a novel view (3rd+4th / quadruplet) on CUB, Quadrupeds, Chairs in the wild. We are able to capture both the shared shapes in categories and instance-specific differences.

3.2.2 Curated Collections

We also examine our method on three real-world datasets that have been curated and used in the literature for the 3D reconstruction problem:

CUB-200-2011 [211]: The CUB dataset consists of 6k images of 200 bird species with annotated foreground masks.

Quadrupeds from ImageNet [41]: The Quadrupeds dataset consists of 25k images of different quadrupeds from ImageNet. Masks are provided by Kulkarni *et al.* [107], who use an off-the-shelf segmentation system [102] and manually filter out the truncated or noisy instances. Quadrupeds consists of multiple 4-legged animal species including buffalo, camels, sheep, dogs, etc. The animals also exhibit rich articulation *e.g.* running, lying, heads up or heads down. This makes the underlying shape variance significantly larger than the CUB dataset.

Chairs in the wild [41, 150, 226]: For chairs in the wild, we combine chairs in PASCAL3D, ImageNet, and Stanford Online Products Dataset to get 2084 images for training and 271 for testing. Masks in [41, 150] are from segmentation systems [25, 102] and those in [226] are from annotations.

Figure 3.6 qualitatively compares our volumetric reconstruction to baselines on the 3 separate real-world datasets. Similar to results on synthetic data, HoloGAN reconstructs only coarse blobby volumes as it does not explicitly consider occupancy or geometric-informed projection. PrGAN collapses to shapes with little variance, since it does not use appearance cues. But the real datasets have noisier foreground masks and textures contain more information. In contrast, we are able to learn the coarse categorical shape just from the foreground images. Our reconstructions also capture subtle differences like the length of bird tails, articulated heads of the quadrupeds, the style of chairs.

Mesh Refinement. Figure 3.8 visualizes our refined meshes from the predicted view and a novel view on these three dataset. We observe that we predict meaningful texture even for invisible regions and that the shape of the mesh also looks plausible from another view. On CUB-200-2011, our method captures the categorical shapes like blobby bodies, beaks and tails while captures subtle shape differences between birds such as the tail length, body width, neck bending, etc. On Quadrupeds, we are able to capture quadrupeds common traits such as torso with one head and front back legs. We can also depict their uniqueness such as the camel hump and longer legs, the tapir having stout neck, the sheep raising up its heads, the horse bending down its neck, *etc.*. On Chairs in the wild, the learned common model differentiates one-leg and four-leg chairs respectively. The four legs and seat can be hallucinated even when occluded. The subtle differences such as a wide or a narrow chair back are also captured. Despite the challenges in the datasets, it is encouraging that our model can capture both, the common shapes and specialized details just by learning from these unannotated image collections.

3.2.3 OpenImages 50 Categories

Finally, the highlight of our model is the ability to scale to images in the wild. We evaluate our model on 50 categories on Open Images including bagel, water tap, hat, *etc.*. The size of each category ranges from 500 to 20k. The foreground masks [7] are from annotation and filtered by a finetuned occlusion classifier. Figure 3.10 visualizes the reconstructed meshes from the predicted view and a novel view. Our method works on a large number of categories, including thin (water taps, saxophone), flat (wheels, surfboards), blobby structures (Christmas trees, vases). We are able to reconstruct shapes with various topology such as bagels, mugs, handbags. The model captures the categorical shapes shared within classes and hallucinates plausible occluded regions (mushroom, mugs). We can also captures details at instance-level, such as the number of wheels of roller-skaters, styles of high-heels, hats, *etc.*.

Integrating on COCO. We additionally show results of our models on COCO [120] without fine-tuning (Figure 3.9). We first detect and segment the objects with off-the-shelf segmentation [102] system. Based on the predicted

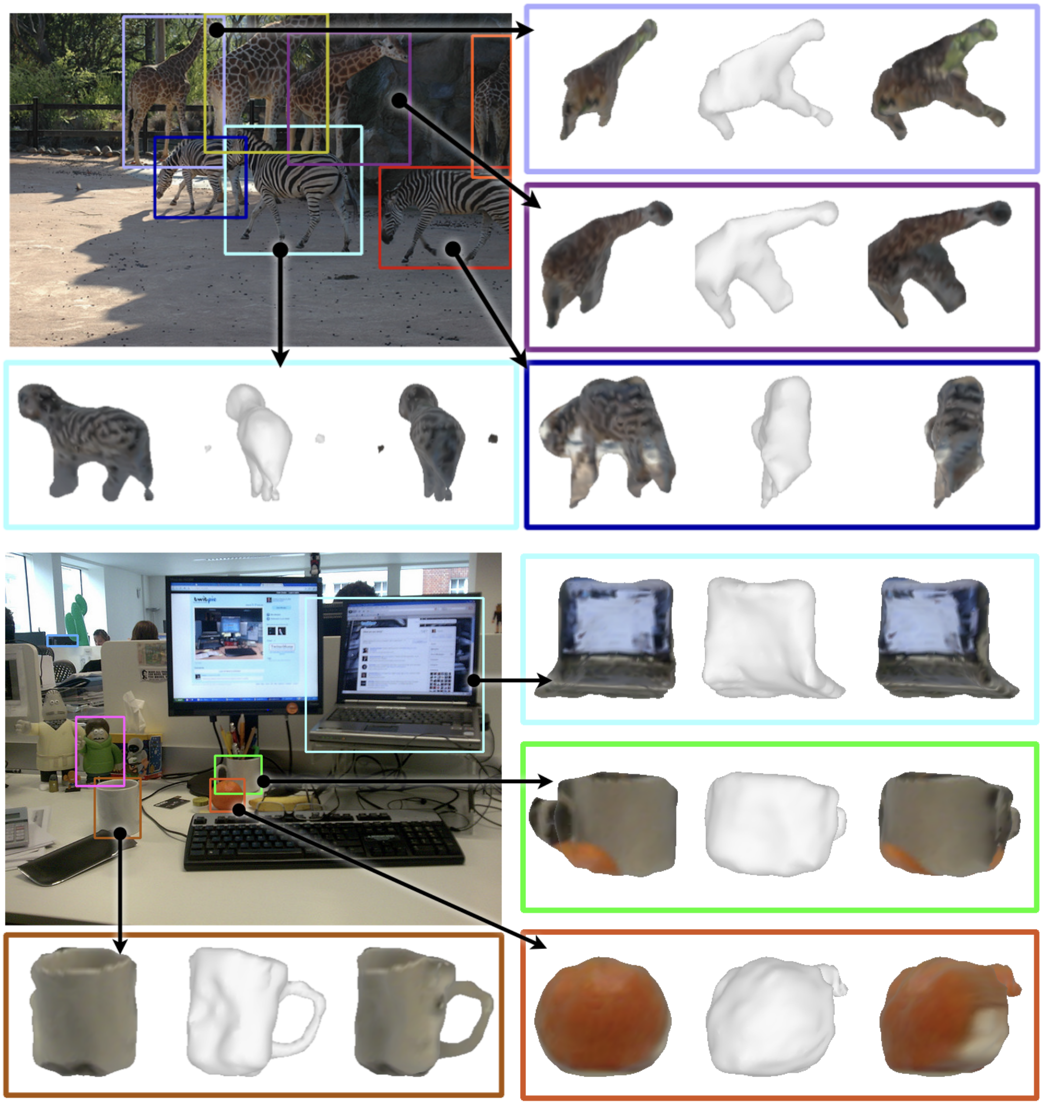


Figure 3.9: Test on COCO: visualization of lifting detection results to meshes via the shelf-supervised models.

classes, we then pass the segmented objects to our category-specific models which are trained on previous datasets. Despite more cluttered scenes and the dataset domain shift, our models can lift the 2D detection to 3D meshes for various categories while preserving instance details.

3.3 Discussion

In this chapter, we presented an approach to predict 3D representations from unannotated images by learning a category-level volumetric prediction followed by instance-level mesh specialization. We found that both are important to infer an accurate 3D reconstruction. While we obtained encouraging results across diverse categories, our approach has several limitations. For example, our rendering model is simplistic and not incorporate lighting during rendering. Thus we cannot easily reason about concave structure. Additionally, while we only examined setups without annotated supervision like mesh templates, our system could potentially incorporate additional (sparse) supervision to improve the reconstruction quality. While these challenges still remain, we believe our work on inferring accurate reconstruction with limited supervising can provide a scalable basis towards the goal of reconstructing generic objects in the wild.

3.4 Implementation Details and Additional Results

3.4.1 Ablation Study

Assumption of viewpoint distribution. We briefly analyze the effect of viewpoint prior. In figure 3.12 we visualize volumetric reconstruction training with different viewpoint prior on the mug category of synthetic data. While our method is robust to some view distribution mismatch, the shapes display artifact (*e.g.* two handles) when the assumed prior is far from the ground-truth viewpoint distribution. It is because different viewpoint distribution may induce different 3D shapes as the adversarial loss matches its projections with the existing image collections. We notice similar artifacts when training on the real datasets (*e.g.* starfish and mugs on OpenImages), as camera pose biases exist by human photographers (*e.g.* front view of starfish or mugs with handles). While we assume azimuth from uniform distribution across all experiments and have achieved some promising results on various categories, we encourage more works to explore the direction of better viewpoint distribution prior.

Robustness against segmentation quality. Our model depends on the segmentation quality, as it is the only supervision. We ablate our model with noisy masks, both qualitatively and quantitatively. The model trained/tested with predictions from [19] (left) or with synthesized noise (mid) performs comparably to using GT, until considerably severe corruption. Our experiments in paper have already suggested that our model is robust to the noise as masks might be truncated, occluded, or corrupted due to prediction error (Fig 3.11 right). We also visualized the masks used in the main paper (Fig 3.11 right). Our experiments suggest that our model is robust to the noise as masks might be truncated, occluded, or corrupted due to prediction error.

3.4.2 Architecture Details

Neural Network Architecture. The encoder is comprised of 4 convolution blocks followed by two heads to output v and z. Each block consists of $Conv(3 \times 3) \rightarrow LeakyReLU$. The feature from the last block is fed to 2 fully-connected layers to get v and is fed to Average Pooling with another fully-connected layer to output z. v is in 2-dim to represent azimuth and elevation while the dimensionality of latent variable z is 128.

The decoder follows StyleGAN [94] to use the latent variable z as a "style" parameters to stylize a constant 256×4^3 feature. Given z, the constant is upsampleed to the implicit 3D feature S_f by a sequence of style blocks. Then S_f is transformed to get the occupancy grid S_o by a $3 \times 3 \times 3$ Deconv layer with Sigmoid activation. Among all of our experiments, our decoder consists of 2 style blocks each of which are built with $Deconv \rightarrow AdaIN \rightarrow LeakyReLU$. The shape of S_f is 64×16^3 and the shape of S_o is 1×32^3.

Training Details. We optimize the losses with Adam [100] optimizer in learning rate 10^{-4}. The learning rate is scheduled to decay linearly after 10k iterations, following prior work [245]. We weight the losses such that they are around the same scale at the start of training. Specifically, we use $\lambda = 10$ for $\mathcal{L}_{pixel} + \mathcal{L}_{perc}$, 1 for $\mathcal{L}_{adv}$ and $\mathcal{L}_{content}$. The volumetric reconstruction network is optimized for 80k. Due to the diverse appearance and data noise on Quadrupeds, we additionally regularize the network by an L2 distance between the predicted voxels and the mean shape of all quadrupeds. The model can still capture the articulation for different instance.

3.4.3 F-score Calculation

In order to calculate F-score – the harmonic mean of recall and precision, the meshes are first converted to point cloud by uniformly sampling from surfaces. The recall is considered as the percentage of ground-truth points whose nearest neighbour in predicted point cloud is within a threshold while the precision is calculated as the other prediction-to-target way.

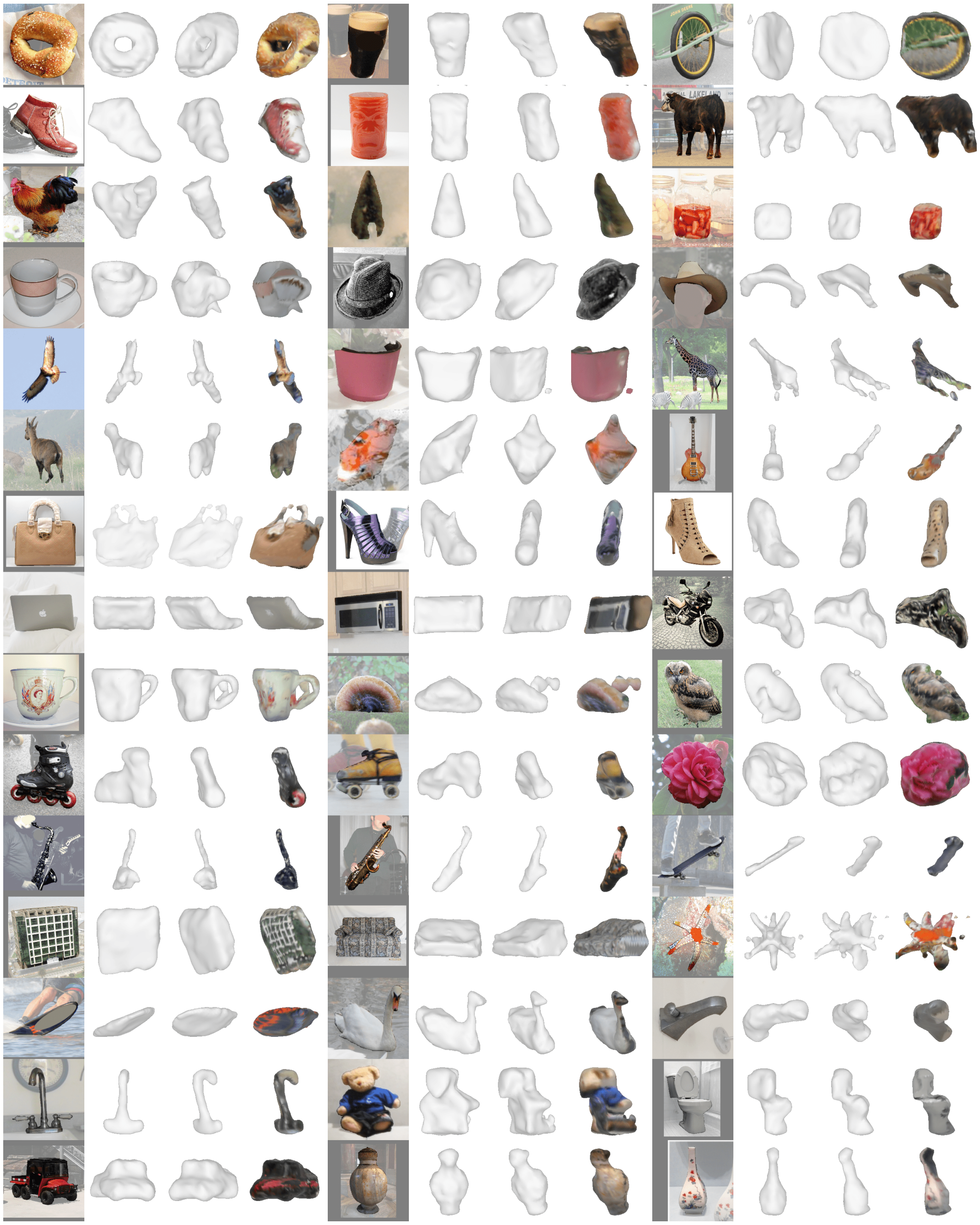


Figure 3.10: Visualizing our reconstructed meshes from the predicted view (2nd in each quadruplet) and a novel view (3rd and 4th in each quadruplet) trained on multiple categories on Open Images.

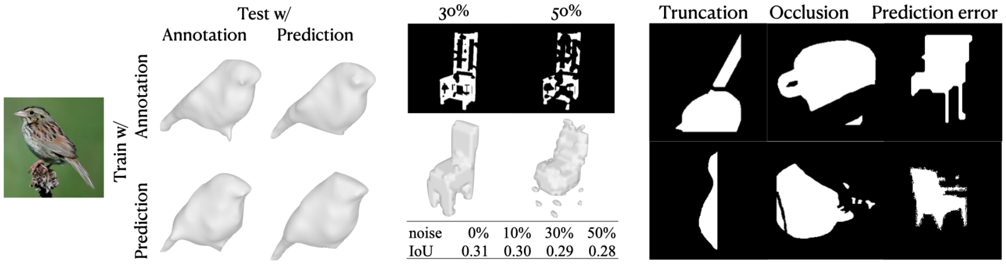


Figure 3.11: Left: qualitative results on CUB replacing annotation with prediction in training and/or test time. Middle: quantitative results on synthetic chairs by adding random noise on masks during both training and inference. Right: masks used to train the models in the paper.

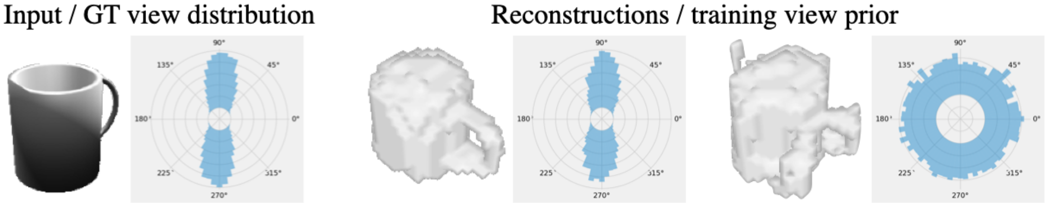

Input / GT view distribution
Reconstructions / training view prior
90°
45°
0°
135°
180°
225°
270°
315°

Figure 3.12: Results on training models with different viewpoint priors.

Chapter 4

Reconstructing Generic Objects in Hands from Single Images

In Chapter 3, we examine scaling up 3D reconstruction for single objects. But those objects are considered in isolation as neither occlusion nor interaction is assumed. Instead, objects are often manipulated by human hands in the real worlds. This interaction naturally introduces contact and occlusions. For example, holding a pen means a stick lying on purlicue and gripped by thumb, index and middle fingers; holding a bowl is placing it on top of an upfacing palm. In the following two chapters, we pursue a *geometric* representation of hand-object interactions for generic objects, from single images (this chapter) to monocular videos (Chapter 5).

Reconstruction of objects in hand in-the-wild is highly challenging and ill-posed due to lack of data, presence of mutual and self-occlusion. Current works [18,52,69,122,201] typically focus on reconstructing objects with known templates, thus reducing the task to 6D pose estimation. We argue that knowing the 3D template of the object as a priori during inference is a strong assumption and prevents these systems from reconstructing unknown objects. Furthermore, they struggle to handle various object shapes in the wild as these templates are rigid and instance-specific. In contrast, our work studies hand-object reconstruction without object templates and instead focuses on reconstructing HOI for novel objects from images.

Our key observation is that hand articulation is driven by the local geometry of the object. Thus, hand articulation provides strong cues for the object in interaction. Fingers curled like fists indicate thin handles in between while open palms are likely to interact with flat surfaces. Instead of treating the hand occlusion as noise to marginalize over, we explicitly consider hand pose as informative cues for the object it interacts with. We operationalize

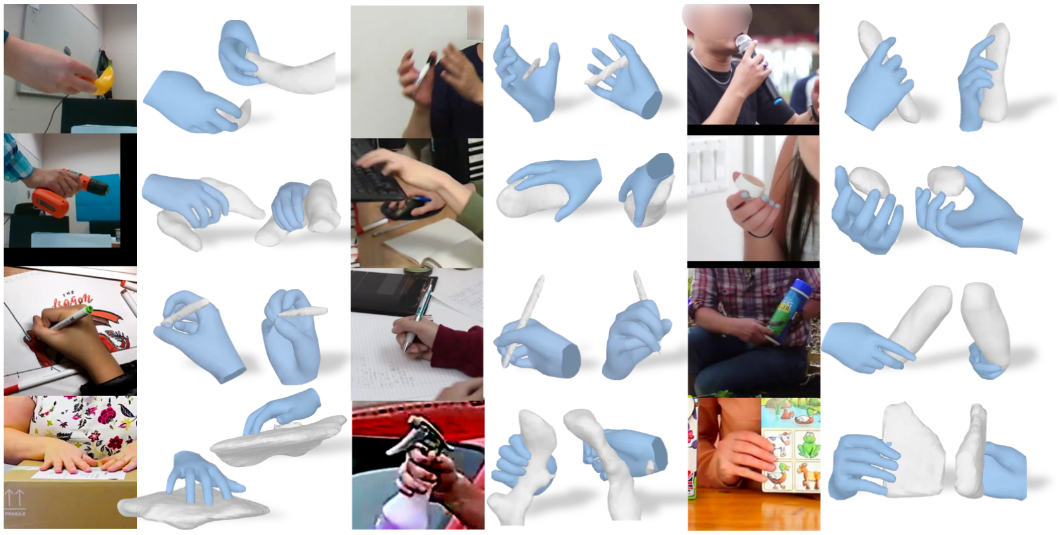


Figure 4.1: Given an RGB image depicting a hand holding an object, we infer the 3D shape of the hand-held object (rendered in the image frame and from a novel view).

this idea by conditionally predicting the object shape based on hand articulation and the input image. Instead of estimating both hand pose and object shape jointly, we leverage advances in hand pose reconstruction to estimate hand pose first. Given the inferred articulated hand along with the input image, our approach then reconstructs the object in a normalized hand-centric coordinate frame.

We evaluate our method across three datasets including synthetic and real-world benchmarks and compare ours with prior explicit and implicit HOI reconstruction methods that infer the shape of unknown objects independent of hand pose. Our articulation-conditioned object shape prediction consistently outperforms prior works by large margins and can reconstruct various objects in a wide range of shapes. We also analyze how our model benefits from articulation-aware coordinates. Lastly, we show that the initial hand pose estimation could be further improved by encouraging interaction between the predicted hand and the object.

4.1 Method

Given an image depicting a hand holding an object, we aim to reconstruct the 3D shape of the underlying object. Our key insight is that the hand articula-

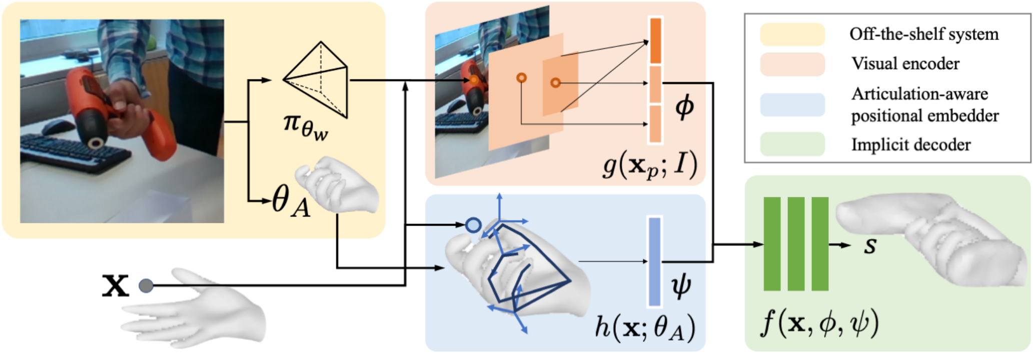


Figure 4.2: Given an image of a hand-held object, we first use an off-the-shelf system to estimate hand articulation θ_A and the camera pose π_w. With the predicted articulated hand along with the image, the object shape is reconstructed by an implicit network. For each query point $\mathbf{x}$ in canonical hand wrist frame, it is transformed to image space $\mathbf{x}_p$ to get visual feature $\phi = g(\mathbf{x}_p, I)$. In parallel, we also encode its articulation-aware representation $\psi = h(\mathbf{x}; \theta)$. Then we use an implicit decoder to predict signed distance value $s = f(\mathbf{x}, \phi, \psi)$.

tion is predictive of the object shape within it, for example, fingers pinching together indicate a thin stick-like structure between them. We operationalize this by explicitly conditioning the inference of object shape on the (predicted) hand articulation.

As shown in Fig 4.2, we first use an off-the-shelf system to estimate hand articulation and predict the camera transformation that projects the canonical articulated hand to the image coordinates. Given the predicted hand along with the input image, we then infer the object shape via an articulation-conditioned reconstruction network. This network is implemented as a pointwise implicit function [154] that maps a query 3D point to a signed distance from the object surface, and the zero-level set of this function can be extracted as the object surface [130]. Instead of predicting this 3D shape in the image coordinate frame, our implicit reconstruction network infers it in a normalized frame around the hand wrist. This allows the network to learn relations between the hand articulation and object shape that are invariant to global transformations.

More formally, given an input image I, we first infer the underlying the hand pose θ and the camera pose π. Then, for any point $\mathbf{x}$ in the normalized wrist frame, the object inference model takes in the query point with the image and predicts its signed distance function s. More specifically, the projection of the point to image coordinates is used to obtain corresponding visual features $\phi = g(\pi(\mathbf{x}); I)$. In parallel, we also encode its position relative to each hand joint to extract an articulation-aware representation $\psi = h(\mathbf{x}; \theta)$. The point-wise visual feature and articulation embedding are then used by an implicit decoder to predict signed distance value $s = f(\phi, \psi)$ at the query point x.

4.1.1 Hand Reconstruction

As explained in Section 2.1, we use an off-the-shelf system [180] to estimate hand articulation and associated camera pose from an input image. A point in the canonical wrist frame where hand palm always faces upwards is projected to the image space by $\mathbf{x}_p = \pi_{\theta_w}(\mathbf{x}) = KT_{\theta_w}\mathbf{x}$. K is the camera intrinsic matrix and T_{θ_w} is the global rigid transformation of the hand.

4.1.2 Articulation-conditioned SDF

Given the predicted hand articulation θ_A, and camera matrix π_{θ_w}, our articulation-conditioned object shape inference network takes an additional input image I to output a signed distance field of the object. For a query point in the wrist coordinate frame, the point-wise network takes into account the query's corresponding visual feature from a visual encoder and its relative position to each joint from an articulation-aware positional embedder. The visual feature and the embedding are then passed to an implicit decoder along with the query to predict the signed distance.

Visual encoder. The visual encoder first extracts the image feature pyramid at different resolutions. For a query 3D point in the wrist frame, the visual encoder projects it to the image coordinate and compute global and local feature from the pyramid. The global part allows us to reason about global context and generate more coherent object shapes. For example, realizing the object is a bottle helps the network to generate a cylinder shape. The local feature allows the prediction more consistent with the visual observation [183,235].

The backbone of the visual encoder is implemented as ResNet [74]. The global feature is a linear combination of the averaged conv5 feature. The local feature for each point is an interpolated feature at image coordinate from where it is projected by the predicted camera $\phi[\pi_{\theta_w}(\mathbf{x})]$, where ϕ denotes resnet feature and $\phi[x]$ represents a bilinear sample of the feature at a 2D location x. The local feature sampling is implemented for every layer of the feature pyramid $\phi_{1,\ldots,5}[x]$. It allows the model to draw visual cues with various resolutions and receptive fields.

Articulation positional embedder. Our key idea is that the hand pose is predictive of the object shape it interacts with. This is especially informative and complements the visual cues for reconstructing hand-held objects that are often occluded. We explicitly encode hand pose information for the query point via the articulation embedder. To do this, one naive way is to simply use identity mapping $\psi = \theta_A$. However, this representation is not robust to hand prediction error as we show in ablation. Furthermore, it is not trivial for the network to relate the reconstruction metric space with the hand pose joint space. For example, if a point is within 2mm from both index and thumb, it is very likely to have some object passing through. To better capture the structure of the problem, we encode the hand pose information by the position of the query points relative to the articulated joints.

More specifically, the articulation embedder takes as input an articulation parameter θ_A and a point position in wrist frame X to output the articulation-aware encoding $\psi = h(X; \theta_A)$. The encoding is a concatenation of the coordinates relative to every joint. Given the articulation parameter θ_A, we run forward kinematics to derive transformation $T(\theta_A) : \mathbb{R}^3 \to \mathbb{R}^{45}$ that maps a point in wrist frame to each joint coordinate. The 15 joint coordinates are position encoded [209] and concatenated together as the final representation $h(\mathbf{x}; \theta_A) = \gamma(T(\theta_A)\mathbf{x})$ where γ is the positional encoder. For more details please refer to appendix.

Implicit decoder. The decoder maps the query points with visual feature and articulation embedding to a signed distance value $f(\mathbf{x}, \phi, \psi) = s$. These two representations ϕ, ψ are concatenated together and passed along with the query point to the decoder. The decoder simply follows the design in DeepSDF [154] which consists of 8 layers of MLP with a skip layer.

4.1.3 Training

To learn our articulation conditioned neural implicit field, we rely on a training dataset where we assume known hand pose and 3D shape of the object in the image frame. We preprocess the data by sampling points inside and outside of the object around the hand to calculate the ground-truth SDF. 95% of the points are sampled around the surface of the objects and others are sampled uniformly in the space. During training, the network optimizes to match the predicted SDF to the ground-truth at the sampled points with the eikonal term as a regularizer.

$$\mathcal{L} = \|s - \hat{s}\| + \lambda(\|\nabla s\| - 1)^2$$

After the network is learned, at inference time, we do not require knoledge of the object 3D shape in the canonical frame a priori which is a major limitation of most most prior works.

4.1.4 Refining hand pose

While our work primarily focuses on object reconstruction conditioned on a predicted hand pose, this initial pose prediction, while reliable, is not perfect. As object reconstruction also leverages visual cues, our insight is that it can provide complementary information to further refine the predicted hand pose. During inference, we show that the predicted hand pose and object shape can therefore be further (jointly) optimized by enforcing physical plausibility – by encouraging contact while discouraging intersection.

We optimize the articulation pose parameters with respect to these two interaction terms, which can be naturally incorporated with an SDF representation. To discourage intersection between the hand and object, we penalize if the points on the hand surface are predicted to have negative SDF values by the object reconstruction model. Following prior work [73], we encourage hand-object contact for specifically defined regions – if the surface points in these contact regions are near the object surface, they are encouraged to come even closer.

$$\begin{aligned}\min_{\theta} \sum_{\mathbf{x} \in \mathcal{H}} & \|\max(-f(\mathbf{x}), 0)\| + \\ & \sum_{\mathbf{x} \in \mathcal{C}} \max(\|\min(f(\mathbf{x}) - \tau, 0)\| - \epsilon, 0)\end{aligned}$$

Note that the SDF f is conditioned on articulation thus it is also a function of the hand pose θ. As we refine the hand pose, the SDF of the object also changes accordingly. One could continuously update the SDF every time the pose is updated but we use a simpler solution that fixes the SDF during hand pose optimization and only update it once using the final optimized pose θ.

	ObMan				HO3D				MOW			
	F-5	F-10	CD	Vol	F-5	F-10	CD	Vol	F-5	F-10	CD	Vol
HO [73]	0.23	0.56	**0.64**	8.64	0.11	0.22	4.19	9.44	0.03	0.06	49.8	25.6
GF [96]	0.30	0.51	1.39	1.84	0.12	0.24	4.96	6.31	0.06	0.13	40.1	**8.82**
Ours	**0.42**	**0.63**	1.02	**1.74**	**0.28**	**0.50**	**1.53**	**4.77**	**0.13**	**0.24**	**23.1**	19.4

Table 4.1: Quantitative results for object reconstruction error using F-score $(5mm, 10mm)$, Chamfer distane (mm) and intersection volume (cm^3). We compare our method with prior works [73, 96] on Obman, HO3D, MOW datasets.

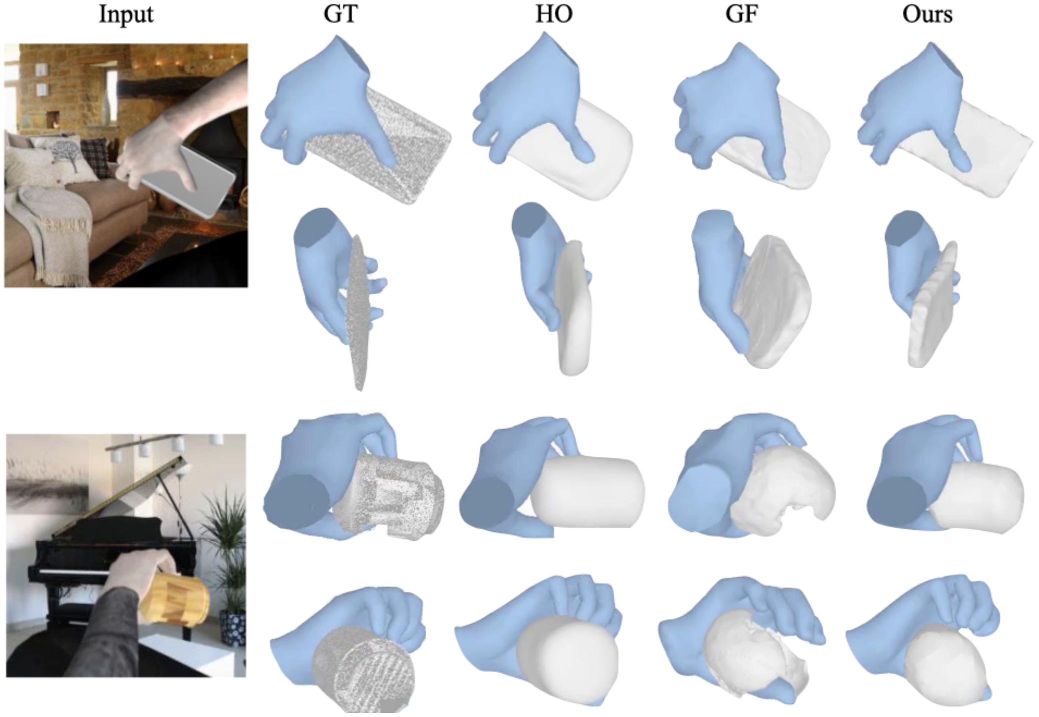


Figure 4.3: Visualizing reconstruction from our method and two baselines [73,96] on ObMan dataset from the image frame and a novel view.

4.2 Experiments

We compare our method with two model-free baselines [73, 96] on three datasets – one synthetic, and two real-world. We show that our approach outperforms baselines across these datasets, both in terms of object reconstruction and modeling hand-object interaction. We further analyze the benefit from explicitly considering hand pose and the benefit from our particular form of articulation-aware positional encoding. Lastly, we show that our reconstructed hand-held object could further refine the initial hand pose estimation and improve hand-object interaction.

Datasets and Setup. We evaluate our method across three datasets.

- ObMan [73] is a synthetic dataset that consists as 8 categories of 2772 objects from 3D warehouse [23]. The grasps are automatically generated by GraspIt [139], resulting in a total of 21K grasps. The grasped objects are rendered over random backgrounds using Blender. We follow the standard splits where there is no overlap between the objects used in training and testing.

- HO-3D [68] is a real-world video dataset consisting of 103k annotated images capturing 10 subjects interacting with 10 common YCB objects [247]. The ground- truth is annotated using multi-camera reconstruction pipelines. To test on more shapes, we create a custom split by holding out one video sequence per object as test set. Please refer to the appendix for more details.

- MOW [19] dataset consists of a curated set of 442 images, spanning 121 object templates, collected from in-the-wild hand-object interaction datasets [36, 187]. It is more diverse in terms of both appearance and object shape compared to the HO3D dataset, but only provides approximate ground-truth. These object shape and hand pose annotations are obtained via a single-frame optimization-based method [19]. We use 350 randomly selected examples for training and the remaining 92 for evaluation.

For the ObMan dataset, we use the hand pose predictor from Hasson *et al.* [73] as this system is specifically trained on this synthetic dataset. For HOI and MOW, we use the FrankMocap [180] system that is trained on multiple real-world datasets. Since HOI data with ground truth in the real world is scarce and lacks diversity in terms of object shape and appearance, we initialize our method and baselines with models pretrained on ObMan and finetune them on HO3D and MOW datasets.

Evaluation metrics. We evaluate the quality of both, object reconstruction and the relation between object and hand. To evaluate the reconstruction quality, we first extract a mesh from the predicted SDF. Following prior works, we then evaluate the object reconstruction by reporting Chamfer distance (CD), but also report the F-score at 5mm and 10mm thresholds as

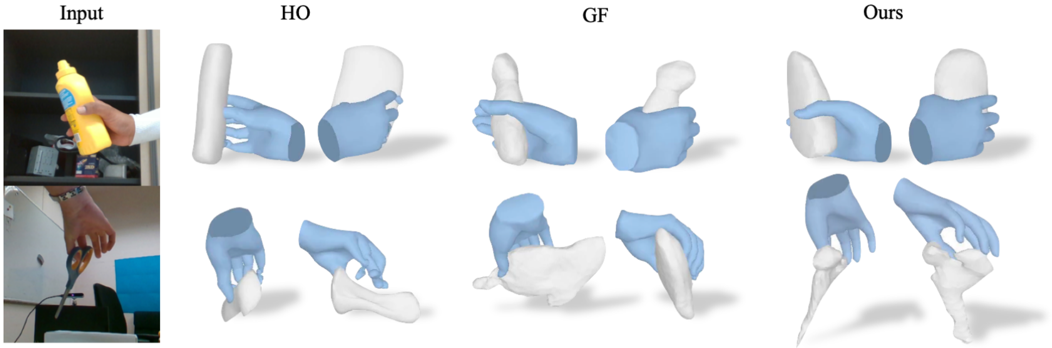


Figure 4.4: Visualizing reconstruction of our method and two baselines [73, 96] on HO3D dataset in the image frame and from a novel view.

Chamfer distance is more vulnerable to outliers [200]. Another desirable property for HOI reconstruction is that the interpenetration between the hand and the object should be minimal, we also report the intersection volume between two meshes (in cm^3) as a measure of understanding the relations between the hand and object.

Baselines. While most prior approaches require a known object template, recent work by Hasson *et al.* (HO) [73] and Karunratanakul *et al.* (GF) [96] can tackle the same task as ours – inferring the shape of a generic object from a single interaction image. HO jointly regresses MANO parameters to estimate hand pose and reconstructs the object in the camera frame. It is based on Atlas-Net [63], and deforms a sphere to infer the object mesh. Closer to our approach, GF is also based on a point-wise implicit network. It takes an image as input and outputs an implicit field that maps a point in the camera frame to a signed distance to both, the hand and object, while also predicting hand part labels.

Our approach differs from these baselines in three main aspects. First, both prior approaches infer object shape independent of the predicted hand pose, while we formulate hand-held object reconstruction as conditional inference. Second, these baselines encode the visual information only via a global feature while we additionally use pixel-aligned local features. Third, while both baselines reconstruct objects in the camera frame, we predict them in a normalized wrist frame with articulation-aware positional encoding. Note that while our approach predicts 3D in a hand-centric frame, the evaluations are all performed in the image coordinate frame for fair comparison (using the predicted hand pose to transform our prediction).

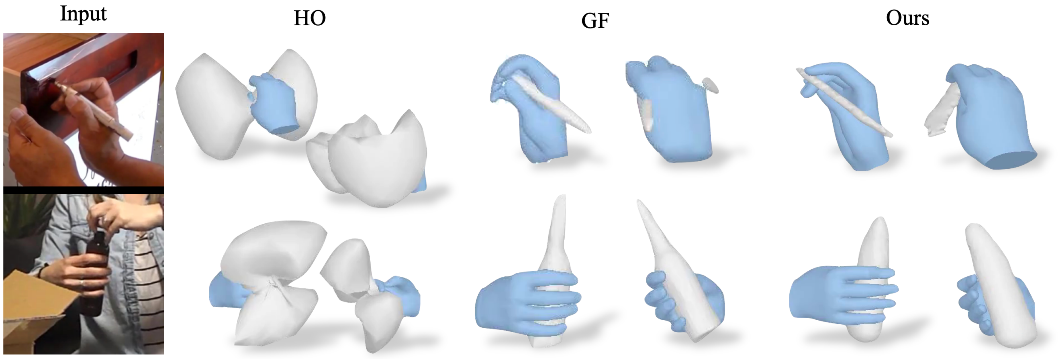


Figure 4.5: Visualizing reconstruction of our method and two baselines [73, 96] on MOW dataset in the image frame and from a novel view.

4.2.1 Results on ObMan

We visualize the reconstructed objects and the corresponding hand poses in Figure 4.3. While baselines can predict the coarse shape of the object, they typically lose sharp details such as the corner of the phone and sometimes miss part of the surface of the object. This may be because they only use a global feature of the image that loses spatial resolution. In contrast, our method reconstructs shape that better aligns with the visual inputs from the original view and hallucinate the invisible part of the objects occluded by hands.

This is also empirically reflected in quantitative results reported in Table 4.1. We outperform baselines by a large margin on F-score. Our improvement over baselines is particularly significant on the smaller threshold, indicating that our method is better to reconstruct local shape. In terms of Chamfer distance, ours is better than GF that is also based on implicit fields. Ours is not as good as HO in Chamfer distance probably because HO explicitly trains to minimize Chamfer distance with a regularizer on edge length which discourages large displacements from a sphere thus producing fewer outliers.

4.2.2 Evaluation on real-world datasets

We visualize the reconstruction in the image frame and a novel view in Figure 4.4 and Figure 4.5. GF can predict blobby cylinders but the reconstructed objects lack details in shape such as around the neck of the mustard bottle, and sometimes reconstructs a different object shape such as predicting boxes instead of scissors. In contrast, our method is able to generate diverse object shape more accurately including boxes, power drills, bottles, pens, cup, spray bottles etc.

train set	F-5 ↑	F-10 ↑	CD ↓
ObMan	0.14	0.27	4.36
MOW	**0.15**	**0.30**	**4.09**

Table 4.2: **Cross-dataset generalization:** we report quantitative results on HO3D for models pretrained on ObMan and MOW.

Zero-shot transfer to HO3D. We also directly evaluate models that are only trained on ObMan and MOW datasets and report their reconstruction results on HO3D dataset. Both models without finetuning still outperform baselines trained on HO3D dataset. Interestingly, even though the MOW dataset only consists of 350 training images, which is significantly less compared to 21K images from the synthetic dataset, learning from MOW still helps cross-dataset generalization. It indicates the importance of diversity for in-the-wild training. Please see the qualitative result in the appendix.

4.2.3 Ablations

Importance of articulation conditioning. We analyze how hand articulation conditioning helps hand-held object reconstruction by constructing a variant of our method that only conditions on pixel-aligned image features. This approach is analogous to the one proposed by Saito *et al.* [183] where human 3D shape is inferred by a pixel-aligned implicit network. Table 4.3 reports results of this variant that do not explicitly consider hand articulation and we observe that the object reconstruction degrades by a large margin while the intersection volume also doubles. This suggests that hand information provides a strong cue that is complementary to visual inputs. Figure 4.6 visualizes comparison between ours and the variant where our method can better respect hand-object physical relations such as objects do not penetrate the hands and the area around fingertips are likely to be in contact with objects.

Representation of hand articulation matters for generalization. To represent articulation information, a natural alternative to our proposed articulation-aware positional encoding is to simply concatenate the query point with the MANO pose parameter θ_A, *i.e.* $\bar{h}(\mathbf{x}; \theta_A) = [\mathbf{x}, \theta_A]$. The result in Table 4.4 shows that although it performs comparably when provided with ground-truth hand pose parameters, it degrades significantly when with predicted

		F-5 ↑	F-10 ↑	CD ↓	Vol ↓
ObMan	Ours	**0.42**	**0.63**	**1.02**	**1.74**
	Ours w/o Art.	0.37	0.56	1.89	3.93
HO3D	Ours	**0.33**	**0.58**	**0.93**	**4.77**
	Ours w/o Art.	0.27	0.48	1.18	6.30
MOW	Ours	**0.13**	**0.24**	**23.1**	19.4
	Ours w/o Art.	0.10	0.19	29.0	**17.3**

Table 4.3: **Analysis of articulation-conditioning:** we report quantitative results of object error in F-score, Chamfer distane (CD), intersection volume on 3 datasets and compare ours with the ablation that only consider visual feature.

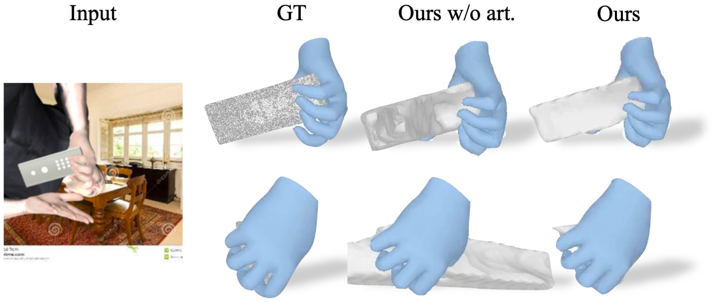


Figure 4.6: Visualizing reconstruction of hand-held object with or without explicitly considering hand pose.

hand pose despite that we perform jitter augmentation on hand articulation for both methods. More interestingly, it performs even worse than the variant without articulation-aware encoding. This indicates that the object shape overfits to pose parameters which are constant within one example. In contrast, our articulation-aware positional encoding generalizes better.

Robustness against hand prediction quality. We use hand poses corrupted by different levels of noise, either from Gaussian or more structured prediction noise. For the latter, we linearly interpolate (and even extrapolate) the true poses and off-the-shelf predictions. Our method still outperform baselines even when the predicted hand pose is *with twice more error* (Tab 4.5 and Fig 4.7).

Test-time refinement improves hand pose. The object reconstruction above is obtained by direct feed-forward prediction. We then show that our articulation-conditioned object shape can in turn refine the initial hand pose estimation

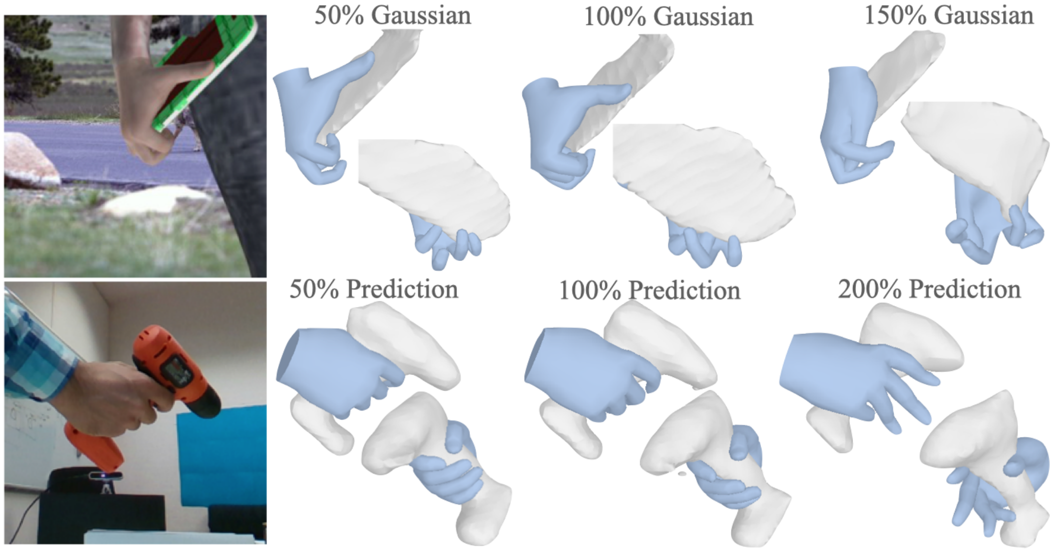


Figure 4.7: Top: Object reconstruction given hand pose corrupted by Gaussian on Obman dataset. Bottom: Object reconstruction given hand pose corrupted by prediction error on HO3D dataset.

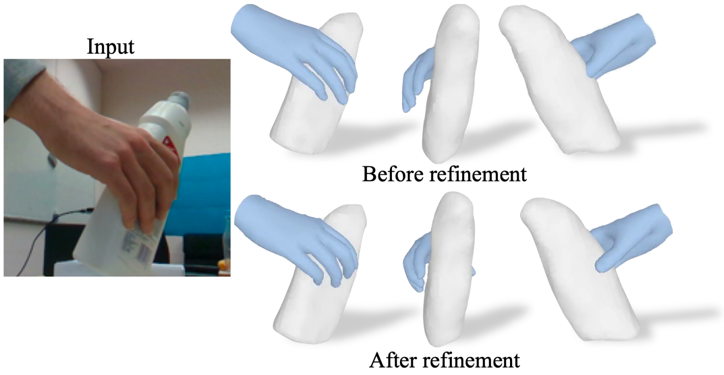


Figure 4.8: Visualizing hand-object reconstruction before and after test-time refinement in the image frame and from two novel views.

Method	F5 ↑	F10 ↑	CD ↓	Vol↓
Art.-aware PE*	**0.49**	**0.70**	**0.92**	1.73
Pose param.*	0.46	0.66	1.25	**1.44**
Art.-aware PE	**0.42**	**0.63**	**1.02**	**1.74**
Pose param.	0.23	0.42	1.82	2.57
W/o art.	0.37	0.56	1.89	3.93

Table 4.4: **Analysis of articulation-aware encoding:** We compare different ways to incorporate hand articulation: articulation-aware positional encoding and pose parameters. Star indicates reconstruct object shape given ground truth hand articulation.

	Noise	ObMan			HO3D		
	Level	F5 ↑	F10 ↑	CD ↓	F5 ↑	F10 ↑	CD ↓
Gaussian	50% σ	0.40	0.63	1.01	0.28	0.50	1.51
	100% σ	0.31	0.53	1.40	0.25	0.46	1.68
	150% σ	0.24	0.42	1.94	0.22	0.42	1.93
Prediction	50%	0.46	0.67	0.96	0.29	0.52	1.48
	100% *	0.42	0.63	1.02	0.28	0.50	1.53
	200%	0.35	0.56	1.28	0.26	0.47	1.67
Baselines	HO	0.23	0.56	0.64	0.11	0.22	4.19
	GF	0.30	0.51	1.39	0.12	0.24	4.96
GT	0%	0.49	0.70	0.92	0.30	0.53	1.46

Table 4.5: Error analysis against hand pose noise. σ is the average prediction error. * marks our unablated method.

and improve the HOI quality. We report end point error (EPE in mm) for each joint on Obman dataset in Table 4.6. To evaluate HOI quality, we report intersection volume along with simulation displacement of the object. We follow prior works [73,96,205] to pass the HOI reconstruction to a simulator and report how much the object slips from hand after running simulation for a fixed amount of time.

As shown in Table 4.6, both object and hand reconstruction improve after test-time refinement. The simulation displacement of the object drops with less intersection region. When ground truth hand articulation is provided, the object error and simulation displacement continue to improve. Figure 4.8 visualizes one example before and after refinement. Four finger tips are attracted to object surface while the thumb is pushed out of the object.

	F-5↑	CD↓	Vol↓	Sim↓	EPE↓
ours w/o rf	0.17	1.02	1.74	3.32	8.9
ours w rf	0.17	**1.00**	**1.28**	**3.00**	**8.7**
ours w GT pose	0.20	0.92	1.73	2.44	–

Table 4.6: **Test-time refinement.** We report object error, intersection volume, simulation displacement and hand error before and after test-time refinement.

4.3 Discussion

In this chapter, we propose a method to infer implicit 3D shape of generic objects in hand. We explicitly treat predicted hand pose as a cue for object inference via an articulation-aware positional encoding. We have shown that this complements visual cues, especially when the hand-held object is occluded. While the results are encouraging, there are several limitations. For example, our work cannot be directly adapted to reconstructing dynamic grasps from videos where object consistency given varying articulation is required. Additionally, we require 3D ground-truth for training and it would be interesting to extend it with differentiable rendering techniques. Despite these challenges, we believe that our work on reconstructing hand-held generic objects takes an encouraging step towards understanding HOI for in-the-wild videos.

4.4 Implementation Details and Additional Results

4.4.1 Camera conversion

The off-the-shelf system predicts a weak perspective camera with a scale factor s and 2D translation t_x, t_y. One can transform the point via the global hand rotation and translation and then project it via the predicted camera s, t_x, t_y.

$$sT_{\theta_w}X + (t_x, t_y)$$

We found that a full perspective camera help to account for large perspective effect. Therefore, we convert the weak perspective camera to a full perspective one by translating the final mesh by an offset $(t_x, t_y, f/s)$. In summary, we project a query point in the wrist frame to the image by

$$\pi_{\theta_w}(X) = K[T_{\theta_w}X + (t_x, t_y, f/s)]$$

4.4.2 Coordinate Transformation

Our articulation embedder takes as input an articulation parameter θ_A and a point position in wrist frame X to output the articulation-aware encoding $\psi = h(X; \theta_A)$. The encoding is a concatenation of the coordinates relative to every joint. Given the articulation parameter θ_A, we run forward kinematics to derive transformation $T(\theta_A) : \mathbb{R}^3 \to \mathbb{R}^{45}$ that maps a point in wrist frame to each joint coordinate.

The transformation between wrist to one joint T_j is computed by forward kinematics chain. Consider one bone that connects joint j to its child i (*e.g.* index proximal phalanx). The transformation matrix from this joint frame to its child joint frame would be

$$T_{ji} = \begin{pmatrix} R(\theta_j) & t_{ji} \\ 0 & 1 \end{pmatrix}$$

where t_j is the bone length pre-defined in MANO models. Then the transformation from wrist to any joint is the product of every transformation in the kinematic chain $T_j = T_{wi} \cdot T_{ik} \cdot \ldots T_{lj}$. The coordinate of the queried point relative to the joint becomes $^jX =^j T^w_w X$.

4.4.3 Training

We train our model using Adam optimizer with learning rate $1e - 4$ on 8 GPUs. The batch size is 64. We train our model on ObMan for 200 epochs and finetune it on HO3D and RHOI for 50k iterations respectively. The coefficient of eikonal term is 0.1.

4.4.4 HO3D dataset split

HO-3D [68] is a real-world video dataset consisting of 103k annotated images capturing 10 subjects interacting with 10 common YCB objects [247]. The original train-test splits are created by partitioning the interaction sequences. Sequences in the original test set involve only 4 objects of which three appear in train set (bleach cleanser, mustard bottle, meat can) and all of them are cuboidal shape. To test on more non-trivial shapes like power drill, we create a custom split by holding out one video sequence per object as test set. We list our sequences for train and test set in Table 4.7.

Objects	Test Sequences	Train Sequence
010_potted_meat_can	GPMF10	MPM14, GPMF13, MPM12, GPMF12, MPM11, GPMF11, MPM13, MPM10, GPMF14
021_bleach_cleanser	ABF10	SB11, SB12, ABF11, ABF13, SB10, ABF12, ABF14, SB13, SB14
019_pitcher_base	AP10	AP11, AP14, AP13, AP12
003_cracker_box	MC1	MC2, MC6, MC5, MC4
006_mustard_bottle	SM1	SM5, SM2, SM4, SM3
004_sugar_box	SS1	ShSu12, SiS1, SS2, ShSu14, ShSu13, SS3, ShSu10
035_power_drill	MDF10	MDF12, MDF14, MDF11, ND2, MDF13
011_banana	BB10	BB12, SiBF10, SiBF14, SiBF11, SiBF12, BB13, BB11, SiBF13, BB14
037_scissors	GSF10	GSF13, GSF12, GSF14, GSF11
025_mug	SMu1	SMu41, SMu42, SMu40

Table 4.7: Our customized split on HO3D dataset.

4.4.5 Qualitative Results

We provide more qualitative results rendered in the image frame and from another view in this PDF and video results when moving camera around the object in the zipped website.

Figure 4.9 visualizes reconstruction from our method and two baselines [73, 96] on ObMan dataset from the image frame and a novel view.

Figure 4.10 visualizes reconstruction from our method and two baselines [73, 96] on HO3D dataset from the image frame and a novel view.

Figure 4.11, 4.12 visualizes reconstruction from our method and two baselines [73, 96] on RHOI dataset from the image frame and a novel view.

Figure 4.13 visualizes reconstruction of hand-held object with or without explicitly considering hand pose on ObMan, HO3D and RHOI.

Figure 4.14 visualizes reconstruction of hand-held object from our models that only trained on ObMan and RHOI datasets.

Figure 4.15 visualizes hand-object reconstruction before and after test-time refinement in the image frame and from two novel views.

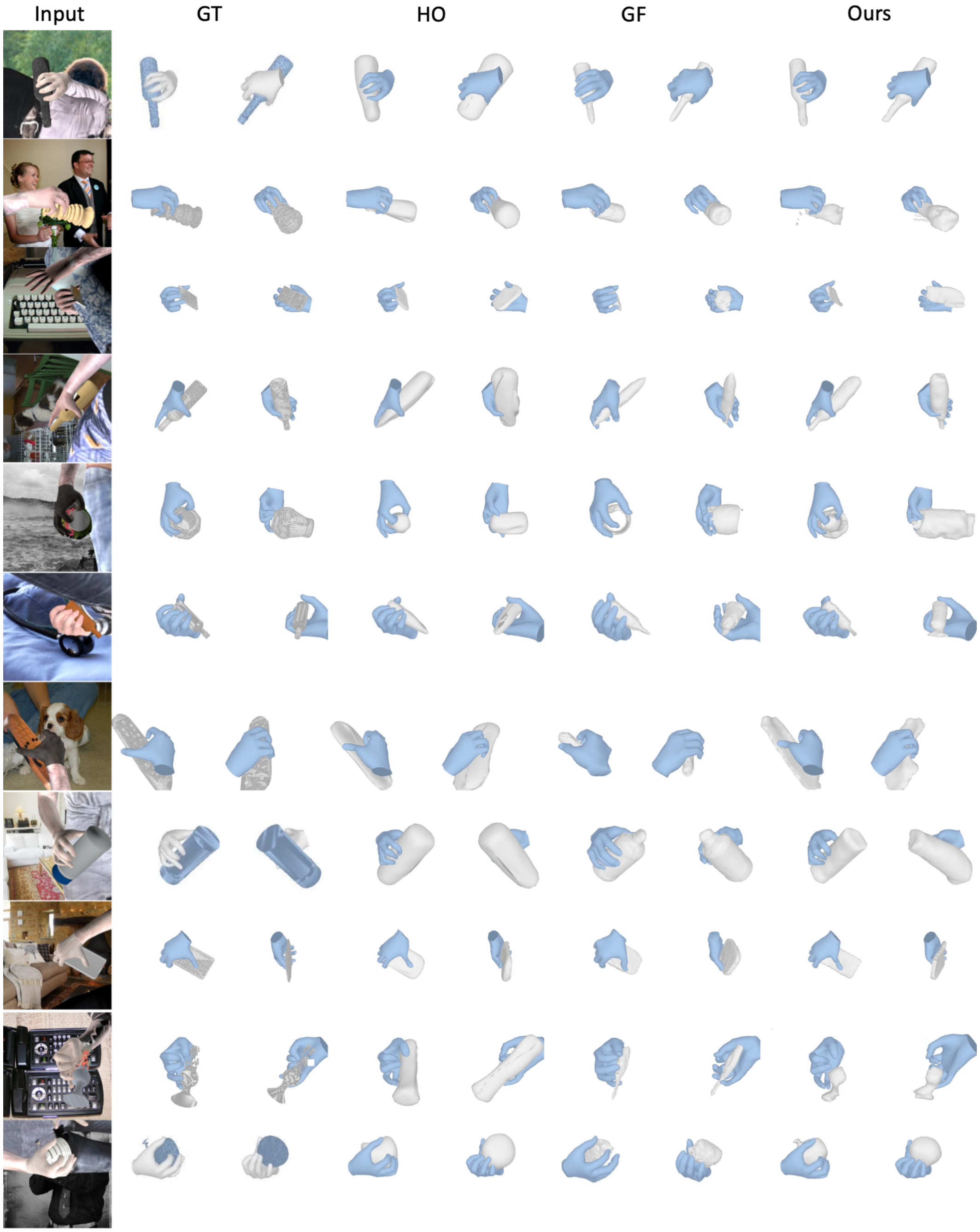


Figure 4.9: Visualizing reconstruction from our method and two baselines [73,96] on ObMan dataset from the image frame and a novel view.

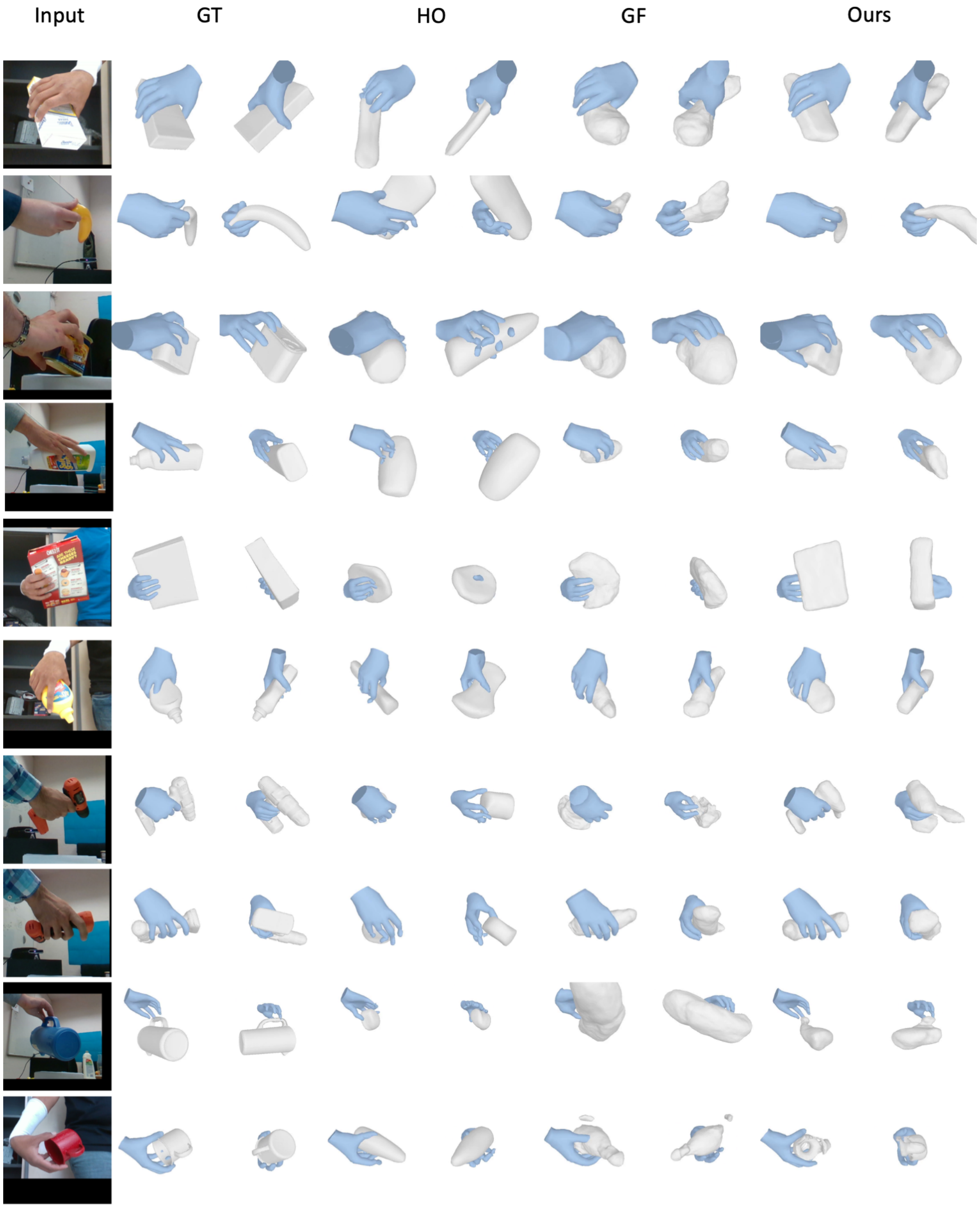


Figure 4.10: Visualizing reconstruction from our method and two baselines [73,96] on HO3D dataset from the image frame and a novel view.

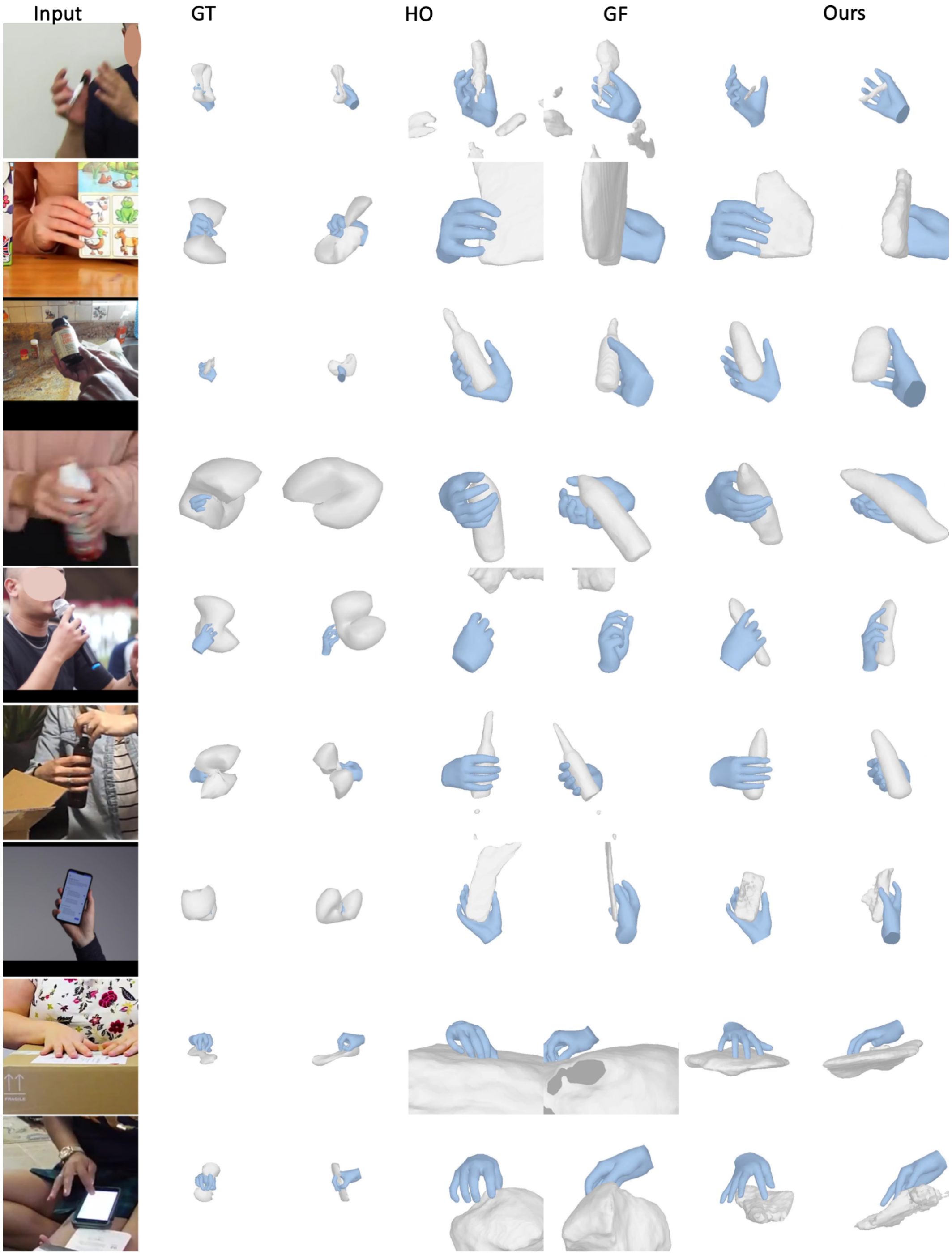


Figure 4.11: Visualizing reconstruction from our method and two baselines [73,96] on RHOI dataset from the image frame and a novel view.

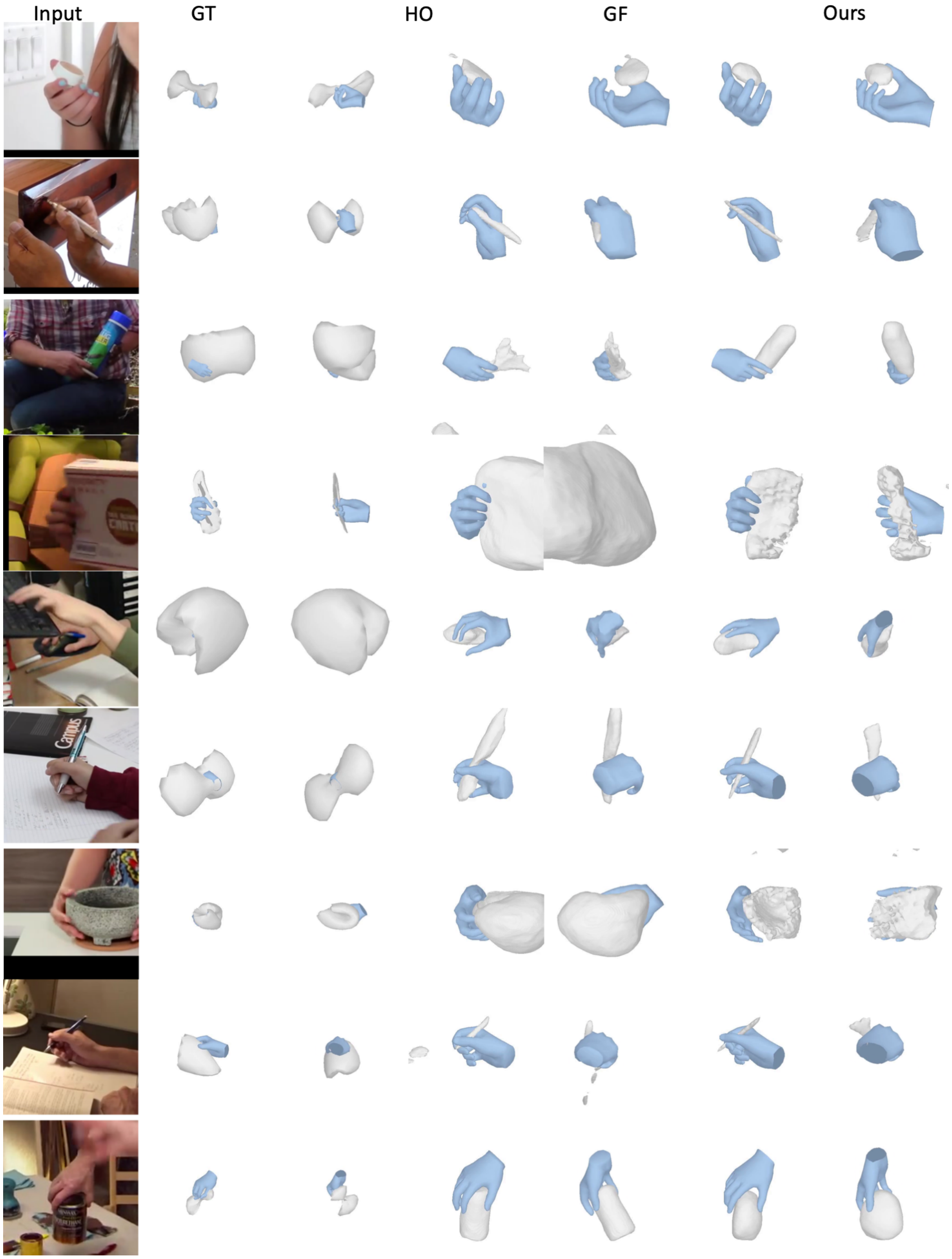


Figure 4.12: Visualizing reconstruction from our method and two baselines [73,96] on RHOI dataset from the image frame and a novel view.

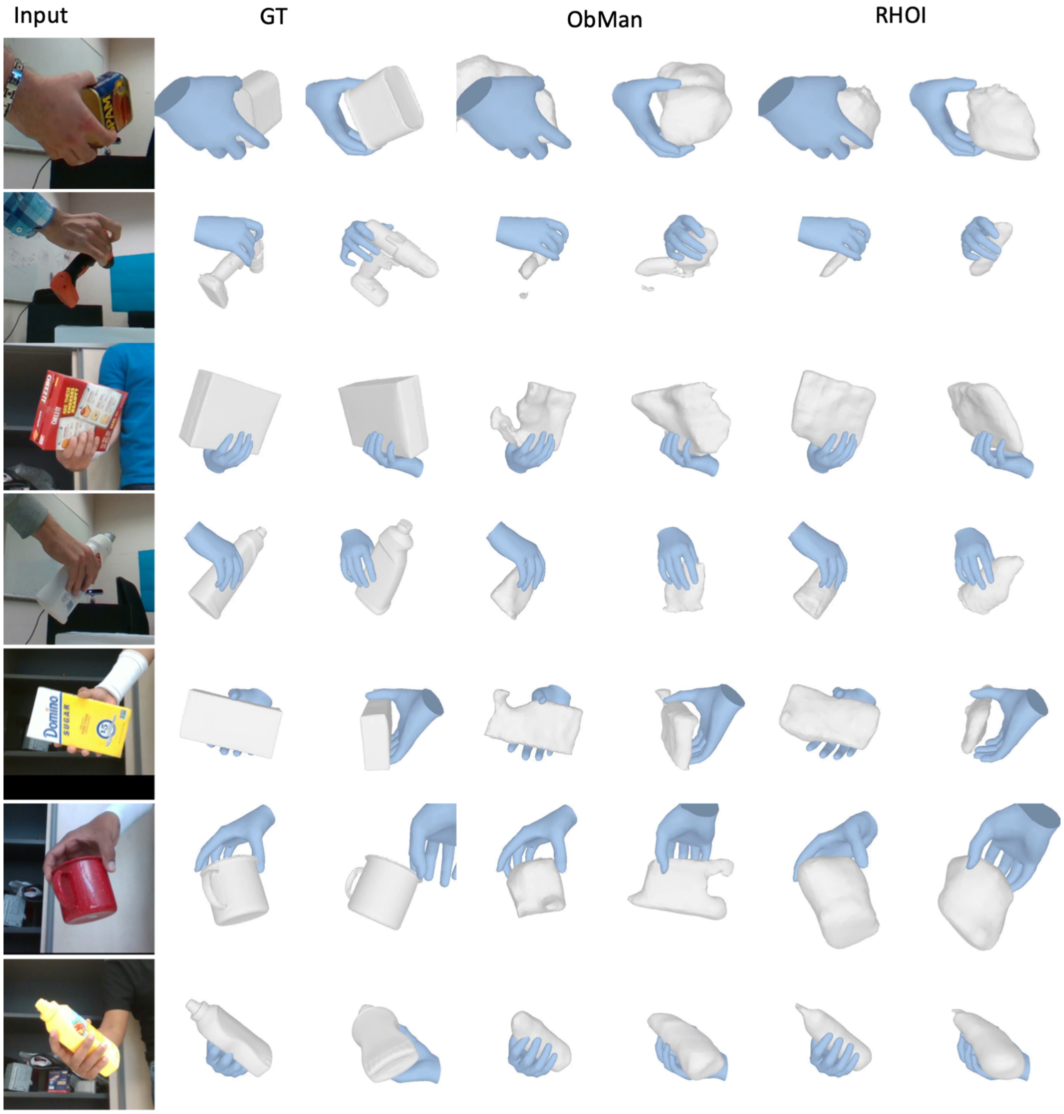


Figure 4.13: **Cross-dataset generalization:** we show quantitative results on HO3D for models pretrained on ObMan and RHOI.

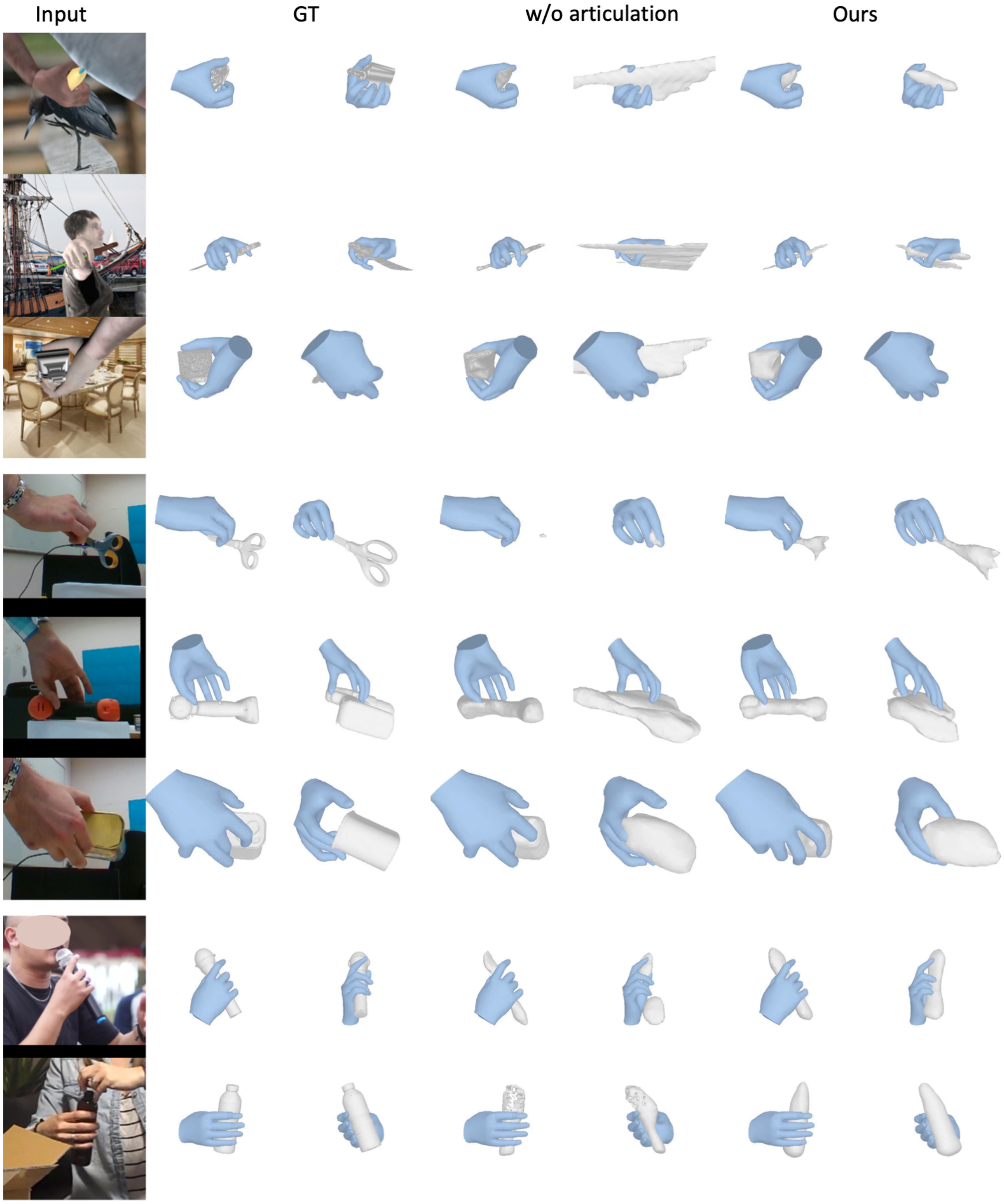


Figure 4.14: Visualizing reconstruction of hand-held object with or without explicitly considering hand pose on ObMan, HO3D and RHOI.

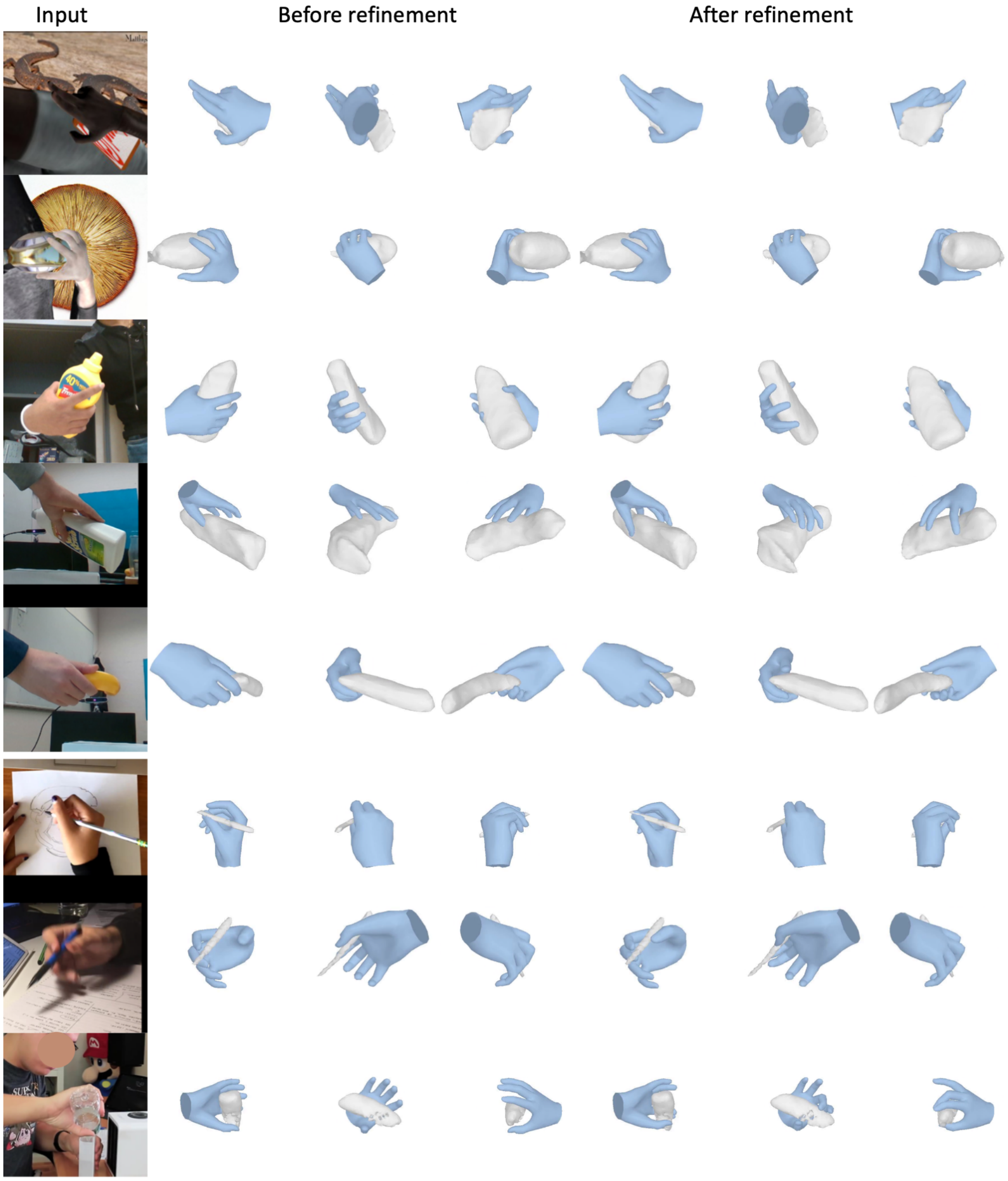


Figure 4.15: Visualizing hand-object reconstruction before and after test-time refinement in the image frame and from two novel views.

Chapter 5

Reconstructing Generic Objects in Hands from Monocular Videos

In previous chapter, we present a method to reconstruct Hand-Object Interactions (HOI) from single images by learning a data-driven prior between hand pose and object geometry. Although single-view 3D reconstruction approaches [73, 96, 232] can leverage data-driven techniques to hallucinate unobserved part to reconstruct HOI images, these approaches cannot obtain precise reconstructions given the fundamentally limited nature of the single-view input.

On the other hand, prior video-based HOI reconstruction methods primarily exploit multi-view cues and rely on purely geometry-driven optimization for reconstruction [85, 218]. As a result, these methods are suited for in-hand scanning where a user carefully presents exhaustive views of the object of interest, but they are not applicable to our setting as aspects of the object may typically be unobserved.

In this chapter, we combine the benefits of both worlds to explore understanding everyday HOI in 3D from casual video clips. Specifically, given a short clip of a human interacting with a rigid object, our approach can infer the shape of the underlying object as well as its (time-varying) relative transformation w.r.t. an articulated hand (see Fig. 5.1 for sample results).

Towards enabling accurate reconstruction given short everyday interaction clips, our approach unifies the data-driven and the geometry-driven techniques. Akin to the prior video-based reconstruction methods, we frame the reconstruction task as that of optimizing a video-specific temporal scene representation. However, instead of purely relying on geometric reprojection errors, we also incorporate data-driven priors to guide the optimization. In particular, we learn a 2D diffusion network which models the dis-

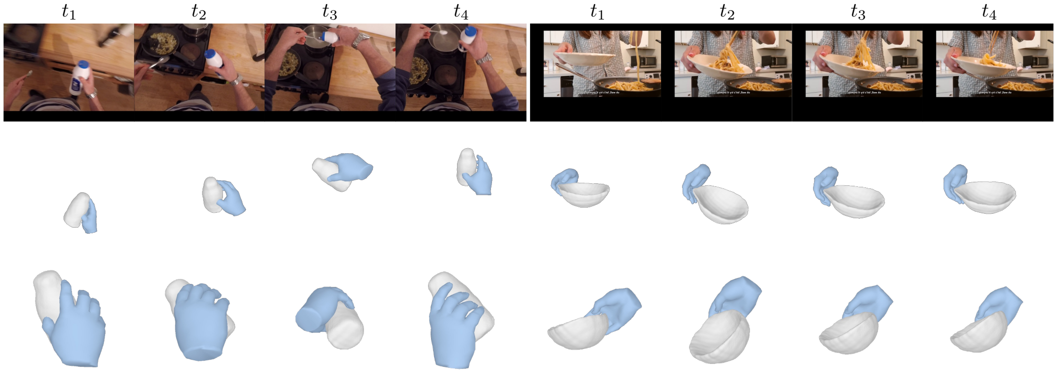


Figure 5.1: Given a video clip depicting a hand-object interaction, we infer the underlying 3D shape of both the hand and the object. **Top:** sampled input frames; **Middle:** reconstruction in the image frame; **Bottom:** reconstruction from a novel view.

tribution over plausible (geometric) object renderings conditioned on estimated hand configurations. Inspired by recent applications in text-based 3D generation [118, 161], we use this diffusion model as a generic data-driven regularizer for the video-specific 3D optimization.

We empirically evaluate our system across several first-person hand-object interaction clips from the HOI4D dataset [127], and show that it significantly improves over both prior single-view and multi-view methods. To demonstrate its applicability in more general settings, we also show qualitative results on arbitrary interaction clips from YouTube, including both first-person and third-person clips.

5.1 Preliminary: Diffusion Models

Diffusion models [82] are a family of generative models. An advantage of diffusion models is that they allow computing log-likelihood gradients via score distillation [161, 213] and thus can be used as foundation generative priors in multiple domains like image generation [169,175], 3D object generation [92,118,161], novel-view synthesis [121,134], human motion [95,202], video generation [190], *etc.*.

Specifically, view-conditioned diffusion models like DreamFusion [161], and Magic3D [118] have demonstrated the potential of diffusion models in optimizing 3D scenes using conditioned text prompts. On the other hand, approaches like NeRDi [40] and RealFusion [134] focus on 3D reconstruction from images.

Meanwhile, recent image-conditioned generative models achieve impressive results on various image translation tasks such as image editing [6, 88, 135], style transfer [113, 157]. But without further design, the edits of end-to-end methods mostly modify textures and style, but preserve structures, or insert new content to user-specified regions [3, 169, 182].

In the following chapters, we apply diffusion models to hand-object interactions. In Chapter 5, we leverage geometry-based information to reconstruct 3D models, which found to be beneficial in terms of generalizing to novel scenes under distinct RGB appearances. In Chapter 6, we focus on affordance synthesis where both layout (structure) and appearance are automatically reasoned about. The repurposed diffusion model show significant generalization due to its large-scale pretraining data. In Chapter 7, we use diffusion model to learn a unified prior for 3D hand-object interactions and apply it to both tasks of HOI reconstruction and grasp synthesis.

To be self-contained, we briefly explain the fundamentals of diffusion models in the remaining section. Diffusion models learn to generate samples from a data distribution $p(\mathbf{x})$ by sequentially transforming samples from a tractable distribution $p(\mathbf{x}_T)$ (*e.g.*, Gaussian distribution). There are two processes in diffusion models: 1) a forward noise process $q(\mathbf{x}_t|\mathbf{x}_{t-1})$ that gradually adds a small amount of noise and degrades clean data samples towards the prior Gaussian distribution; 2) a learnable backward denoising process $p(\mathbf{x}_{t-1}|\mathbf{x}_t)$ that is trained to remove the added noise. The backward process is implemented as a neural network. During inference, a noise vector $\mathbf{x}_T$ is sampled from the Gaussian prior and is sequentially denoised by the learned backward model [194, 195].

The training of a diffusion model can be treated as training a denoising autoencoder for L2 loss [210] at various noise levels, *i.e.*, denoise $\mathbf{x}_0$ for different $\mathbf{x}_t$ given t. We adopt the widely used loss term in Denoising Diffusion Probabilistic Models (DDPM) [82, 194], which reconstructs the added noise that corrupted the input samples. Specifically, we use the notation $\mathcal{L}_{\text{DDPM}}[\mathbf{x}; \mathbf{c}]$ to denote a DDPM loss term that performs diffusion over $\mathbf{x}$ but is also conditioned on $\mathbf{c}$ (that are not diffused or denoised):

$$\mathcal{L}_{\text{DDPM}}[\mathbf{x}; \mathbf{c}] = \mathbb{E}_{(\mathbf{x},\mathbf{c}),\epsilon\sim\mathcal{N}(0,I),t}\|\mathbf{x} - D_\theta(\mathbf{x}_t, t, \mathbf{c})\|_2^2, \tag{5.1}$$

where $\mathbf{x}_t$ is a linear combination of the data $\mathbf{x}$ and noise ϵ, and D_θ is a denoiser model that takes in the noisy data $\mathbf{x}_t$, time t and condition $\mathbf{c}$. This also covers the unconditional case as we can simply set $\mathbf{c}$ as some null token like $\varnothing$ [83].

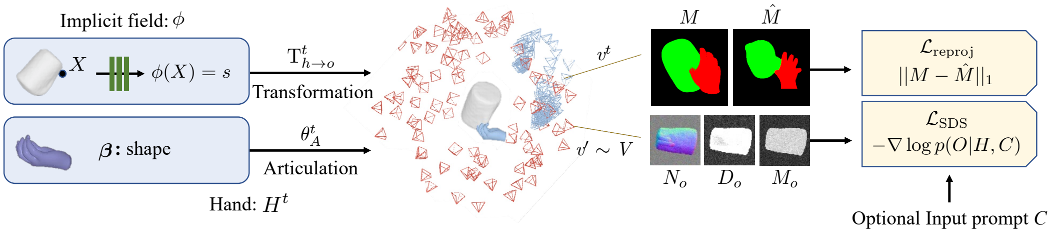


Figure 5.2: **Method Overview :** We model the HOI scene (middle) by a time-persistent implicit field ϕ for the object, hand meshes H^t parameterized by hand shape β, hand articulation θ^t_A, along a time-varying rigid transformation $T^t_{h\to o}$ for object pose. We register the cameras in the hand frame. We optimize a video-specific scene representation using reprojection loss from the original view and diffusion distillation loss from a novel view v'.

5.2 Method

Given a monocular video of a hand interacting with a rigid object, we aim to reconstruct the underlying hand-object interaction, *i.e.*, the 3D shape of the object, its pose in every frame, along with per-frame hand meshes and camera poses. We frame the inference as per-video optimization of an underlying 3D representations. While the multiple frames allow leveraging multi-view cues, they are not sufficient as the object of interests is often partially visible in everyday video clips, due to limited viewpoints and mutual occlusion. Our key insight is to incorporate both view consistency across multiple frames and a data-driven prior of the HOIs geometry. The learned interaction prior captures both category priors, *e.g.* mugs are generally cylindrical, and hand priors, *e.g.* pinched fingers are likely to hold thin handles. We train a conditional diffusion model for the prior that guides the HOI to be reconstructed during per-video optimization.

More specifically, given a monocular video $\hat{I}^t$ with corresponding hand and object masks $\hat{M}^t \equiv (\hat{M}^t_h, \hat{M}^t_o)$, we aim to optimize a HOI representation (Sec. 5.2.1) that consists of a time-persistent implicit field $\boldsymbol{\phi}$ for the rigid object, a time-varying morphable mesh for the hand H^t, the relative transformation between hand and object $T^t_{h\to o}$, and time-varying camera poses $T^t_{c\to h}$. The optimization objective consists of two terms (Sec. 5.2.3): a reprojection error from the estimated original viewpoint and data-driven prior term that encourages the object geometry to appear more plausible given category and hand information when looking from another viewpoint. The prior is implemented as a diffusion model conditioned on a text prompt C about the category and renderings of the hand $\pi(H)$ with geometry cues (Sec. 5.2.2). It denoises the rendering of the object $\pi(O)$ and backpropagates the gradient to the 3D HOI representation by score distillation sampling (SDS) [161].

5.2.1 HOI Scene Representation

Implicit field for the object. The rigid object is represented by a time-persistent implicit field $\boldsymbol{\phi}$ that can handle unknown topology and has shown promising results when optimizing for challenging shapes [215, 228, 230]. For every point in the object frame, we use multi-layer perceptrons to predict the signed distance function (SDF) to the object surface, $s = \boldsymbol{\phi}(\boldsymbol{X})$.

Time-varying hand meshes. We use a pre-defined parametric mesh model MANO [178] to represent hands across frames. The mesh can be animated by low-dimensional parameters and thus can better capture more structured motions, *i.e.* hand articulation. We obtain hand meshes H^t in a canonical hand wrist frame by rigging MANO with a 45-dim pose parameters $\boldsymbol{\theta}^t_A$ and 10-dim shape parameters $\boldsymbol{\beta}$, *i.e.* $H^t = \text{MANO}(\boldsymbol{\theta}^t_A, \boldsymbol{\beta})$. The canonical wrist frame is invariant to wrist orientation and only captures finger articulations.

Composing to a scene. Given the time-persistent object representation $\boldsymbol{\phi}$ and a time-varying hand mesh H^t, we then compose them into a scene at time t such that they can be reprojected back to the image space from the cameras. Prior works [68, 71, 158] typically track 6D object pose directly in the camera frame $T_{c\to o}$ which requires an object template to define the object pose. In our case, since we do not have access to object templates, the object pose in the camera frame is hard to estimate directly. Instead, we track object pose with respect to hand wrist $T^t_{h\to o}$ and initialize them to identity. It is based on the observation that the object of interest usually moves together with the hand and undergoes "common fate" [189]. A point in the rigid object frame can be related to the predicted camera frame by composing the two transformations, camera-to-hand $T^t_{c\to h}$ and hand-to-object $T^t_{h\to o}$. For notation convention, we denote the implicit field transformed to the hand frame at time t as $\boldsymbol{\phi}^t(\cdot) \equiv \boldsymbol{\phi}(T_{h\to o}(\cdot))$. Besides modeling camera extrinsics, we also optimize for per-frame camera intrinsics $\boldsymbol{K}^t$ to account for zoom-in effect, cropping operation, and inaccurate intrinsic estimation.

In summary, given a monocular video with corresponding masks, the parameters to be optimized are

$$\boldsymbol{\phi}, \boldsymbol{\beta}, \boldsymbol{\theta}^t_A, T^t_{h\to o}, T^t_{c\to h}, \boldsymbol{K}^t \tag{5.2}$$

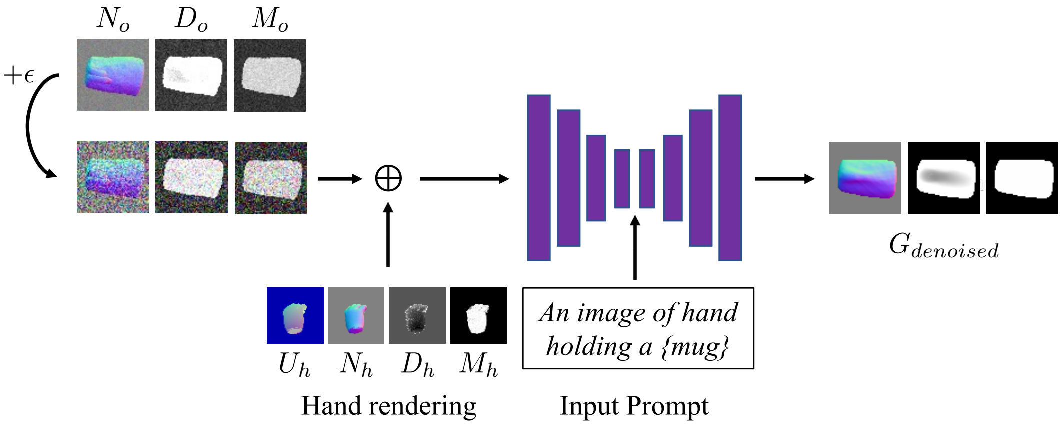


Figure 5.3: **Geometry-informed Diffusion Method:** Our diffusion model takes in a noisy geometry rendering of the object, the geometry rendering of the hand, and a text prompt, to output the denoised geometry rendering of objects.

Differentialble Rendering. To render the HOI scene into an image, we separately render the object (using volumetric rendering [230]) and the hand (using mesh rendering [125, 156]) to obtain geometry cues. We then blend their renderings into HOI images by their rendered depth.

Given an arbitrary viewpoint v, both differentiable renders can render geometry images including mask, depth, and normal images, *i.e.* $G_h \equiv (M_h, D_h, N_h), G_o \equiv (M_o, D_o, N_o)$ To compose them into a semantic mask M_{HOI} that is later used to calculate the reprojection loss, we softly blend the individual masks by their predicted depth. Similar to blending two-layer surfaces of mesh rendering, the final semantic masks can be computed by alpha blending: $M = B(M_h, M_o, D_h, D_o)$. Please refer to supplementary material for the full derivation of the blending function B.

5.2.2 Data-Driven Prior for Geometry

When observing everyday interactions, we do not directly observe all aspects of the object because of occlusions and limited viewpoint variability. Despite this, we aim to reconstruct the 3D shape of the full object. To do so, we rely on a data-driven prior that captures the likelihood of a common object geometry given its category and the hand interacting with it $p(\boldsymbol{\phi}^t|H^t, C)$. More specifically, we use a diffusion model which learns a data-driven distribution over geometry rendering of objects given that of hands and category.

$$\log p(\boldsymbol{\phi}^t | H^t, C) \approx \mathbb{E}_{v \sim V} \log p(\pi(\boldsymbol{\phi}^t; v) | \pi(H^t; v), C) \tag{5.3}$$

where $v \sim V$ is a viewpoint drawn from a prior distribution, C as category label and π as rendering function. Since this learned prior only operates in geometry domain, there is no domain gap to transfer the prior across daily videos with complicated appearances. We first pretrain this diffusion model with large-scale ground truth HOIs and then use the learned prior to guide per-sequence optimization (Sec. 5.2.3).

Learning prior over a-modal HOI geometry. As explained in 5.1, diffusion models are a class of probabilistic generative models that gradually transform a noise from a tractable distribution (Gaussian) to a complex (e.g. real image) data distribution. They are supervised to capture the likelihood by de-noising corrupted images (Equation 5.1).

In our case, as shown in Fig. 5.3, the diffusion model denoises the a-modal geometry rendering of an object given text prompt and hand. Additionally, the diffusion model is also conditioned on the rendering of uv-coordinate of MANO hand U_h because it can better disambiguate if the hand palm faces front or back. More specifically, the training objective is $\mathcal{L}_{\text{diff}} = \mathcal{L}_{\text{DDPM}}[G_o; C, G_h, U_h]$. The text prompt comes from a text template: "an image of a hand holding {*category*}".

Implementation Details. When we train the diffusion model with the rendering of ground truth HOI, we draw viewpoints with rotation from the uniform distribution in SO(3) . We use the backbone of a text-to-image model [149] with cross attention and modify it to diffuse 5-channel geometry images (3 for normal, 1 for mask and 1 for depth). We initialize the weights from the image-conditioned diffusion model [149] pretrained with large-scale text-image pairs. The additional channels in the first layer are loaded from the average of the pretrained weights.

5.2.3 Reconstructing Interaction Clips in 3D

After learning the above interactions prior, at inference time when given a short monocular clip with semantic masks of hand and object, we optimize a per-sequence HOI representation to recover the underlying hand-object interactions. We do so by differentiable rendering of the 3D scene representation from the original views and from random novel views. The optimization objectives consist of the following terms.

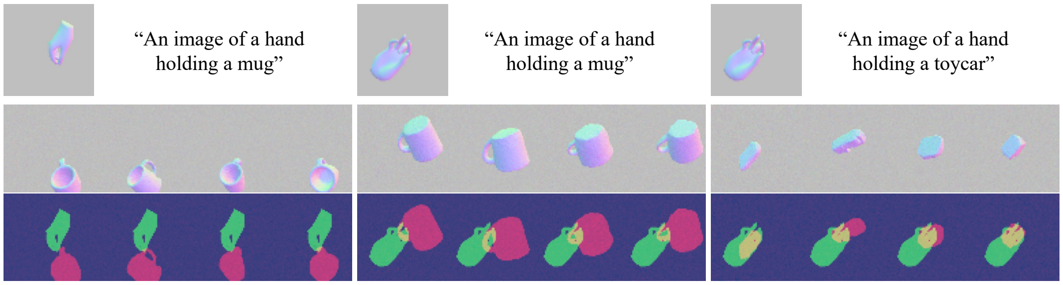

“An image of a hand holding a mug”
“An image of a hand holding a mug”
“An image of a hand holding a toycar”

Figure 5.4: **Generations from conditional diffusion model:** Given the geometry rendering of hand G_h (only showing surface normals) and a text prompt C, we visualize 4 different generations from the diffusion model. Middle row shows the generated surface normal of the objects and bottom row visualizes the generated object masks overlayed on the given hand masks. Note the left and middle column share the same text condition while middle and right column share the same hand condition.

Reprojection error. First, the HOI representation is optimized to explain the input video. We render the semantic mask of the scene from the estimated cameras for each frame and compare the rendering of the semantic masks (considering hand-object occlusion) with the ground truth masks: $\mathcal{L}_{\text{reproj}} = \sum_t \|M^t - \hat{M}^t\|_1$

Learned prior guidance. In the meantime, the scene is guided by the learned interactino prior to appear more likely from a novel viewpoint following Scored Distillation Sampling (SDS) [161]. SDS treats the output of a diffusion model as a critic to approximate the gradient step towards more likely images without back-propagating through the diffusion model for compute efficiency:

$$\mathcal{L}_{SDS} = \mathbb{E}_{v,\epsilon,i}[w_i\|\pi(\phi^t) - \hat{G}_o^i\|_2^2] \tag{5.4}$$

where $\hat{G}_o^i$ is the reconstructed signal from the pre-trained diffusion model. Please refer to relevant works [134, 161] or supplmentary for full details.

Other regularization. We also include two regularization terms: one Eikonal loss [62] that encourages the implicit field ϕ to be a valid distance function $\mathcal{L}_{\text{eik}} = \|\nabla_X \phi^2 - 1\|^2$, and another temporal loss that encourages the hand to move smoothly with respect to the object $\mathcal{L}_{\text{smooth}} = \sum_t \|T^t_{h\to o}H^t - T^{t-1}_{h\to o}H^{t-1}\|_2^2$

Initialization and training details. While the camera and object poses are learned jointly with object shape, it is crucial to initialize them to a coarse po-

Table 5.1: **Comparison with baselines:** Quantitative results for object reconstruction error using F1@5mm and F1@10mm scores and Chamfer Distance (mm). We compare our method with prior works HHOR [85] and iHOI [232] on the HOI4D dataset.

	Mug			Bottle			Kettle			Bowl			Knife			ToyCar			Mean		
	$F@5$	$F@10$	CD	$F@5$	$F@10$	CD	$F@5$	$F@10$	CD	$F@5$	$F@10$	CD	$F@5$	$F@10$	CD	$F@5$	$F@10$	CD	$F@5$	$F@10$	CD
HHOR [85]	0.18	0.37	6.9	0.26	0.56	3.1	0.12	0.30	11.3	0.31	0.54	4.2	**0.71**	0.93	0.6	0.26	0.59	1.9	0.31	0.55	4.68
iHOI [232]	0.44	0.71	2.1	0.47	0.77	1.5	0.21	0.45	6.3	0.38	0.64	3.1	0.33	0.68	2.8	0.66	0.95	0.5	0.42	0.70	2.73
Ours	**0.67**	**0.86**	**1.0**	**0.62**	**0.92**	**0.7**	**0.47**	**0.73**	**1.6**	**0.68**	**0.93**	**0.6**	0.66	**0.96**	**0.6**	**0.81**	**0.98**	**0.3**	**0.65**	**0.90**	**0.79**

sition [119]. We use FrankMocap [179], an off-the-shelf hand reconstruction system, to initialize the hand parameters, camera-to-hand transformations, and camera intrinsic. More specifically, FrankMocap predicts finger articulation $\boldsymbol{\theta}_A^t$, wrist orientation $\boldsymbol{\theta}_w^t$, and a weak perspective camera. The last two are used to compute camera-to-hand transformation and intrinsics of a full perspective camera. See appendix for derivation. We initialize the object implicit field to a coarse sphere [230] and the object poses $T_{h\to o}^t$ to identity such that the initial object is roughly round hand palm.

The per-frame hand pose estimation sometimes fails miserably in some challenging frames due to occlusion and motion blur. We run a lightweight trajectory optimization on wrist orientation to correct the catastrophic failure. The optimization objective encourages smooth joint motion across frames while penalizing the difference to the per-frame prediction, *i.e.* $\mathcal{L} = \|H(\boldsymbol{x}^t) - H(\hat{\boldsymbol{x}}^t)\| + \lambda\|H(\boldsymbol{x}^{t+1}) - H(\boldsymbol{x}^t)\|$ where λ is 0.01. Please see appendix for full details.

5.3 Experiments

We first train the diffusion model on the egocentric HOI4D [127] dataset and visualize its generations in Section 5.3.1. Then, we evaluate the reconstruction of hand-object interactions quantitatively and qualitatively on the held-out sequences and compare DiffHOI with two model-free baselines (Section 5.3.2). We then analyze the effects of both category-prior and hand-prior respectively, ablate the contribution from each geometry modality, and analyze its robustness to initial prediction errors (Section 5.3.3). In Section 5.3.4, we discuss how DiffHOI compares with other template-based methods. Lastly, in Section 5.3.5, we show that our method is able to reconstruct HOI from in-the-wild video clips both in first-person and from

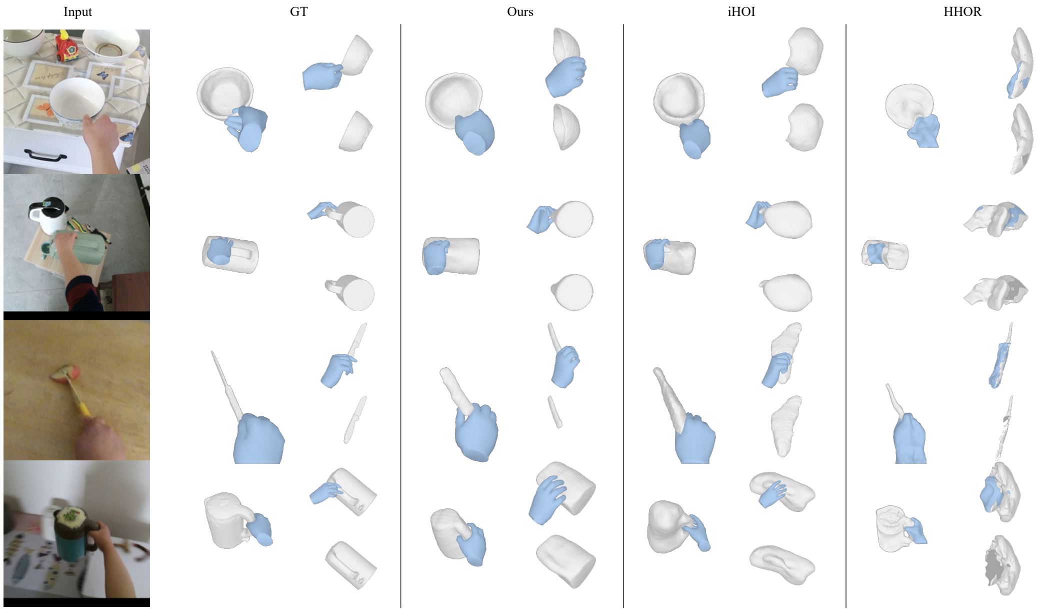


Figure 5.5: **Qualitative evaluation on HOI4D:** We show reconstruction by our method (DiffHOI) along with two baselines [85,232] in the image frame (left) and another novel view with (top right) or without (bottom right) hand. Please see project website for reconstruction videos.

third-person view.

Dataset and Setup. HOI4D is an egocentric dataset consisting of short video clips of hand interacting with objects. It is collected under controlled environments and recorded by head-wear RGBD cameras. Ground truth is provided by fitting 6D pose of scanned objects to the RGBD videos. We use all of the 6 rigid object categories in portable size (mug, bottle, kettle, knife, toy car, bowl). To train the diffusion model, we render one random novel viewpoint for each frame resulting in 35k training points. We test the object reconstruction on held-out instances, two sequences per category. All of baselines and our method use the segmentation masks from ground truth annotations and the hand poses from the off-the-shelf prediction system [179] if required.

For in-the-wild dataset, we test on clips from EPIC-KITCHENS [35] videos and casual YouTube videos downloaded from the Internet. The segmentation masks are obtained using an off-the-shelf video object segmentation system [26].

5.3.1 Visualizing Data-Driven Priors

We show conditional generations by the pre-trained diffusion model in Fig. 5.10. Given the geometry rendering of hand (only visualizing surface normal), as well as a text prompt, we visualize 4 different generations from the diffusion model. Middle row shows the generated surface normal of the object and bottom row visualizes the generated object masks overlayed on top of the given hand mask, for a better view of the hand-object relations. Note that left and middle column condition on the same text prompts while middle and right column conditions on the same hand pose. Please see appendix for additional examples and visualizations of all modalities.

The generated object match the category information in the prompt while the generations are diverse in position, orientation, and size. Yet, all of the hand-object interactions are plausible, *e.g.* different generated handles all appear at the tip of the hand. Comparing middle and right examples, different category prompts lead to different generations given the same hand rendering. With the same prompt but different hands (left and middle), the generated objects flip the orientation accordingly. In summary, Fig. 5.10 indicates that the learned prior is aware of both the hand prior and the category-level prior hence being informative to guide the 3D reconstruction from clips.

5.3.2 Comparing Reconstructions of HOI4D

Evaluation Metric. We evaluate the object reconstruction errors. Following prior works [67, 85], we first align the reconstructed object shape with the ground truth by Iterative Closest Point (ICP), allowing scaling. Then we compute Chamfer distance (CD), F-score [200] at $5mm$ and $10mm$ and report mean over 2 sequences for each category. Chamfer distance focuses on the global shapes more and is affected by outliers while F-score focuses on local shape details at different thresholds [200].

Baselines. While few prior works tackle our challenging setting – 3D HOI reconstruction from casual monocular clips without knowing the templates, the closest works are two template-free methods from Huang *et al.* [85] (HHOR) and Ye *et al.* [232] (iHOI).

HHOR is proposed for in-hand scanning. It optimizes a deformable semantic implicit field to jointly model hand and object. HHOR captures the dynamics by a per-frame warping field while no prior is used during optimization. iHOI is a feed-forward method and reconstructs 3D objects from

Table 5.2: **Analysis of the effect of data-driven priors:** Quantitative results on HOI4D for object reconstruction error in the object-centric frame ($F@5$, $F@10$, CD) and for hand-object alignment in the hand frame (CD_h). We compare our method with ablations that does not use prior, or use other variants of diffusion models that only conditions on hand or category.

	$F@5$	$F@10$	CD	CD_h
No prior	0.47	0.73	2.7	**37.0**
Hand prior	0.39	0.65	2.8	55.0
Category prior	0.56	0.87	1.6	85.2
Ours	**0.62**	**0.91**	**0.8**	48.7

Table 5.3: **Ablation without surface normal, mask and depth:** Quantitative results on the HOI4D dataset for object reconstruction error using mean F1 scores (5mm, 10mm), CD in object frame and for hand-object alignment using CD in hand frame ($\mathrm{CD_h}$). We compare our method with other ablations that do not distill normals, masks, and depths respectively.

	$F@5$	$F@10$	CD	CD_h
− normal	0.37	0.57	4.5	282.6
− mask	0.57	0.84	1.2	106.7
− depth	0.66	0.93	0.7	49.6
Ours	**0.70**	**0.93**	**0.7**	**41.9**

single-view images by learning the hand prior between hand poses and object shapes. The method does not leverage category-level prior and do not consider time-consistency of shapes. We finetune their pretrained model to take in segmentation masks. We evaluate their result by aligning their predictions with ground truth for each frame and report the average number across all frames.

Results. We visualize the reconstructed HOI and object shapes from the image frame and a novel viewpoint in Fig. 5.5. HHOR generates good-looking results from the original view but actually degenerates to a flat surface since it does not incorporate any prior knowledge besides the visual observation. It also cannot decompose the hand and the object on the unobserved side of the scene because HHOR distinguishes them by per-point classification predicted from the neural field, which does not get gradient

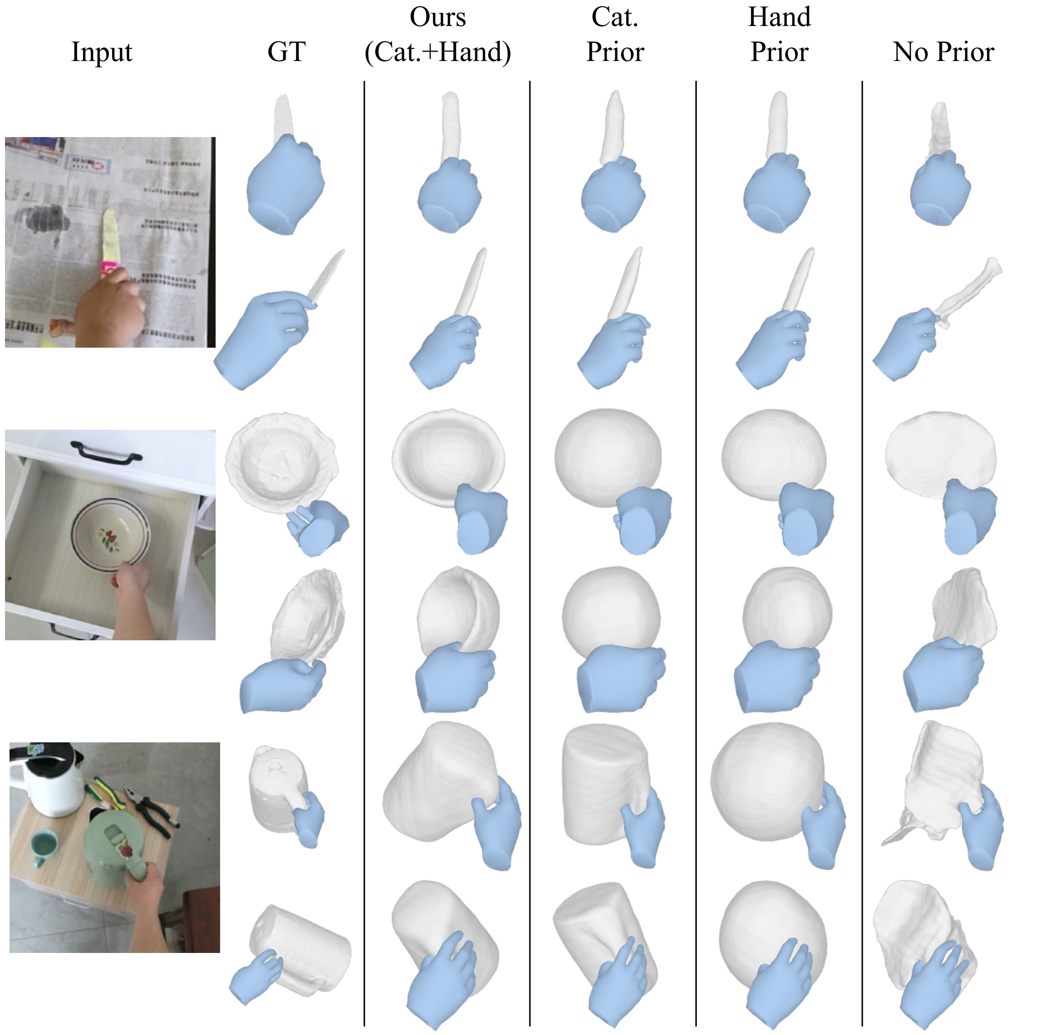


Figure 5.6: Visualizing HOI reconstruction comparisons of our method with other variants of diffusion models that only incorporate category prior, hand prior, and no prior. (Top: image frame, bottom: novel view)

from the observations. iHOI reconstructs better object shapes and interactions but it is not very accurate as it cannot aggregate information across different frames. Its prediction is not time consistent either (better visualized as videos). In contrast, we are able to reconstruct time-persistent object shapes with time changing hand poses. The reconstructed object is more accurate, *e.g.* knife blade is thinner and the kettle body is more cylindrical.

This is consistent with quantitative results in Tab. 5.1. HHOR generally

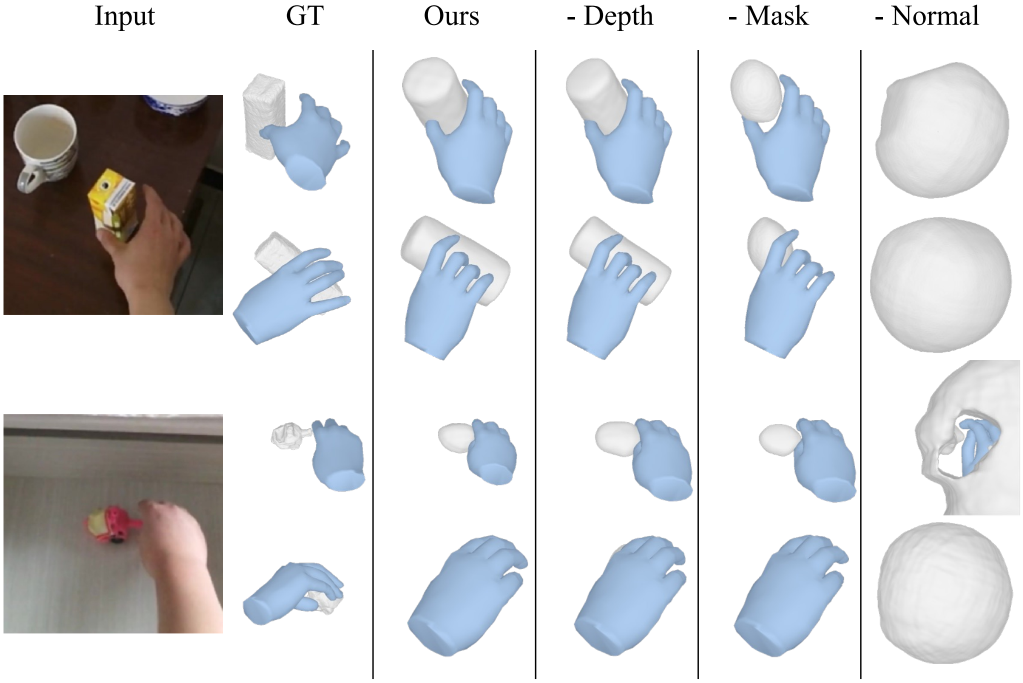


Figure 5.7: **Ablation Study:** Visualizing HOI reconstruction comparison of our method and variants that do not distill on depth, mask, and normals. (Top: image frame, bottom: novel view)

performs unfavorably except for knife category. While iHOI performs better, its quality is limited by only relying on information from a single frame. DiffHOI outperforms the baseline methods by large margins in most sequences and performs the best on all three metrics for mean values.

5.3.3 Ablation Studies

We ablate our system carefully to analyze the contribution of each component. Besides the object reconstruction errors in the aligned object-centric frame, we further evaluate the hand-object *arrangement* by reporting the Chamfer distance of objects in hand frame, *i.e.* $CD_h \equiv CD(T^t_{o\to h}O, \hat{T}^t_{o\to h}\hat{O}))$. We only report mean value in the main paper. Please refer to supplementary for category-wise results.

How does each learned prior help? We analyze how the category and hand priors affect reconstruction by training two more diffusion models conditioned only on text-prompt or hand renderings respectively. We also com-

Table 5.4: **Error analysis against hand pose noise:** * marks our unablated method. Numbers in parentheses are per-frame prediction errors before optimization.

	Object Reconstruction			Hand Estimation	
	F@5 ↑	F@10 ↑	CD ↓	MPJPE ↓	AUC ↑
GT	0.68	0.91	0.75	–	–
Prediction*	0.62	0.91	0.77	26.9(28.4)	0.49(0.47)
Pred. Error ×2	0.63	0.87	1.01	40.7(44.6)	0.31(0.27)

pare with the variant without optimizing $\mathcal{L}_{\text{SDS}}$ (no prior). As reported quantitatively, we find that *category prior helps object reconstructions while hand prior helps hand-object relation* (*Tab. 5.2*). And combining them both results in best performance.

We highlight an interesting qualitative result of reconstructing the bowl in Fig. 5.6. Neither prior can reconstruct the concave shape on its own – the hand pose alone is not predictive enough of the object shape while only knowing the object to be a bowl cannot make the SDS converge to a consensus direction that the bowl faces. Only knowing *both* can the concave shapes be recovered. This example further highlights the importance of both priors.

Which geometry modality matters more for distillation? Next, we investigate how much each geometry modality (mask, normal, depth) contributes when distilling them into 3D shapes. Given the same pretrained diffusion model, we disable one of the three input modalities in optimization by setting its weight on $\mathcal{L}_{\text{SDS}}$ to 0.

As visualized in Fig. 5.7, the surface normal is the most important modality. Interestingly, the model collapses if not distilling surface normals and even performs worse than the no-prior variant. Without distillation on masks, the object shape becomes less accurate probably because binary masks predict more discriminative signals on shapes. Relative depth does not help much with global object shape but it helps in aligning detailed local geometry (F@5) and aligning the object to hand (F@10).

How robust is the system to hand pose prediction errors? We report the object reconstruction performance when using GT vs predicted hand pose in Tab. 5.4, and find that our system is robust to some prediction error. Moreover, even if we artificially degrade the prediction by doubling the error, our performance remains better than the baselines (Tab. 5.1). We also report the hand pose estimation metrics and find that our optimization improves the initial predictions (in parentheses).

Table 5.5: **Comparison with template-based baseline:** Quantitative results on the HOI4D dataset for object reconstruction error in the object-centric frame ($F@5, F@10, CD$) and for hand-object alignment (CD_h). We compare our method with HOMAN [71] with the ground truth template (-GT), with random templates from the training split (and reporting the average), and with furthest template from the ground truth (-furthest).

	$F@5 \uparrow$	$F@10 \uparrow$	$CD \downarrow$	$CD_h \downarrow$
HOMAN-GT	1.00	1.00	0.00	84.3
HOMAN-average	0.76	0.94	0.48	120.9
HOMAN-furthest	0.49	0.78	1.33	157.9
Ours(DiffHOI)	0.62	0.91	0.78	48.7

5.3.4 Comparing with Template-Based Methods

We compare with HOMAN [71], a representative template-based method that optimizes object 6D poses and hand articulations with respect to reprojection error and multiple interaction objectives including contact, intersection, distance, relative depth, temporal smoothness, *etc.*.

We show quantitative and qualitative results in Tab. 5.5 and 5.8. Note that evaluating HOMAN in terms of object reconstruction is equivalent to evaluating templates since the objects are aligned in the object-centric frame. We first report the average object reconstruction errors when optimizing with different templates from training sets. While the gap indicates potential room to improve object shapes for template-free methods, DiffHOI is favorable over some templates in the training set. Nevertheless, when evaluating the objects in the hand frame, DiffHOI outperforms HOMAN by a large margin. The numbers along with visualizations in Fig. 5.8 indicate that template-based methods, even when optimizes with multiple objectives to encourage interactions, still struggle to place objects in the context of hands, especially for subtle parts like handles. Furthermore, optimizing with random templates degrades CD_h significantly, highlighting the inherent drawbacks of template-based methods to demand the accurate templates.

5.3.5 Reconstructing In-the-Wild Video Clips

Lastly, we show that our method can be directly applied to more challenging video clips. In Fig. 5.9 top, we compare between our method and iHOI [232]. iHOI predicts reasonable shapes from the front view but fails on transparent objects like the plastic bottle since it is never trained on such appearance. In contrast, we transfer better to in-the-wild sequences as the learned prior

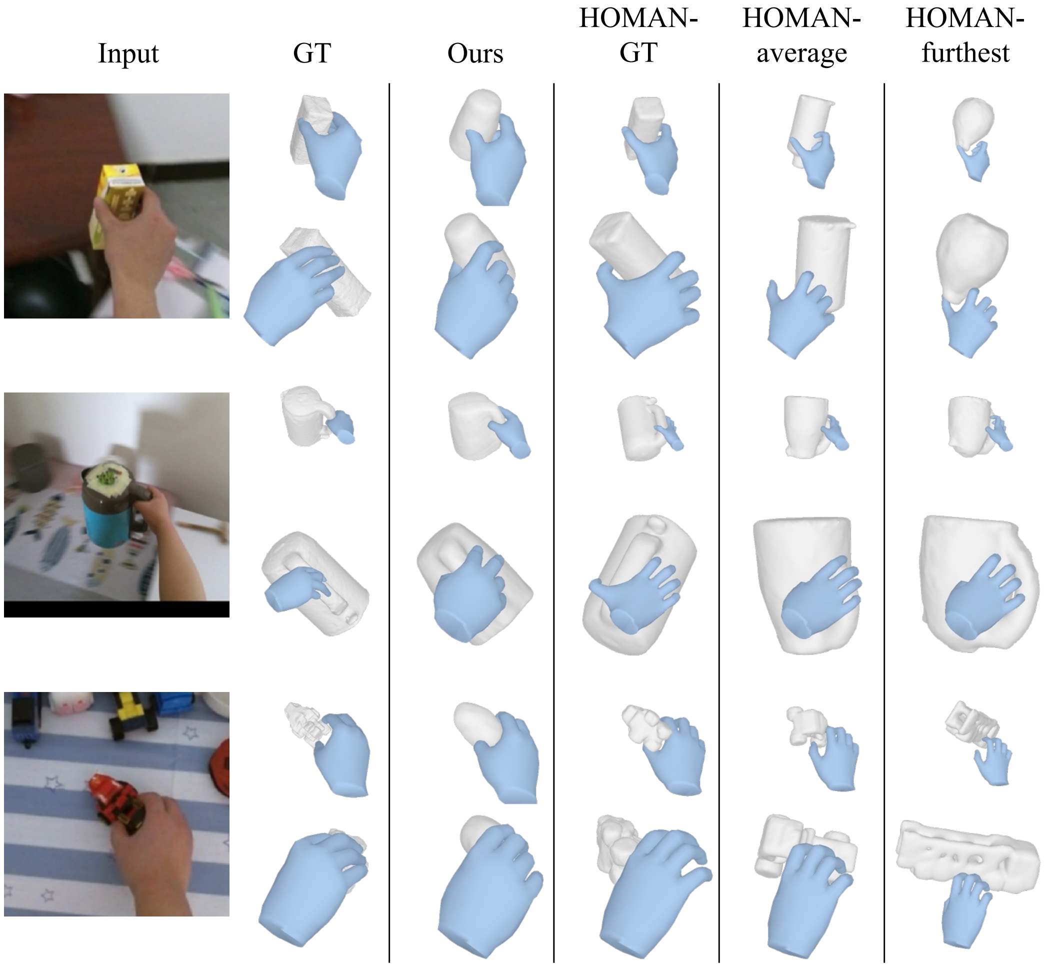


Figure 5.8: **Comparing with template-based method:** We show reconstruction in the image frame (top) and from a novel view (bottom) by our method and HOMAN [71] when provided with ground-truth templates, a random template, and the most dissimilar template in the training split.

only take on geometry cues. In Fig. 5.9 bottom, we visualize more results from our method. By incorporating learned priors, our method is robust to mask prediction inaccuracy, occlusion from irrelevant objects (the onion occludes knife blade), truncation of the HOI scene (bowl at the bottom left), *etc.*. Our method can also work across ego-centric and third-person views since the learned prior is trained with uniformly sampled viewpoints. The reconstructed shapes vary from thin objects like knives to larger objects like kettles.

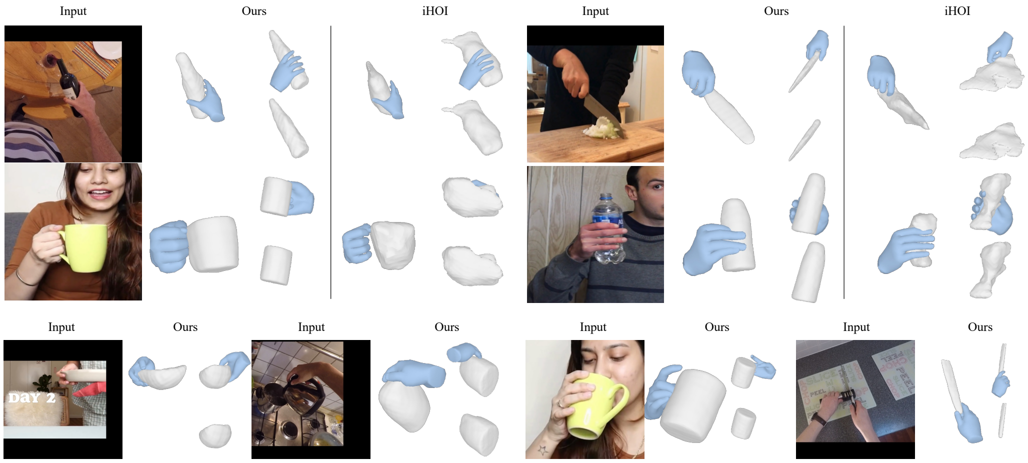


Figure 5.9: Comparing reconstructions of our method and the iHOI baseline [232] on 8 in-the-wild video clips taken from the Internet (Left: image frame, top right: novel view HOI, and bottom right: novel view object-only).

5.4 Discussion

In this work, we propose a method to reconstruct hand-object interactions without any object templates from daily video clips. Our method is the first to tackle this challenging setting. We represent the HOI scene by a model-free implicit field for the object and a model-based mesh for the hand. The scene is optimized with respect to re-projection error and a data-driven geometry prior that captures the object shape given category information and hand poses. Both of these modules are shown as critical for successful reconstruction. Despite the encouraging results, there are several limitations: the current method can only handle small hand-object motions in short video clips up to a few (∼5) seconds. Despite the challenges, we believe that our work takes an encouraging step towards a holistic understanding of human-object interactions in everyday videos.

5.5 Implementation Details and Additional Results

In the supplementary materials, we provide more implementation details and experimental results. We discuss the details of differentiable rendering of the HOI scene representation (Sec. 5.5.1), network architectures (Sec. 5.5.2), scored distillation sampling of the pretrained diffusion model (Sec. 5.5.3), and initialization details (Sec. 5.5.5). We also describe how to get 2D seg-

Table 5.6: **Full ablation results of object reconstruction:** Quantitative results for object reconstruction error using F1@5mm and F1@10mm scores and Chamfer Distance (mm). We compare our method with variants that do not optimize per-frame object poses (Sec.5.5.8), blend hand and object masks in a hard way (Sec.5.5.9), or do not distill certain geometry modality (Sec. 4.2, Tab. 4)

	Mug			Bottle			Kettle			Bowl			Knife			ToyCar			Mean		
	$F@5$	$F@10$	CD	$F@5$	$F@10$	CD	$F@5$	$F@10$	CD	$F@5$	$F@10$	CD	$F@5$	$F@10$	CD	$F@5$	$F@10$	CD	$F@5$	$F@10$	CD
wo learning pose	0.66	0.88	0.8	0.27	0.82	1.3	0.26	0.65	2.1	0.74	0.97	0.4	0.57	0.95	0.7	0.75	0.97	0.4	0.54	0.83	0.89
hard blending	0.66	0.88	0.8	0.42	0.85	1.0	0.36	0.72	2.6	0.49	0.85	1.1	0.71	0.96	0.5	0.78	0.99	0.3	0.57	0.83	0.97
− mask	0.54	0.81	1.3	0.24	0.53	2.5	0.47	0.80	2.0	0.76	0.98	0.4	0.72	0.97	0.5	0.67	0.98	0.5	0.57	0.81	1.06
− normal	0.50	0.77	1.4	0.20	0.43	3.9	0.21	0.43	6.1	0.38	0.63	4.0	0.11	0.24	11.1	0.83	0.96	0.4	0.37	0.55	3.46
− depth	0.62	0.89	0.8	0.79	0.98	0.3	0.53	0.84	1.6	0.55	0.90	0.8	0.72	0.96	0.5	0.77	0.97	0.4	0.66	0.89	0.74
Ours	0.66	0.89	0.8	0.84	0.99	0.3	0.71	0.88	1.1	0.50	0.86	0.9	0.74	0.98	0.4	0.75	0.97	0.4	**0.70**	**0.89**	**0.69**

Table 5.7: **Full ablation results of HOI alignment:** Quantitative results for hand-object alignment using Chamfer distance (mm) in hand frame ($\mathrm{CD_h}$). We compare our method with variants that do not optimize per-frame object poses (Sec.5.5.8), blend hand and object masks in a hard way (Sec.5.5.9), or do not distill certain geometry modality (Sec. 4.2, Tab. 4).

	Mug	Bottle	Kettle	Bowl	Knife	ToyCar	Mean
wo opt. obj pose	32.1	11.1	41.2	103.7	111.2	52.7	58.67
hard blending	26.2	14.7	96.2	177.0	80.7	72.9	77.96
− mask	30.6	17.8	48.4	418.0	80.3	45.0	106.69
− normal	489.9	319.8	69.4	172.8	299.8	344.1	282.62
− depth	19.3	9.6	134.2	53.5	35.0	49.6	50.20
Ours	20.4	12.2	45.2	92.8	53.8	27.0	**41.92**

mentation masks from in-the-wild clips (Sec. 5.5.4). Then, we show generation by the diffusion model (Sec. 5.5.6), full quantitative results reported in the main paper (Sec. 5.5.7). Furthermore, we also show supporting evidence that optimizing per-frame object poses (Sec. 5.5.8) and soft blending (Sec. 5.5.9) are both important for better performance.

5.5.1 Differentiable Rendering

Given an HOI scene representation at a certain time t consisting of an implicit field for the object and a mesh for the hand, we use differentiable volumetric renderer [230] and mesh renderer [125, 156] to get their masks (M_o, M_h) and depth (D_o, D_h). In order to supervise them with reprojection loss with respect to the ground truth semantic masks, we blend hand and object masks by their predicted depths to obtain the rendered semantic masks $M \equiv B(M_h, M_o, D_h, D_o)$.

The soft blending is computed as expected light transported to the cameras, similar to blending two-layer surfaces of in mesh rendering [156]. More specifically, denote m_h, d_h, m_o, d_o as the value at pixel (i, j), *e.g.* $m_h \equiv M_h[i, j]$. For any pixel (i, j), the blended value is computed as

$$m = B(m_h, m_o, d_h, d_o) = \frac{\sum_{k=0,1} w_k l_k}{\sum_{k=0,1} w_k + w_{bg}} \tag{5.5}$$

where subscript k denotes the sorted value of hand and object according to the predicted depth; l_k is the one-hot semantic label (all 0 for background). w_k is the weight computed from depth:

$$w_k = m_k \exp \frac{z_k - \max_{k,i,j} Z_k[i, j]}{\gamma}, z_k = m_k \frac{d^{\text{far}} - d_k}{d^{\text{far}} - d^{\text{near}}} \tag{5.6}$$

We show in Sec. 5.5.9 that soft blending (with loss in semantic masks) is important for better results and performs favorably to the alternative (hard blending with ordinal depth loss [71,236]).

5.5.2 Network Architectures and Training Details (Sec. 3.1 3.2)

Implicit field. We use Multi-Layer Perceptron (MLPs) to implement the neural implicit surface of the object ϕ. We borrow the architecture in the original VolSDF [230] and reduce the network capacity to half as we find it to suffice. More specifically, we stack four-layer blocks of which each is a linear layer with channel dim 64 followed by a SoftPlus activation. We apply positional encoding to the queried point X with 6 frequencies.

Conditional diffusion models. The backbone of the conditional diffusion model is based on the architecture of the text-to-image inpainting model [149]. More specifically, it is a 16-layer UNet with cross attentions and skip layers. The text condition along with the diffusion step embedding is passed to the bottleneck of the UNet and is fused with the image feature by cross-attention. The text prompt is encoded as CLIP tokens [168].

Details of training diffusion model. We train the diffusion model with batch size 8, learning rate $1e-4$. We use AdamW [131] optimizer with weight decay 0.01 and train for $500k$ iterations. We use linear noise schedule [175].

Details of optimizing HOI scene. We follow the training setup in a reimplementation [1] of the original paper [230]. We optimize the scene with 1024 rays per step, and set initial learning rate $5e-4$ with exponential learning rate scheduler. We use Adam [99] optimizer and optimize for $50k$ iterations per scene. Within a batch, we bias the sampled pixels from the background, hand, and object region with probability $0.35, 0.35, 0.3$ and linearly interpolate the probability to $0.1, 0.1, 0.8$ in order to spend more effective computation on the object of interest, same as HHOR [85]. In the first 100 warm-up iterations, we turn off SDS and only optimize for the reprojection loss and other regularization terms. This will make the optimization more stable.

5.5.3 Score Distillation Sampling (Sec. 3.3)

With the pretrained diffusion model, we follow DreamFusion [161] to distill the learned prior to the 3D representation. The main idea is to let the diffusion model denoise the corrupted renderings and treats the denoised output as 'ground truth'. More specifically, at each optimization step, we randomly sampled a viewpoint with random rotation from $SO(3)$ and random camera distance. Then, we render the geometry renderings G_o, G_h from the given viewpoint in resolution 64x64. Next, we corrupt the geometry rendering of the object with some noise $G_o^i = \sqrt{\bar{\alpha}_i} G_o + \sqrt{1-\bar{\alpha}_i}\epsilon$ ($\bar{\alpha}$ is the noise scheduling, ϵ is a gaussian noise) and pass it through the diffusion model along with the geometry rendering of the hand and text prompt.

$$\hat{G}_o^i = D_\psi(G_o^i | G_h, C) \tag{5.7}$$

We set the classifier-free guidance scale to 4, which is different from the original paper where a small guidance scale cannot converge. It is probably because 2D observations provide stronger cues than text thus leading to easier convergence.

5.5.4 Obtaining hand-object masks for in-the-wild clips.

While we provide ground truth segmentation masks to all methods on HOI4D, we obtain the segmentation masks by off-the-shelf prediction systems [26, 103, 187] for in-the-wild clips. More specifically, we first use a hand-object interaction detector [187] to detect the location of the hand and the active object in the first frame. Then, given the detected bounding boxes, we use PointRend [103] to get the corresponding masks. Next, we pass the masks of

[1] `https://github.com/ventusff/neurecon`

interest in the first frame to a video object segmentation system STCN [26] and obtain the tracked masks in every frame.

To automatically filter out the clips with undesirable segmentation quality, we run the STCN to track forward and backward in time and calculate the Intersection over Union (IoU) between the initial masks and the masks after tracking back. We use clips with IoU higher than 40% for both hand and object masks.

5.5.5 Initialization with Off-the-Shelf Predictions (Sec. 3.3.)

We use an off-the-shelf hand reconstruction [180] to estimate initial camera poses $T^t_{c\to h}$, hand shape parameter β, and hand articulation θ^t_A. The off-the-shelf system predicts per-frame 10-dim hand shape parameters β^t, 48-dim hand poses θ^t, and a weak perspective camera s^t, t^t_x, t^t_y. We take the average of shape parameters across all frames to initialize the hand shape parameter. Among the 48-dim predicted hand pose, we use the 45-dim finger articulation θ^t_A to initialize hand articulation parameter while use the remaining 3-dim wrist orientation θ_w as the rotation component of camera pose $T^t_{c\to h}$. The translation component is computed by converting the predicted weak-perspective camera to a full-perspective camera (we use a pinhole camera with a focal length of 1 and the principal point at the center of the frame following Zhang *et al.* [236]). This is to handle large perspective effects, which are common in daily videos of indoor scenes. Given focal length *f* and principal points p_x, p_y, the translation component then becomes $l^t = ((t^t_x - p_x)/s^t, (t^t_y - p_y)/s^t, f/s^t)$. To put them together, the initial camera pose in the hand frame is initialized as:

$$T^t_{c\to h} = [R^t|l^t] = [\text{Rot}(\theta^t_w)| \begin{pmatrix} (t^t_x - p_x)/s^t \\ (t^t_y - p_y)/s^t \\ f/s^t \end{pmatrix}] \tag{5.8}$$

5.5.6 Results of diffusion model generation

We show some conditional generations by the pre-trained diffusion model in Fig. 5.10. Given the geometry rendering of hand (i) of which row 1-4 visualize surface normal, depth, mask, and uv coordinate, as well as a text prompt with category information, we visualize 5 different generations (ii-vi) from the diffusion model. Row 1-3 in col ii-vi shows the generated geometry rendering of the object, and row 4 visualizes overlayed hand and object masks

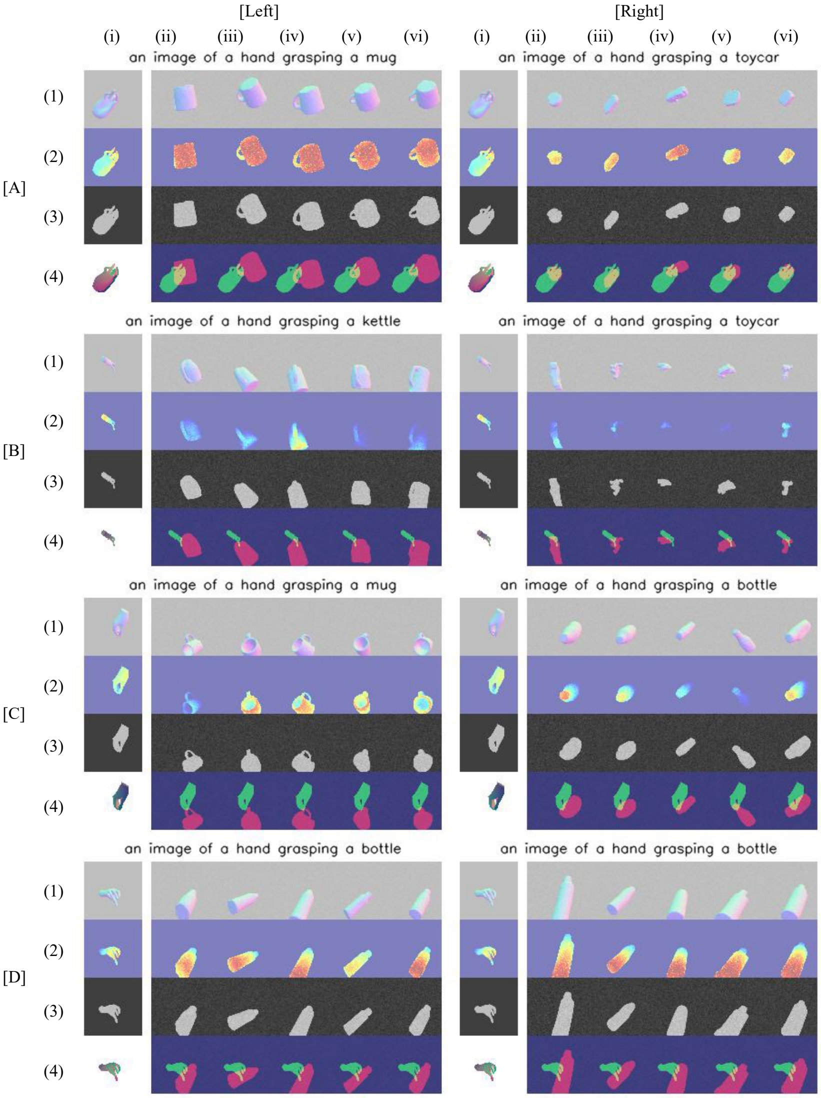

[Left]
[Right]
(i)
(ii)
(iii)
(iv)
(v)
(vi)
an image of a hand grasping a mug
an image of a hand grasping a toycar
an image of a hand grasping a kettle
an image of a hand grasping a toycar
an image of a hand grasping a mug
an image of a hand grasping a bottle
an image of a hand grasping a bottle
an image of a hand grasping a bottle
[A]
[B]
[C]
[D]
(1)
(2)
(3)
(4)

Figure 5.10: **Generations from conditional diffusion model.** Given the geometry rendering of hand (i) (row 1-4 visualizing surface normal, depth, mask, and uv coordinate), as well as a text prompt with category information, we visualize 5 different generations (ii-vi) from the diffusion model. Row 1-3 in col ii-vi shows the generated geometry rendering of the object, and row 4 visualizes overlayed hand and object masks for a better view of the hand-object relations. All examples on the left use the ground truth paired hand and category information while each example to its right uses another random category but remain hand the same.

for a better view of the hand-object relations, *i.e.* our model does not output (ii-vi 4). All examples on the left use the ground truth pairs of hand and category information while each example to its right uses another random category but remains hand the same.

As shown in the figure, the generated object matched the category information in the prompt while the generations are diverse in position, orientation, and size. Yet, all of the hand-object interactions are realistic, *e.g.* different generated kettle/mug handles all appear at the tip of the hand. Comparing left and right examples, different category prompts lead to different generations given the same hand rendering. With the same prompt but different hands, the generated objects also change appearance accordingly. For example, in the subfigure [Left A,C], the handles appear at the left when the hand approaches from the left and vice versa.

Fig. 5.10 indicates that the learned prior is aware of both the hand prior and the category-level prior hence being informative to guide the 3D reconstruction from clips.

5.5.7 Category-wise results in ablations (Tab. 4)

In Tab. 4 in the main paper, we only report mean value across all categories due to space limits. We provide quantitative results across all categories in Tab. 5.6 (object reconstruction) and Tab. 5.7 (HOI alignment).

5.5.8 Ablation: Optimizing vs Fixing Object Pose.

While we observe that the pose of the object in contact relative to hands $T^t_{h \rightarrow o}$ does not change much, we still optimize per-frame object poses to account for potential relative motion. As reported in Tab. 5.6, 5.7 and shown on the project page, allowing changing pose across time improves the performance.

5.5.9 Ablation: Soft Blending

Our method obtains the final HOI semantic masks by soft blending hand and object rendering as a weighted sum of the labels where the weight depends on their predicted depth. The alternative way is to select the label of the front surface and apply additional ordinal depth loss. This is common in optimizing the interactions of two template meshes [71, 236]. As shown in the qualitative results on the webpage, the alternative method generates less desirable hand-object relations as the hand intersects with the object. It is consistent with quantitative results in Tab. 5.6 and 5.7.

Chapter 6

Predicting Hand-Object Interactions via Image Synthesis

While previous chapters have studied reconstructing the ongoing interactions, it does not address the question of what are the possible interactions that could happen. Consider the bottles, bowls and cups shown in the left column of Figure 6.1. How might a human hand interact with such objects? Not only is it easy to imagine, from a single image, the types of interactions that might occur (*e.g.*, 'grab/hold'), and the interaction locations that might happen (*e.g.* 'handle/body'), but it is also quite natural to hallucinate—in vivid detail— several ways in which a hand might contact and use the objects. This ability to predict and hallucinate hand-object-interactions (HOI) is critical to functional understanding of a scene, as well as to visual imitation and manipulation.

Can current computer vision algorithms do the same? On the one hand, there has been a lot of progress in image generation, such as synthesizing realistic high-resolution images spanning a wide range of object categories [109, 184] from human faces to ImageNet classes. Newer diffusion models such as Dall-E 2 [169] and Stable Diffusion [175] can generate remarkably novel images in diverse styles. In fact, highly-realistic HOI images can be synthesized from simple text inputs such as "a hand holding a cup" [169, 175].

On the other hand, however, such models fail when conditioned on an image of a particular object instance. Given an image of an object, it remains an extremely challenging problem to generate realistic human object interaction. Solving this problem requires (at least implicitly) an understanding of physical constraints such as collision and force stability, as well as modeling the semantics and functionality of objects — the underlying affor-

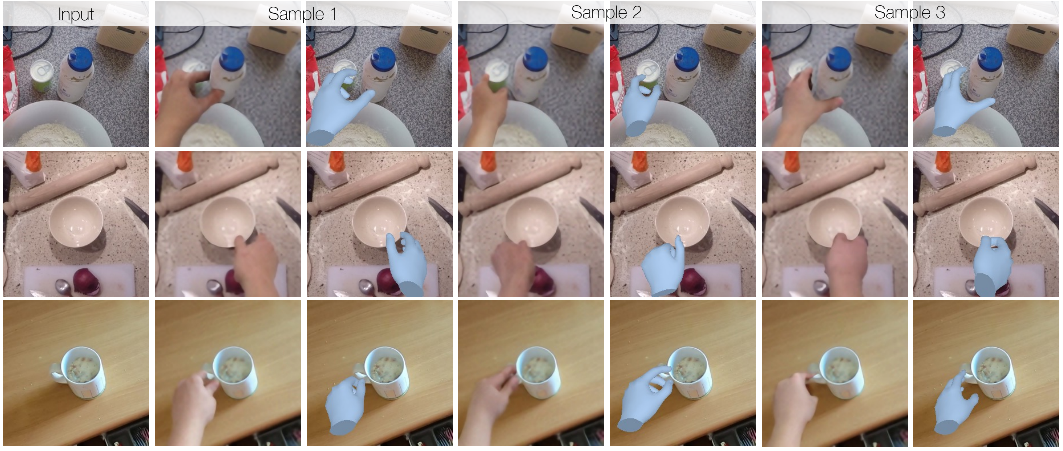


Figure 6.1: Given a single RGB image of an object (first column), we synthesize plausible images of hand-object interactions from which feasible 3D hand poses can be directly extracted (remaining columns).

dances [54]. For example, the hand should prefer to grab the kettle handle but avoid grabbing the knife blade. Furthermore, in order to produce visually plausible results, it also requires modeling occlusions between hands and objects, their scale, lighting, texture, *etc.*.

In this chapter, we propose a method for interaction synthesis that addresses these issues using diffusion models. In contrast to a generic image-conditioned diffusion model, we build upon the classic idea of disentangling *where* to interact (*layout*) from *how* to interact (*content*) [64, 79]. Our key insight is that diverse interactions largely arise from hand-object layout, whereas hand articulations are driven by local object geometry. For example, a mug can be grasped by either its handle or body, but once the grasping location is determined, the placement of the fingers depends on the object's local surface and the articulation will exhibit only subtle differences. We operationalize this idea by proposing a two-step stochastic procedure: 1) a *LayoutNet* that generates 2D spatial arrangements of hands and objects, and 2) a *ContentNet* that is conditioned on the query object image and the sampled HOI layout to synthesize the images of hand-object interactions. These two modules are both implemented as image-conditioned diffusion models.

We evaluate our method on HOI4D and EPIC-KITCHEN [37, 128]. Our method outperforms generic image generation baselines, and the extracted hand poses from our HOI synthesis are favored in user studies against baselines that are trained to directly predict hand poses. We also demonstrate surprisingly robust generalization ability across datasets, and we show that

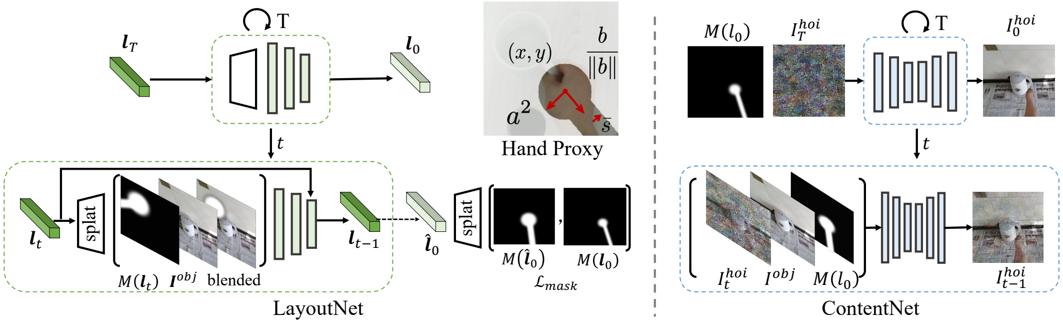


Figure 6.2: The proposed method consists of two image-conditioned diffusion models: LayoutNet and ContentNet. Given an object image, we first use LayoutNet (left) to predict a HOI spatial arrangement $\boldsymbol{l}_0$. For every diffusion step, the LayoutNet splats the noisy layout parameter into image space, concatenates it with the object image and their blending, and predicts the denoised layout. We apply the diffusion loss in the splatted 2D space $\mathcal{L}_{\text{mask}}$. Then the ContentNet (right) takes in the predicted layout along with the object image to synthesize an HOI image. The two modules are connected by the articulation-agnostic hand proxy (middle top).

our model can quickly adapt to new hand-object-interactions with only a few examples. Lastly, we show that our proposed method enables editing and guided generation from partially specified layout parameters. This allows us to reuse heatmap prediction from prior work [47, 146] and to generate consistent hand sizes for different objects in one scene.

Our main contributions are summarized below: 1) we propose a two-step method to synthesize hand-object interactions from an object image, which allows affordance information extracted from it; 2) we use inpainting techinuqes to supervise the model with paired real-world HOI and object-only images and propose a novel data augmentation method to alleviate overfit to artifacts; and 3) we show that our approach generates realistic HOI images along with plausible 3D poses and generalizes surprisingly well on out-of-distribution scenes. 4) We also highlight several applications that would benefit from such a method.

6.1 Method

Given an image of an object, we aim to synthesize images depicting plausible ways of a human hand interacting with it. Our key insight is that this multimodal process follows a coarse-to-fine procedure. For example, a mug can either be held by its handle or body, but once decided, the hand articulation is largely driven by the local geometry of the mug. We operationalize this idea by proposing a two-step stochastic approach as shown in Fig 6.2.

We first use a LayoutNet to predict plausible spatial arrangement of the object and the hand (Sec 6.1.1). The LayoutNet predicts hand proxy that abstracts away appearance and explicitly specifies 2D location, size and approaching direction of a grasp. This abstraction allows global reasoning of hand-object relations and also enables users to specify the interactions. Then, given the predicted hand proxy and the object image, we synthesize a plausible appearance of an HOI via a ContentNet (Sec 6.1.2). This allows the network to implicitly reason about 3D wrist orientation, finger placement, and occlusion based on the object's local shape. We use conditional diffusion models for both networks to achieve high-quality layout and visual content. The synthesized HOI image is realistic such that a feasible 3D hand pose can be directly extracted from it by an off-the-shelf hand pose reconstruction model (Sec 6.2.2).

Both networks are based on diffusion models introduced in Section 5.1. To supervise the system, we need pixel-aligned pairs of HOI images and object-only images that depict the exact same objects from the exact same viewpoints with the exact same lighting. We obtain such pairs by inpainting techniques that remove humans from HOI images. We further propose a novel data augmentation to prevent the trained model from overfitting to the inpainting artifacts (Sec 6.1.3).

6.1.1 LayoutNet: predicting where to grasp

Given an object image $\mathbf{I}^{obj}$, the LayoutNet aims to generate a plausible HOI layout $\boldsymbol{l}$ from the learned distribution $p(\boldsymbol{l}|\mathbf{I}^{obj})$. We follow the diffusion model regime that sequentially denoises a noisy layout parameter to output the final layout. For every denoising step, the LayoutNet takes in the (noisy) layout parameter along with the object image and denoises it sequentially, *i.e.* $\boldsymbol{l}_{t-1} \sim \phi(\boldsymbol{l}_{t-1}|\boldsymbol{l}_t, \mathbf{I}^{obj})$. We splat the layout parameter onto the image space to better reason about 2D spatial relationships to the object image and we further introduce an auxiliary loss term to train diffusion models in the layout parameter space.

Layout parameterization. Hands in HOI images typically appear as hands (from wrist to fingers) with forearms. Based on this observation, we introduce an articulation-agnostic hand proxy that only preserves this basic hand structure. As shown in Fig 6.2, the layout parameter consists of hand palm size a^2, location x, y and approaching direction $\arctan(b_1, b_2)$, *i.e.* $\boldsymbol{l} := (a, x, y, b_1, b_2)$. The ratio of hand palm size and forearm width $\bar{s}$ remains a constant that is set to the mean value over the training set. We obtain the ground truth parameters from hand detection (for location and size) and hand/forearm segmentation (for orientation).

Predicting Layout. The diffusion-based LayoutNet takes in a noisy 5-parameter vector $\boldsymbol{l}_t$ with the object image and outputs the denoised layout vector $\boldsymbol{l}_{t-1}$ (we define $l_0 = l$). To better reason about the spatial relation between hand and object, we splat the layout parameter into the image space $M(\boldsymbol{l}_t)$. The splatted layout mask is then concatenated with the object image and is passed to the diffusion-based LayoutNet. We splat the layout parameter to 2D by the spatial transformer network [89] that transforms a canonical mask template by a similarity transformation.

DDPM loss for layout. One could directly train the LayoutNet with the DDPM loss (Eq. 5.1) in the layout parameter space: $\mathcal{L}_{para} := \mathcal{L}_{\text{DDPM}}[\boldsymbol{l}; \mathbf{I}^{obj}]$. However, when diffusing in such a space, multiple parameters can induce an identical layout, such as a size parameter with opposite signs or approaching directions that are scaled by a constant. DDPM loss in the parameter space would penalize predictions even if they guide the parameter to a equivalent one that induce the same layout masks as the ground truth. As the downstream ContentNet only takes in the splatted masks and not their parameters, we propose to directly apply the DDPM loss in the splatted image space (see appendix for details):

$$\mathcal{L}_{mask} = \mathbb{E}_{(\boldsymbol{l}_0, \mathbf{I}^{obj}), \epsilon \sim \mathcal{N}(0,I), t} \| M(\boldsymbol{l}_0) - M(\hat{\boldsymbol{l}}_0) \|_2^2. \tag{6.1}$$

where $\hat{\boldsymbol{l}}_0 := D_\theta(\boldsymbol{l}_t, t, \mathbf{I}^{obj})$ is the output of our trained denoiser that takes in the current noisy layout $\boldsymbol{l}_t$, the time t and the object image $\mathbf{I}^{obj}$ for conditioning.

In practice, we apply losses in both the parameter space and image spaces $\mathcal{L}_{mask} + \lambda \mathcal{L}_{para}$ because when the layout parameters are very noisy in the early diffusion steps, the splatted loss in 2D alone is a too-weak training signal.

Network architecture. We implement the backbone network as a UNet with cross-attention layers and initialize it from the pretrained diffusion model [149]. The model takes in images with seven channels as shown in Fig 6.2: 3 for the object image, 1 for the splatted layout mask and another 3 that blends the layout mask with object image. The noisy layout parameter attends spatially to the feature grid from the UNet's bottleneck and spit out the denoised output.

Guided layout generation. The LayoutNet is trained to be conditioned on an object image only but the generation can be guided with additional conditions at test time without retraining. For example, we can condition the network to generate layouts such that their locations are at certain places *i.e.* $\boldsymbol{l} \sim p(\boldsymbol{l}_0|\mathbf{I}^{obj}, x = x_0, y = y_0)$. We use techniques [195] in diffusion models that hijack the conditions after each diffusion steps with corresponding noise levels. This guided diffusion enables user editing and HOI synthesis for scenes with a consistent hand scale (Sec. 7.1.2). Please refer to the appendix for LayoutNet implementation details.

6.1.2 ContentNet: predicting how to grasp

Given the sampled layout $\boldsymbol{l}$ and the object image $\mathbf{I}^{obj}$, the ContentNet synthesizes a HOI image $\mathbf{I}^{hoi}$. While the synthesized HOI images should respect the provided layout, the generation is still stochastic because hand appearance may vary in shape, finger articulation, skin colors, *etc.*. We leverage the recent success of diffusion models in image synthesis and formulate the articulation network as a image-conditioned diffusion model. As shown in Fig 6.2, at each step of diffusion, the network takes in channel-wise concatenation of the noisy HOI image, the object image and the splatted mask from the layout parameter and outputs the denoised HOI images $D_\phi(\mathbf{I}_t^{hoi}, t, [\mathbf{I}^{obj}, M(\boldsymbol{l})])$.

We implement the image-conditioned diffusion model in the latent space [174, 191, 207] and finetune it from the inpainting model that is pre-trained on large-scale data. The pretraining is beneficial as the model has learned the prior of retaining the pixels in unmask region and hallucinate to fill the masked region. During finetuning, the model further learns to respect the predicted layout, *i.e.*, retaining the object appearance if not occluded by hand and synthesizing hand and forearm appearance depicting finger articulation, wrist orientation, etc.

6.1.3 Constructing Paired Training Data

To train such a system, we need pairs of object-only images and HOI image. These pairs need to be pixel-aligned except for the hand regions. One possible way is to use synthetic data [32,72] and render their 3D HOI scene with and without hands. But this introduces domain gap between simulation and the real-world thus hurts generalization. We instead follow a different approach.

As shown in Fig 6.3, we first extract object-centric HOI crops from egocentric videos with 80% square padding. Then we segment the hand regions to be removed and pass them to the inpainting system [149] to hallucinate

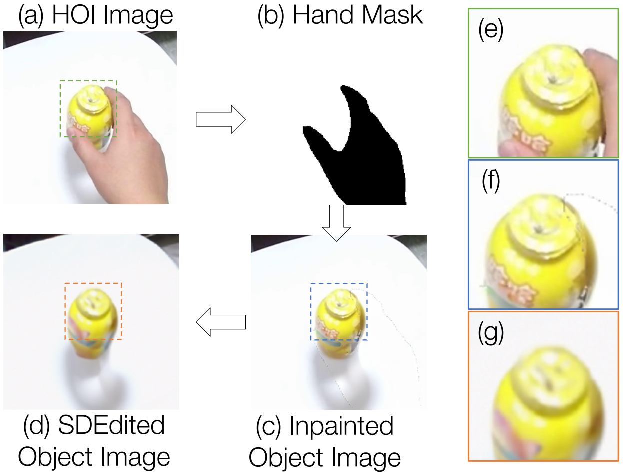


Figure 6.3: **Paired Data Generation:** Given an HOI image, we first segment out hand (b) and remove it by inpainting (c). Then we use SDEdit [136] to reduce inpainting artifact (d). As inpainting introduce discrepancy between mask and unmasked region (f) while SDEdit undesirably modifies the unmasked object region, we mix up *both* object image sets in training.

the objects behind hands. The inpainter is trained on millions of data with people filtered out therefore it is suitable for our task.

Data Augmentation. Although the inpainting generates impressive object-only images, it still introduces editing artifacts, which the networks can easily overfit to [237], such as sharp boundary and blurriness in masked regions. We use SDEdit [135] to reduce the discrepancy between the masked and unmasked regions. SDEdit first adds a small amount of noise (we use 5% of the whole diffusion process) to the given image and then denoises it to optimize overall image realism. However, although the discrepancy within images reduces, the unmasked object region is undesirably modified and the overall SDEdited images appear blurrier. In practice, we mix up the object-only images with and without SDEdit for training.

We collect all data pairs from HOI4D [128]. After some automatic sanity filtering (such as ensuring hands are removed), we generate 364k pairs of object-only images and HOI-images in total. We call the dataset HO3Pairs (Hand-Object interaction and Object-Only Pairs). We provide details and more examples of the dataset in the appendix.

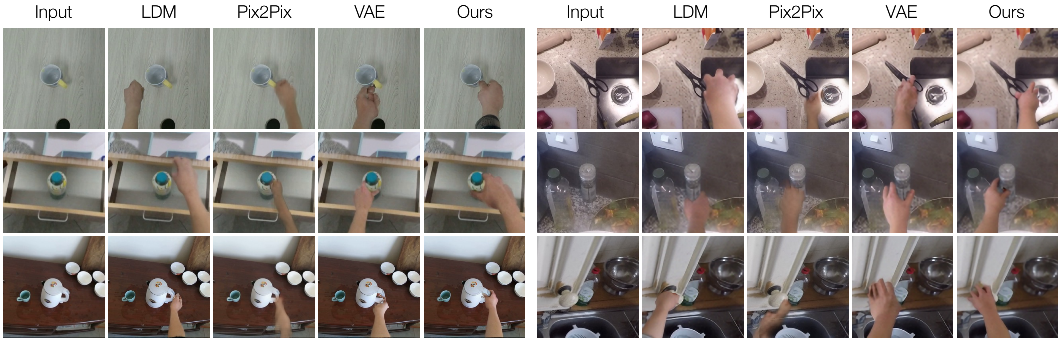


Figure 6.4: Visualizing HOI synthesis from our method and three baselines [88,101,174] on HOI4D (left) and EPIC-KITCHEN dataset (right).

Table 6.1: Quantitative results for HOI synthesis using contact recall, FID score, and a user study on the HOI4D and EPIC-KITCHEN datasets. We compare our method with prior works [88,101,174].

Method	HOI4D dataset												EPIC-KITCHEN dataset		
	Contact Recall(%)										FID	User Study	Contact Recall	FID	User Study
	Kettle	Knife	TrashCan	Chair	Mug	Bowl	ToyCar	Laptop	Bottle	mean					
LDM [174]	82.67	72.28	83.33	82.08	66.67	78.10	88.00	62.00	87.22	64.44	105.26	27.5	76.56	118.15	23.3
Pix2Pix [88]	79.50	70.26	82.50	76.88	68.50	79.64	89.00	63.00	85.42	73.02	107.09	15.5	70.00	125.62	13.3
VAE-ContentNet [101]	91.00	78.95	91.50	85.63	73.00	90.00	94.00	69.00	90.00	83.49	**98.19**	23.0	82.03	**115.86**	27.9
Ours	91.00	84.21	97.00	88.75	60.00	92.86	96.00	72.00	91.67	**87.14**	99.00	**34.0**	**86.56**	117.22	**35.4**

6.2 Experiments

We train our model on the contructed HO3Pairs dataset, evaluate it on the HOI4D [128] dataset and show zero-shot generalization to the EPIC-KITCHEN [37] dataset. We evaluate both the generated HOI images and the extracted 3D poses. For image synthesis, we compare with conditional image synthesis baselines and show that our method generates more plausible hands in interaction. Beyond 2D HOI image synthesis, we compare the extracted 3D poses with prior works that directly predict 3D hand poses. Furthermore, we show several applications enabled by the proposed HOI synthesis method, including few-shot adaptation, image editing by layout, heatmap-guided prediction and integrating object affordance with the scene.

Datasets Instead of testing with inpainted object images, we evaluate our model on the real object-only images cropped from the frames without hands. The goal is to prevent models from cheating by overfitting to the inpainting artifacts, as justified in the ablations below.

The HOI4D dataset is an egocentric video dataset recording humans in a lab environment interacting with various objects such as kettles, bottles, laptops, *etc.*. The dataset provides manual annotations of hand and object masks, action labels, object categories, instance ID, and ground truth 3D hand poses. We train and evaluate on 10 categories where full annotations are released. For each category, we hold out 5 object instances for evaluation. In total, we collect 126 testing images.

The EPIC-KITCHEN dataset displays more diverse and cluttered scenes. We construct our test set by randomly selecting 10 frames from each video clip. We detect and crop out objects without hands [224]. In total, we collect 500 object-only images for testing.

6.2.1 Evaluating Image Synthesis

Evaluation Metrics. We evaluate HOI generation using three metrics. First, we report the FID score [80, 185], which is widely used for image synthesis that measures the distance between two image sets. We generate 10 samples for every input and calculate FID with 1000 HOI images extracted from the test sets. We further evaluate the physical feasibility of the generated hands by the contact recall metric — it computes the ratio of the generated hands that are in the "in-contact" state by an off-the-shelf hand detector [187]. We also carry out user studies to evaluate their perceptual plausibility. Specifically, we present two images from two randomly selected methods to users and ask them to select the more plausible one. We collect 200 (for HOI4D) and 240 (for EPIC-KITCHEN) answers and report the likelihood of the methods being chosen.

Baselines. We compare our method with three strong image-conditional synthesis baselines. 1) *Latent Diffusion Model* (*LDM*) [174] is one of the state-of-the-art generic image generation models that is pre-trained with large-scale image data. We condition the model on the object image and finetune it on HO3Pair dataset. This baseline jointly generates both layout and appearance with one network. 2) *Pix2Pix* [88] is commonly used for pose-conditioned human/hand synthesis [20, 143]. We modify the model to condition on the generated layout masks that are predicted from our LayoutNet. 3) *VAE* [101] is a widely applied generative model in recent affordance literature [50, 115, 239]. This baseline uses a VAE with ResNet [75] as backbone to predict a layout parameter. The layout is then passed to our ContentNet to generate images.

Results. We visualize the generated HOI images in Fig 6.4. Pix2Pix typically lacks detailed finger articulation. While LDM and VAE generate more realistic hand articulations than Pix2Pix, the generated hands sometimes do not make contact with the objects. The hand appearance near the contact

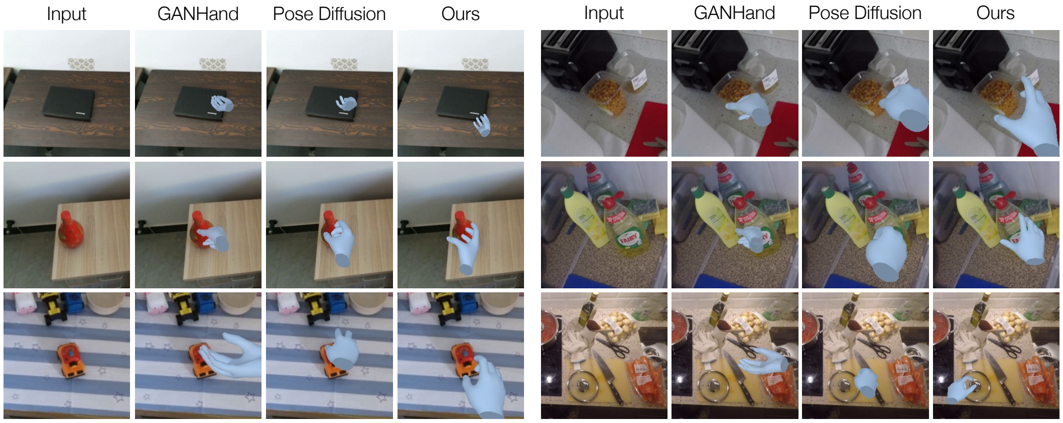


Figure 6.5: Visualizing 3D affordance prediction from our method, GANHand [32] and diffusion model [174] that directly predicts 3D pose on HOI4D (left) and EPIC-KITCHEN dataset (right).

region is less realistic. In some cases, LDM does not add hands at all to the given object images. In contrast, our model can generate hands with more plausible articulation and the synthesized contact regions are more realistic. This is consistent with the quantitative results in Tab 6.1. While we perform comparably to the baselines in terms of the FID score, we achieve the best in terms of contact recall. The user study shows that our results are favored the most. This may indicate that humans perceive interaction quality as a more important factor than general image synthesis quality.

Generalizing to EPIC-KITCHEN. Although our model is trained only on the HOI4D dataset with limited scenes and relatively clean backgrounds, our model can generalize to the EPIC-KITCHEN dataset without any finetuning. In Fig 6.4, the model also generalizes to interact with unseen categories such as scissors and cabinet handles. Tab 6.1 reports similar trends: performing best in contact recall, comparably well in image synthesis and is favored the most by users.

Ablation: Data Augmentation. Tab 6.2 shows the benefits of data augmentation to prevent overfitting. Without any data augmentation, the model performs well on the inpainted object images but catastrophically fails on the real ones. When we add aggressive common data augmentations like Gaussian blur and Gaussian noise, the performance improves. Training on SDEdited images further boosts the performance. The results also justify the use of real object images as test set since evaluating on the inpainted object images may not reflect the real performance.

Table 6.2: **Analysis of data augmentation**: contact recall (CR%) and FID score on the real and the inpainted object image set of HOI4D and comparisons of ours with the ablations of excluding aggressive common data augmentation (CmnAug) or SDEdit [136].

		Real Obj Img		Inpainted Img	
CmnAug	SDEdit	CR	FID	CR	FID
		39.37	113.93	89.05	89.38
✓		79.52	99.12	93.81	89.01
✓	✓	87.14	99.00	94.29	88.50

Table 6.3: User study for 3D affordance prediction on HOI4D and EPIC-KITCHEN dataset. We compare our method with GANHand [32] and a diffusion model that directly predicts 3D poses.

Method	HOI4D	EPIC
GANHand [32]	23.8	23.53
3D Pose Diffusion	27.9	34.1
Ours	**48.2**	**42.4**

Ablation: LayoutNet Design. We analyze the benefits from our LayoutNet design by reporting contact recall. The LayoutNet predicts more physically feasible hands by taking in the splatted layout masks instead of the 5-parameter layout vector (87.14% vs 78.10%). Moreover, the contact recall drops to 83.96% when the diffusion loss in Sec 6.1.1 is removed, verifying its contribution to the LayoutNet.

6.2.2 Evaluating Extracted 3D Hand Poses

Thanks to the realism of the generated HOI images, 3D hand poses can be directly extracted from them by an off-the-shelf hand pose estimator [181]. We conduct a user study to compare the 3D poses extracted from our HOI images against methods that directly predict 3D pose from object images. We present the rendered hand meshes overlaid on the object images to users and are asked to select the more plausible one. In total, we collected 400 and 380 answers from users for HOI4D and EPIC-KITCHEN, respectively.

Baselines. While most 3D hand pose generation works require 3D object meshes as inputs, a recent work by Corona *et al.* (GANHand) [32] can hallucinate hand poses from an object image. Specifically, they first map the object image to a grasp type [48] with the predefined coarse pose and

Table 6.4: **Few-shot Adaption:** Quantitative results using contact recall when finetuning the proposed HOI synthesis model and a pretrained inpainting model with 32 samples from new categories.

	bucket	scissors	stapler	mean
w HOI pretrain	92.0	95.0	70.0	85.7
w/o HOI pretrain	90.0	68.8	34.0	64.3

then regress a refinement on top. We finetune their released model on the HO3Pairs datasets with the ground truth 3D hand poses. We additionally implement a diffusion model baseline that sequentially diffuses 3D hand poses. The architecture is mostly the same as the LayoutNet but the diffused parameter is increased to 51 (48 for hand poses and 3 for scale and location) and the splatting function is replaced by the MANO [178] layer that renders hand poses to image. See the appendix for implementation details.

Results. As shown in Fig 6.5, GANHand [32] predicts reasonable hand poses for some objects but fails when the grasp type is not correctly classified. The hand pose diffusion model sometimes generates infeasible hand poses like acute joint angles. Our model is able to generate hand poses that are compatible with the objects. Furthermore, while previous methods typically assume right hands only, our model can automatically generate both left and right hands by implicitly learning the correlation between approaching direction and hand sides. The qualitative performance is also supported by the user study in Tab 6.3.

6.2.3 Application

We showcase several applications that are enabled by the proposed method for hand-object-image synthesis.

Few-shot Adaptation. In Tab 6.4, we show that our model can be quickly adapted to a new HOI category with as few as 32 training samples. We initialize both LayoutNet and ContentNet from our HOI4D-pretrained checkpoints and compare it with the baseline model that was pre-trained for inpainting on a large-scale image dataset [174]. We finetune both models on 32 samples from three novel categories in HOI4D and test with novel instances. The baseline model adapts quickly on some classes, justifying our reasons to finetune our model from them—generic large-scale image pretraining indeed already learns good priors of HOI. Furthermore, our HOI synthesis model performs even better than the baseline.

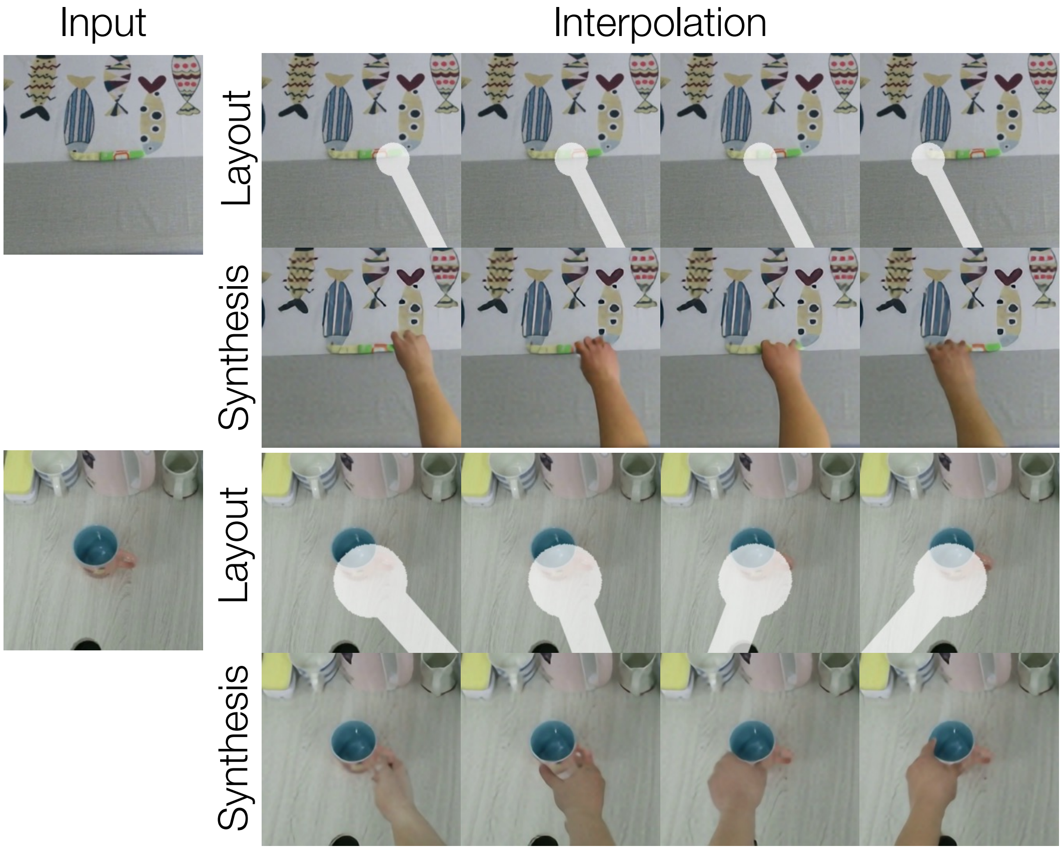


Figure 6.6: **Layout Editing**: Visualizing HOI synthesis when the conditioned layouts gradually change location and orientation.

Layout Editing. The layout representation allows users to edit and control the generated hand's structure. As shown in Fig 6.6, while we gradually change the layout's location and orientation, the synthesized hand's appearance changes accordingly. As the approaching direction to the mug changes from right to left, the synthesized fingers change accordingly from pinching around the handle to a wider grip around the mug's body.

Heatmap-Guided Synthesis. As shown in Sec 6.1.1, our synthesized HOI images can be conditioned on a specified location without any retraining. This not only allows users to edit with just keypoints, but also enables our model to utilize contact heatmap predictions from prior works [47, 146]. In Fig 6.7, we sample points from the heatmaps and conditionally generate layouts and HOI images which further specifies *how* to interact at the sampled location.

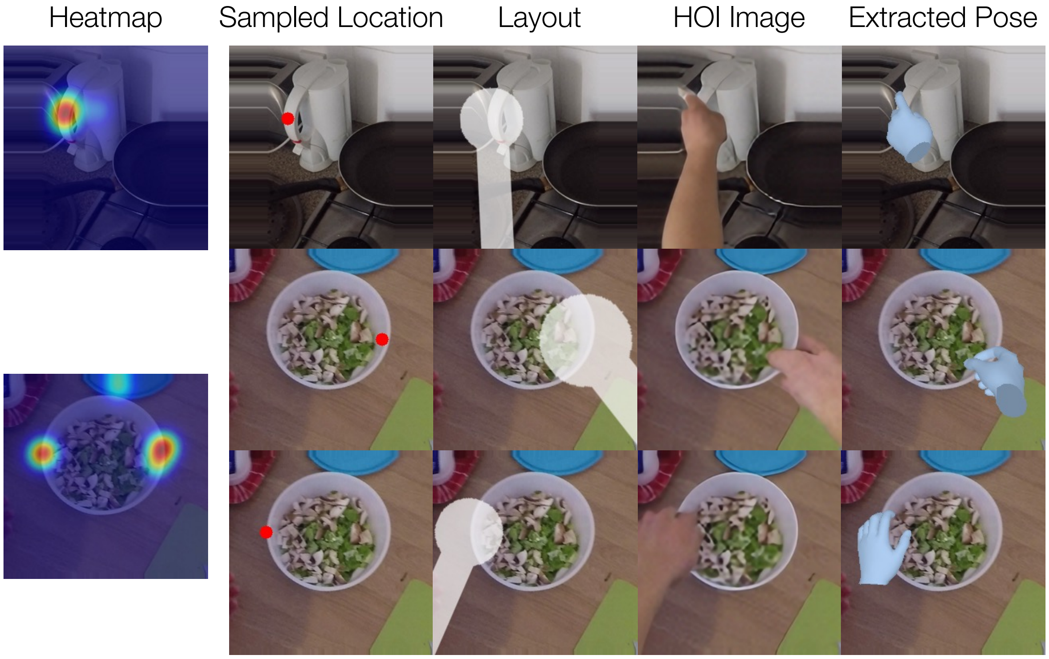


Figure 6.7: **Heatmap-guided synthesis:** Given a heatmap, LayoutNet is guided to generate layout at the sampled location, from which HOI images are synthesized and 3D poses are extracted.

Integration to scene. We integrate our object-centric HOI synthesis to scene-level affordance prediction. While the layout size is predicted relative to each object, hands for different objects in one scene should exhibit consistent scale. To do so, we first specify one shared hand size for each scene and calculate the corresponding relative sizes in each crops (we assume objects at similar depth and thus sizes can be transformed by crop sizes, although more comprehensive view conversions can be used). The LayoutNet is conditioned to generate these specified sizes with guided generation techniques (Sec 6.1.1). Fig 6.8 shows the extracted hand meshes from each crops transferred back to the scene.

6.3 Discussion

In this chapter, we propose to synthesize hand-object interactions from a given object image. We explicitly reason about *where* to interact and *how* to interact by LayoutNet and ContentNet. Both of them are implemented as

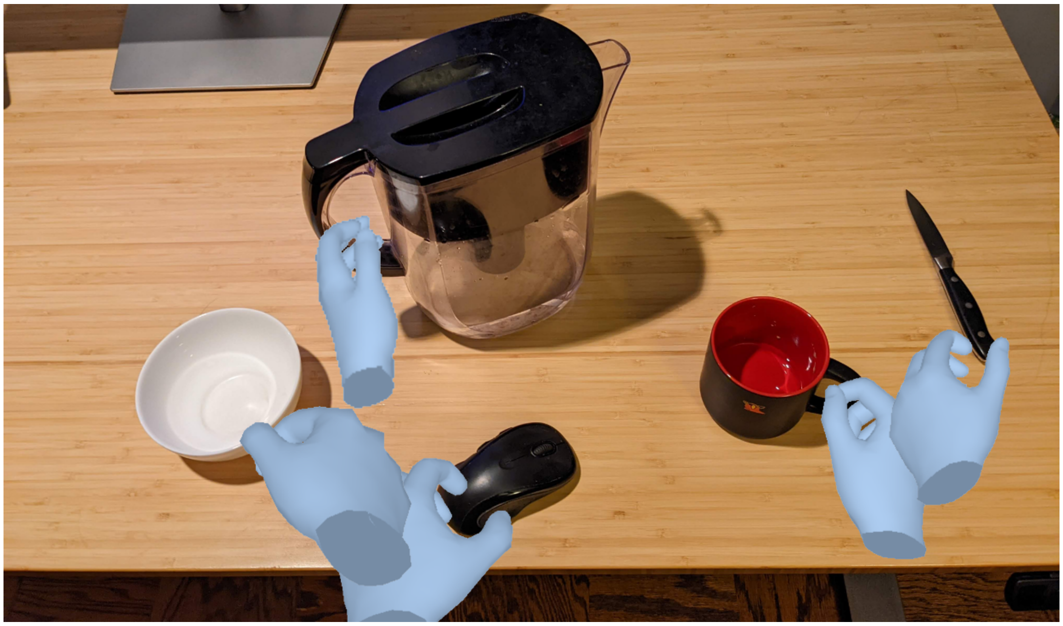


Figure 6.8: **Scene-level Integration:** Given a cluttered scene, we detect each object and synthesize its interactions individually. Each object's layout scale is guided to appear in the same size when transferred back to the scene.

diffusion models to achieve controllable and high-quality visual results. The synthesized HOI images enable a shortcut to more plausible 3D affordance via reconstructing hand poses from them. Although the generation quality and the consistency between the extracted 3D poses and images can be further improved, we believe that HOI synthesis along with our proposed solution opens doors for many promising applications and contributes towards the general goal of understanding human interactions in the wild.

6.4 Appendix

6.5 Implementation Details and Additional Results

In the appendix, we provide more implementation details and more qualitative results. We discuss the details of articulation-agnostic hand proxy and how to apply DDPM loss in the image space for training the LayoutNet (Sec. 6.5.1). We also present ablations on ContentNet(Sec. 6.5.2). We further show: (i) the paired data construction method being robust, in Sec. 6.5.3,

(ii) baseline implementations details in Sec. 6.5.4, (iii) details of integrating our approach to scene-level affordance prediction in Sec. 6.5.5. Finally, we discuss the limitation of our approach (Sec. 6.5.6), and show more qualitative results in Sec. 6.5.7. **Visual results are also included in the video.**

6.5.1 LayoutNet (Sec 6.1.1)

Layout parameters. As mentioned in Sec 6.1.1, we parameterize the layout as (x, y, a, b_1, b_2), where x, y is the location, a^2 is size, and b_1, b_2 are unnormalized approaching direction parameters. For training the LayoutNet, we obtain the ground truth parameters from off-the-shelf 2D hand prediction systems. The size and location comes from the predicted bounding box of a hand detector [187], which typically defines the hand region up to the wrist. The orientation is calculated from hand segmentation whose region is typically defined as the entire hand region, including hand and forearm. The approaching direction is calculated as the first principal component of a hand mask that centers on the location of the palm of the predicted hand.

We splat the layout parameters onto 2D via the spatial transformer network [89] that transforms a canonical mask template by a similarity transformation. The 2D similarity transformation is determined from the layout parameters. More formally,

$$T_l = \begin{pmatrix} sR & t \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} a^2\hat{b}_1 & -a^2\hat{b}_2 & x \\ a^2\hat{b}_2 & a^2\hat{b}_1 & y \\ 0 & 0 & 1 \end{pmatrix},$$

where $\hat{b}_1, \hat{b}_2$ is the normalized vector of b_1, b_2.

The lollipop-shape template in the canonical space is implemented with its circle being an isometric 2D Gaussian with a standard deviation of 1 and its rectangle being a 1D Gaussian with a standard deviation $\bar{s} = 2$. The width of the rectangle is calculated from the training data as the average ratio of the widths of forearms and palms.

DDPM loss on mask. In Equation 5.1, we write the DDPM loss in terms of reconstructing clean samples. In practice, we follow prior works [149, 169, 174] that reconstruct the added noise ϵ as

$$\mathcal{L}_{\text{DDPM}}^{\text{noise}} = \mathbb{E}_{x,\epsilon\sim\mathcal{N}(0,I),t}\|\epsilon - \epsilon_\theta(x_t, t)\|_2^2.$$

The estimated clean sample $\hat{l}_0$ is connected with the estimated noise by $\hat{l}_0 = \frac{1}{\sqrt{1-\bar{\alpha}_t}} l_t - \frac{\sqrt{\bar{\alpha}_t}}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta$, where $\alpha_t, \bar{\alpha}_t$ represent the noise schedule for each diffusion time step.

We train the LayoutNet with a weighted sum of the parameter loss $\mathcal{L}_{\text{para}}$ for esitmating the noise term ϵ, and a mask loss $\mathcal{L}_{\text{mask}}$ for estimating the clean sample term $\hat{l}_0$. The hyperparamter λ is set to 10.

Guided layout generation. LayoutNet inherits properties from diffusion models that can be guided to generate samples with additional constraints at test time. We follow Song *et al.* [194]. After each diffusion steps, we hijack the additional constraints with corresponding noise levels for the next diffusion step.

More specifically, instead of passing in the network's output $\boldsymbol{x}_t$ from the previous time step, we hijack it with $\boldsymbol{x}_t \leftarrow \tilde{\boldsymbol{x}}_t\boldsymbol{m} + \boldsymbol{x}_t(1 - \boldsymbol{m})$, where $\boldsymbol{m}$ is the indicator mask of the given condition $\tilde{\boldsymbol{x}}_0$. The unspecified constraints in $\tilde{\boldsymbol{x}}_0$ are set to 0. $\tilde{\boldsymbol{x}}_t$ represents the additional constraint with corresponding noise level, *i.e.* $\sqrt{1 - \bar{\alpha}_t}\tilde{\boldsymbol{x}}_0 + \sqrt{\bar{\alpha}_t}\boldsymbol{\epsilon}$.

6.5.2 ContentNet (Sec 6.1.2)

The goal of ContentNet is to generate high-resolution (256^2) realistic HOI images conditioned on the predicted layout and the input object image. We tried two different approaches commonly used in diffusion models [149, 174] as backbones for the ContentNet. One way (called ours/AffordDiff-LDM) is to follow Rombach *et al.* [175], as described in our main paper, that implements the ContentNet in the latent space where images of size 256^2 are compressed to 3-dimensional features of size 64^2 by a fixed pretrained autoencoder. The other way (called ours/AffordDiff-GLIDE) is to follow Nichol *et al.* [149] that uses a cascaded diffusion model that first generates images of size 64^2 and then upsamples them by a factor of 4.

All of the quantitative results in our main paper, including the user studies and all ablations, are based on Afford-LDM. AffordDiff-GLIDE is better in terms of contact recall $(90.8\%$ vs $87.1\%)$ while AffordDiff-LDM is significantly better in terms of FID score $(99.0$ vs $121.6)$. We find that AffordDiff-LDM generates less blurry results and the hand texture appears sharper and more realistic. In comparison, we find AffordDiff-GLIDE perceptually preferred because AffordDiff-GLIDE generates more realistic, though blurrier, finger articulations. The qualitative results in the main paper on EPIC-KITCHEN dataset (Fig 1 and Fig4 right in the main paper) show Afford-GLIDE. However, we provide the qualitative comparison of Afford-LDM with baselines in Fig 6.9 and Fig 6.10 of the appendix. We further provide a comparison of these two variants in Fig 6.15 of the appendix.

6.5.3 Constructing Paired Training Data (Section 6.1.3)

Cropping Details. We crop all objects with 80% squared padding before resizing such that objects (hands) appear in similar (different) sizes. The model learns the priors of their relative scales, *e.g.*, a hand to grasp a kettle appears much smaller than that of a mug (Fig 4).

We show that the proposed method to obtain pixel-aligned pairs of HOI and object-only images is robust and can also be applied to more cluttered images. When there is more than one hand in the HOI image, we randomly select one to remove. We show results of applying our data construction method on the HOI4D (Fig 6.9) and the EPIC-KITCHEN (Fig 6.10) datasets.

6.5.4 Baselines Implementation

Pix2Pix [88] (Sec4.1) We modify the official Pix2Pix implementation[1]. Given the predicted layout and the provided object image, we concatenate them channel-wise and pass them through 6 blocks of ResNet to output HOI images. The discriminator takes in the concatenation the of the object-only image, the splatted layout image, and generated HOI image and learns to discriminate between the real and fake domains. We tried batchnorm and instancenorm and found that batchnorm generated better results in general but has some black holes if the background statistics deviate from that of the training set.
VAE [101] (Sec4.1) VAE is notoriously known for being hard to balance for both generation variance and reconstruction quality. We sweep hyperparameters of the KL divergence loss's weights from $1, 1e{-}1, 1e{-}2, 1e{-}3, 1e{-}4$ and use $1e - 3$ as it produces the highest contact recall.
GANHand [32] (Sec4.2) GANHand is originally proposed both to predict 3D MANO hands for images of YCB objects [16] and to optimize physical plausibility with respect to the known or reconstructed 3D shapes of YCB objects. We compare our method with their sub-network for grasp prediction from RGB images (blue branch in their original paper, Fig 4). The sub-network takes in the object's identity, the desk plane equation and the object's center in 3D space, in addition to the object image. Since these are not available in the HOI4D dataset, we set them to zeros. We apply an additional reconstruction loss for 3D hand joints, MANO hand parameters and camera parameters. We finetune the network from the public checkpoints for another 10k iterations.

[1] https://github.com/junyanz/pytorch-CycleGAN-and-pix2pix

6.5.5 Scene Integration

We integrate our object-centric HOI synthesis to scene-level affordance prediction. We first detect the objects in the scene and then expand the detected bounding box's size with the same pad ratio (0.8 of the original object size). However, when the scene is crowded, the extended object crops may include other objects thus distracting the layout generation. We instead crop the object with the detected bounding box and pad the cropped object with boundary values. This allows the network to generate hand interaction only for the object of interest.

6.5.6 Limitation and Failure Cases

Although it is encouraging that the proposed model can perform zero-shot generalization to the EPIC-KITCHEN dataset, the proposed method inherits limited generalization capabilities from general learning-based algorithms. The proposed model will fail when the object image's appearance deviates too much from the training set, *e.g.* for too cluttered scenes, extreme lighting, very large objects (like a fridge) or very small objects (like a pin), *etc.*. The current model also cannot generate hands entering from the top of the frame or generate hands from a third-person's view due to the bias in the training set. These limitations require training with more diverse data. Additionally, the consistency of the hand's appearance and of the extracted hand poses can be further improved.

6.5.7 Qualitative Results

Fig 6.9 shows more examples of the constructed paired training data. We train all the models with a uniform mixture of inpainted and SDEdited object images.

Fig 6.10 shows that the proposed paired data construction is robust and can be applied to the EPIC-KITCHEN dataset.

Fig 6.11 shows more comparisons of the generated HOI images by the proposed method (LDM-version as reported in tables) and other image synthesis baselines [88, 101, 174] on the HOI4D dataset.

Fig 6.12 shows more comparisons of the generated HOI images by the proposed method (LDM-version as reported in tables) and other image synthesis baselines [88, 101, 174] on EPIC-KITCHEN dataset.

Fig 6.13 shows more comparisons of the extracted 3D hand pose obtained by the proposed method and other 3D affordance baselines [32, 174] on the

HOI4D dataset.

Fig 6.14 shows more comparisons of the extracted 3D hand pose obtained by the proposed method and other 3D affordance baselines [32, 174] on the EPIC-KITCHEN dataset.

Fig 6.15 shows an ablation study on comparison of the LDM and GLIDE version of our model on HOI4D and EPIC-KITCHEN datasets.

Fig 6.16 shows more layout editing results.

Fig 6.17 shows more results of heatmap-guided synthesis.

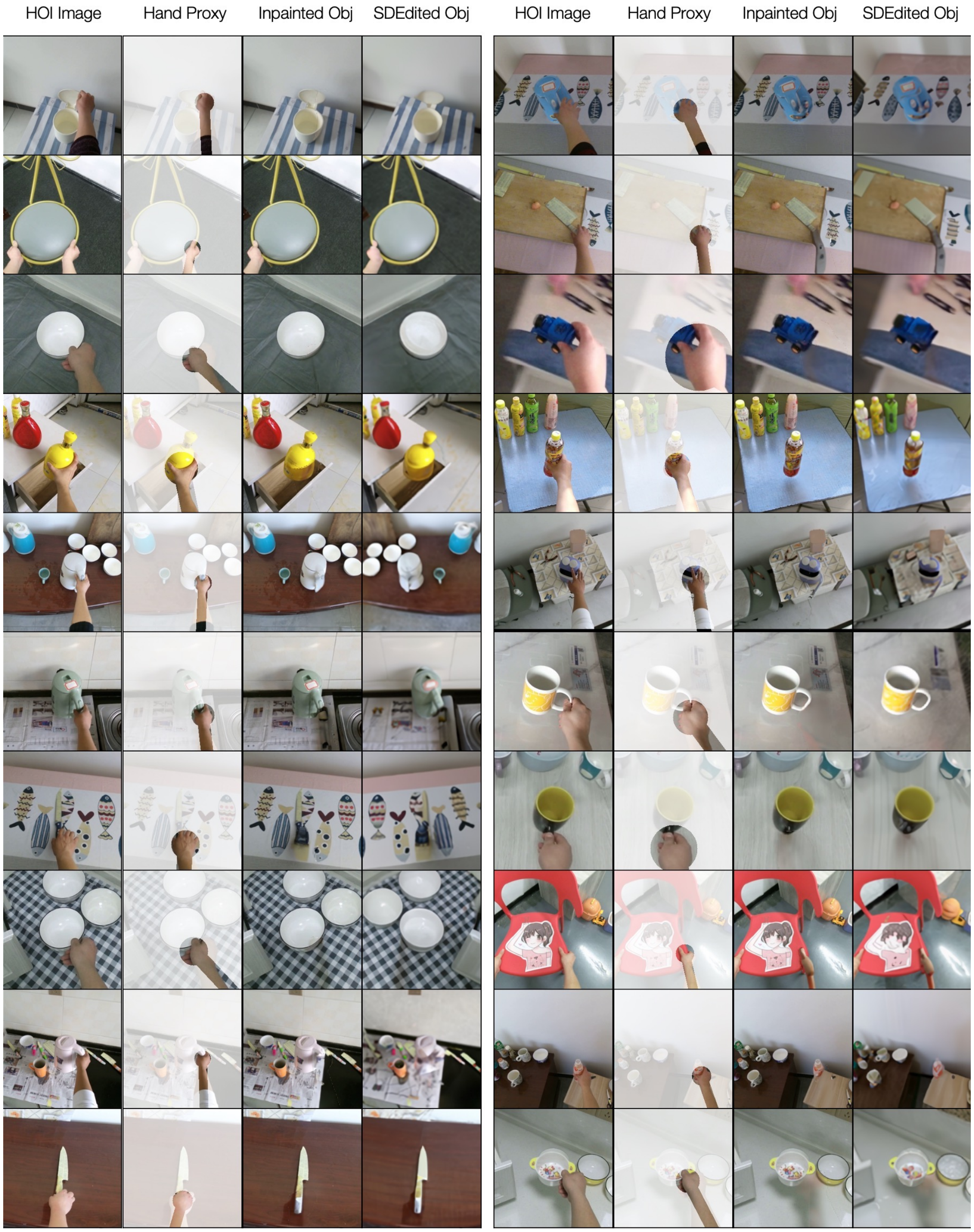


Figure 6.9: Visualizing more examples of the constructed paired training data. We train all the models with a mixture of inpainted and SDEdited object images.

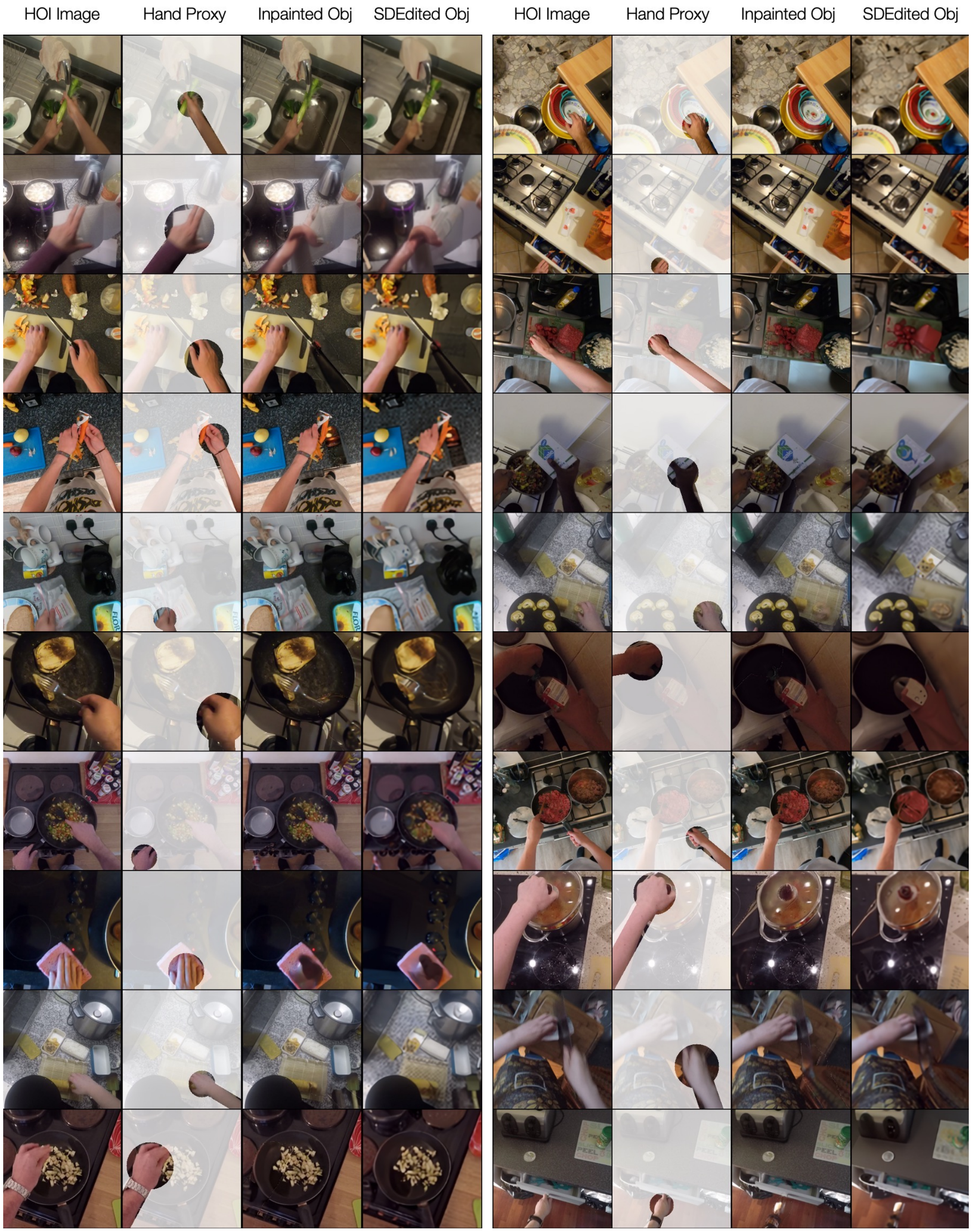


Figure 6.10: Visualizing the proposed paired data construction applied to EPIC-KITCHEN.

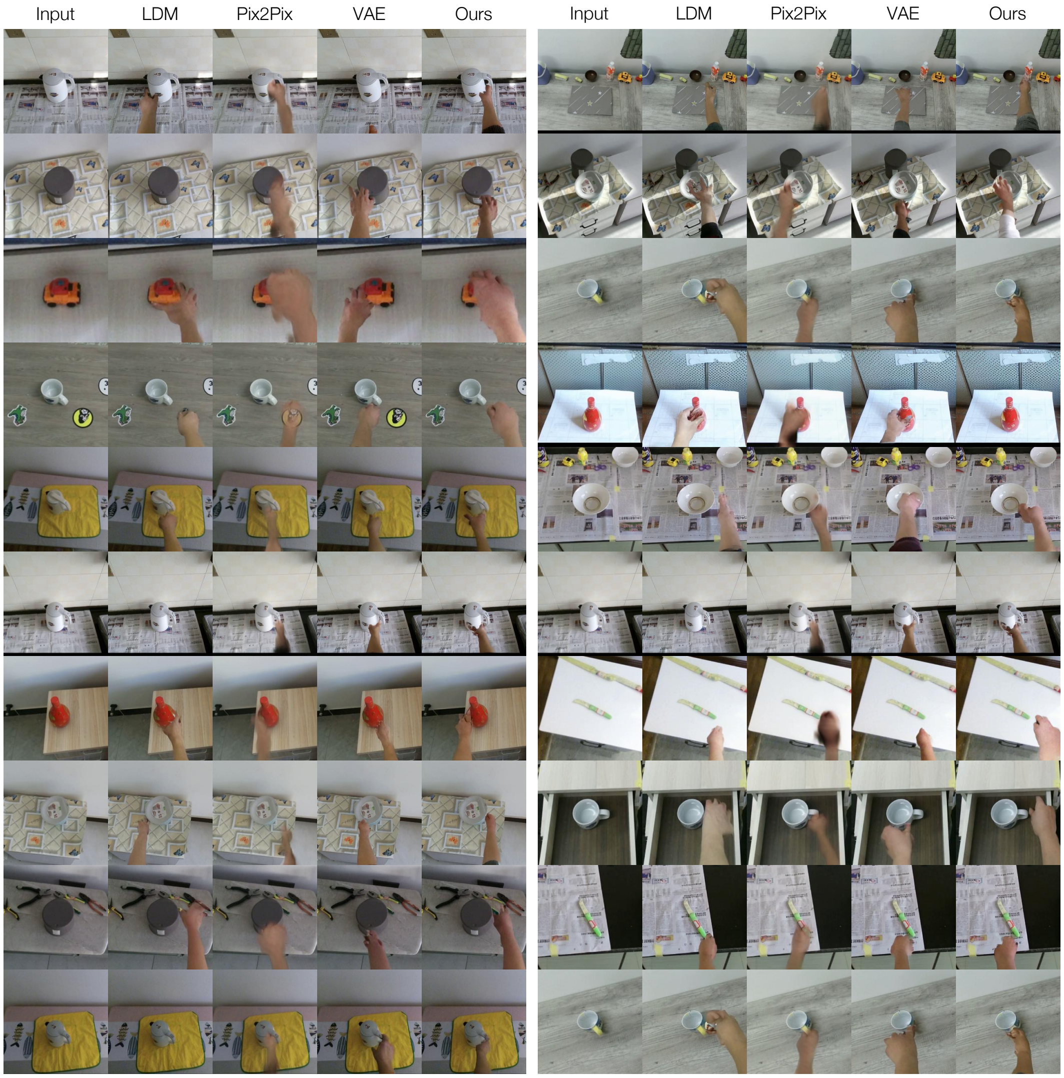


Figure 6.11: Visualizing more comparisons of the generated HOI images from the proposed method and other image synthesis baselines [88,101,174] on the HOI4D dataset.

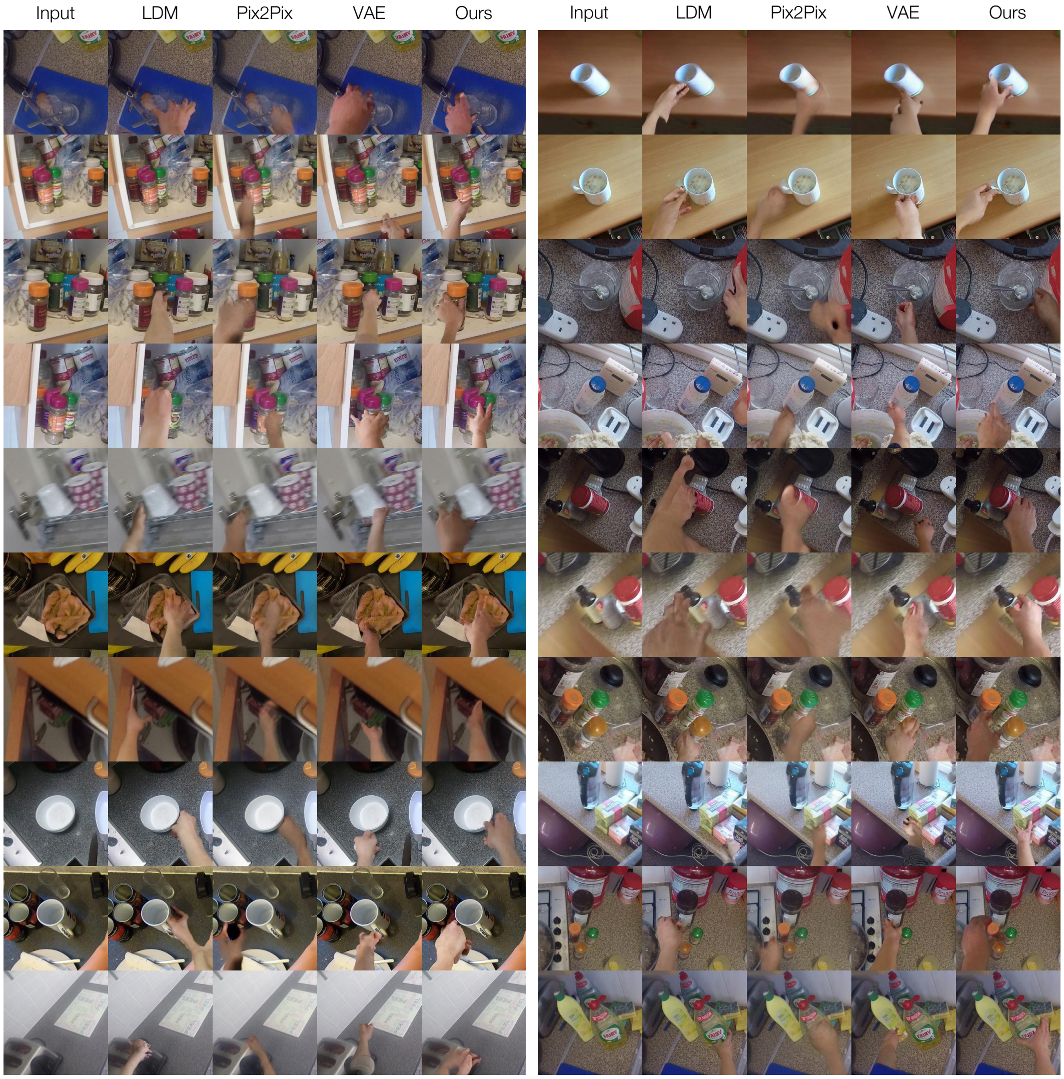


Figure 6.12: Visualizing more comparisons of the generated HOI images from the proposed method and other image synthesis baselines [88,101,174] on the EPIC-KITCHEN dataset.

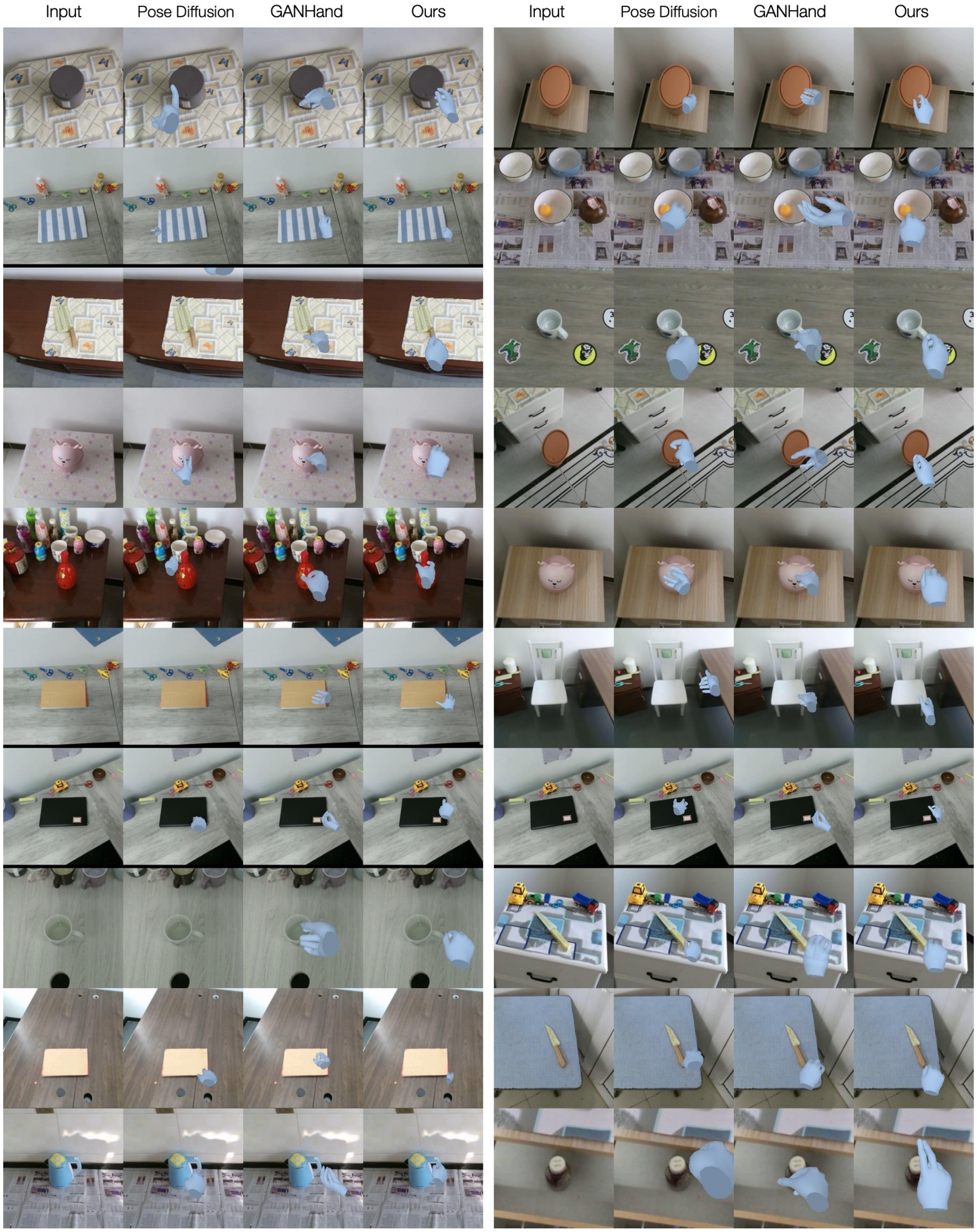


Figure 6.13: Visualizing more comparisons of the extracted 3D hand pose from the proposed method and other 3D affordance baselines [32, 174] on the HOI4D dataset.

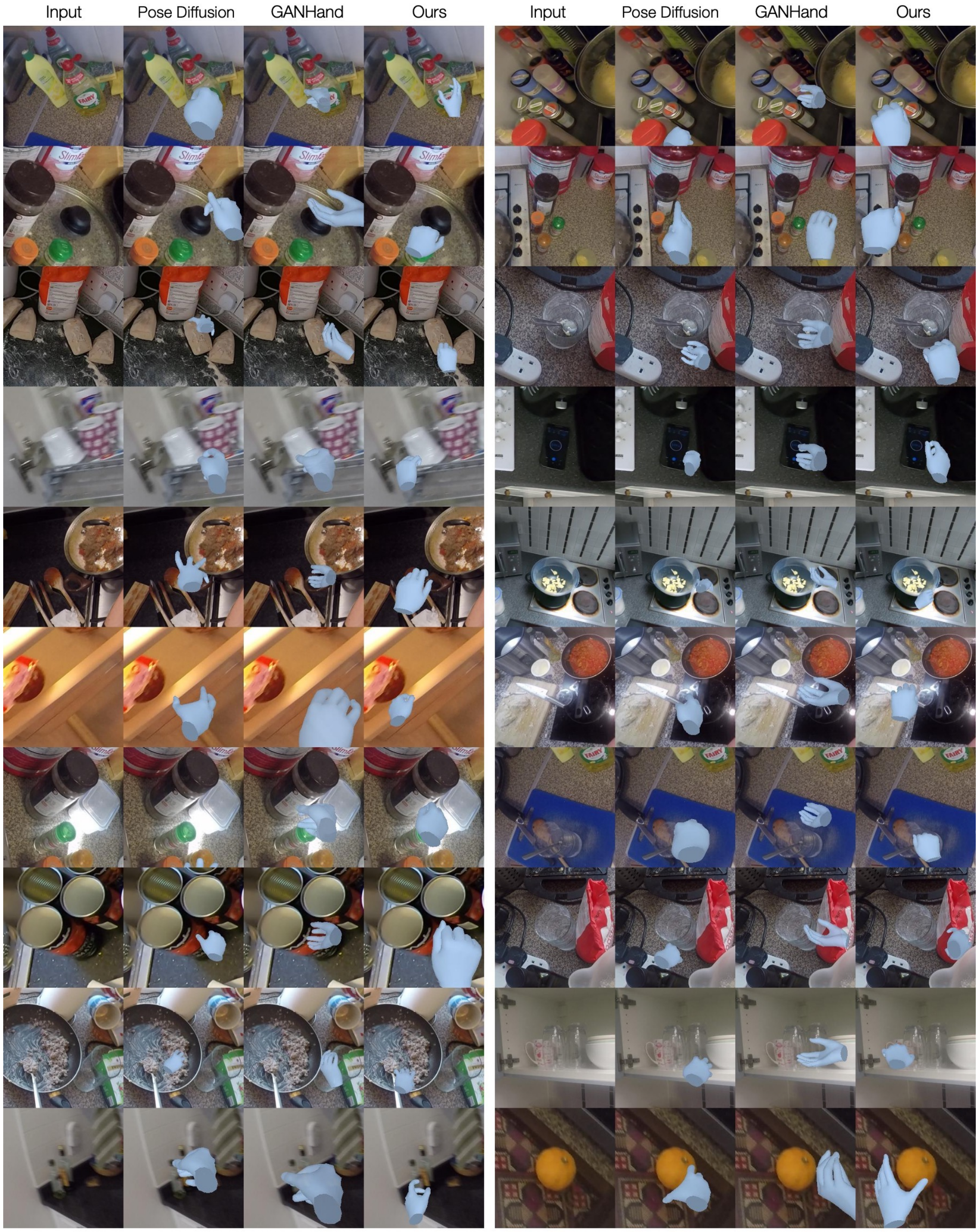


Figure 6.14: Visualizing more comparisons of the extracted 3D hand pose from the proposed method and other 3D affordance baselines on the EPIC-KITCHEN dataset.

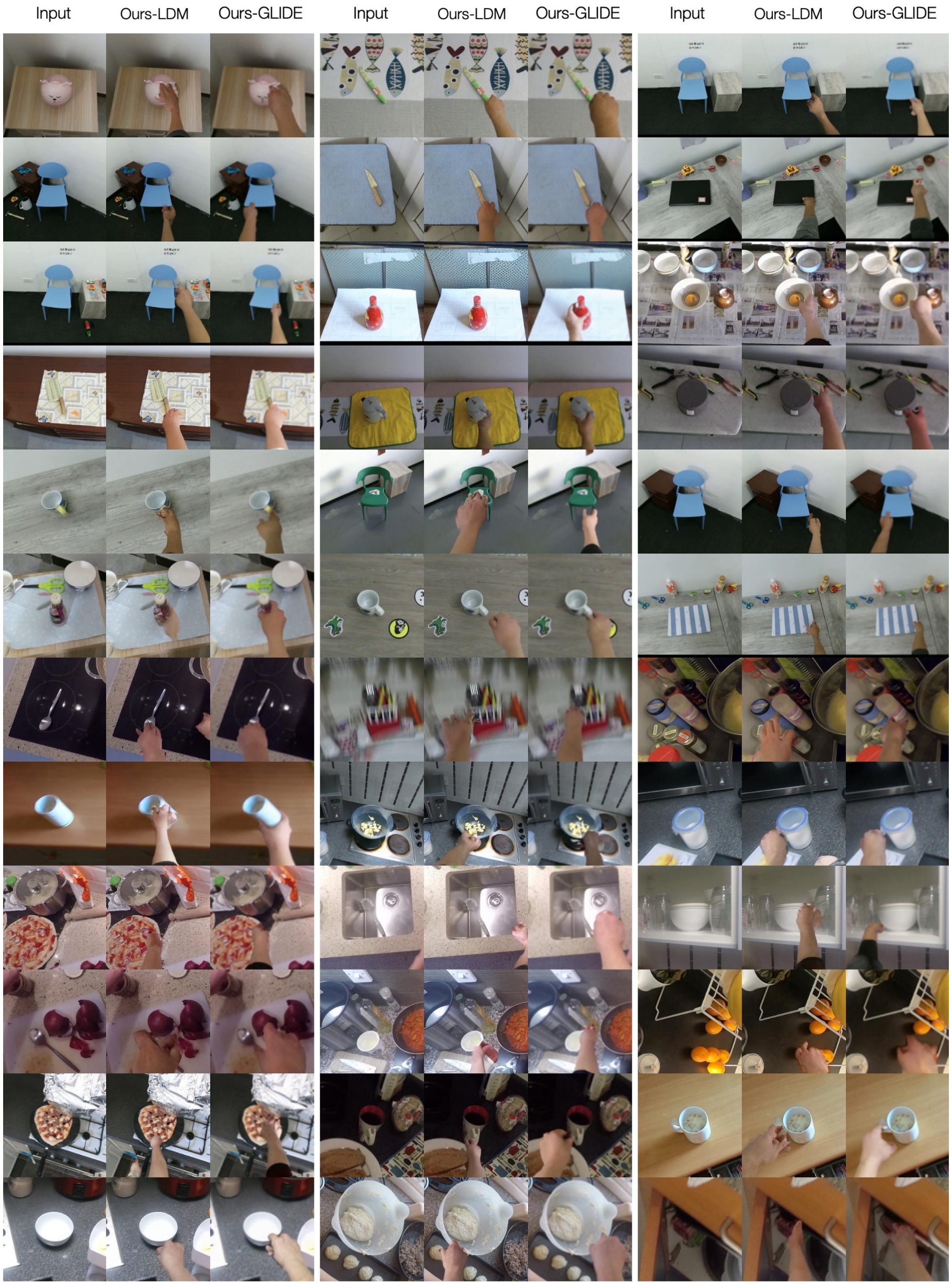


Figure 6.15: Visualizing the ablation of ContentNet for its LDM-based and GLIDE-based implementations (Sec 6.5.2).

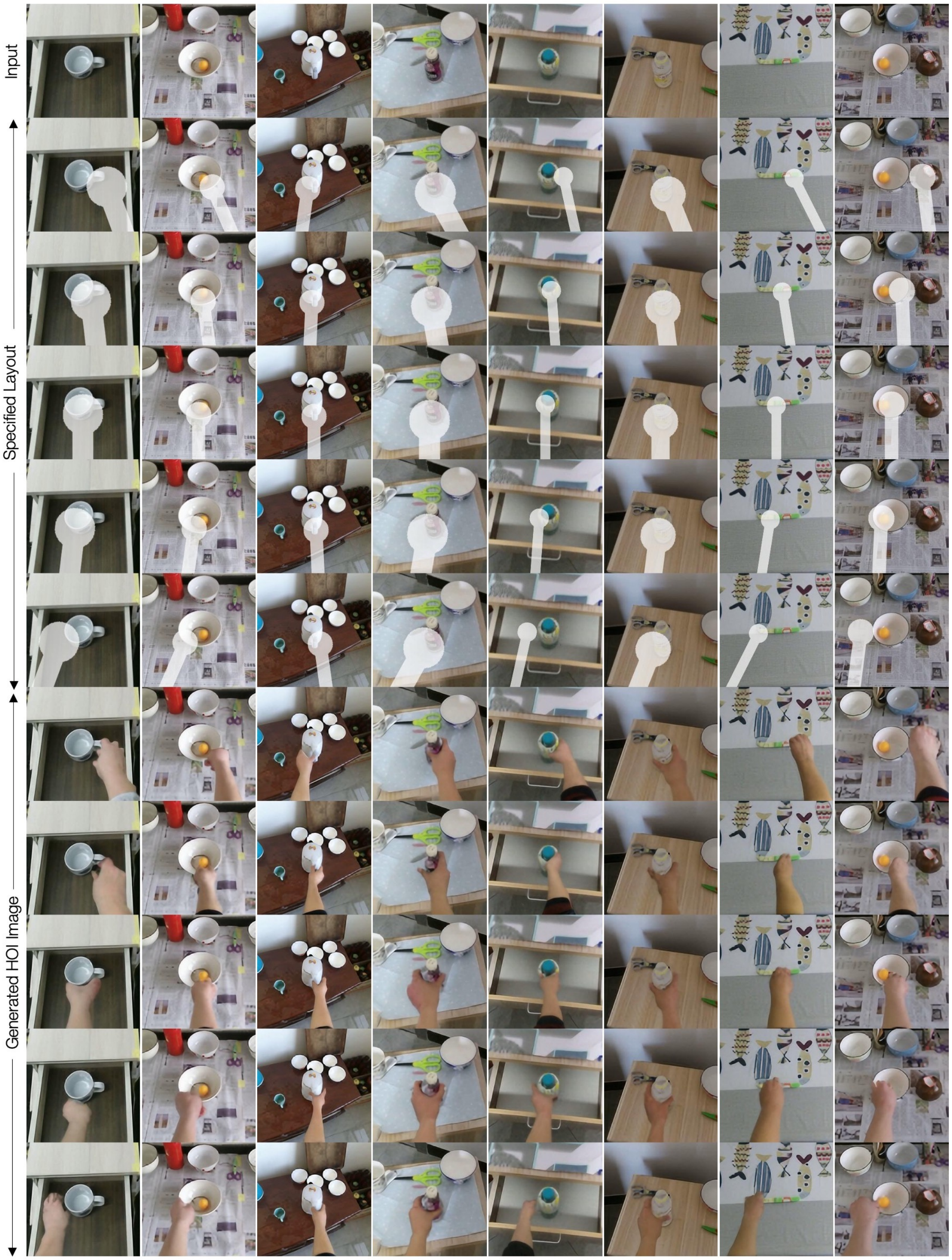


Figure 6.16: Visualizing more layout editing results.

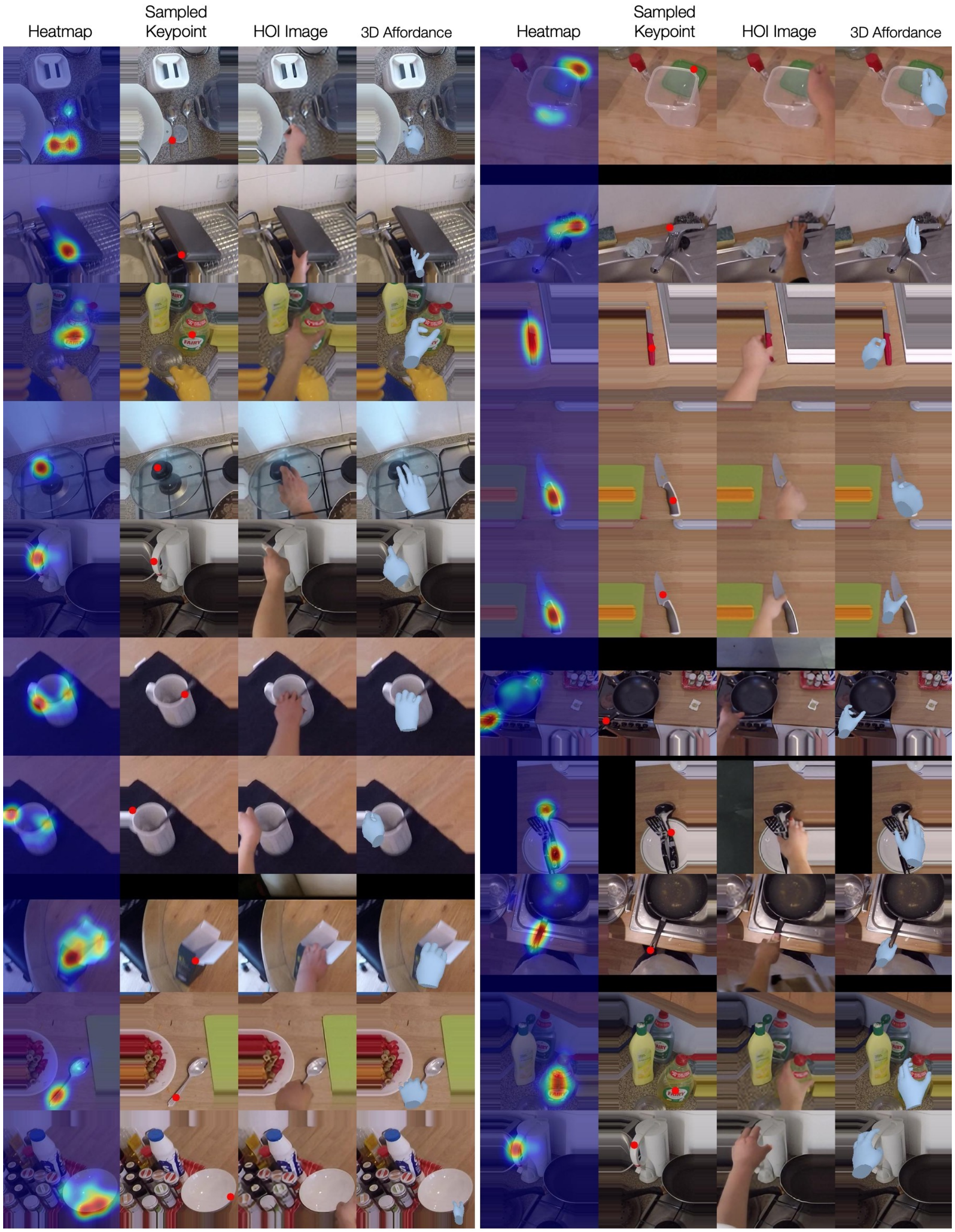


Figure 6.17: Visualizing more results of heatmap-guided synthesis.

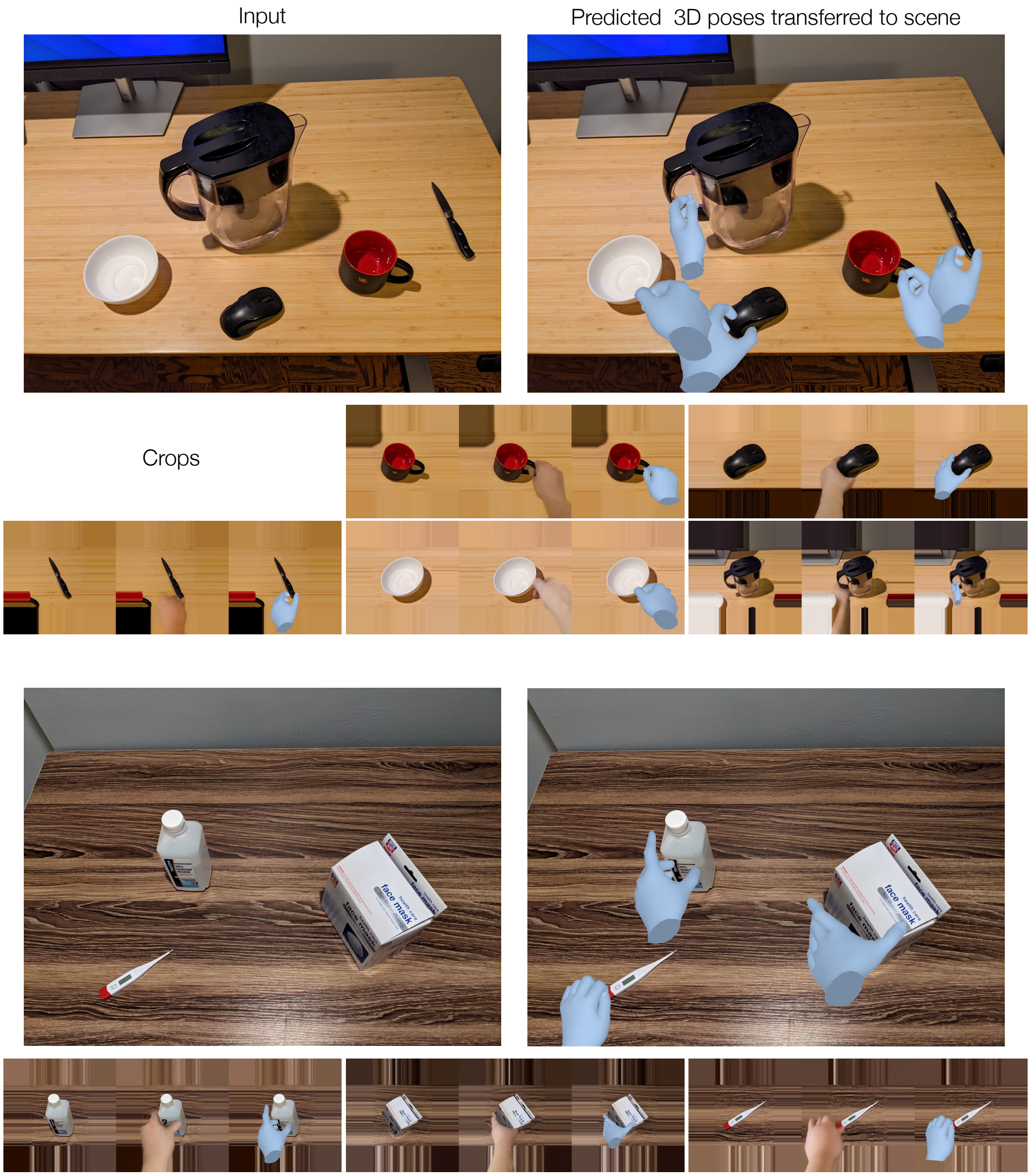


Figure 6.18: Visualizing more scene integration results with the individual prediction from crops.

Chapter 7

Generative Hand-Object Prior for Reconstruction and Prediction

While we have explored data-driven approach to the problem of reconstruction (Chapter 3, 4, 5) and prediction (Chapter 6) separately, in this chapter, we explore if a single generic data-driven prior can be used to aid both reconstruction and prediction. The insight is that knowing the general 3D structure of hand-object interaction can be beneficial for both tasks. For example, imagine holding a bottle, or a knife, or a pair of scissors. Not only can you picture the differing shapes of these objects *e.g.* a cylindrical bottle or a flat knife, but you can also easily envision the *varying* configurations your hand *would* adopt when interacting with each of them. Even though the form of these hand-object interactions may vary widely depending on factors such as geometry (*e.g.* we will hold a pen and a pan rather differently), or intent (*e.g.* passing a knife vs. using it to cut), we humans can effortlessly picture such interactions with everyday objects in our daily lives. In this chapter, our goal is to build a computational system that can similarly generate plausible hand-object configurations.

Specifically, we learn a denoising diffusion-based generative model that captures the joint distribution of both hand and object during interaction in 3D. Given a category-conditioned description *e.g.* 'a hand holding a plate', our generative model can synthesize both, plausible object shape as well as the relative configuration and articulation of the human hand (see Fig. 7.1 top). A key question we address is that what are good HOI *representations* for the model. While objects shapes are typically described via spatial (signed) distance fields, human hands are commonly modeled via a parametric mesh controlled by an articulation variable. Instead of modeling these disparate representations in our generative model, we propose a homogeneous HOI

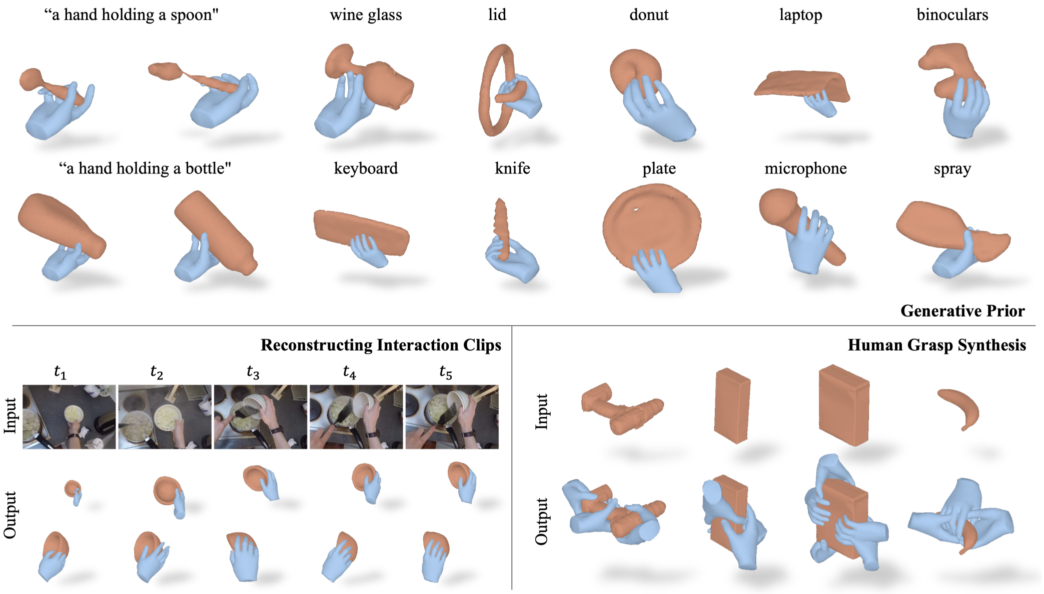


Figure 7.1: G-HOP can generate plausible hand-object interactions across a wide variety of objects (top). The learned generative prior can also guide inference for tasks such as reconstructing everyday interaction clips and synthesizing human grasps given object meshes.

representation and show that this allows learning a 3D diffusion model that jointly generates the hand and object.

In addition to enabling synthesis of diverse plausible hand and object shapes, our diffusion model can also serve as a generic prior to aid inference across tasks where such a representation is a desired output. For example, the ability to reconstruct or predict interactions is of central importance for robots aiming to learn from humans, or virtual assistant trying to aid them. We consider two well-studied tasks along these lines: i) reconstructing 3D hand-object shapes from everyday interaction clips, and ii) synthesizing plausible human grasps given an arbitrary object mesh. To leverage the learned generative model as a prior for inference, we note that our diffusion model allows computing the (approximate) log-likelihood gradient given any hand-object configuration. We incorporate this in an optimization framework that combines the prior likelihood-based guidance with task-specific objectives (*e.g.* video reprojection error for reconstruction) or constraints (*e.g.* known object mesh for synthesis) for inference.

While understanding hand-object interactions is an increasingly popular

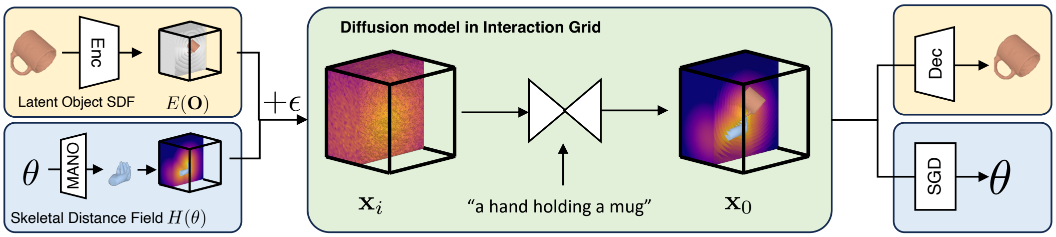


Figure 7.2: **Method Overview of Generative Hand-Object Prior:** Hand-object interactions are represented as interaction grids within the diffusion model. This interaction grid concatenates the (latent) signed distance field for object and skeletal distance field for the hand. Given a noisy interaction grid and a text prompt, our diffusion model predicts a denoised grid. To extract 3D shape of HOI from the interaction grid, we use decoder to decode object latent code and run gradient descent on hand field to extract hand pose parameters.

research area, real-world datasets capturing such interactions in 3D are still sparse. We therefore aggregate 7 diverse real-world interaction datasets resulting in long-tailed collection of interactions across 157 object categories, and train a shared model across these. To the best of our knowledge, our work represents the first such generative model that can jointly generate both, the hand and object, and we show that it allows synthesizing diverse hand-object interactions across categories. Moreover, we also empirically evaluate the prior-guided inference for the tasks of video-based reconstruction and human grasp synthesis, and find that our learned prior can help accomplish both these tasks, and even improve over task-specific state-of-the-art methods.

7.1 Method

We first seek to model the joint distribution of the geometry of hand-object interactions $p(\mathbf{O}, \mathbf{H}|\mathbf{C})$ where $\mathbf{C}$ is the text of an object category. We use a diffusion model Ψ (Section 5.1) to learn this generative prior, and propose a spatial interaction grid representation for learning (Sec. 7.1.1). We then apply this learned prior to guide reconstruction from monocular video clips and human grasp synthesis (Sec. 7.1.2). For both tasks, we frame inference as test-time optimization that combines task-specific constraints/objectives with score "distillation" from the pre-trained diffusion model.

7.1.1 Generative Hand-Object Prior

In this work, we propose 'interaction grids' as a homogeneous HOI representation that allows the diffusion models to effectively reason about the 3D hand-object interactions. Specifically, an interaction grid (Fig. 7.2) is a concatenation of a latent signed distance value grid representing the object $E(\mathbf{O})$ and a 'skeletal distance' field based grid parameterized by 3D hand pose $H(\boldsymbol{\theta})$, *i.e.* $\mathbf{x} \equiv (E(\mathbf{O}), H(\boldsymbol{\theta}))$. We model the interaction grid in a normalized hand-centric frame, where the hand palm always faces upwards. The hand-centric frame more effectively captures the inherent structures of interaction common to various objects, such as grasping handles, regardless of whether the object is a kettle or a power drill [232].

Latent Object Signed Distance Field. We use a signed distance field (SDF) grid to capture object details. As the memory grows cubically with grid resolution, we follow prior works to use a VQ-VAE [208] to compress high-resolution SDF grids into lower-dimension object latent. $\mathbf{z} = E(\mathbf{O}), \mathbf{O} = D(\mathbf{z})$. Note that when training the autoencoder, the object SDF grids are also transformed into hand-centric frame.

Skeletal distance field for Parametric Hand. While there is consensus on how to represent objects, it is unclear what is a good representation of hand during interaction. Many prior works generate hand/human shape by diffusing in the compact pose parameter space [95,202] but we find this space not ideal when we diffuse it jointly with objects latent grids (see supplementary) probably because the diffusion model cannot easily to reason about spatial interactions using this heterogeneous representation (1D articulation vector and 3D SDF grid). Instead, we propose to represent hand in a pose-parameterized distance field $H(\boldsymbol{\theta})$. It is a 15-channel 3D grid that encodes the distance to each joint. $H(\boldsymbol{\theta})[u, v, w]_{i=1:15} \equiv \|\mathbf{X}_{[u,v,w]} - J_i\|_2^2$. This skeletal distance field can be converted from pose parameter space and vice versa by leveraging differentiable parametric mesh model MANO [178]. Specifically, MANO takes in the pose parameter and outputs joint position $J_i(\boldsymbol{\theta})$ to compute the skeletal field. To recover pose parameter $\boldsymbol{\theta}$ from a skeletal distance field, we run gradient decent on pose parameter to minimize the distance between the induced field and the given field, $\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}}(H(\boldsymbol{\theta}) - \hat{H}) + w\|\boldsymbol{\theta}\|_2^2$.

Denoising Diffusion Model. In training, the diffusion model takes in a text embedding and a noisy 3D interaction grid $\mathbf{x}_i$ and is supervised to restore the clean grid $\hat{\mathbf{x}}_0$.

$$\mathcal{L}_{\text{DDPM}}[\mathbf{x}; \mathbf{C}] = \mathbb{E}_{i,\epsilon\sim\mathcal{N}(\mathbf{0},\mathbf{I})} w_i \|\hat{\mathbf{x}}_0 - \Psi(\mathbf{x}_i, i, \mathbf{C})\|_2^2 \tag{7.1}$$

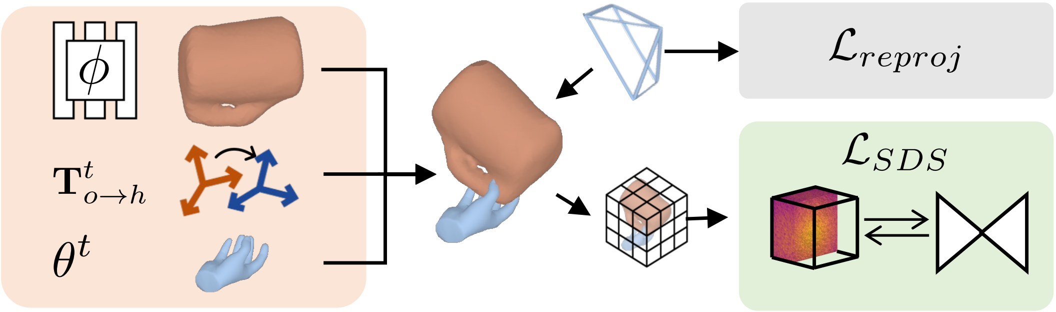


Figure 7.3: **Reconstructing Interaction Clips:** We parameterize HOI scene as object implicit field, hand pose, and their relative transformation (left). The scene parameters are optimized with respect to the SDS loss on extracted interaction grid and reprojection loss (right).

The object distance field is in resolution 64^3 and the VQ-VAE downsamples the resolution to 16^3 which is then concatenated with the hand skeletal field. We implement the diffusion model as 3D-UNet with three 3D convolution blocks. The text prompt is encoded by CLIP [168] text encoder and is passed to the 3D-UNet by cross-attention at each block.

7.1.2 Prior-guided Reconstruction and Generation

Given the learned generative prior, we leverage it for both HOI reconstruction and human grasp synthesis. The inference in both tasks is performed via test-time optimization which is guided by distilling the learned prior. We use score distillation sampling (SDS [161,213]) to approximate log-probability gradients of interaction grids $\mathbf{x}$ from the diffusion model. Specifically, to guide the grid to be more plausible at every optimization step, we corrupt the current interaction grid $\mathbf{x}$ by a certain amount of noise and let diffusion model denoise it. The discrepancy between this denoised prediction and the current estimate can be be used an objective to obtain log-likelihood gradients:

$$\nabla_{\mathbf{x}} \log p(\mathbf{x}) \approx \nabla_{\mathbf{x}} L_{SDS}[\mathbf{x}] = \mathbb{E}_{\epsilon,i}[w_i(\mathbf{x} - \hat{\mathbf{x}}_i)] \tag{7.2}$$

In the following section, we will show that both reconstruction and grasp synthesis can leverage the common optimization frameworks by instantiating task-specific parameters and constraints.

Reconstructing Interaction Clips

Given a video clip depicting a hand interacting with a rigid object, we aim to reconstruct the underlying 3D shape of the hand and the object. We follow DiffHOI [233] which performs inference via a optimizing 3D scene representation with respect to a reprojection term and a data-driven prior term. Instead of their 2D diffusion prior which can only guide object shape inference, we substitute our learned joint 3D generative prior and show that it leads to improved performance for video-based reconstruction.

Scene Parameters and Rendering. We adopt a similar representation as DiffHOI [233], which decomposes the HOI scene into three parts: i) a time-persistent object signed distance field represented by an implicit neural network $\phi(\cdot)$; ii) time-varying hand pose parameters $\boldsymbol{\theta}^t$, and iii) the relative poses $\mathbf{T}^t_{o\to h}$ between them. This scene representation can be rendered into 2D masks $\mathbf{I}^t$ by differentiably compositing renderings of the volumetric object and hand mesh.

Prior-Guided Reconstruction. Different from DiffHOI, our data-driven prior is in 3D space instead of 2D. Furthermore, our prior also models the hand pose rather than use it as a condition, and can thus also provide gradients to guide hand pose optimization. Specifically, to regularize the 3D representation, we query the 3D volume in the hand-centric frame to get interaction grid for each frame and pass the grid to the pre-trained diffusion model, *i.e.* $\mathbf{x}^t = (E(\phi(\mathbf{T}^{t-1}_{o\to h} X_{grid})), H(\boldsymbol{\theta}^t))$, where X_{grid} is the coordinate of the queried volume. Other losses are similar to [233]: the reprojection term is computed in the mask space $\mathcal{L}_{reproj} = \|\mathbf{I}^t - \hat{\mathbf{I}^t}\|$; other regularization include Eikonal loss and temporal smoothness.

The optimization converges faster than previous work, perhaps because the prior in 3D provides stronger supervision. Specifically, we optimize 15000 iterations for each video clips which takes about an hour (which is 85% faster than DiffHOI [233]).

Synthesizing Plausible Human Grasps

Given an object mesh M_o, we aim to synthesize human grasps for the object. Formally, this corresponds to sampling from the conditional distribution $p(\mathbf{H}|\mathbf{O}, \mathbf{C})$. While our diffusion model captures the joint distribution of hand and object, it does not allow sampling human grasp directly given an object. Instead, we obtain plausible grasps via a test-time optimization approach to seek grasping modes while constraining the object to match the

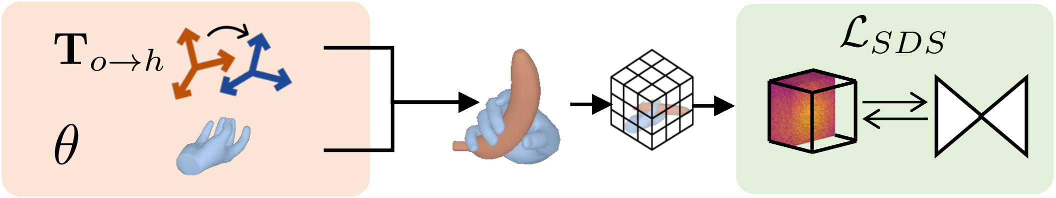


Figure 7.4: **Grasp Synthesis:** We parameterize human grasps via hand articulation parameters and the relative hand-object transformation (left). These are optimized with respect to SDS loss by converting grasp (and known shape) to interaction grid (right).

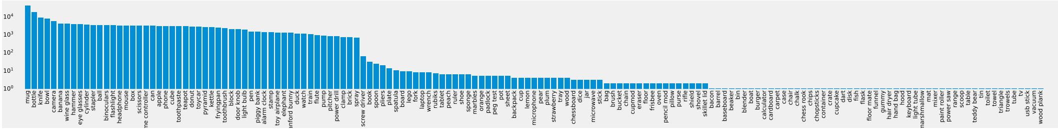


Figure 7.5: **Dataset Statistics:** number of training samples for each category when training our generative prior. Zoom in for better view.

input. We also provide a mechanism to rank the generation by measuring consensus between diffusion model and the grasp synthesis.

Grasp Parameters. We parameterize a human grasp by the relative pose of the hand with respect to the object $\mathbf{T}_{o \to h}$, along with its articulation $\boldsymbol{\theta}$. We initialize hand articulation to a mean configuration while initializing relative pose with a random orientation and translation.

Optimization. In order to use diffusion model to guide grasp synthesis, we first convert the object mesh into SDF grid G_o, which is then transformed

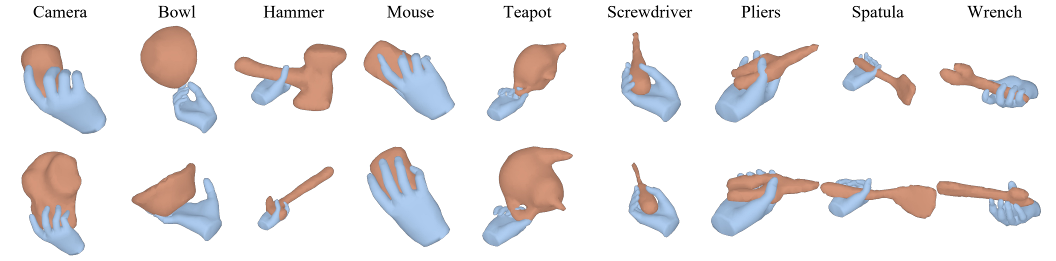


Figure 7.6: **Generations from Generative Hand-Object Prior:** Given a text prompt (only showing class label), we visualize two generated interactions from G-HOP . Categories are sorted from most common to least common in training (left to right). Generations are diverse in terms of object shape such as teapots, hand articulation such as mouse, and use intent like hammer.

from the object-centric to the diffusion model coordinate (hand-centric) by the relative pose $\mathbf{T}_{o\to h}$, *i.e.* $\mathbf{x} = (E(\mathbf{T}_{o\to h}G_o), H(\boldsymbol{\theta}))$. We optimize the relative pose along with hand articulation for 500 iterations by maximizing the interaction likelihood from Eq. 7.2, *i.e.* $\log p(\boldsymbol{x}(\boldsymbol{T}_{o\to h}, \boldsymbol{\theta}))$. To account for accuracy loss when converted to low-resolution grids, we refine the predicted hand with the original mesh to encourage surface contact and penalize mesh collision. We show in supplementary that the distillation provides a good initialization for the mesh refinement while surface refinement further improves contact and grasp stability.

Ranking Grasps. The proposed approach to grasp synthesis is stochastic due to different initialization and the stochastic distillation process. Thus diverse grasps can be sampled. Furthermore, many applications like robotic manipulation would also want to know how plausible each grasp is. We also propose a mechanism to evaluate the sampled grasp. We approximate the likelihood upper bound [81] by averaging SDS loss across different time steps i:

$$s(\theta, \mathbf{T}_{o\to h}) = -\sum_{i=1}^{T} w_i \|\mathbf{x}(\boldsymbol{\theta}, \mathbf{T}_{o\to h})) - \hat{\mathbf{x}}_i(\epsilon)\|_2^2 \tag{7.3}$$

Intuitively, this measures the agreement between the prediction and the denoised output from the diffusion model, which indicates the distance of the current grasp to a plausible mode. We observe that this score provides a consistent and meaningful ranking across different samples.

7.2 Experiments

We train the generative prior on a collection of HOI datasets. We first show data distribution on this dataset collection and then visualize samples from the learned generative prior (Sec 7.2.1). In Sec. 7.2.2, we show that the learned prior benefits the task of reconstructing interaction clips. Our method outperforms other reconstruction baselines on HOI4D and we also show reconstruction of in-the-wild videos. In Sec. 7.2.3, we evaluate human grasps that are synthesized by directly applying our learned prior. We compare G-HOP to other baselines on two datasets and conduct user study to show that human grasp synthesized by ours is the most preferred one.

Training Data. We train our diffusion model on a combination of several world datasets including [13,24,33,127,199,229], using their annotated 3D meshes of hand and objects. The name of categories across datasets are

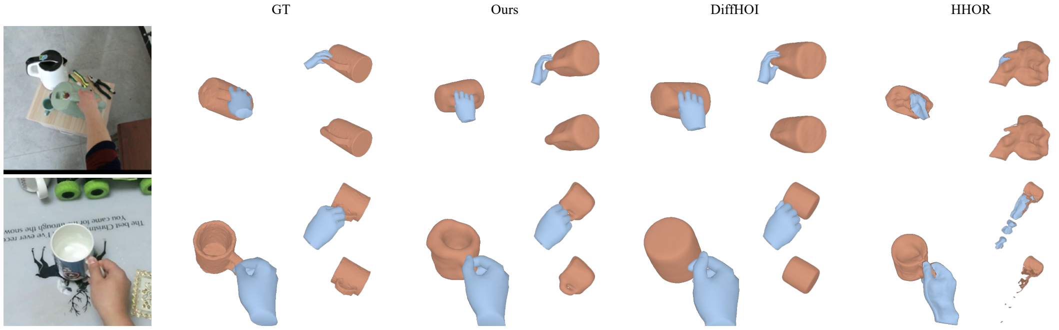


Figure 7.7: **Qualitative Evaluation on HOI4D:** We show reconstruction by G-HOP and two other video reconstruction baselines [85, 233] in the image frame (left) and from another view with (top right) or without (bottom right) reconstructed hand. Please see our project page for reconstruction videos from all methods.

not standardized so we manually map synonyms or different formats to the same word (*e.g.* cellphone, iphone → phone, doorknob, door_knob → door knob). In total, we reduce 362 different words to 155 classes. All training data were converted into SDF grids, in hand-centric frame, with a resolution of 64^3 and spanning 30cm in all directions.

7.2.1 Visualizing Data-Driven Prior

We visualize the number of training samples per class in Fig. 6.3. The data is extremely unbalanced and follows a long-tail distribution. Classes with most training samples like mug consist of more than 10k grasps while few-shot classes such as skillet lid consist of fewer than 100 grasps.

In Fig. 7.6, we visualize hand-object interactions generated from the learned generative prior. We show 3 samples in different rows for each class. The classes from left to right are sorted by the training size from more to less. We see that the generated objects vary in shape. For example, different cameras display various lengths of lens. The generated samples are also diverse in terms of ways to hold them. Some hammers are held by handles and some are held by heads (for hand-over). We also find that the model can generate diverse and plausible samples on few-shot classes (shown on the right side).

7.2.2 Reconstructing Interaction clips

Setup and Evaluation Metrics. We evaluate interaction reconstruction on the HOI4D dataset. HOI4D is an egocentric dataset recording people inter-

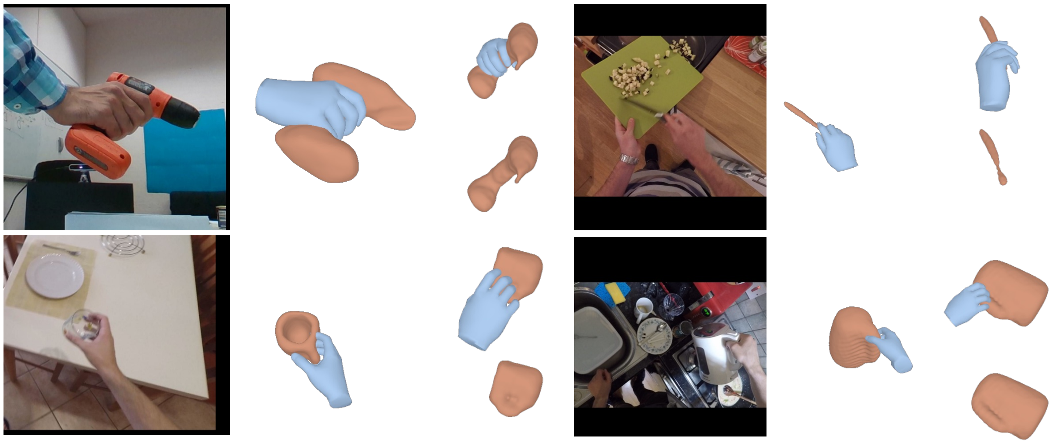


Figure 7.8: **In-the-Wild Reconstruction:** reconstruction on interaction clips from novel datasets [35, 68].

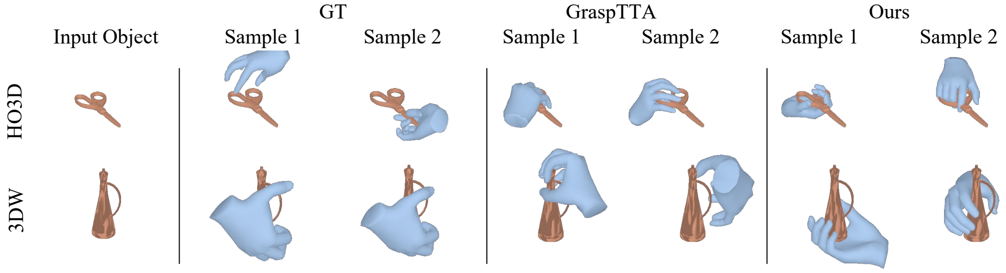


Figure 7.9: **Visualizing Grasp Generations:** Given an object mesh (left) from HO3D or ObMan, we sample two grasps from each method.

acting with different objects. We use the same split as DiffHOI [233] that consists of 2 video clips for all portable rigid object categories. The objects in the test set are held out from the train set. We evaluate three aspects of the output: object reconstruction error, hand reconstruction error (MPJPE, AUC), and hand-object alignment (CD$_h$). Following prior works [85, 233], we align the object reconstruction with the ground truth by scaled Iterative Closest Points (ICP) and report F-score at 5mm, 10mm, and Chamfer distance in the aligned space. To evaluate the relation between hand and object, we report Chamfer distance of objects in hand-centric frame $CD_h \equiv CD(\mathbf{T}^t_{o\rightarrow h}O, \hat{\mathbf{T}}^t_{o\rightarrow h}\hat{O})$.

Baselines. We compare with three other template-free baselines that tackle reconstruction from casual monocular interaction clips.

i) *iHOI* [232] is a single-view 3D reconstruction method that learns to map from image feature and hand articulation to in-hand object shape. The model is finetuned on HOI4D and reconstruction is evaluated per-video frame.

Table 7.1: **Comparing HOI reconstruction:** object error (F@5mm, F@10mm, CD), hand-object alignment CD$_h$, and hand error (MPJPE, AUC) on HOI4D. We compare G-HOP with baselines and also ablate if reconstruction benefits from priors in the 3D space or from joint modeling hand and object.

	Object Error			Align	Hand Error	
	F5↑	F10↑	CD↓	CD$_h$ ↓	MPJPE↓	AUC↑
iHOI [232]	0.42	0.70	2.7	27.1	1.19	0.76
HHOR [85]	0.31	0.55	4.7	165.4	-	-
DiffHOI [233]	0.62	0.91	0.8	48.7	1.12	0.78
G-HOP	**0.76**	**0.97**	**0.4**	**18.4**	**1.05**	**0.79**
G-HOP(Cond)	0.66	0.92	0.7	19.3	1.14	0.77

ii) *HHOR* [85] optimizes a hand-object field with respect to the input video without any data-driven prior.
iii) *DiffHOI* [233] is closest to our work. The main difference is that the prior in their work takes hand pose as input thus modeling the *conditonal* probability $p(\pi(O)|\pi(H), C)$. Additionally, their prior is an image-based diffusion model instead of a 3D diffusion model.
iv) *G-HOP* (*Cond*) is our ablated models that is conditioned on hand pose and text prompt (same as DiffHOI but with 3D backbone). It aims to disentangle the effect of upgrading the prior from 2D to 3D from modeling joint instead of conditional probability.
For fair comparison, our diffusion model for HOI4D evaluation only trains on HOI4D train split. All other experiments use the model trained on all datasets.

Results. We visualize reconstructions from different methods in Fig. 7.7 in the image frame and from a novel viewpoint. HHOR, which does not leverage data-driven learning, struggles with unobserved regions and outputs degenerate solutions as shown from the novel view. While iHOI reconstructs better shapes for each frame, there are not temporally consistent (shown in supplementary video) and it cannot benefit from multi-view cues. In comparison, DiffHOI reconstructs temporally consistent and more realistic results, but the reconstructed shape is relatively coarse. For instance, the kettle handle is merely a bump on top of a cylinder and the reconstruction does not reflect the concavity of the mug. In contrast, the reconstruction from

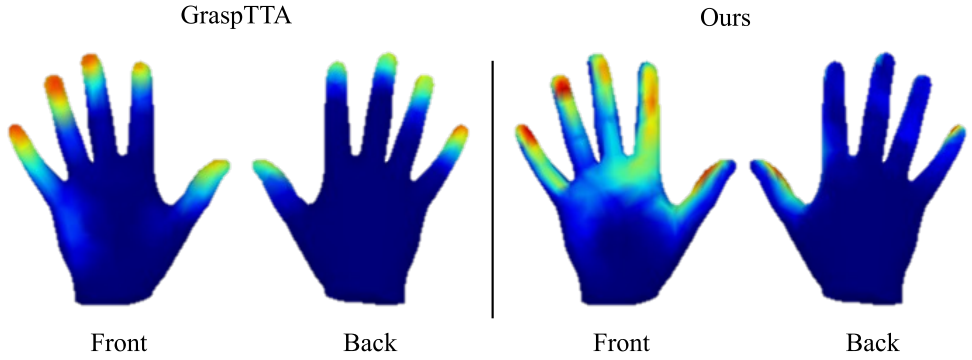


Figure 7.10: **Contact Map on Hand:** We visualize contact probability on hand over all generated samples from G-HOP and GraspTTA [90] on the HO3D dataset.

G-HOP captures more details of object shape. In the bottom row, it even captures the space between the handle and the cup body. The visualization is consistent with the quantitative results in Tab. 7.1. Furthermore, we also find that the hand pose reconstruction also improves since the prior in G-HOP can also guide hand pose as well.

Ablations. Comparing with the ablated 3D conditional model (Tab. 7.1), we find that upgrading 2D prior to 3D improves object reconstruction significantly but does not improve hand reconstruction much. Joint modeling leads to better hand pose, which can in return improve object reconstruction further. Interestingly, we also find that the variant that *jointly* models HOI in *image* space performs even worse than DiffHOI. See appendix (2D joint prior) for further discussion.

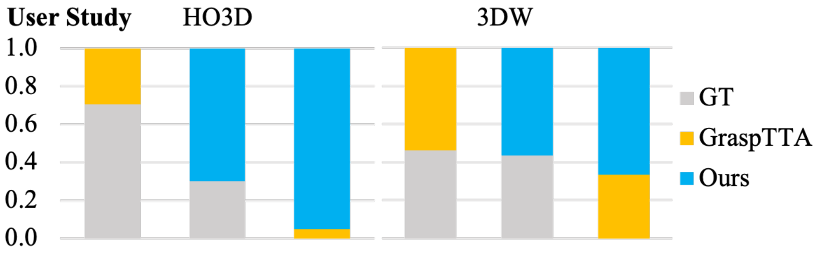


Figure 7.11: **Comparison with Baselines:** Preference percentages from users for pairwise method comparison on HO3D and ObMan.

Table 7.2: **Comparison with Baselines:** We compare our synthesized human grasps against GraspTTA [90] and annotated grasps provided by datasets (GT) on HO3D and ObMan. We report table the intersection between meshes, displacement distance in simulation, and hand contact ratio and area.

		Intersection			Disp.	Contact	
		maxD$\downarrow$	avgD$\downarrow$	vol$\downarrow$	avg $\downarrow$	ratio$\uparrow$	area$\uparrow$
HO3D	GT	**1.32**	0.37	6.16	2.32	0.95	0.15
	GraspTTA	2.44	0.61	**5.25**	2.89	1.00	0.23
	G-HOP	1.84	**0.31**	11.46	**0.95**	**1.00**	**0.23**
ObMan	GT*	0.98	0.74	**1.70**	1.57	1.00	0.12
	GraspTTA	0.87	0.58	5.56	**1.54**	**1.00**	0.18
	G-HOP	**0.74**	**0.51**	17.40	1.85	0.93	**0.25**

7.2.3 Synthesizing Plausible Grasps

Setup and Evaluation Metrics. We evaluate human grasp synthesis on two datasets [68,73]. HO3D is a real-world HOI dataset whose objects come from the YCB dataset [15], which has appeared in our training data. To test the generalization ability to novel objects, we also evaluate on a subset of 3D Warehouse used in Hasson *et al.* [73] (ObMan). It is a synthetic dataset that our prior has never seen in training. Following prior work [90,96], we evaluate grasp quality by 1) the amount of intersection between hands and objects (mean volume, maximum and mean depth), 2) the displacement of objects when placed into simulation [34], and 3) the contact hand region (ratio and area, where ratio is the percentage of grasps that have non-zero contact area). There is a trade-off between contact/simulation displacement and intersection. While the metrics characterize the grasp quality, no single metric alone is conclusive on grasp synthesis. So we also conducted a user study. We show users two human grasps randomly chosen from two methods and ask them to select their preferred one. We collected 440 and 380 answers from 22/19 users on HO3D and ObMan accordingly.

Baselines. We compare with baseline GraspTTA [90] which is trained on in-domain data (ObMan with annotated grasps). It learns to generate contact maps on hand and object which are then optimized along with hand pose be self-consistent during test time. We also compare with ground truth annotation in both datasets. While Grasping Fields [96] is also a representa-

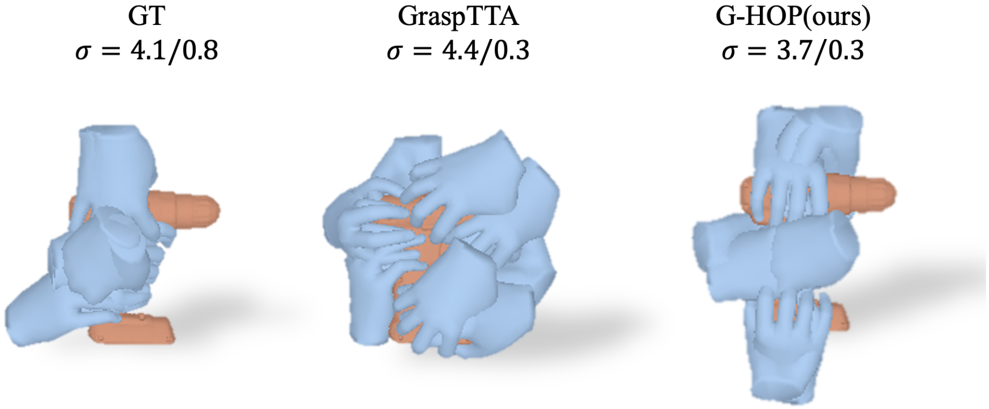


Figure 7.12: **Grasp Diversity:** 10 random grasps of a power drill. Although GraspTTA generates more diverse grasps, some of them are not plausible as they disregard object functions.

tive method for grasp generation, their evaluation setup assumes a known object pose relative to the hand unlike ours, and randomizing this relative pose significantly affects their performance. We detail this further and report our results under their evaluation setting in supplementary.

Results. Fig. 7.9 visualizes two human grasp synthesis from each method for a given object. Annotated grasps (GT) in two datasets display different grasping styles. Semi-automatically generated grasps [73] sometimes do not look natural and tend to "over-grasp" as they are generated to maximize stability. GraspTTA is trained on the same dataset and shows similar over-grasp behavior while our grasps appear more natural. In contrast, G-HOP grasps objects from different directions while all of the synthesized hands make contact with the objects.

Grasp Diversity. We calculate the mean of standard deviations of hand vertices σ from 100 generations per object in the object/hand-centric frame on HO3D in Fig. 7.12. All methods show comparable diversity in the object-centric frame but both methods can improve on the diversity of finger articulation. Note that standard deviation on its own is not a good metric as diverse samples may be implausible or ignore object affordance as visualized.

Grasp Characteristic. Fig. 7.10 visualizes the overall contact probability on hand across all generated grasps. The contact region of GraspTTA is centered at fingertips and (implausibly) even at the nail region shown on the back of the hand. Contact regions from G-HOP are distributed on both fingers and palm, which is more consistent with how humans use their hands [11].

Table 7.3: **Ranking Grasps:** plausibility on HO3D over all grasps, along with the top and bottom 10% grasps ranked by G-HOP.

	maxD↓	avgD↓	vol↓	disp ↓	ratio↑	area↑
G-HOP (top 10%)	**1.74**	**0.31**	**10.57**	**0.71**	1.00	0.22
G-HOP (all)	1.84	0.31	11.46	0.95	1.00	0.23
G-HOP (bottom 10%)	1.87	0.33	13.11	1.41	1.00	0.23

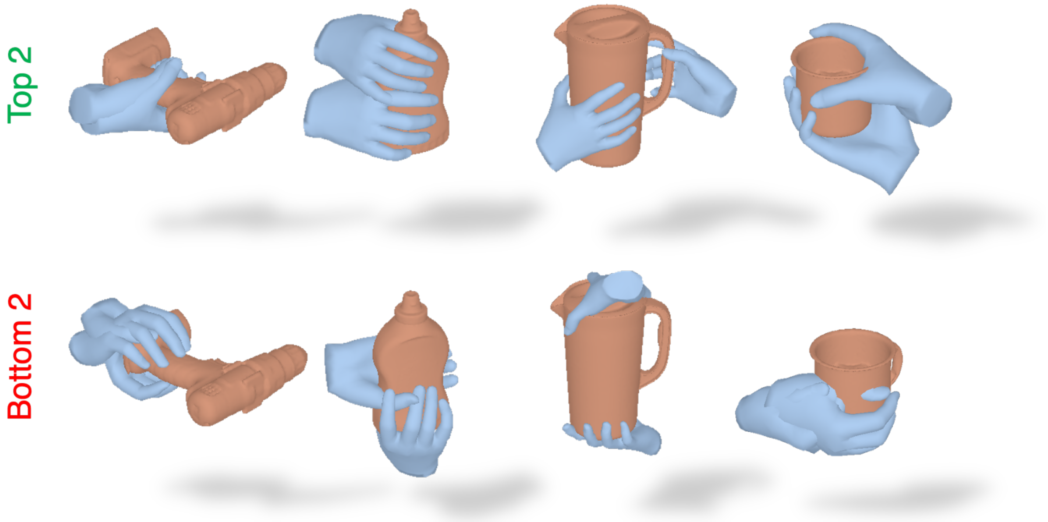


Figure 7.13: **Ranking Grasps:** We visualize grasps with two highest scores (top) and two lowest scores (bottom) among 100 generated grasps from G-HOP.

Tab. 7.2 also reflects the same characteristics. Although G-HOP has higher intersection volume, it has lowest average intersection depth and largest contact area. It also achieves the best performance in terms of grasp stability on HO3D and comparable results on out-of-domain ObMan objects. In user studies, G-HOP is preferred against all methods on both datasets, even when comparing with ground-truth.

Ranking Grasps. Finally, we show that the proposed grasp score yields meaningful grasp ranking. In Fig. 7.13, we visualize top 2 and bottom 2 grasps out of 100 generations from G-HOP, evaluated by the proposed evaluation method. The ranking matches human's common sense. For example, power drills are often held in the middle; narrow side of bottles is often held upwards. Physically infeasible grasps are ranked low such as hands penetrating the mug. Furthermore, the worst two grasps out of 100 are still reasonable in most cases. Note that all the grasps we show to users are randomly chosen for fair comparison. Quantitatively, top-ranked grasps in

Tab. 7.3 show reduced simulation displacement and less intersection, validating our ranking approach's efficacy.

7.3 Discussion

In this chapter, we propose a method to jointly generate 3D shape of HOI given an object category. Our method is the first to generate HOI across such diverse categories. The learned prior G-HOP can serve as generic prior for relevant tasks like reconstructing interaction clips and human grasp synthesis, and we find that it leads to better performance than current task-specific baselines. Despite the encouraging results, we are aware of several limitations: current method requires category information as input which may prevent the model from further scaling up; there is no explicit mechanism to guarantee contact; and the model is still not at a scale comparable to generative models in other domains due to limited training data. Nevertheless, we believe that our work takes an encouraging step towards scaling up a general understanding of hand-object interactions.

7.4 Implementation Details and Additional Results

In the supplementary materials, we provide more implementation details and experimental results on the generative hand-object prior, prior-guided reconstruction, as well as prior-guided grasp synthesis. We discuss network architecture (Sec. 7.4.1), effect of hand representation (Sec. 7.4.1), how to extract hand pose from skeletal distance field (Sec. 7.4.1), and the text prompt we used (Sec. 7.4.1). Then, we show implementation details in reconstructing interaction clips and per-category results in Sec. 7.4.2. Furthermore, we analyze the effect of mesh refinement in grasp synthesis and discuss comparison with prior work Grasping Field [96] in Sec. 7.4.3.

7.4.1 Hand-Object Prior

Network Architecture

We use the same network architecture of latent autoencoder and 3D UNet diffusion model backbone as in SDFusion [27]. The 3D UNet backbone consists of several residual blocks. Each block is a stack of GroupNorm layer [223], non-linear activation [44], and 3D convolutional layer, with optional cross attention layer to time embedding and text embedding. We provide an overview of network details and hyperparameters of our model in Tab. 7.4.

	G-HOP
z-shape	$16^3 \times 3$
$\vert\mathcal{Z}\vert$	8196
Input Channel	3 + 15
Diffusion Steps	1000
Noise Schedule	linear
Channels	64
Number of Blocks	3
Attention resolutions	4,2
Channel Multiplier	1,2,3
Number of Heads	8
Transformers Depth	1
Batch Size	64
Iterations	500k
Learning Rate	1e-4

Table 7.4: **Network architecture for G-HOP.**

Ablating Skeletal Distance Field

Many previous works [95,202] learn a diffusion model in the compact hand/human pose parameter space. We try to represent hand shape by hand pose parameters but find that this pose space is not optimal for jointly diffusing hand pose and objects in interaction. More specifically, the ablated method (pose parameter space) uses the same architecture as the main model except for the (noisy) pose parameter is passed via cross-attention layer instead of concatenating skeletal distance field to the object latent grid. We also search hyperparameters such as weights in DDPM loss to balance diffusing hand pose and diffusing object latent. We visualize the best ablated model in Fig. 7.14 in comparison with our proposed model that represent hand shape by skeletal distance field. The diffusion model with pose parameter space struggles to generate plausible hand articulation together with objects. This is probably because the diffusion model is hard to reason about interaction in the heterogeneous space (1D hand pose and 3D object grids).

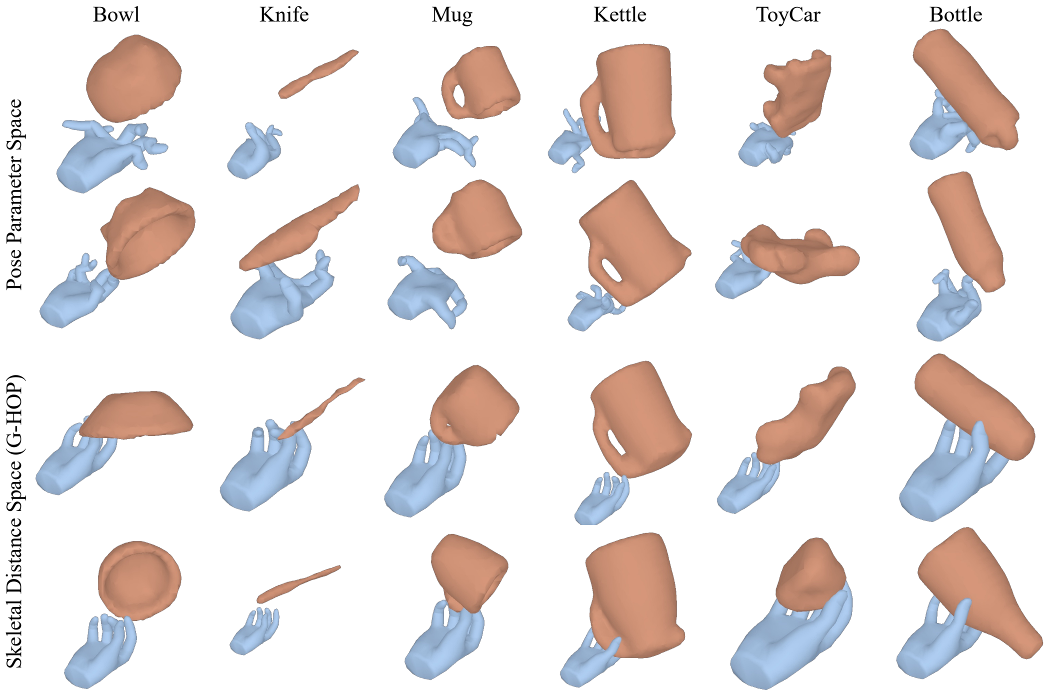


Figure 7.14: **Comparing Hand Representation in Generative Hand-Object Prior:** Top 2 rows show the diffusion model that represents hand shape as pose parameters; bottom 2 rows show the diffusion model (ours) that represents hand shape as skeletal distance field. The homogeneous grid space is easier for the network to reason about interaction.

Hand Pose from Skeletal Distance Field

Our proposed diffusion model generates skeletal distance field, from which hand pose parameters can be extracted. Given a target skeletal distance field $\hat{H}$, we optimize hand pose $\boldsymbol{\theta}$ such that its induced field is closer to the target, *i.e.* $\boldsymbol{\theta}* = \arg\min_{\boldsymbol{\theta}}(H(\boldsymbol{\theta}) - \hat{H})^2 + w\|\boldsymbol{\theta}\|_2^2$. We set w to 1e-5 and optimizes for 1000 steps with Adam optimizer [100] with learning rate 1e-2.

Text Prompt Template

We use the template "a hand holding a {*category*}" to convert category into a text prompt. In addition, we find that appending additional category attribute like size and shape beneficial when we scale up the number of category (see results in Sec. 7.4.2). It may be because attributes help to transfer information between categories with similar shapes but distinct semantics,

Table 7.5: **Comparing Object Error of HOI Reconstruction on HOI4D.**

	Mug			Bottle			Kettle			Bowl			Knife			ToyCar			**mean**		
	F5↑	F10↑	CD↓	F5↑	F10↑	CD↓	F5↑	F10↑	CD↓	F5↑	F10↑	CD↓	F5↑	F10↑	CD↓	F5↑	F10↑	CD↓	F5↑	F10↑	CD↓
iHOI [232]	0.44	0.71	2.1	0.47	0.77	1.5	0.21	0.45	6.3	0.38	0.64	3.1	0.33	0.68	2.8	0.66	0.95	0.5	0.42	0.70	2.7
HHOR [85]	0.18	0.37	6.9	0.26	0.56	3.1	0.12	0.30	11.3	0.31	0.54	4.2	0.71	0.93	0.6	0.26	0.59	1.9	0.31	0.55	4.7
DiffHOI [233]	0.64	0.86	1.0	0.54	0.92	0.7	0.43	0.77	1.5	0.79	0.98	0.4	0.50	0.95	0.8	0.83	0.99	0.3	0.62	0.91	0.8
G-HOP	0.62	0.93	0.7	0.93	1.00	0.2	0.64	0.96	0.6	0.66	0.96	0.5	0.91	0.99	0.2	0.78	0.98	0.3	0.76	0.97	0.4
G-HOP(Cond)	0.57	0.87	1.0	0.74	0.98	0.4	0.46	0.83	1.3	0.47	0.84	1.1	0.95	1.00	0.1	0.74	0.98	0.4	0.66	0.92	0.7
G-HOP(2D)	0.54	0.80	1.3	0.26	0.58	2.5	0.46	0.85	1.1	0.35	0.57	6.4	0.21	0.68	1.9	0.79	0.97	0.3	0.43	0.74	2.3

e.g. pens and spoons are all thin sticks. We use LLM [151] to generate attribute automatically. We list text prompt we used in Tab. 7.11.

Table 7.6: **Comparing Hand Error of HOI Reconstruction on HOI4D.**

	Mug		Bottle		Kettle		Bowl		Knife		ToyCar		**mean**	
	MPJPE↓	AUC↑	MPJPE↓	AUC↑	MPJPE↓	AUC↑	MPJPE↓	AUC↑	MPJPE↓	AUC↑	MPJPE↓	AUC↑	MPJPE↓	AUC↑
iHOI [232]	1.10	0.78	1.09	0.78	1.11	0.78	1.23	0.76	1.39	0.72	1.20	0.76	1.19	0.76
DiffHOI [233]	1.06	0.79	1.01	0.80	1.07	0.79	1.21	0.76	1.33	0.73	1.04	0.79	1.12	0.78
G-HOP	1.02	0.80	0.97	0.81	0.98	0.81	1.09	0.78	1.20	0.76	1.02	0.80	1.05	0.79
G-HOP(Cond)	1.08	0.78	1.06	0.79	1.09	0.79	1.18	0.76	1.34	0.73	1.11	0.78	1.14	0.77
G-HOP(2D)	1.10	0.78	0.97	0.81	1.06	0.79	1.24	0.75	1.24	0.75	1.07	0.79	1.11	0.78

Table 7.7: **Comparing Hand-Object Alignment ($\mathbf{CD}_h$ ↓) of HOI Reconstruction on HOI4D.**

	Mug	Bottle	Kettle	Bowl	Knife	ToyCar	**mean**
iHOI [232]	19.7	13.9	35.9	49.3	21.9	21.6	27.1
HHOR [85]	229.1	172.0	100.4	50.1	185.1	255.8	165.4
DiffHOI [233]	18.1	15.3	42.2	101.8	91.6	23.3	48.7
G-HOP	12.4	9.7	41.8	26.2	13.2	7.5	18.4
G-HOP(Cond)	10.2	6.9	40.7	10.4	39.1	8.5	19.3
G-HOP(2D)	14.5	33.6	61.6	71.0	141.7	38.8	60.2

Table 7.8: **Additional Ablation Studies of HOI reconstruction:** We report object error (F@5mm, F@10mm, CD), hand-object alignment CD$_h$, and hand error (MPJPE, AUC) on HOI4D. We analyze the effect of other implementation details, including dynamic noise thresholding and choice of text prompt templates.

	Object Error			Align	Hand Error	
	F5↑	F10↑	CD↓	CD$_h$ ↓	MPJPE↓	AUC↑
G-HOP	0.76	0.97	0.4	18.4	1.05	0.79
$U_b = 0.25$	0.69	0.95	0.5	50.0	1.01	0.80
$U_b = 0.75$	0.49	0.76	4.0	48.1	1.06	0.79
G-HOP (G) w/ attr	0.65	0.92	0.7	17.8	1.06	0.79
G-HOP (G) wo/ attr	0.61	0.89	0.8	24.6	1.04	0.79

7.4.2 Reconstructing Interaction Clips

Following prior work [233], we evaluate reconstruction on two sequences per category on HOI4D. We report mean performance per category in terms of object error (Tab. 7.5), hand error (Tab. 7.6), and their alignment (Tab. 7.7). In addition to baselines and ablations reported in main paper, we also analyze the effect of other implementation details as follows:

Dynamic Noise Threshold. The amount of injected noise in SDS has large impact on the guidance effect. We find that thin structures are better captured when adding a smaller noise while thick structure are better captured when adding larger noise. We use an adaptive noise scheduler that dynamically adjusts the maximum amount of noise $U_b, i \sim \mathcal{U}[U_a, U_b]$ based on the current object shape. More specifically, it is a linear interpolation based on minimal object SDF value in the current representation, *i.e.*

$$U_b = \frac{s - s_{\min}}{s_{\max} - s_{\min}} U_{b\max} + (1 - \frac{s - s_{\min}}{s_{\max} - s_{\min}}) U_{b\min}$$
$$s = clamp(\min O[X_{grid}], s_{\min}, s_{\max})$$

In our experiment, we set $U_{b\,\max} = 0.75, U_{b\,\min} = 0.25, s_{\min} = -0.2, s_{\max} = -0.01$. As reported in Tab. 7.8, our dynamic noise threshold leads to better performance than constant noise threshold.

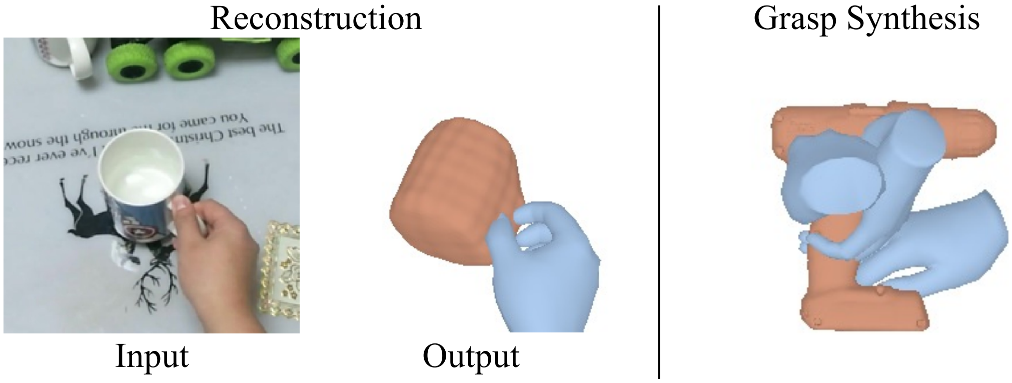


Figure 7.15: **2D Joint Prior (DiffHOI-J):** reconstruction and grasp synthesis results guided by 2D joint prior.

Scaling Up Number of Categories. For fair comparison, we use the diffusion model that only trains on HOI4D dataset to reconstruct interaction clips. In Tab. 7.8,, we also compare with the generalist model (G-HOP (G)) that trains on all seven datasets. Note that we use G-HOP (G) in all other experiments. We find that adding attribute to text prompt helps when scaling up to more categories. While G-HOP (G) leads to a bit worse reconstruction performance on the HOI4D dataset than the specialist which is trained only on HOI4D, it still outperforms other baselines.

2D Joint Prior. We trained a joint prior version of DiffHOI, or a 2D version of G-HOP $p(\pi(O), \pi(H)|C)$. Interestingly, we find that this cannot effectively guide grasp synthesis or reconstruction (Fig. 7.15, Tab. 7.5-7.7). It performs even worse than DiffHOI [233], perhaps because it is harder to learn the distribution over object, hand, and rendering viewpoints (unlike DiffHOI where the 'conditioning' informs about the hand and viewpoint).

7.4.3 Grasp Synthesis

Comparison with Grasping Field. Grasping Field [96] is a representative method that uses a conditional VAE to generate hand surface distance field given an object point cloud. Their evaluation setup generates grasps for known object pose with respect to hand. We evaluate G-HOP under their setup by only optimizing hand articulation while keeping the relative pose as the given ground truth. We denote this setting with known object

Table 7.9: **Comparison with Baselines:** We compare human grasp synthesis along with prior work GF [96]. * denotes GF's evaluation setting with known object pose.

		Intersection			Disp.	Contact	
		max D ↓	avg D ↓	vol ↓	avg ↓	ratio ↑	area ↑
ObMan	GF [96]*	0.56	0.44	6.05	2.07	0.89	0.06
	G-HOP*	0.97	0.70	6.39	2.03	1.00	0.13
	GF [96]	0.79	0.64	43.35	1.82	1.00	0.09
	G-HOP	0.74	0.51	17.40	1.85	0.93	0.25

Table 7.10: **Effect of Refinement:** We report human grasp synthesis before and after mesh refinement. G-HOP† denotes generated grasps before mesh refinement.

		Intersection			Disp.	Contact	
		max D ↓	avg D ↓	vol ↓	avg ↓	ratio ↑	area ↑
ObMan	G-HOP†	0.74	0.57	8.25	3.87	0.82	0.12
	G-HOP	0.74	0.51	17.40	1.85	0.93	0.25
HO3D	G-HOP†	1.84	0.31	11.46	0.95	1.00	0.23
	G-HOP	2.42	0.68	7.55	2.48	0.99	0.20

pose as *. G-HOP also benefits from well initialized object poses as contact ratio increases to 100%. Our contact area reduces probably because the given object pose are obtained from GT grasps that uses more finger tips and this makes the human hand palm harder to make contact. We also show that randomizing the relative pose (our evaluation setup) significant affects their performance, as visualized in Fig. 7.16. Note that GF gets large intersection volume but less intersection depth. This is because the latter is only calculated on each hand vertices inside of the object. For example, in the second

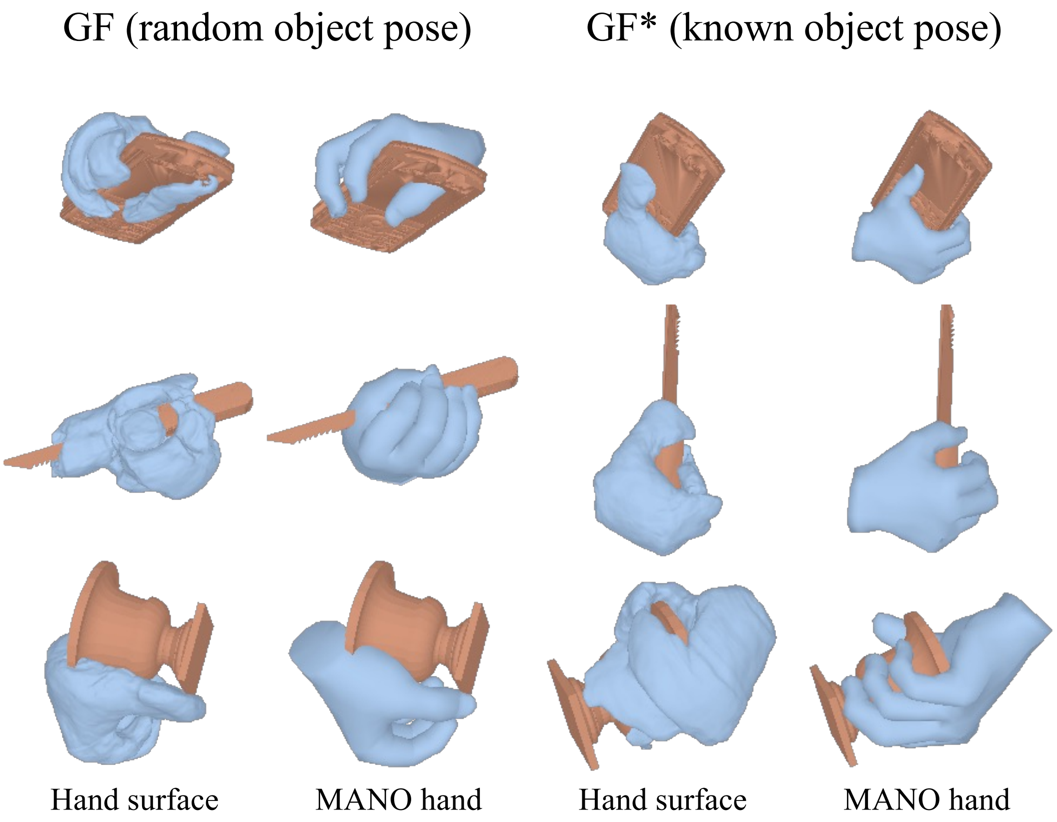


Figure 7.16: GF assumes known object pose when evaluating. Randomizing object pose affects their performance.

row of Fig. 7.16, the knife penetrates hand, leading to high volume. But the maximum intersection depth for each hand vertices is less than the thickness of the knife.

Effect of Refinement After optimizing human grasps with respect to SDS loss using object SDF grid, we also do a light-weight mesh refinement by replacing the object SDF grid with the original mesh. It is to account for loss of accuracy during mesh conversion. We use the same objectives in previous work [73, 232] that encourage contact and discourage penetration. We denote the generated grasps before mesh refinement as † and report its performance on two datasets in Tab. 7.10. Even without mesh refinement, the generated grasps also have large contact area and less displacement in simulation. The refinement process can adjust hand pose to further improves

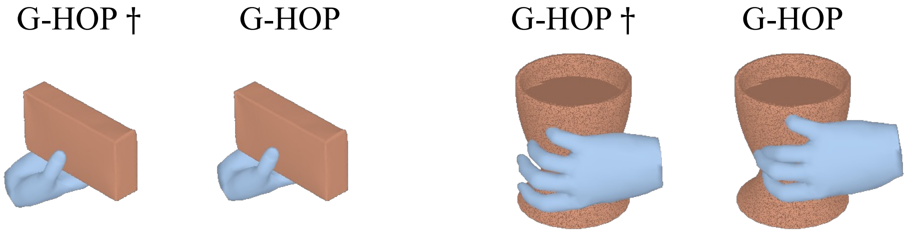
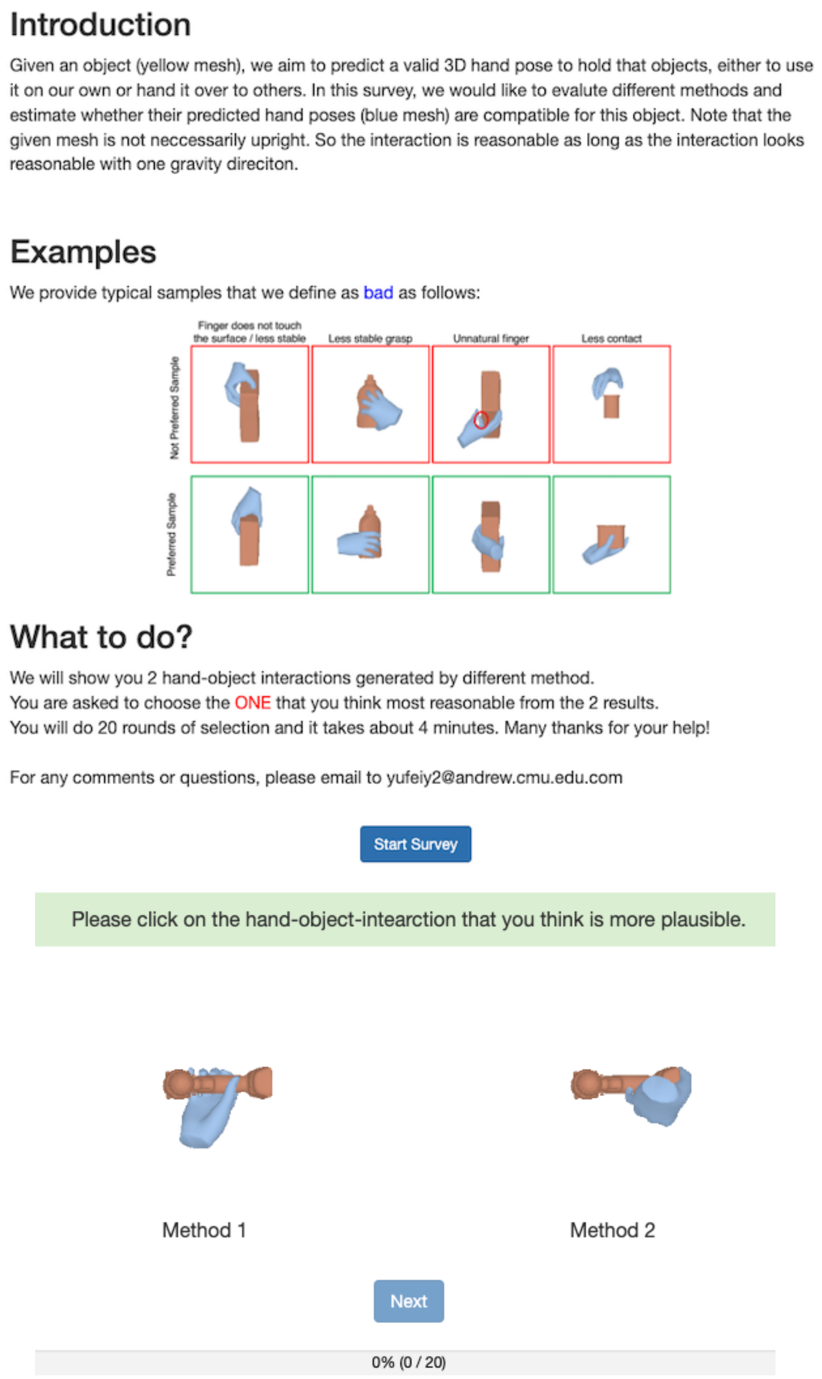


Figure 7.17: **Effect of Mesh Refinement:** We visualize synthesized grasps before (G-HOP†) and after (G-HOP) refinement.

the contact and grasp stability.

User Study Interface

Fig. 7.18 shows the user interface for evaluating the generated grasps. Users are presented two grasps visualized from different view angles as gif and are asked to choose the more plausible grasps.



Introduction

Given an object (yellow mesh), we aim to predict a valid 3D hand pose to hold that objects, either to use it on our own or hand it over to others. In this survey, we would like to evalute different methods and estimate whether their predicted hand poses (blue mesh) are compatible for this object. Note that the given mesh is not neccessarily upright. So the interaction is reasonable as long as the interaction looks reasonable with one gravity direciton.

Examples

We provide typical samples that we define as bad as follows:

What to do?

We will show you 2 hand-object interactions generated by different method.
You are asked to choose the ONE that you think most reasonable from the 2 results.
You will do 20 rounds of selection and it takes about 4 minutes. Many thanks for your help!

For any comments or questions, please email to yufeiy2@andrew.cmu.edu.com

Please click on the hand-object-intearction that you think is more plausible.

Figure 7.18: **User Study Interface:** We visualize user study interface including the user instruction page and the survey page.

Table 7.11: We provide list of class names and their attributes used in the text prompt. The class names are manually merged across different datasets while the attributes are automatically generated by large language model [151].

Class	Attribute
plate	medium, flat, circular
baseboard	big, long, rectangular
stamp	small, flat, square
laptop	big, flat, rectangular
funnel	medium, conical
spatula	medium, flat, elongated
pear	small, pear shaped
lemon	small, oval
stick	varies, cylindrical, long
cylinder	varies, cylindrical
mug	medium, cylindrical, handle attached
flute	medium, cylindrical, long
shield	big, curved, oval or round
floor	big, flat, rectangular or irregular
mouse	medium, oval, handheld
fish	varies, animal shaped
screw driver	medium, cylindrical, elongated
pen	small, cylindrical, elongated
hair dryer	medium, elongated, handheld
burger	medium, cylindrical, stacked layers
paint roller	medium, cylindrical, handheld
power saw	big, elongated, handheld or standalone
bottle	medium, cylindrical, narrow neck
pump	varies, mechanical, various shapes
flask	medium, cylindrical or conical, narrow neck
sheet	big, flat, rectangular
hand bag	medium, varies, handle attached
stapler	medium, rectangular, handheld
gummy	small, animal or object shaped
fork	small, elongated, tines at one end
wood	varies, solid, various shapes
chopsticks	small, cylindrical, elongated
strawberry	small, heart-shaped
cupmod	medium, cylindrical, handle attached
spray	medium, cylindrical, nozzle at top
crate	big, cuboid, open structure

Continued on next column

Continued from previous column

Class	Attribute
microwave	big, rectangular, box-like
headphone	medium, round or oval, worn over ears
apple	small, round, stem at top
backpack	big, varies, straps attached
brick	medium, rectangular, solid
wood plank	big, flat, rectangular
tv	big, flat, rectangular
rubiks	small, cubical, multicolored faces
carpet	big, flat, rectangular or oval
container	varies, solid, various shapes
lego	small, rectangular or square, connecting knobs
jar	medium, cylindrical or oval, lid on top
oven	big, box-like, door at front
mixer	big, varies, mechanical
train	big, cylindrical, long
teddy bear	medium, animal shaped, soft
chess rook	small, cylindrical, castle-shaped top
binoculars	medium, cylindrical, two lenses
pencil mod	small, cylindrical, elongated
knife	medium, flat, sharp edge
tin	medium, cylindrical or rectangular, lid on top
light tube	medium, cylindrical, elongated
ball	small, spherical
cupcake	small, cylindrical, rounded top
spoon	small, oval or round, handle attached
chalk	small, cylindrical, elongated
light bulb	small, round, screw base
case	varies, box-like, lid or zipper
peg test	varies, varies, testing equipment
piggy bank	medium, animal shaped, slot on top
kettle	medium, rounded, spout and handle
wrench	medium, elongated, adjustable jaw
bacon	small, flat, elongated
purse	medium, varies, handle or strap
boat	big, elongated, hollow
disk	small, flat, circular
game controller	medium, ergonomic, buttons and joysticks
keyboard	medium, flat, rectangular
trowels	medium, flat, handle attached
shovel	big, flat, long handle
eye glasses	small, oval or round, frame with lenses
stanford bunny	small, animal shaped, 3D model
camera	medium, box-like, lens at front

Continued on next column

Continued from previous column

Class	Attribute
rifle	big, elongated, barrel and stock
can	small, cylindrical, lid on top
range	big, flat or box-like, knobs and burners
toy airplane	small, aerodynamic, wings attached
cube	varies, cubical
tablet	medium, flat, rectangular
teapot	medium, rounded, spout and handle
chair	big, varies, seat and backrest
beaker	small, cylindrical, pouring lip
plum	small, round, pit inside
triangle	varies, triangular
barrel	big, cylindrical, hollow
cup	small, cylindrical, handle attached
toothpaste	small, cylindrical, tube-shaped
bag	varies, varies, handle or strap
pyramid	varies, pyramidal
dice	small, cubical, numbered faces
ruler	small, flat, rectangular
scissors	small, paired blades, handles
clamp	small, C or G shaped, screw mechanism
phone	medium, flat, rectangular
marbles	small, spherical, glass or clay
dart	small, conical, pointed tip
calculator	medium, flat, rectangular
duck	varies, animal shaped
chain	varies, interlinked, metal
bucket	medium, cylindrical, handle attached
peach	small, round, pit inside
donut	small, cylindrical, hole in center
flashlight	medium, cylindrical, light at one end
sponge	small, soft, varies
mat	medium, flat, rectangular or oval
cardboard	varies, flat, rectangular
scoop	small, semi-spherical, handle attached
block	varies, solid, cuboidal
pliers	medium, paired jaws, handles
board	big, flat, rectangular
shoe	medium, foot-shaped, footwear
floor mate	varies, flat, used for cleaning
brush	varies, bristles attached, handle
alarm clock	small, circular or square, time display
hood	big, curved, worn over head
pot	medium, cylindrical, handle attached

Continued on next column

Continued from previous column

Class	Attribute
chessboard	medium, square, 8x8 squares
pillow	medium, soft, rectangular
power drill	medium, cylindrical, elongated
marshmallow	small, cylindrical or cubic, soft
bowl	medium, round, hollow
tube	varies, cylindrical, hollow
frisbee	medium, flat, circular
hammer	medium, heavy head, handle attached
toothbrush	small, bristles at end, handle
toycar	small, car shaped, wheels attached
elephant	big, animal shaped
tray	medium, flat, raised edges
box	varies, cuboidal, lid or flaps
book	medium, flat, rectangular
skillet lid	medium, flat or domed, handle on top
table	big, flat, supported by legs
banana	small, curved, elongated
padlock	small, rounded or square, shackle on top
bin	big, cylindrical or cuboidal, open top
blender	medium, cylindrical, mechanical
pitcher	medium, cylindrical, handle and spout
toilet	big, bowl-shaped, plumbing fixture
wine glass	small, stemmed, conical
towel	big, flat, rectangular
vacuum	big, cylindrical, mechanical
chips	small, flat, round or oval
orange	small, round, citrus fruit
microphone	small, cylindrical, handheld
usb stick	small, rectangular, electronic
door knob	small, round, mounted on door
fryingpan	medium, flat, round
watch	small, round, straps attached
eraser	small, rectangular or cylindrical, soft

Concluded

Chapter 8

Conclusions

In this thesis, we explored methods that learn to infer possible 3D representations of everyday hand-object interactions from visual inputs. Although we have made progress in pushing the boundaries of perceiving everyday human interaction, especially for generic unknown objects, there are still many open problems to be addressed.

First, in the thesis, we primarily focus on hand interactions with rigid objects of portable size in quasi-static scenarios, where humans reach out to grasp objects in order to hold them firmly. However, human-hand interactions are highly dynamic. These interactions, including hand pose and even object states, keep changing depending on the task progress [91]. For example, making dumplings involves a series of distinct dynamics and deformations of the dough. To model such interactions, we need to develop methods that can capture the fine motion of hands and moments of contact, as well as to investigate better 3D representations that can respect the dynamics and deformations of the physical process.

Second, the work presented in the thesis involves only a single object and is limited to short interactions up to a few seconds. However, human interactions come from a much larger context, engaged with multiple objects of interest. For example, making a pour-over coffee takes a few minutes and requires interactions with multiple objects such as a filter, grinder, kettle, cup,*etc.*, as well as interactions among these entities. Manually specifying the interactions and their preconditions becomes intractable. It may involve some hierarchy representation to model the long-term interactions at different levels of granularity and leveraging common sense from large language models as a high-level planner to navigate through the hierarchy [241,242].

Finally, it is a very exciting direction to apply the learned interaction priors to other scientific domains such as robotics and biomechanics. In robotics, it is promising to use the learned human priors to guide robot manipulations, especially for tasks that require human-like dexterity [4,28,38,104,133,167,212]. In biomechanics, we would like to provide scalable methods for healthcare applications such as automatically monitoring hand usage for patients with hand injuries, and assisting rehabilitation process. On the other side, we also need to draw inspiration from these fields to further improve the vision systems, such as co-designing the vision and manipulation policy for robotics [166], and leveraging well-studied biomechanics models [43] to capture human hands better.

Bibliography

[1] I. Akhter and M. J. Black. Pose-conditioned joint angle limits for 3d human pose reconstruction. In *CVPR*, 2015.

[2] D. Antotsiou, G. Garcia-Hernando, and T.-K. Kim. Task-oriented hand motion retargeting for dexterous manipulation imitation. In *ECCV Workshops*, 2018.

[3] O. Avrahami, D. Lischinski, and O. Fried. Blended diffusion for text-driven editing of natural images. In *CVPR*, 2022.

[4] S. Bahl, A. Gupta, and D. Pathak. Human-to-robot imitation in the wild. *RSS*, 2022.

[5] S. Bambach, S. Lee, D. J. Crandall, and C. Yu. Lending a hand: Detecting hands and recognizing activities in complex egocentric interactions. In *ICCV*, 2015.

[6] O. Bar-Tal, D. Ofri-Amar, R. Fridman, Y. Kasten, and T. Dekel. Text2live: Text-driven layered image and video editing. In *ECCV*, 2022.

[7] R. Benenson, S. Popov, and V. Ferrari. Large-scale interactive object segmentation with human annotators. In *CVPR*, 2019.

[8] A. Bicchi and V. Kumar. Robotic grasping and contact: A review. In *ICRA*, 2000.

[9] F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero, and M. J. Black. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In *ECCV*, 2016.

[10] A. Boukhayma, R. d. Bem, and P. H. Torr. 3d hand shape and pose from images in the wild. In *CVPR*, 2019.

[11] S. Brahmbhatt, A. Handa, J. Hays, and D. Fox. Contactgrasp: Functional multi-finger grasp synthesis from contact. In *IROS*, 2019.

[12] S. Brahmbhatt, C. Tang, C. D. Twigg, C. C. Kemp, and J. Hays. Contactpose: A dataset of grasps with object contact and hand pose. In *ECCV*, 2020.

[13] S. Brahmbhatt, C. Tang, C. D. Twigg, C. C. Kemp, and J. Hays. Contactpose: A dataset of grasps with object contact and hand pose. In *ECCV*, 2020.

[14] M. Cai, K. M. Kitani, and Y. Sato. Understanding hand-object manipulation with grasp types and object attributes. In *Robotics: Science and Systems*, 2016.

[15] B. Calli, A. Singh, A. Walsman, S. Srinivasa, P. Abbeel, and A. M. Dollar. The ycb object and model set: Towards common benchmarks for manipulation research. In *ICAR*, 2015.

[16] B. Calli, A. Walsman, A. Singh, S. Srinivasa, P. Abbeel, and A. M. Dollar. Benchmarking in manipulation research: The ycb object and model set and benchmarking protocols. *arXiv*, 2015.

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

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

[19] Z. Cao, I. Radosavovic, A. Kanazawa, and J. Malik. Reconstructing hand-object interactions in the wild. In *ICCV*, 2021.

[20] C. Chan, S. Ginosar, T. Zhou, and A. A. Efros. Everybody dance now. In *ICCV*, 2019.

[21] E. R. Chan, C. Z. Lin, M. A. Chan, K. Nagano, B. Pan, S. De Mello, O. Gallo, L. J. Guibas, J. Tremblay, S. Khamis, et al. Efficient geometry-aware 3d generative adversarial networks. In *CVPR*, 2022.

[22] A. X. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015.

[23] A. X. Chang, T. A. Funkhouser, L. J. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su, J. Xiao, L. Yi, and F. Yu. Shapenet: An information-rich 3d model repository. *ArXiv*, 2015.

[24] Y.-W. Chao, W. Yang, Y. Xiang, P. Molchanov, A. Handa, J. Tremblay, Y. S. Narang, K. Van Wyk, U. Iqbal, S. Birchfield, J. Kautz, and D. Fox. DexYCB: A benchmark for capturing hand grasping of objects. In *CVPR*, 2021.

[25] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *TPAMI*, 2017.

[26] H. K. Cheng, Y.-W. Tai, and C.-K. Tang. Rethinking space-time networks with improved memory coverage for efficient video object segmentation. In *NeurIPS*, 2021.

[27] Y.-C. Cheng, H.-Y. Lee, S. Tulyakov, A. G. Schwing, and L.-Y. Gui. Sdfusion: Multimodal 3d shape completion, reconstruction, and generation. In *CVPR*, 2023.

[28] C. Chi, Z. Xu, C. Pan, E. Cousineau, B. Burchfiel, S. Feng, R. Tedrake, and S. Song. Universal manipulation interface: In-the-wild robot teaching without in-the-wild robots. *arXiv preprint arXiv:2402.10329*, 2024.

[29] H. Choi, G. Moon, and K. M. Lee. Pose2mesh: Graph convolutional network for 3d human pose and mesh recovery from a 2d human pose. In *ECCV*, 2020.

[30] C. B. Choy, D. Xu, J. Gwak, K. Chen, and S. Savarese. 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In *ECCV*, 2016.

[31] K. Connolly and M. Dalgleish. The emergence of a tool-using skill in infancy. *Developmental Psychology*, 1989.

[32] E. Corona, A. Pumarola, G. Alenya, F. Moreno-Noguer, and G. Rogez. Ganhand: Predicting human grasp affordances in multi-object scenes. In *CVPR*, 2020.

[33] E. Corona, A. Pumarola, G. Alenya, F. Moreno-Noguer, and G. Rogez. Ganhand: Predicting human grasp affordances in multi-object scenes. In *CVPR*, 2020.

[34] E. Coumans and Y. Bai. Pybullet, a python module for physics simulation for games, robotics and machine learning. `http://pybullet.org`, 2016–2021.

[35] D. Damen, H. Doughty, G. M. Farinella, S. Fidler, A. Furnari, E. Kazakos, D. Moltisanti, J. Munro, T. Perrett, W. Price, and M. Wray. Scaling egocentric vision: The epic-kitchens dataset. In *ECCV (ECCV)*, 2018.

[36] D. Damen, H. Doughty, G. M. Farinella, S. Fidler, A. Furnari, E. Kazakos, D. Moltisanti, J. Munro, T. Perrett, W. Price, and M. Wray. Scaling egocentric vision: The epic-kitchens dataset. *ECCV*, 2018.

[37] D. Damen, H. Doughty, G. M. Farinella, S. Fidler, A. Furnari, E. Kazakos, D. Moltisanti, J. Munro, T. Perrett, W. Price, and M. Wray. Scaling egocentric vision: The epic-kitchens dataset. In *ECCV*, 2018.

[38] S. Dasari, A. Gupta, and V. Kumar. Learning dexterous manipulation from exemplar object trajectories and pre-grasps. In *ICRA*, 2023.

[39] M. Deitke, R. Liu, M. Wallingford, H. Ngo, O. Michel, A. Kusupati, A. Fan, C. Laforte, V. Voleti, S. Y. Gadre, et al. Objaverse-xl: A universe of 10m+ 3d objects. *NeurIPS*, 2024.

[40] C. Deng, C. M. Jiang, C. Qi, X. Yan, Y. Zhou, L. J. Guibas, and D. Anguelov. Nerdi: Single-view nerf synthesis with language-guided diffusion as general image priors. *ArXiv*, 2022.

[41] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

[42] P. Dhariwal and A. Nichol. Diffusion models beat gans on image synthesis. *NeurIPS*, 2021.

[43] S. F. Duncan, C. E. Saracevic, and R. Kakinoki. Biomechanics of the hand. *Hand clinics*, 2013.

[44] S. Elfwing, E. Uchibe, and K. Doya. Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. *Neural networks*, 2018.

[45] G. ElKoura and K. Singh. Handrix: animating the human hand. In *SCA*, 2003.

[46] Z. Fan, O. Taheri, D. Tzionas, M. Kocabas, M. Kaufmann, M. J. Black, and O. Hilliges. Articulated objects in free-form hand interaction. *ArXiv*, 2022.

[47] K. Fang, T.-L. Wu, D. Yang, S. Savarese, and J. J. Lim. Demo2vec: Reasoning object affordances from online videos. In *CVPR*, 2018.

[48] T. Feix, J. Romero, H.-B. Schmiedmayer, A. M. Dollar, and D. Kragic. The grasp taxonomy of human grasp types. *THMS*, 2015.

[49] V. Florence, J. J. Corso, and B. Griffin. Robot-supervised learning for object segmentation. In *ICRA*, 2020.

[50] D. F. Fouhey, V. Delaitre, A. Gupta, A. A. Efros, I. Laptev, and J. Sivic. People watching: Human actions as a cue for single view geometry. In *ECCV*, 2012.

[51] M. Gadelha, S. Maji, and R. Wang. 3d shape induction from 2d views of multiple objects. In *3DV*, 2017.

[52] G. Garcia-Hernando, S. Yuan, S. Baek, and T.-K. Kim. First-person hand action benchmark with rgb-d videos and 3d hand pose annotations. In *CVPR*, 2018.

[53] L. Ge, Z. Ren, Y. Li, Z. Xue, Y. Wang, J. Cai, and J. Yuan. 3d hand shape and pose estimation from a single rgb image. In *CVPR*, 2019.

[54] J. J. Gibson. The ecological approach to the visual perception of pictures. *Leonardo*, 1978.

[55] R. Girdhar, D. F. Fouhey, M. Rodriguez, and A. Gupta. Learning a predictable and generative vector representation for objects. In *ECCV*, 2016.

[56] R. Girdhar and K. Grauman. Anticipative video transformer. In *ICCV*, 2021.

[57] G. Gkioxari, J. Malik, and J. Johnson. Mesh r-cnn. In *ICCV*, 2019.

[58] S. Goel, A. Kanazawa, and J. Malik. Shape and viewpoint without keypoints. In *ECCV*, 2020.

[59] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *NeurIPS*, 2014.

[60] M. Goyal, S. Modi, R. Goyal, and S. Gupta. Human hands as probes for interactive object understanding. In *CVPR*, 2022.

[61] P. Grady, C. Tang, C. D. Twigg, M. Vo, S. Brahmbhatt, and C. C. Kemp. Contactopt: Optimizing contact to improve grasps. *CVPR*, 2021.

[62] A. Gropp, L. Yariv, N. Haim, M. Atzmon, and Y. Lipman. Implicit geometric regularization for learning shapes. In *ICML*, 2020.

[63] T. Groueix, M. Fisher, V. G. Kim, B. C. Russell, and M. Aubry. A papier-mâché approach to learning 3d surface generation. In *CVPR*, 2018.

[64] A. Gupta, S. Satkin, A. A. Efros, and M. Hebert. From 3d scene geometry to human workspace. In *CVPR*, 2011.

[65] H. Hamer, J. Gall, T. Weise, and L. Van Gool. An object-dependent hand pose prior from sparse training data. In *CVPR*, 2010.

[66] H. Hamer, K. Schindler, E. Koller-Meier, and L. Van Gool. Tracking a hand manipulating an object. In *ICCV*, 2009.

[67] S. Hampali, T. Hodan, L. Tran, L. Ma, C. Keskin, and V. Lepetit. In-hand 3d object scanning from an rgb sequence. *ArXiv*, 2022.

[68] S. Hampali, M. Rad, M. Oberweger, and V. Lepetit. Honnotate: A method for 3d annotation of hand and object poses. In *CVPR*, 2020.

[69] Y. Hasson, B. Tekin, F. Bogo, I. Laptev, M. Pollefeys, and C. Schmid. Leveraging photometric consistency over time for sparsely supervised hand-object reconstruction. In *CVPR*, 2020.

[70] Y. Hasson, G. Varol, I. Laptev, and C. Schmid. Towards unconstrained joint hand-object reconstruction from rgb videos. In *ArXiv*, 2021.

[71] Y. Hasson, G. Varol, C. Schmid, and I. Laptev. Towards unconstrained joint hand-object reconstruction from rgb videos. In *3DV*, 2021.

[72] Y. Hasson, G. Varol, D. Tzionas, I. Kalevatykh, M. J. Black, I. Laptev, and C. Schmid. Learning joint reconstruction of hands and manipulated objects. In *CVPR*, 2019.

[73] Y. Hasson, G. Varol, D. Tzionas, I. Kalevatykh, M. J. Black, I. Laptev, and C. Schmid. Learning joint reconstruction of hands and manipulated objects. In *CVPR*, 2019.

[74] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *CVPR*, 2016.

[75] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[76] P. Henderson and V. Ferrari. Learning single-image 3d reconstruction by generative modelling of shape, pose and shading. *IJCV*, 2019.

[77] P. Henderson, V. Tsiminaki, and C. H. Lampert. Leveraging 2d data to learn textured 3d mesh generation. In *CVPR*, 2020.

[78] P. Henzler, N. J. Mitra, and T. Ritschel. Escaping plato's cave: 3d shape from adversarial rendering. In *ICCV*, 2019.

[79] T. Hermans, J. M. Rehg, and A. Bobick. Affordance prediction via learned object attributes. In *ICRA Workshop*, 2011.

[80] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *NeurIPS*, 2017.

[81] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. *ArXiv*, 2020.

[82] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 2020.

[83] J. Ho and T. Salimans. Classifier-free diffusion guidance. *NerIPS Workshop*, 2022.

[84] Y. Hong, K. Zhang, J. Gu, S. Bi, Y. Zhou, D. Liu, F. Liu, K. Sunkavalli, T. Bui, and H. Tan. Lrm: Large reconstruction model for single image to 3d. *arXiv*, 2023.

[85] D. Huang, X. Ji, X. He, J. Sun, T. He, Q. Shuai, W. Ouyang, and X. Zhou. Reconstructing hand-held objects from monocular video. In *SIGGRAPH Asia*, 2022.

[86] Y. Huang, M. Cai, Z. Li, and Y. Sato. Predicting gaze in egocentric video by learning task-dependent attention transition. In *ECCV*, 2018.

[87] U. Iqbal, P. Molchanov, T. B. J. Gall, and J. Kautz. Hand pose estimation via latent 2.5 d heatmap regression. In *ECCV*, 2018.

[88] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017.

[89] M. Jaderberg, K. Simonyan, A. Zisserman, et al. Spatial transformer networks. In *NeurIPS*, 2015.

[90] H. Jiang, S. Liu, J. Wang, and X. Wang. Hand-object contact consistency reasoning for human grasps generation. In *ICCV*, 2021.

[91] L. A. Jones and S. J. Lederman. *Human hand function*. 2006.

[92] H. Jun and A. Nichol. Shap-e: Generating conditional 3d implicit functions. *arXiv*, 2023.

[93] A. Kanazawa, S. Tulsiani, A. A. Efros, and J. Malik. Learning category-specific mesh reconstruction from image collections. In *ECCV*, 2018.

[94] T. Karras, S. Laine, and T. Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019.

[95] K. Karunratanakul, K. Preechakul, S. Suwajanakorn, and S. Tang. Guided motion diffusion for controllable human motion synthesis. In *ICCV*, 2023.

[96] K. Karunratanakul, J. Yang, Y. Zhang, M. J. Black, K. Muandet, and S. Tang. Grasping field: Learning implicit representations for human grasps. In *3DV*, 2020.

[97] H. Kato and T. Harada. Learning view priors for single-view 3d reconstruction. In *CVPR*, 2019.

[98] J.-S. Kim and J.-M. Park. Physics-based hand interaction with virtual objects. In *ICRA*, 2015.

[99] D. P. Kingma, J. A. Ba, and J. Adam. A method for stochastic optimization. arxiv 2014. *arXiv preprint arXiv:1412.6980*, 2020.

[100] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *ICLR*, 2014.

[101] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[102] A. Kirillov, Y. Wu, K. He, and R. Girshick. Pointrend: Image segmentation as rendering. *CVPR*, 2020.

[103] A. Kirillov, Y. Wu, K. He, and R. Girshick. Pointrend: Image segmentation as rendering. In *CVPR*, 2020.

[104] M. Kokic, D. Kragic, and J. Bohg. Learning task-oriented grasping from human activity datasets. *RA-L*, 2020.

[105] C. Kong and S. Lucey. Deep non-rigid structure from motion. In *ICCV*, 2019.

[106] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NeurIPS*, 2012.

[107] N. Kulkarni, A. Gupta, D. Fouhey, and S. Tulsiani. Articulation-aware canonical surface mapping. In *CVPR*, 2020.

[108] N. Kulkarni, A. Gupta, and S. Tulsiani. Canonical surface mapping via geometric cycle consistency. In *ICCV*, 2019.

[109] N. Kumari, R. Zhang, E. Shechtman, and J.-Y. Zhu. Ensembling off-the-shelf models for gan training. In *CVPR*, 2022.

[110] A. Kuznetsova, H. Rom, N. Alldrin, J. Uijlings, I. Krasin, J. Pont-Tuset, S. Kamali, S. Popov, M. Malloci, A. Kolesnikov, T. Duerig, and V. Ferrari. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *IJCV*, 2020.

[111] A. B. L. Larsen, S. K. Sønderby, H. Larochelle, and O. Winther. Autoencoding beyond pixels using a learned similarity metric. *ICML*, 2015.

[112] Y. J. Lee and K. Grauman. Predicting important objects for egocentric video summarization. *IJCV*, 2015.

[113] X. L. Li, J. Thickstun, I. Gulrajani, P. Liang, and T. B. Hashimoto. Diffusion-lm improves controllable text generation. *CoRR*, 2022.

[114] X. Li, S. Liu, K. Kim, S. De Mello, V. Jampani, M.-H. Yang, and J. Kautz. Self-supervised single-view 3d reconstruction via semantic consistency. In *ECCV*, 2020.

[115] X. Li, S. Liu, K. Kim, X. Wang, M.-H. Yang, and J. Kautz. Putting humans in a scene: Learning affordance in 3d indoor environments. In *CVPR*, 2019.

[116] Y. Li, J. L. Fu, and N. S. Pollard. Data-driven grasp synthesis using shape matching and task-based pruning. *TVCG*, 2007.

[117] Z. Li, S. Niklaus, N. Snavely, and O. Wang. Neural scene flow fields for space-time view synthesis of dynamic scenes. In *CVPR*, 2021.

[118] C.-H. Lin, J. Gao, L. Tang, T. Takikawa, X. Zeng, X. Huang, K. Kreis, S. Fidler, M.-Y. Liu, and T.-Y. Lin. Magic3d: High-resolution text-to-3d content creation. *arXiv preprint arXiv:2211.10440*, 2022.

[119] C.-H. Lin, W.-C. Ma, A. Torralba, and S. Lucey. Barf: Bundle-adjusting neural radiance fields. In *ICCV*, 2021.

[120] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014.

[121] R. Liu, R. Wu, B. Van Hoorick, P. Tokmakov, S. Zakharov, and C. Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. *CVPR*, 2023.

[122] S. Liu, H. Jiang, J. Xu, S. Liu, and X. Wang. Semi-supervised 3d hand-object poses estimation with interactions in time. In *CVPR*, 2021.

[123] S. Liu, S. Tripathi, S. Majumdar, and X. Wang. Joint hand motion and interaction hotspots prediction from egocentric videos. In *CVPR*, 2022.

[124] S. Liu, Y. Zhou, J. Yang, S. Gupta, and S. Wang. Contactgen: Generative contact modeling for grasp generation. In *ICCV*, 2023.

[125] S. Liu, T. Li, W. Chen, and H. Li. Soft rasterizer: A differentiable renderer for image-based 3d reasoning. In *ICCV*, 2019.

[126] S. Liu, S. Saito, W. Chen, and H. Li. Learning to infer implicit surfaces without 3d supervision. In *NeurIPS*, 2019.

[127] Y. Liu, Y. Liu, C. Jiang, K. Lyu, W. Wan, H. Shen, B. Liang, Z. Fu, H. Wang, and L. Yi. Hoi4d: A 4d egocentric dataset for category-level human-object interaction. In *CVPR*, pages 21013–21022, June 2022.

[128] Y. Liu, Y. Liu, C. Jiang, K. Lyu, W. Wan, H. Shen, B. Liang, Z. Fu, H. Wang, and L. Yi. HOI4D: A 4D egocentric dataset for category-level human-object interaction. In *CVPR*, 2022.

[129] S. Lombardi, T. Simon, J. Saragih, G. Schwartz, A. Lehrmann, and Y. Sheikh. Neural volumes: Learning dynamic renderable volumes from images. *TOG*, 2019.

[130] W. E. Lorensen and H. E. Cline. Marching cubes: A high resolution 3d surface construction algorithm. *SIGGRAPH*, 1987.

[131] I. Loshchilov and F. Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[132] J. Mahler, M. Matl, V. Satish, M. Danielczuk, B. DeRose, S. McKinley, and K. Goldberg. Learning ambidextrous robot grasping policies. *Science Robotics*, 2019.

[133] P. Mandikal and K. Grauman. Dexvip: Learning dexterous grasping with human hand pose priors from video. In *CoRL*, 2022.

[134] L. Melas-Kyriazi, C. Rupprecht, I. Laina, and A. Vedaldi. Realfusion: 360° reconstruction of any object from a single image. In *Arxiv*, 2023.

[135] C. Meng, Y. He, Y. Song, J. Song, J. Wu, J.-Y. Zhu, and S. Ermon. SDEdit: Guided image synthesis and editing with stochastic differential equations. In *ICLR*, 2022.

[136] C. Meng, Y. Song, J. Song, J. Wu, J.-Y. Zhu, and S. Ermon. Sdedit: Image synthesis and editing with stochastic differential equations. *ICLR*, 2022.

[137] L. Mescheder, M. Oechsle, M. Niemeyer, S. Nowozin, and A. Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *CVPR*, 2019.

[138] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 2021.

[139] A. T. Miller and P. K. Allen. Graspit! a versatile simulator for robotic grasping. *IEEE Robotics & Automation Magazine*, 2004.

[140] A. Mittal, A. Zisserman, and P. H. Torr. Hand detection using multiple proposals. In *BMVC*, 2011.

[141] K. Mo, L. J. Guibas, M. Mukadam, A. Gupta, and S. Tulsiani. Where2Act: From pixels to actions for articulated 3D objects. In *ICCV*, 2021.

[142] F. Mueller, F. Bernard, O. Sotnychenko, D. Mehta, S. Sridhar, D. Casas, and C. Theobalt. Ganerated hands for real-time 3d hand tracking from monocular rgb. In *CVPR*, 2018.

[143] F. Mueller, F. Bernard, O. Sotnychenko, D. Mehta, S. Sridhar, D. Casas, and C. Theobalt. GANerated hands for real-time 3D hand tracking from monocular RGB. In *CVPR*, 2018.

[144] F. Mueller, M. Davis, F. Bernard, O. Sotnychenko, M. Verschoor, M. A. Otaduy, D. Casas, and C. Theobalt. Real-time pose and shape reconstruction of two interacting hands with a single depth camera. *TOG*, 2019.

[145] T. Müller, A. Evans, C. Schied, and A. Keller. Instant neural graphics primitives with a multiresolution hash encoding. *TOG*, 2022.

[146] T. Nagarajan, C. Feichtenhofer, and K. Grauman. Grounded human-object interaction hotspots from video. In *ICCV*, 2019.

[147] T. Nagarajan, Y. Li, C. Feichtenhofer, and K. Grauman. Ego-topo: Environment affordances from egocentric video. In *CVPR*, 2020.

[148] T. Nguyen-Phuoc, C. Li, L. Theis, C. Richardt, and Y.-L. Yang. Hologan: Unsupervised learning of 3d representations from natural images. In *ICCV*, 2019.

[149] A. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. McGrew, I. Sutskever, and M. Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.

[150] H. Oh Song, Y. Xiang, S. Jegelka, and S. Savarese. Deep metric learning via lifted structured feature embedding. In *CVPR*, 2016.

[151] OpenAI. Gpt-4 technical report, 2023.

[152] J. Pan, C. C. Ferrer, K. McGuinness, N. E. O'Connor, J. Torres, E. Sayrol, and X. Giro-i Nieto. Salgan: Visual saliency prediction with generative adversarial networks. *arXiv*, 2017.

[153] P. Panteleris, I. Oikonomidis, and A. Argyros. Using a single rgb frame for real time 3d hand pose estimation in the wild. In *WACV*, 2018.

[154] J. J. Park, P. Florence, J. Straub, R. Newcombe, and S. Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *CVPR*, 2019.

[155] K. Park, U. Sinha, J. T. Barron, S. Bouaziz, D. B. Goldman, S. M. Seitz, and R. Martin-Brualla. Nerfies: Deformable neural radiance fields. *ICCV*, 2021.

[156] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *NeurIPS*, 2019.

[157] O. Patashnik, Z. Wu, E. Shechtman, D. Cohen-Or, and D. Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *ICCV*, 2021.

[158] A. Patel, A. Wang, I. Radosavovic, and J. Malik. Learning to imitate object interactions from internet videos. *arXiv:2211.13225*, 2022.

[159] G. Pavlakos, D. Shan, I. Radosavovic, A. Kanazawa, D. Fouhey, and J. Malik. Reconstructing hands in 3d with transformers. In *CVPR*, 2024.

[160] A. Pemasiri, K. N. Thanh, S. Sridharan, and C. Fookes. Im2mesh gan: Accurate 3d hand mesh recovery from a single rgb image. *arXiv*, 2021.

[161] B. Poole, A. Jain, J. T. Barron, and B. Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv*, 2022.

[162] D. Prattichizzo and J. C. Trinkle. Grasping. *Springer handbook of robotics*, 2016.

[163] W. Price, C. Vondrick, and D. Damen. Unweavenet: Unweaving activity stories. In *CVPR*, 2022.

[164] A. Pumarola, E. Corona, G. Pons-Moll, and F. Moreno-Noguer. D-nerf: Neural radiance fields for dynamic scenes. *arXiv preprint arXiv:2011.13961*, 2020.

[165] S. Purushwalkam, T. Ye, S. Gupta, and A. Gupta. Aligning videos in space and time. In *ECCV*, 2020.

[166] H. Qi, B. Yi, S. Suresh, M. Lambeta, Y. Ma, R. Calandra, and J. Malik. General in-hand object rotation with vision and touch. In *CoRL*, 2023.

[167] Y. Qin, Y.-H. Wu, S. Liu, H. Jiang, R. Yang, Y. Fu, and X. Wang. Dexmv: Imitation learning for dexterous manipulation from human videos. In *ECCV*, 2022.

[168] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

[169] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen. Hierarchical text-conditional image generation with CLIP latents. *arXiv*, 2022.

[170] J. Reizenstein, R. Shapovalov, P. Henzler, L. Sbordone, P. Labatut, and D. Novotny. Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In *ICCV*, 2021.

[171] L. G. Roberts. *Machine perception of three-dimensional solids*. PhD thesis, Massachusetts Institute of Technology, 1963.

[172] G. Rogez, M. Khademi, J. Supančič III, J. M. M. Montiel, and D. Ramanan. 3d hand pose detection in egocentric rgb-d images. In *ECCV*, 2014.

[173] G. Rogez, J. S. Supancic, and D. Ramanan. Understanding everyday hands in action from rgb-d images. In *ICCV*, 2015.

[174] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022.

[175] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022.

[176] J. Romero, H. Kjellström, and D. Kragic. Hands in action: real-time 3d reconstruction of hands in interaction with objects. In *ICRA*, 2010.

[177] J. Romero, D. Tzionas, and M. J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ToG*, 2017.

[178] J. Romero, D. Tzionas, and M. J. Black. Embodied hands: Modeling and capturing hands and bodies together. *SIGGRAPH Asia*, 36(6), November 2017.

[179] Y. Rong, T. Shiratori, and H. Joo. Frankmocap: A monocular 3d whole-body pose estimation system via regression and integration. In *ICCV Workshops*, 2021.

[180] Y. Rong, T. Shiratori, and H. Joo. Frankmocap: Fast monocular 3d hand and body motion capture by regression and integration. *ICCV Workshop*, 2021.

[181] Y. Rong, T. Shiratori, and H. Joo. Frankmocap: Fast monocular 3d hand and body motion capture by regression and integration. *ICCVW*, 2021.

[182] C. Saharia, W. Chan, H. Chang, C. Lee, J. Ho, T. Salimans, D. Fleet, and M. Norouzi. Palette: Image-to-image diffusion models. In *SIGGRAPH*, 2022.

[183] S. Saito, Z. Huang, R. Natsume, S. Morishima, A. Kanazawa, and H. Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. *ICCV*, 2019.

[184] A. Sauer, K. Schwarz, and A. Geiger. Stylegan-xl: Scaling stylegan to large diverse datasets. In *SIGGRAPH*, 2022.

[185] M. Seitzer. pytorch-fid: FID Score for PyTorch. `https://github.com/mseitzer/pytorch-fid`, August 2020.

[186] F. Sener, D. Chatterjee, D. Shelepov, K. He, D. Singhania, R. Wang, and A. Yao. Assembly101: A large-scale multi-view video dataset for understanding procedural activities. *CVPR*, 2022.

[187] D. Shan, J. Geng, M. Shu, and D. F. Fouhey. Understanding human hands in contact at internet scale. In *CVPR*, 2020.

[188] D. Shan, R. Higgins, and D. Fouhey. Cohesiv: Contrastive object and hand embedding segmentation in video. *NeurIPS*, 2021.

[189] D. Shan, R. Higgins, and D. Fouhey. Cohesiv: Contrastive object and hand embedding segmentation in video. *NeurIPS*, 2021.

[190] U. Singer, A. Polyak, T. Hayes, X. Yin, J. An, S. Zhang, Q. Hu, H. Yang, O. Ashual, O. Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *ICLR*, 2023.

[191] A. Sinha, J. Song, C. Meng, and S. Ermon. D2c: Diffusion-decoding models for few-shot conditional generation. *NeurIPS*, 2021.

[192] P. Sinha and E. Adelson. Recovering reflectance and illumination in a world of painted polyhedra. In *ICCV*, 1993.

[193] V. Sitzmann, M. Zollhöfer, and G. Wetzstein. Scene representation networks: Continuous 3d-structure-aware neural scene representations. In *NeurIPS*, 2019.

[194] J. Song, C. Meng, and S. Ermon. Denoising diffusion implicit models. In *ICLR*, 2021.

[195] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole. Score-based generative modeling through stochastic differential equations. In *ICLR*, 2021.

[196] E. H. Spriggs, F. De La Torre, and M. Hebert. Temporal segmentation and activity classification from first-person sensing. In *CVPRW*, 2009.

[197] S. Sridhar, F. Mueller, A. Oulasvirta, and C. Theobalt. Fast and robust hand tracking using detection-guided optimization. In *CVPR*, 2015.

[198] S. Sridhar, F. Mueller, M. Zollhöfer, D. Casas, A. Oulasvirta, and C. Theobalt. Real-time joint tracking of a hand manipulating an object from rgb-d input. In *ECCV*, 2016.

[199] O. Taheri, N. Ghorbani, M. J. Black, and D. Tzionas. Grab: A dataset of whole-body human grasping of objects. In *ECCV*, 2020.

[200] M. Tatarchenko, S. R. Richter, R. Ranftl, Z. Li, V. Koltun, and T. Brox. What do single-view 3d reconstruction networks learn? In *CVPR*, 2019.

[201] B. Tekin, F. Bogo, and M. Pollefeys. H+ o: Unified egocentric recognition of 3d hand-object poses and interactions. In *CVPR*, 2019.

[202] G. Tevet, S. Raab, B. Gordon, Y. Shafir, D. Cohen-or, and A. H. Bermano. Human motion diffusion model. In *ICLR*, 2023.

[203] S. Tulsiani, T. Zhou, A. A. Efros, and J. Malik. Multi-view supervision for single-view reconstruction via differentiable ray consistency. In *CVPR*, 2017.

[204] S. Tulsiani, T. Zhou, A. A. Efros, and J. Malik. Multi-view supervision for single-view reconstruction via differentiable ray consistency. *TPAMI*, 2019.

[205] D. Tzionas, L. Ballan, A. Srikantha, P. Aponte, M. Pollefeys, and J. Gall. Capturing hands in action using discriminative salient points and physics simulation. *IJCV*, 2016.

[206] D. Tzionas and J. Gall. 3d object reconstruction from hand-object interactions. In *ICCV*, 2015.

[207] A. Vahdat, K. Kreis, and J. Kautz. Score-based generative modeling in latent space. *NeurIPS*, 2021.

[208] A. Van Den Oord, O. Vinyals, et al. Neural discrete representation learning. *NeurIPS*, 2017.

[209] A. Vaswani, N. M. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. *NeurIPS*, 2017.

[210] P. Vincent. A connection between score matching and denoising autoencoders. *Neural computation*, 23(7), 2011.

[211] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

[212] C. Wang, H. Shi, W. Wang, R. Zhang, L. Fei-Fei, and C. K. Liu. Dexcap: Scalable and portable mocap data collection system for dexterous manipulation. *arXiv preprint arXiv:2403.07788*, 2024.

[213] H. Wang, X. Du, J. Li, R. A. Yeh, and G. Shakhnarovich. Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation. In *CVPR*, 2023.

[214] N. Wang, Y. Zhang, Z. Li, Y. Fu, W. Liu, and Y.-G. Jiang. Pixel2mesh: Generating 3d mesh models from single rgb images. In *ECCV*, 2018.

[215] P. Wang, L. Liu, Y. Liu, C. Theobalt, T. Komura, and W. Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *arXiv preprint arXiv:2106.10689*, 2021.

[216] X. Wang, Y. Ye, and A. Gupta. Zero-shot recognition via semantic embeddings and knowledge graphs. In *CVPR*, 2018.

[217] D. Warehouse. https://3dwarehouse.sketchup.com/. Accessed: 2019-11-15.

[218] B. Wen, J. Tremblay, V. Blukis, S. Tyree, T. Müller, A. Evans, D. Fox, J. Kautz, and S. Birchfield. Bundlesdf: Neural 6-dof tracking and 3d reconstruction of unknown objects. In *CVPR*, 2023.

[219] O. Wiles and A. Zisserman. Silnet: Single-and multi-view reconstruction by learning from silhouettes. *BMVC*, 2017.

[220] J. Wu, C. Zhang, T. Xue, W. T. Freeman, and J. B. Tenenbaum. Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. In *NeurIPS*, 2016.

[221] S. Wu, C. Rupprecht, and A. Vedaldi. Unsupervised learning of probably symmetric deformable 3d objects from images in the wild. In *CVPR*, 2020.

[222] Y.-H. Wu, J. Wang, and X. Wang. Learning generalizable dexterous manipulation from human grasp affordance. *CoRL*, 2022.

[223] Y. Wu and K. He. Group normalization. In *ECCV*, 2018.

[224] Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo, and R. Girshick. Detectron2. `https://github.com/facebookresearch/detectron2`, 2019.

[225] W. Xian, J.-B. Huang, J. Kopf, and C. Kim. Space-time neural irradiance fields for free-viewpoint video. In *CVPR*, 2021.

[226] Y. Xiang, R. Mottaghi, and S. Savarese. Beyond pascal: A benchmark for 3d object detection in the wild. In *WACV*, 2014.

[227] X. Yan, J. Yang, E. Yumer, Y. Guo, and H. Lee. Perspective transformer nets: Learning single-view 3d object reconstruction without 3d supervision. In *NeurIPS*, 2016.

[228] G. Yang, M. Vo, N. Neverova, D. Ramanan, A. Vedaldi, and H. Joo. Banmo: Building animatable 3d neural models from many casual videos. In *CVPR*, 2022.

[229] L. Yang, K. Li, X. Zhan, F. Wu, A. Xu, L. Liu, and C. Lu. OakInk: A large-scale knowledge repository for understanding hand-object interaction. In *CVPR*, 2022.

[230] L. Yariv, J. Gu, Y. Kasten, and Y. Lipman. Volume rendering of neural implicit surfaces. *NeurIPS*, 2021.

[231] Y. Ye, D. Gandhi, A. Gupta, and S. Tulsiani. Object-centric forward modeling for model predictive control. In *CoRL*, 2019.

[232] Y. Ye, A. Gupta, and S. Tulsiani. What's in your hands? 3d reconstruction of generic objects in hands. In *CVPR*, 2022.

[233] Y. Ye, P. Hebbar, A. Gupta, and S. Tulsiani. Diffusion-guided reconstruction of everyday hand-object interaction clips. In *ICCV*, 2023.

[234] Y. Ye, M. Singh, A. Gupta, and S. Tulsiani. Compositional video prediction. In *ICCV*, 2019.

[235] A. Yu, V. Ye, M. Tancik, and A. Kanazawa. pixelnerf: Neural radiance fields from one or few images. *CVPR*, 2021.

[236] J. Y. Zhang, S. Pepose, H. Joo, D. Ramanan, J. Malik, and A. Kanazawa. Perceiving 3d human-object spatial arrangements from a single image in the wild. In *ECCV*, 2020.

[237] L. Zhang, T. Wen, J. Min, J. Wang, D. Han, and J. Shi. Learning object placement by inpainting for compositional data augmentation. In *ECCV*, 2020.

[238] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.

[239] S. Zhang, Y. Zhang, Q. Ma, M. J. Black, and S. Tang. PLACE: Proximity learning of articulation and contact in 3D environments. In *3DV*, 2020.

[240] X. Zhang, Q. Li, H. Mo, W. Zhang, and W. Zheng. End-to-end hand mesh recovery from a monocular rgb image. In *ICCV*, 2019.

[241] Z. Zhao, W. S. Lee, and D. Hsu. Large language models as commonsense knowledge for large-scale task planning. *NeurIPS*, 2024.

[242] H. Zhen, X. Qiu, P. Chen, J. Yang, X. Yan, Y. Du, Y. Hong, and C. Gan. 3d-vla: A 3d vision-language-action generative world model. *arXiv preprint arXiv:2403.09631*, 2024.

[243] Y. Zhou, M. Habermann, W. Xu, I. Habibie, C. Theobalt, and F. Xu. Monocular real-time hand shape and motion capture using multi-modal data. In *CVPR*, 2020.

[244] Z. Zhou and S. Tulsiani. Sparsefusion: Distilling view-conditioned diffusion for 3d reconstruction. In *CVPR*, 2023.

[245] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, 2017.

[246] C. Zimmermann and T. Brox. Learning to estimate 3d hand pose from single rgb images. In *ICCV*, 2017.

[247] B. Çalli, A. Singh, A. Walsman, S. S. Srinivasa, P. Abbeel, and A. M. Dollar. The ycb object and model set: Towards common benchmarks for manipulation research. *ICAR*, 2015.