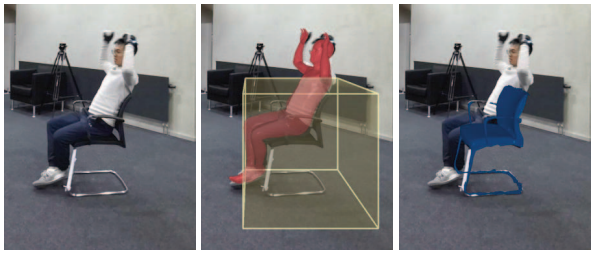
Reconstruction of 3D Interaction Models from Images using Shape Prior

Mehrshad Mirmohammadi[1] Parham Saremi[1] Yen-Ling Kuo[2] Xi Wang [3]
[1] Sharif University of Technology [2]University of Virginia [3]ETH Zürich

Abstract

We investigate the reconstruction of 3D human-object interactions from images, encompassing 3D human shape and pose estimation as well as object shape and pose estimation. To address this task, we introduce an autoregressive transformer-based variational autoencoder capable of learning a robust shape prior from extensive 3D shape datasets. Additionally, we leverage the reconstructed 3D human body as supplementary features for object shape and pose estimation. In contrast, prior methods only predict object pose and rely on shape templates for shape prediction. Experimental findings on the BEHAVE dataset underscore the effectiveness of our proposed approach, achieving a 40.7cm Chamfer distance and demonstrating the advantages of learning a shape prior.



(a) (b) (c)

Figure 1. **Reconstructing 3D human-object interaction models from images.** We propose a framework that reconstructs 3D models of a human interacting with an object. (a) An RGB image is taken as input. (b) Our method first reconstructs a 3D human using the off-the-shelf method and predicts a 3D bounding box of the object of interest. (c) Using a learned shape prior, our method reconstructs the object shapes and poses given the constraints posed by the 3D humans.

1. Introduction

Modeling 3D human-object interactions is an important task because humans live in a natural environment and actively engage with objects in their environment. Being able to understand how humans interact with scenes will further help us in gaining an understanding of how they perform a particular task. A successful development of methods that can model human-object interactions will be central to advancing research in human-centered AI and in advancing the state-of-the-art in computer vision, computer graphics, and human-computer interaction.

Despite decades of research on human body modeling, human-object interaction is still a challenging task, partially due to the absence of large-scale 3D data. Developing an approach that can extract useful information from largely available 2D images holds the promise to provide more accessible and generalizable ways to obtain rich diversity in terms of objects and interaction types.

We introduce a new approach that reconstructs 3D human-object interaction models from images using a shape prior learned from a large amount of 3D shapes. Existing works either require manual selection of 3D object templates [46] or assume that the object shape is known [37], and therefore do not scale well towards real-world data. Instead, we propose a transformer-based variational autoencoder that leverages a shape prior learned through training on existing large datasets of 3D shapes. Our method takes images as input and learns to reconstruct 3D human-object interaction models by jointly reasoning about their shapes and poses in a shared space (see Fig. 1). Our approach builds on the insight that 1) body poses can be reliably reconstructed via off-the-shelf methods; 2) body poses are highly predictive of the object categories that humans interact with, 3) and human and object shape and pose can be jointly optimized when leveraging the underlying interaction between humans and objects. More specifically, we parametrize the objects by implicit signed distance fields (SDFs) and learn a generic distribution over 3D shapes that can be used as a prior at interference time, rendering the requirement of known object templates unnecessary. The estimated 3D body is encoded as additional features, providing spatial constraints.

We evaluate our method on the BEHAVE [2] dataset and compare it against state-of-the-art methods. Our experimental results demonstrate that our proposed method achieves competitive performance, even when tackling more generalized and challenging tasks. Moreover, our approach significantly enhances reconstruction performance

by effectively leveraging the learned shape prior when provided with accurate 3D bounding boxes.

2. Related Work

Prior work on human-object interaction modeling has explored ways to reconstruct 3D objects and humans separately, and recently, start to reconstruct them jointly by considering their interaction.

Reconstruct 3D human from single images. Many existing methods for estimating 3D human poses and shapes (HPS) directly predict the SMPL parameters from images [24]. SMPLify [3] is the first method that fits the SMPL model to the detected 2D keypoints. Lassner et al. [23] extend the method by considering both silhouettes and keypoints during fitting. Recently, many deep neural network-based architectures aim to regress human poses and shapes from pixels [9, 13, 20, 29, 41, 48]. To deal with the lack of annotations of in-the-wild images for HPS, methods like HMR [20] employ a reprojection loss of keypoints as weak supervision and SPEC [22] estimates camera parameters to improve reconstruction by camera calibration. Several approaches also perform coarse-to-fine or iterative refinement of HPS estimation. TetraTSDF [29] regresses a coarse SMPL first and then builds the outer shell based on the SMPL model. DeepHuman [48] performs parametric estimation first and then refines the normal map. Considering the granularity of features, PyMaf [44] creates a mesh pyramid feature from the input image and iteratively improves the meshes. THUNDR [41] adopts transformers for iterative refinement. Body parts can provide information to adjacent parts for reconstruction. HoloPose [14] proposes a part-based architecture for parameter regression. PARE [21] explores a soft-attention mechanism and guides the attention by visible parts, which improves the estimation of the occluded parts. Beyond the parametric body models, several approaches such as IPNET [1] and ICON [38] use implicit functions (IFs) to represent fine shape details and varied topology. They combine parametric models and IFs to leverage the best of both worlds.

Reconstruct 3D object from single images. Estimating the 3D poses and shapes of objects from a single RGB image [16] is a challenging task as there are shape ambiguities given a single object view. Several representations are explored and used to reconstruct 3D objects including voxels [8, 12, 30, 35], point clouds [11, 25, 36], meshes [32, 33], implicit 3D surfaces such as SDFs [19, 27, 39] and UDFs [7], or function space of 3D surface [26]. To improve reconstruction, previous works also consider various priors. TARS3D [10] learns a category-specific prior that represents the topology of different object categories. UNICORN [28] explicitly adds a loss enforcing consistency between instances having similar shapes or textures. In Ye et al. [40], they use hand articulation as prior as it is highly relevant to the shape of the hand-held objects. But many of these approaches generate only one shape for input or do category-specific optimization. To improve generalizability, AutoSDF [27] and 3DILG [43] model the distribution over 3D shapes to generate multiple plausible outputs. ss3d [31] pretrains a reconstruction model using multi-view renderings of synthetic data, allowing the model to benefit from the common structure across categories. With the success of generative models, several approaches use GAN-based models [47], denoising diffusion-based models [6, 42], or NeRF-based models [5] for 3D shape generation.

Reconstruct 3D interaction from single images. Prior works use the 3D scene information [15] and contacts heuristics [46] between humans and objects to reconstruct their 3D spatial arrangements. PROX [15] models human-scene interactions by considering that a) two objects cannot interpenetrate each other and that b) physical interactions require contact. However, their approach requires 3D scene information. PHOSA [46], on the other side, reconstructs 3D humans and objects separately. It then applies hand-crafted heuristics, such as manually defined object contact regions, to improve the 3D reconstruction of both humans and objects. Nevertheless, it is not possible to define all the possible human-object contact regions in advance. Most recent works try to remove the hand-crafted heuristics by using datasets designed explicitly for this purpose [2, 17] or querying large language models to retrieve possible contact pairs [34]. In particular, CHORE [37] uses the BEHAVE [2] dataset to design the first end-to-end learning-based approach to jointly reconstruct 3D humans, objects and contacts from a single image. They assume a know object mesh template as input. In contrast, our approach directly estimates the 3D object mesh from the input image. Wang et al. [34] use a two-stage optimization technique and infer the action type from the human pose and use it to query a language model to recover contact information between humans and objects. CHAIRS [18] avoid using heuristics for human-object contacts by learning an interaction prior. However, this model is trained on a dataset that only includes sittable objects(e.g. chairs, sofas, stools, and benches), not allowing generalization to new objects. Finally, neural dome [45] exploits multi-view images of the same scene to mitigate problems such as occlusion and shape ambiguities. Nevertheless, obtaining multi-view images in real-world applications is challenging, limiting the generalization ability of this kind of approach.

3. Method

We introduce a method for detailed human-object interaction reconstruction from single images. See Fig. 2 for a schematic overview of our method. Reconstructing 3D interaction models from RGB images is a challenging task. Solving it requires 1) accurately estimating humans and ob-

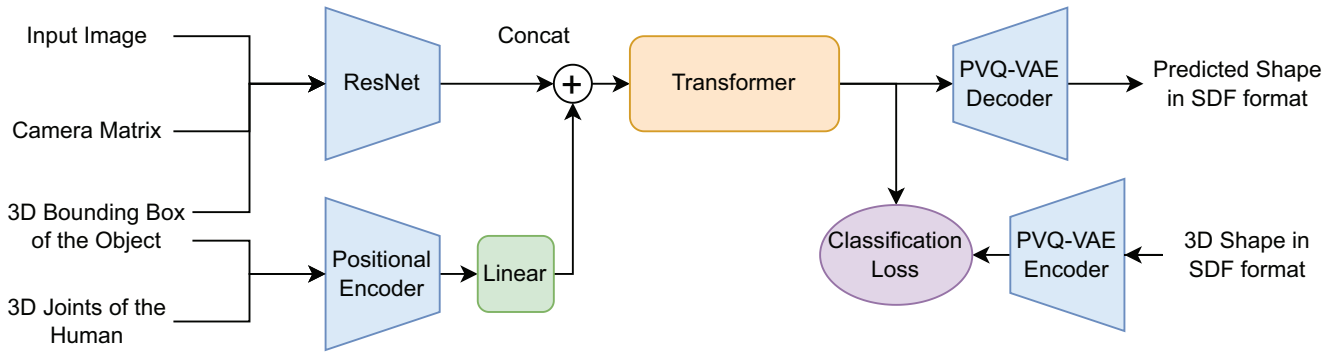


Figure 2. **Overview of our proposed method.** Our method aims to reconstruct the interior SDF values of a 3D bounding box by leveraging both the input image and the predicted human pose. This involves subdividing the bounding box into cells and extracting visual and human-pose encodings for each cell. The core of our approach lies in the implementation of an Autoregressive Transformer architecture, enabling us to predict representations for these cells. These representations are then decoded using the PVQ-VAE decoder module [27]. It's worth noting that the transformer predicts the vector quantized representation for each cell, thus we approach this task in a classification setting.

jects from images with arbitrary backgrounds, which includes estimating both shapes and poses and 2) reconstructing detailed physically plausible 3D interaction models that match the image observations. In contrast to prior works that rely on off-the-shelf object detection and segmentation tools or knowledge of object shapes, we solve the task directly in 3D space. To achieve this, we learn a shape prior for 3D shapes from large 3D data of shapes and reconstruct the 3D interaction model by leveraging the learned shape prior. This allows us to deal with diverse object shapes and categories.

We first jointly estimate the 3D bounding boxes of both humans and objects (Sec. 3.1). We then reconstruct a 3D human mesh using an off-the-shelf approach and use it as a constraint for reconstructing the object shape and pose (Sec. 3.2).

3.1. Estimate 3D bounding boxes

Recognizing objects in 3D from a single image is a fundamental and challenging task of computer vision. In order for our model to correctly reconstruct the object mesh, it is crucial to first localize it accurately in the 3D space. In particular, our method needs to localize the object interacting with the human. For this reason, we reformulate the object detection task as a 3D human-object interaction (HOI) detection. To accomplish this task, we adapt the off-the-shelf method [4] to detect human-objects interactions. More specifically, we train Cube R-CNN [4] using as supervision only the bounding box of the object interacting with the human. Furthermore, we also train the detection model to predict human bounding boxes to introduce a strong interaction prior. Our primary assumption is that the human location already gives essential cues about the object's position in the 3D space. Consequently, the joint estimation of human and object 3D locations allows the model to reason about their interaction and improves the accuracy of the 3D object location estimation.

3.2. Reconstruct 3D shape and pose

Given an image, we first reconstruct the SMPL body model [24] of the person. We use an off-the-shelf 3D human reconstruction method [21] to regress the body pose and shape parameters, and the 3D joint positions can be derived from the SMPL model.

With the 3D bounding boxes acquired in the previous step, we aim to reconstruct the object confined with the box, with constraints on translation and scale. Inspired by AutoSDF [27], we use patch-wised Vector Quantized Variational AutoEncoder, PVQ-VAE for short, to reconstruct the 3D object. PVQ-VAE learns quantized vector representations for 3D space using signed distance fields (SDF). We divide the 3D bounding box into $k \times k \times k$ cells and encode each cell independently. The embedding vector is then mapped to the nearest representation in the codebook $\mathcal{Z}$ of size $\mathbf{N}$, which is jointly learned during training. In the end, each patch is represented by an integer in the range of [1, ..., $\mathbf{N}$]. A decoder jointly decodes all cells to output the reconstructed 3D object. We utilize a pre-trained PVQ-VAE model on ShapeNet. This enables us to establish a robust 3D shape prior to three-dimensional structures and reconstruct objects even when occlusions are present. This is due to the fact that the occlusion of a portion of the image leads to a less precise representation of the affected patches. Nonetheless, the decoder's training with an extensive range of 3D shapes allows it to extrapolate the 3D shape from the more precise patches, compensating for the occluded regions. In addition, the cell-based formulation allows us to efficiently extract relevant information from images.

More specifically, we encode each cell using ResNet and concatenate them together to have a single feature map as the representation of the image. Motivated by the idea that body poses are predictive of the objects that humans inter-

act with, we consider the relative distances to the human body as additional constraints. Similar to the previous work on hand-object reconstruction [40], we calculate the relative distance between the cell center and the J body joints and encode them with trigonometric functions before passing through a linear layer. In the end, each cell is represented by the combination of the image encoding and the encoding of the human pose.

3.3. Implementation details

We implemented our code in PyTorch and leveraged certain components from the work of AutoSDF [27]. Our training process begins with a preprocessing step, where we prepare all the images along with their corresponding meshes. During this stage, we determine the 3D bounding box for each mesh and compute the Signed Distance Field (SDF) values for the interior of these boxes. Subsequently, we utilize the trained AutoSDF Encoder to encode the interior regions of each bounding box.

The training of our proposed method involves using the calculated bounding boxes and human meshes obtained from ground truth data. For inference, we rely on the predicted bounding boxes generated by Cube R-CNN [4] and the predicted human mesh from PARE [21].

Additionally, it is essential to ensure that all predicted entities exist in the same coordinate system and maintain consistency with each other. While Cube R-CNN [4] takes the camera's projection matrix as input and produces results aligned with this camera, PARE predicts outputs in its coordinate system, using its own camera parameters. Hence, a crucial step involves aligning these two systems together.

To achieve the alignment, we employ optimization methods to determine suitable rotation and translation transformations. These transformations are carefully calculated such that when applied to the output of PARE [21] and subsequently projected using the desired camera matrix, they yield the same projection as the original output obtained with the initial camera parameters. By employing this process, we can effectively align PARE's output with our original camera system without the need for any additional supervision.

4. Experiments

4.1. Dataset

We evaluate our proposed model on the BEHAVE [2] dataset. It captures full-body human-object interactions and consists of multi-view RGB-D video frames of people interacting with objects in diverse ways. The corresponding 3D SMPL body models, object shapes and poses, and the annotated contacts between human bodies and objects are provided in BEHAVE. There are 15k frames in total where humans interact with 20 common objects.

4.2. Evaluation metrics

Following the standard evaluation protocol [2, 34, 37], we first align the reconstructed SMPL models to the ground truth and apply the same alignment transformations to the objects. We then compute the two-way Chamfer distance on objects. All the reported numbers are in centimeters.

4.3. Comparison with state-of-the-art

We compared our method with two state-of-the-art approaches: CHORE [37] and PHOSA [46]. Note that these methods only predict the object's pose and rely on templates for shape prediction. In contrast, our approach overcomes this limitation and excels in predicting both the object's pose and shape, handling a more challenging and general problem. The results in Table 1 show that our method performs competitively with prior methods. Figure 3 shows several examples of predicted 3D bounding boxes and Figure 4 shows several reconstruction results compared to ground truth data across various object categories.

Dataset	Methods	Object Chamfer ↓
BEHAVE[2]	PHOSA [46]	26.62 ± 21.87
	CHORE [37]	10.66 ± 7.71
	Ours	40.7 ± 54.9

Table 1. **Comparison to the state-of-the-art methods on BEHAVE.** We compare our method to state-of-the-art optimisation-based method PHOSA [46] and learning-based method CHORE [37]. Note that templates of object shape are used as input in both PHOSA and CHORE where our approach estimates both object shapes and poses.

4.4. Ablation

B-Box Mode	Visibility	Object Chamfer ↓
GT	All	4.3 ± 3.3
	$> 30\%$	4.1 ± 3.1
	$< 30\%$	7.4 ± 5.0
Predicted	All	40.7 ± 54.9
	$> 30\%$	39.3 ± 54.4
	$< 30\%$	65.5 ± 58.6

Table 2. **Ablation studies on BEHAVE [2] dataset** for different B-Box modes and visibility specific results.

To evaluate the impact of the bounding boxes' qualities on the generated outputs, we evaluated our method using both ground-truth (GT) bounding boxes and Cube R-CNN [4] generated bounding boxes (Predicted). The results are shown in Table 2.

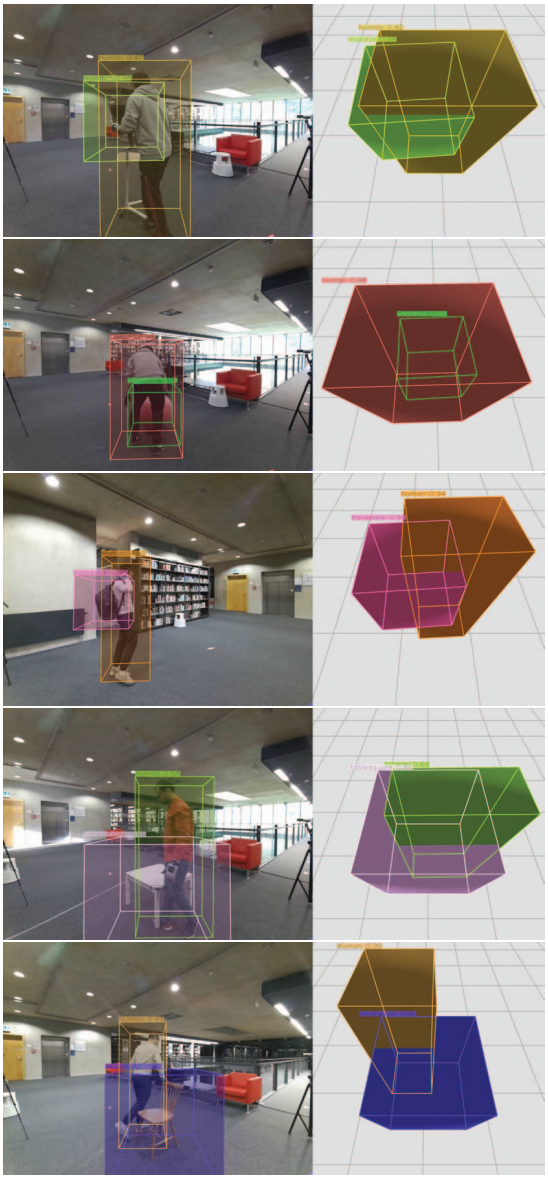


Figure 3. **Examples of predicted 3D bounding boxes on the BEHAVE dataset.** The predicted bounding boxes are projected in the camera view on the left and from the top on the right.

As indicated in the table, utilizing more accurate ground-truth bounding boxes results in a substantial improvement in 3D object reconstruction quality compared to using predicted bounding boxes. Notably, the chamfer distance decreases significantly, from 40.7 cm to 4.3 cm, upon switching from predicted to ground-truth bounding boxes. This observation suggests that our shape prior exhibits sufficient strength in reconstructing object shapes, with most of the error in the Chamfer distance arising from inaccurate pose predictions.

Furthermore, by only considering objects in a specific range of visibility, it is understood that the visibility of the object has a significant impact on the quality of the generated output. Based on Table 2, we can see that by only considering objects with less than 30% visibility, the chamfer distance is increased by 50% compared to only considering objects with visibility greater than 30%. Note that both state-of-the-art methods CHORE [37] and PHOSA [46] struggle to reconstruct anything meaning, or fail completely with images where objects are less than 30% visible. In fact, CHORE skips those images while PHOSA would mostly fails at the first stage to detect the corresponding object.

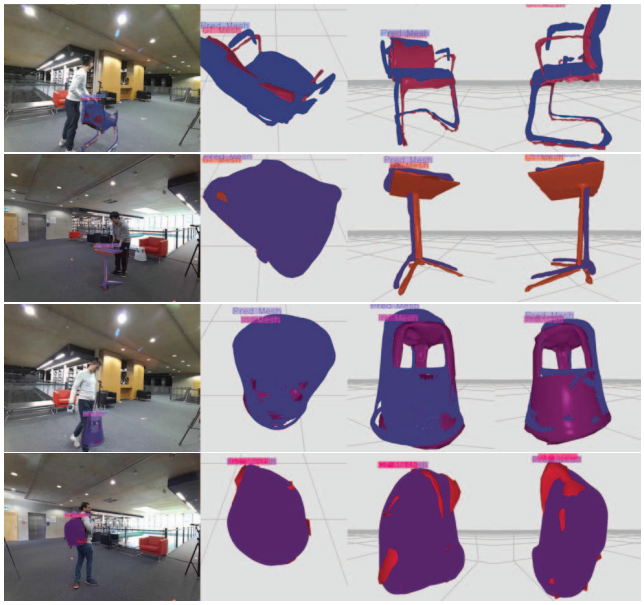


Figure 4. **Examples of reconstructed objects on the BEHAVE dataset.** We compare the reconstructed (Pred) meshes (blue) to the ground-truth (GT) meshes (red). Input images with overlaid objects are shown on the left and on the right we see the 3D objects projected in three different angles.

5. Conclusion

In this work, we present a method to reconstruct 3D human-object interaction models from images. Our method leverages a strong shape prior, learned from large datasets of 3D shapes, for object shape and pose reconstruction. We use neural implicit representations for object reconstruction and consider reconstructed body pose as additional constraints. Experiments on the BEHAVE dataset demonstrate the effectiveness of the proposed method which is capable of reconstructing both 3D humans and objects from images, in contrast to existing works that consider a preselected set of object shapes or assume known object shapes.

Acknowledgements This work was supported by an ETH Zurich Postdoctoral Fellowship and an ETH Career Seed Awards. We thank Alessandro Ruzzi for his help with the experiments.

References

[1] Bharat Lal Bhatnagar, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Combining implicit function learning and parametric models for 3d human reconstruction. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 311–329. Springer, 2020. 2

[2] Bharat Lal Bhatnagar, Xianghui Xie, Ilya A Petrov, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Behave: Dataset and method for tracking human object interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15935–15946, 2022. 1, 2, 4

[3] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*, pages 561–578. Springer, 2016. 2

[4] Garrick Brazil, Julian Straub, Nikhila Ravi, Justin Johnson, and Georgia Gkioxari. Omni3d: A large benchmark and model for 3d object detection in the wild. *arXiv preprint arXiv:2207.10660*, 2022. 3, 4

[5] Hanzhi Chen, Fabian Manhardt, Nassir Navab, and Benjamin Busam. Texpose: Neural texture learning for self-supervised 6d object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4841–4852, 2023. 2

[6] Yen-Chi Cheng, Hsin-Ying Lee, Sergey Tulyakov, Alexander Schwing, and Liangyan Gui. Sdfusion: Multimodal 3d shape completion, reconstruction, and generation. *arXiv preprint arXiv:2212.04493*, 2022. 2

[7] Julian Chibane, Gerard Pons-Moll, et al. Neural unsigned distance fields for implicit function learning. *Advances in Neural Information Processing Systems*, 33:21638–21652, 2020. 2

[8] Christopher B Choy, Danfei Xu, JunYoung Gwak, Kevin Chen, and Silvio Savarese. 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14*, pages 628–644. Springer, 2016. 2

[9] Hai Ci, Mingdong Wu, Wentao Zhu, Xiaoxuan Ma, Hao Dong, Fangwei Zhong, and Yizhou Wang. Gfpose: Learning 3d human pose prior with gradient fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4800–4810, 2023. 2

[10] Shivam Duggal and Deepak Pathak. Topologically-aware deformation fields for single-view 3d reconstruction. *CVPR*, 2022. 2

[11] Haoqiang Fan, Hao Su, and Leonidas J Guibas. A point set generation network for 3d object reconstruction from a single image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 605–613, 2017. 2

[12] Rohit Girdhar, David F Fouhey, Mikel Rodriguez, and Abhinav Gupta. Learning a predictable and generative vector representation for objects. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI 14*, pages 484–499. Springer, 2016. 2

[13] Jia Gong, Lin Geng Foo, Zhipeng Fan, Qiuhong Ke, Hossein Rahmani, and Jun Liu. Diffpose: Toward more reliable 3d pose estimation. *arXiv preprint arXiv:2211.16940*, 2022. 2

[14] Riza Alp Guler and Iasonas Kokkinos. Holopose: Holistic 3d human reconstruction in-the-wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10884–10894, 2019. 2

[15] Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael J Black. Resolving 3d human pose ambiguities with 3d scene constraints. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2282–2292, 2019. 2

[16] Tomáš Hodaň, Martin Sundermeyer, Bertram Drost, Yann Labbé, Eric Brachmann, Frank Michel, Carsten Rother, and Jiří Matas. Bop challenge 2020 on 6d object localization. In *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 577–594. Springer, 2020. 2

[17] Yinghao Huang, Omid Taheri, Michael J Black, and Dimitrios Tzionas. Intercap: Joint markerless 3d tracking of humans and objects in interaction. In *Pattern Recognition: 44th DAGM German Conference, DAGM GCPR 2022, Konstanz, Germany, September 27–30, 2022, Proceedings*, pages 281–299. Springer, 2022. 2

[18] Nan Jiang, Tengyu Liu, Zhexuan Cao, Jieming Cui, Yixin Chen, He Wang, Yixin Zhu, and Siyuan Huang. Chairs: Towards full-body articulated human-object interaction, 2022. 2

[19] Yue Jiang, Dantong Ji, Zhizhong Han, and Matthias Zwicker. Sdfdiff: Differentiable rendering of signed distance fields for 3d shape optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1251–1261, 2020. 2

[20] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7122–7131, 2018. 2

[21] Muhammed Kocabas, Chun-Hao P Huang, Otmar Hilliges, and Michael J Black. Pare: Part attention regressor for 3d human body estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11127–11137, 2021. 2, 3, 4

[22] Muhammed Kocabas, Chun-Hao P. Huang, Joachim Tesch, Lea Müller, Otmar Hilliges, and Michael J. Black. SPEC: Seeing people in the wild with an estimated camera. In *Proceedings International Conference on Computer Vision (ICCV)*, pages 11035–11045. IEEE, Oct. 2021. 2

[23] Christoph Lassner, Javier Romero, Martin Kiefel, Federica Bogo, Michael J. Black, and Peter V. Gehler. Unite the people: Closing the loop between 3d and 2d human representations. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 2

[24] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-

person linear model. *ACM transactions on graphics (TOG)*, 34(6):1–16, 2015. 2, 3

[25] Priyanka Mandikal, KL Navaneet, Mayank Agarwal, and R Venkatesh Babu. 3d-lmnet: Latent embedding matching for accurate and diverse 3d point cloud reconstruction from a single image. *arXiv preprint arXiv:1807.07796*, 2018. 2

[26] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4460–4470, 2019. 2

[27] Paritosh Mittal, Yen-Chi Cheng, Maneesh Singh, and Shubham Tulsiani. Autosdf: Shape priors for 3d completion, reconstruction and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 306–315, 2022. 2, 3, 4

[28] Tom Monnier, Matthew Fisher, Alexei A. Efros, and Mathieu Aubry. Share With Thy Neighbors: Single-View Reconstruction by Cross-Instance Consistency. In *ECCV*, 2022. 2

[29] Hayato Onizuka, Zehra Hayirci, Diego Thomas, Akihiro Sugimoto, Hideaki Uchiyama, and Rin-ichiro Taniguchi. Tetratsdf: 3d human reconstruction from a single image with a tetrahedral outer shell. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2

[30] Shubham Tulsiani, Tinghui Zhou, Alexei A Efros, and Jitendra Malik. Multi-view supervision for single-view reconstruction via differentiable ray consistency. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2626–2634, 2017. 2

[31] Kalyan Alwala Vasudev, Abhinav Gupta, and Shubham Tulsiani. Pre-train, self-train, distill: A simple recipe for supersizing 3d reconstruction. In *Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[32] Nanyang Wang, Yinda Zhang, Zhuwen Li, Yanwei Fu, Wei Liu, and Yu-Gang Jiang. Pixel2mesh: Generating 3d mesh models from single rgb images. In *Proceedings of the European conference on computer vision (ECCV)*, pages 52–67, 2018. 2

[33] Weiyue Wang, Duygu Ceylan, Radomir Mech, and Ulrich Neumann. 3dn: 3d deformation network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1038–1046, 2019. 2

[34] Xi Wang, Gen Li, Yen-Ling Kuo, Muhammed Kocabas, Emre Aksan, and Otmar Hilliges. Reconstructing action-conditioned human-object interactions using commonsense knowledge priors. *arXiv preprint arXiv:2209.02485*, 2022. 2, 4

[35] Jiajun Wu, Chengkai Zhang, Xiuming Zhang, Zhoutong Zhang, William T Freeman, and Joshua B Tenenbaum. Learning shape priors for single-view 3d completion and reconstruction. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 646–662, 2018. 2

[36] Rundi Wu, Yixin Zhuang, Kai Xu, Hao Zhang, and Baoquan Chen. Pq-net: A generative part seq2seq network for 3d shapes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 829–838, 2020. 2

[37] Xianghui Xie, Bharat Lal Bhatnagar, and Gerard Pons-Moll. Chore: Contact, human and object reconstruction from a single rgb image. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part II*, pages 125–145. Springer, 2022. 1, 2, 4, 5

[38] Yuliang Xiu, Jinlong Yang, Dimitrios Tzionas, and Michael J. Black. ICON: Implicit Clothed humans Obtained from Normals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13296–13306, June 2022. 2

[39] Qiangeng Xu, Weiyue Wang, Duygu Ceylan, Radomir Mech, and Ulrich Neumann. Disn: Deep implicit surface network for high-quality single-view 3d reconstruction. *Advances in neural information processing systems*, 32, 2019. 2

[40] Yufei Ye, Abhinav Gupta, and Shubham Tulsiani. What's in your hands? 3d reconstruction of generic objects in hands. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3895–3905, June 2022. 2, 4

[41] Mihai Zanfir, Andrei Zanfir, Eduard Gabriel Bazavan, William T Freeman, Rahul Sukthankar, and Cristian Sminchisescu. Thundr: Transformer-based 3d human reconstruction with markers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12971–12980, 2021. 2

[42] Xiaohui Zeng, Arash Vahdat, Francis Williams, Zan Gojcic, Or Litany, Sanja Fidler, and Karsten Kreis. Lion: Latent point diffusion models for 3d shape generation. *arXiv preprint arXiv:2210.06978*, 2022. 2

[43] Biao Zhang, Matthias Nießner, and Peter Wonka. 3dilg: Irregular latent grids for 3d generative modeling. *arXiv preprint arXiv:2205.13914*, 2022. 2

[44] Hongwen Zhang, Yating Tian, Xinchi Zhou, Wanli Ouyang, Yebin Liu, Limin Wang, and Zhenan Sun. Pymaf: 3d human pose and shape regression with pyramidal mesh alignment feedback loop. In *Proceedings of the IEEE International Conference on Computer Vision*, 2021. 2

[45] Juze Zhang, Haimin Luo, Hongdi Yang, Xinru Xu, Qianyang Wu, Ye Shi, Jingyi Yu, Lan Xu, and Jingya Wang. Neuraldome: A neural modeling pipeline on multi-view human-object interactions, 2022. 2

[46] Jason Y Zhang, Sam Pepose, Hanbyul Joo, Deva Ramanan, Jitendra Malik, and Angjoo Kanazawa. Perceiving 3d human-object spatial arrangements from a single image in the wild. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16*, pages 34–51. Springer, 2020. 1, 2, 4, 5

[47] Xin-Yang Zheng, Yang Liu, Peng-Shuai Wang, and Xin Tong. Sdf-stylegan: Implicit sdf-based stylegan for 3d shape generation. In *Comput. Graph. Forum (SGP)*, 2022. 2

[48] Zerong Zheng, Tao Yu, Yixuan Wei, Qionghai Dai, and Yebin Liu. Deephuman: 3d human reconstruction from a single image. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019. 2