InterTrack: Tracking Human Object Interaction without Object Templates

Xianghui Xie[1,2,3] Jan Eric Lenssen[3] Gerard Pons-Moll[1,2,3]

[1]University of Tübingen, Germany [2]Tübingen AI Center, Germany
[3]Max Planck Institute for Informatics, Saarland Informatic Campus, Germany
https://virtualhumans.mpi-inf.mpg.de/InterTrack/

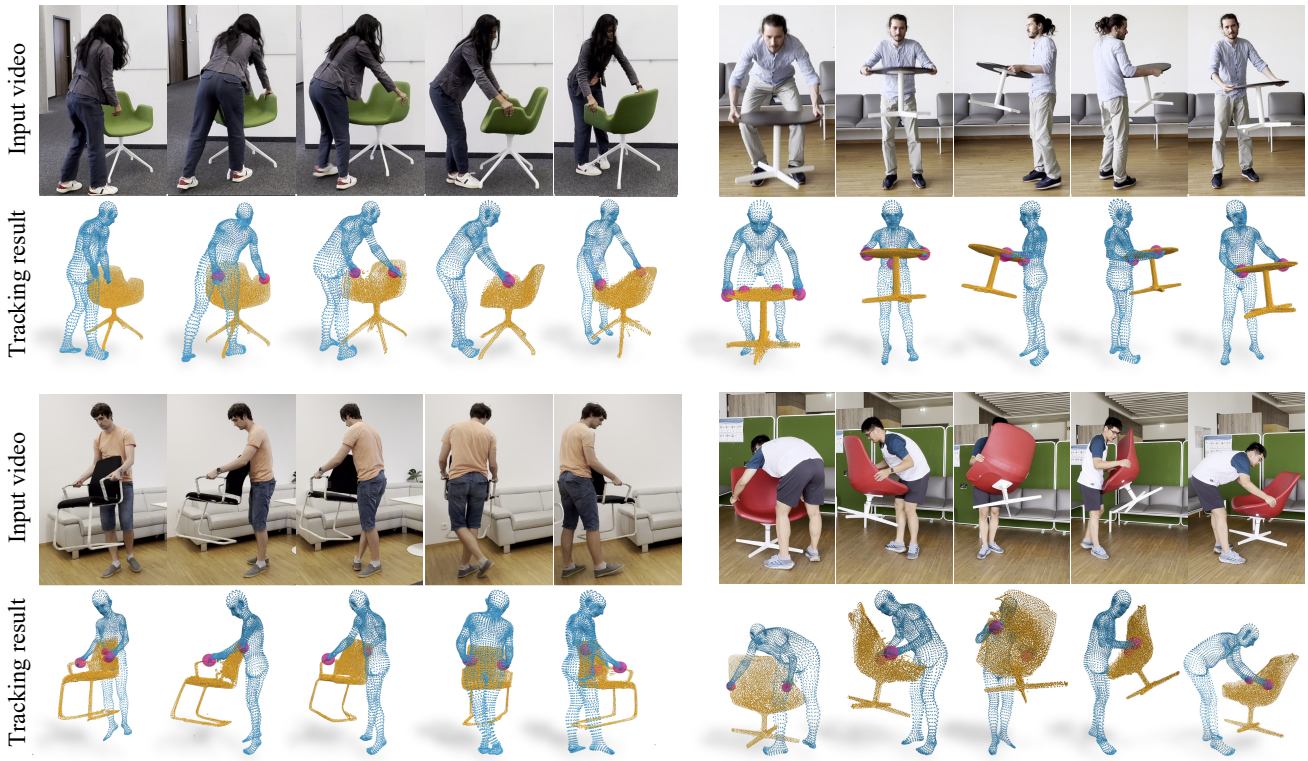

Figure 1. From a monocular RGB video, our method tracks the human and object under occlusion and dynamic motions, without using any object templates. Our method is trained *only* on synthetic data and generalizes well to real-world videos captured by mobile phones.

Abstract

Tracking human object interaction from videos is important to understand human behavior from the rapidly growing stream of video data. Previous video-based methods require predefined object templates while single-image-based methods are template-free but lack temporal consistency. In this paper, we present a method to track human object interaction without any object shape templates. We decompose the 4D tracking problem into per-frame pose tracking and canonical shape optimization. We first apply a single-view reconstruction method to obtain temporally-inconsistent per-frame interaction reconstructions. Then, for the human, we propose an efficient autoencoder to predict SMPL vertices directly from the per-frame reconstructions, introducing temporally consistent correspondence. For the object, we introduce a pose estimator that leverages temporal information to predict smooth object rotations under occlusions. To train our model, we propose a method to generate synthetic interaction videos and synthesize in total 10 hour videos of 8.3k sequences with full 3D ground truth. Experiments on BEHAVE and InterCap show that our method significantly outperforms previous template-based video tracking and single-frame reconstruction methods.

Our proposed synthetic video dataset also allows training video-based methods that generalize to real-world videos. Our code and dataset are publicly released.

1. Introduction

Jointly reconstructing humans and objects is an important task to understand humans and their interaction with the environment. In this paper, we address the problem of tracking human object interaction from a single RGB camera, without any object templates. This is a very challenging task due to depth scale ambiguity, heavy occlusions and dynamic human object motions. Moreover, the human and object poses as well as shapes need to be estimated simultaneously as no prior templates are given.

Earlier work VisTracker [78] pioneered monocular interaction tracking by reasoning about the occluded object using visible frames and human information. However, they rely on pre-defined object templates which limits its applicability to general scenarios. Furthermore, they use per-frame object pose estimator with ad-hoc smoothing to reason occluded objects, which does not fully explore temporal information. More recently, HDM [79] proposed a template-free approach to reconstruct human and object from single images. Their model trained only on synthetic data shows strong generalization ability to real-world images. However, this method fails under heavy occlusions and produces inconsistent shapes across frames.

To address these challenges, we propose **InterTrack**, interaction tracking without object templates. We decompose the 4D video reconstruction problem into per-frame pose estimation and global shape optimization. This decomposition greatly constraints the solution space of 4D reconstruction, making the problem tractable. Specifically, we start with the HDM [79] per-frame reconstructions which provide initial 3D human and object point clouds that are inconsistent across frames. For the human, we propose a simple and efficient autoencoder, **CorrAE**, which directly predicts the SMPL [44] vertices that are aligned with the HDM human reconstructions. This allows us to obtain disentangled SMPL pose and shape parameters for temporally consistent optimization and introduces correspondence over time. For the object, we introduce **TOPNet**, which leverages temporal information to predict object rotation from monocular RGB video. The temporal design allows accurate object pose prediction even under heavy occlusions. With the predicted rotations, we can then optimize the object shape in canonical space as well as per-frame pose transformations, leading to temporally consistent tracking. Last, we jointly optimize human and object based on the predicted contacts, leading to more plausible interactions.

Prior works [77, 78] train their methods on real data and test on same set of object instances, which limits the generalization to new objects [79]. Recent work ProciGen [79] proposed a synthetic interaction dataset with 1M images and 21k different object shapes. However, ProciGen contains only static frames, making it impossible to train video-based methods. To this end, we propose **ProciGen-Video**, a method to generate synthetic interaction videos and we generate 8.5k videos of 10 categories paired with full 3D ground truth. This dataset allows us to train our object pose estimator TOPNet that generalizes to real videos.

We evaluate our method on BEHAVE [6] and InterCap [27] dataset. Experiments show that our method significantly outperforms VisTracker [78] (which requires template) and HDM [79]. Our ablation shows that our CorrAE achieves similar performance compared to SoTA human registration method NICP [46] but is 30 times faster, and our proposed TOPNet works significantly better than prior category-level pose estimator CenterPose [41]. Results also show that pre-training on our ProciGen-Video dataset helps boost the performance and our model trained on synthetic ProciGen-Video generalizes to real videos.

In summary, our key contributions are:

- We propose InterTrack, the first method to track full-body dynamic object interaction from monocular RGB videos without object templates.
- We introduce an efficient autoencoder for human registration that is 30 times faster with comparable performance.
- We propose a video-based object pose estimator that leverages temporal information to predict object rotations even under heavy occlusions.
- We introduce ProciGen-Video, a method to generate synthetic videos for interaction. With this, we create a dataset of 8.5k videos paired with full 3D ground truth.

2. Related Works

General object tracking. Reconstructing and tracking objects from monocular videos has been studied for decades [94]. Early works [49, 50] track objects with TSDF fusion and are further improved with neural networks [71, 83]. They require depth which can be inconvenient hence more recent works track human [31, 34, 73] or learn general articulated objects [86, 87] from RGB videos. Follow up works extend to learn shapes from casual videos [88, 89] or unstructured images in the wild [40, 76]. These methods learn an explicit template shape, the deformation skeleton and skinning weights. Orthogonal to this, NPG [12] and KeyTr [52] represent shapes as basis points and coefficients learned from videos. Despite impressive results, these methods assume single object and cannot handle the compositional shape during human object interaction.

Interaction reconstruction and tracking. Interaction modelling has recently received more and more attention in generation [2, 37, 38, 57, 99, 100, 104] and reconstruction [22, 42, 77, 80, 81, 84, 85, 96, 102]. Hand-object

reconstruction has been well studied and recent methods are able to recover interaction without any templates [58, 92, 93]. Full body object interaction methods usually rely on template objects [9, 28, 48, 68, 74, 77, 97], with only one template-free method [79] trained on large synthetic dataset. 3D interaction from single images is heavily ill-posed hence some methods leverage temporal information to improve robustness, with works that track interaction from multi-view[6, 25, 27, 29, 95] or monocular [30, 62] RGBD cameras. From RGB camera only, researchers leverage photometric consistency [23] or large diffusion model [91] to guide the reconstruction and lots of methods can obtain good object shapes [15, 21, 26, 61, 90] from hand-object interaction. Object shape is less constrained by the human during full-body interaction hence methods rely on pre-scanned point clouds [18, 19] or meshes [78, 82] to track the interaction. Despite being robust to occlusion, the reliance on object templates limits their applicability. In contrast, our method deals with full-body dynamic object interaction and does not require any object templates.

Correspondence estimation. Correspondence is the key for video tracking. For 3D humans, registering common templates like SMPL [44] to 3D scans is a classic problem and has been studied in many works [4, 5, 11, 16, 17, 45, 46]. Most SoTA methods [4, 5, 11, 46] rely on optimization, which is slow for video processing. In contrast, our CorrAE directly predicts SMPL vertices which is more efficient.

For 3D objects, correspondence can be obtained via functional maps [54, 55] or deep neural networks [12, 20, 60, 75, 101, 103]. However, learning-based methods can only process shapes with aligned orientation which does not apply to our objects with arbitrary rotations. Instead, we first estimate the rotations and then optimize the canonical shape.

3. Method

We present InterTrack, an approach to track interacting human and object from monocular video, without object shape templates. This is very challenging as one needs to establish correspondences, and to reason about object pose/shape simultaneously under heavy occlusion and dynamic motion.

Our key idea is to decompose 4D human and object into per-frame poses and global consistent shapes. For the human, we propose a novel autoencoder that directly predicts SMPL [44] vertices from unordered points, allowing us to use the disentangled SMPL pose and shape parameters. For the object, we introduce a video based object pose predictor that estimates temporally consistent object rotations. This enables us to optimize one common shape in canonical space and per-frame object transformations. An overview of our method can be found in Fig. 2.

In this section, we first briefly discuss how do we obtain per-frame 3D reconstruction with HDM [79] in Sec. 3.1. We then introduce our novel CorrAE for human correspondence and optimization in Sec. 3.2 (Fig. 2B), and TOPNet for temporal rotation prediction and shape optimization of objects in Sec. 3.3(Fig. 2C). We then jointly optimize human and object based contacts to obtain plausible interaction (Fig. 2D, Sec. 3.4). Prior datasets either have only limited objects [6, 27] or static frames [79]. To train our video pose estimator, we introduce a method to generate synthetic video dataset for human object interaction (Sec. 3.5).

3.1. Preliminaries

Given an image sequence $\{\mathbf{I}_i, i = 1, ..., T\}$ from monocular RGB camera of a human interacting with an object, we aim at reconstructing temporally consistent 3D human $\mathbf{H}_i$ and object $\mathbf{O}_i$ for each image. We represent each human and object as dense point clouds and obtain a sequence of temporally coherent points where each point has correspondence across frames.

We first apply HDM [79], a SoTA template-free approach to reconstruct interacting human and object for each image. Specifically, given a single RGB image $\mathbf{I}_i$, HDM [79] reconstructs 3D point $\mathbf{P}_i^h \in \mathbb{R}^{N\times 3}$, $\mathbf{P}_i^o \in \mathbb{R}^{N\times 3}$ of human and object respectively. They adopt a conditional generation paradigm which iteratively denoises Gaussian point clouds into clean human and object points, conditioned on the pixel aligned image features for each points.

The HDM reconstruction provides strong prior of the human-object shapes and interaction semantics like contact points. However, the reconstructed points do not have correspondence across frames due to the nature of the stochastic diffusion process. Furthermore, the diffusion model outputs point clouds in a normalized metric space, leading to different sizes of human and object even if they are from the same sequence. We address these problems next.

3.2. Consistent human reconstruction

The goal is to leverage the shape priors from the HDM predicted points that are *unordered* and create a sequence of points with cross-frame consistency. Several previous methods can register the SMPL model [44] into human point clouds [4, 5, 46]. However, these methods use slow optimization which makes it difficult to apply to video data. We propose to use a simple autoencoder, CorrAE $f_{\text{ae}} : \mathbb{R}^{N\times 3} \rightarrow \mathbb{R}^{N_s\times 3}$, which directly outputs the SMPL vertices from unordered points.

Specifically, we use PVCNN [43] as the point cloud encoder which encodes the human points $\mathbf{P}_i^h$ into a 1D latent vector $\mathbf{z}_i^h$. We then stack several MLP layers to decode the latent vector into 3D points $\hat{\mathbf{P}}_i^h \in \mathbb{R}^{3N_s}$, here $N_s = 6890$ is the number of vertices in the SMPL model [44]. Thanks to the regularity of MLP layers, the outputs $\hat{\mathbf{P}}_i^h$ are ordered points [75, 79, 103] and establish correspondence between different human point clouds. Different from [75, 79] that require aligned object shapes, the key to handle hu-

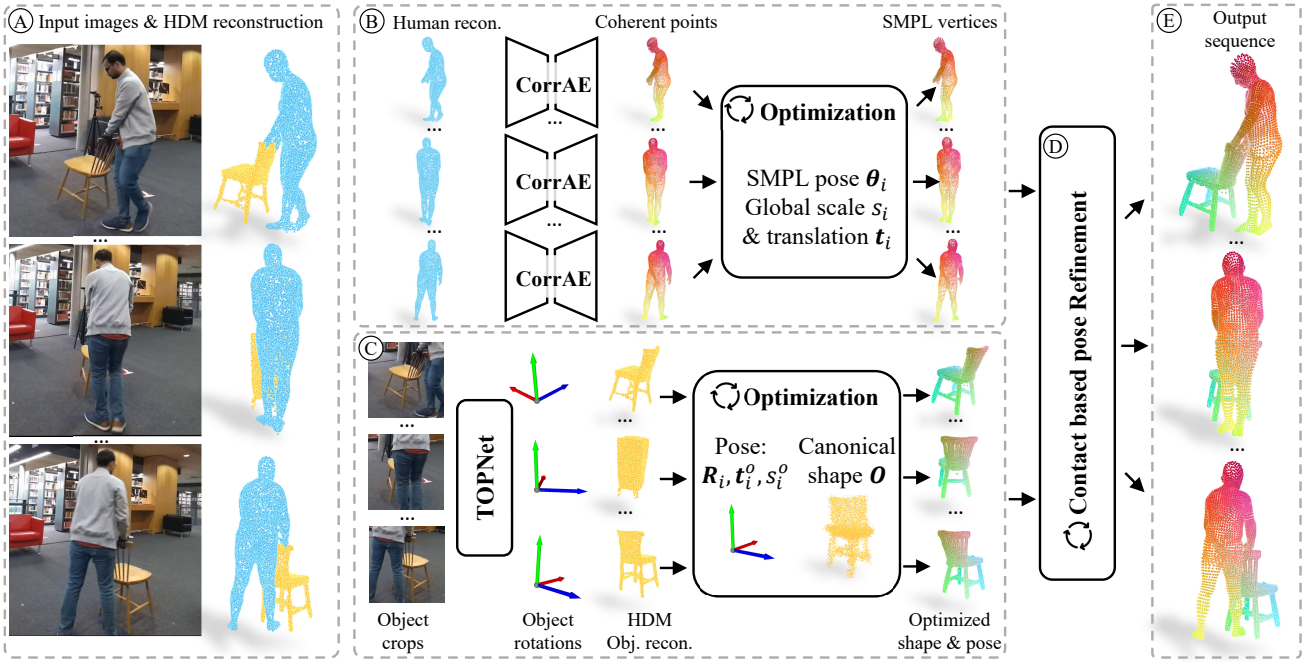

Figure 2. **Method Overview.** Given an image sequence of human object interaction and HDM [79] reconstructions (A), we aim at obtaining coherent tracking of the human and object across frames (E). We first use a simple yet efficient autoencoder CorrAE to obtain coherent humans points and optimize human via the SMPL layer (B, Sec. 3.2). We then use a temporal object pose estimator TOPNet to predict the object rotation, which allows us to optimize a common object shape in canonical space and fine tune pose predictions (C, Sec. 3.3). We then jointly optimize human and object based on contacts to obtain consistent tracking (D, Sec. 3.4).

mans with different orientations is to train the network using Chamfer distance combined with vertex to vertex error:

$$\mathcal{L}_{\text{ae}} = d_{\text{cd}}(f_{\text{ae}}(\mathbf{P}), \mathbf{P}) + \lambda_{\text{v2v}} ||f_{\text{ae}}(\mathbf{P}) - \mathbf{V}_{\text{SMPL}}||_2^2 \quad (1)$$

where $\mathbf{V}_{\text{SMPL}}, \mathbf{P}$ are the vertices and surface samples from SMPL meshes respectively. We train this model on the ground truth SMPL meshes from synthetic ProciGen [79] dataset and find it works well on the reconstructed point clouds from HDM [79]. More importantly, it is much faster than traditional optimization-based method [46], see Tab. 5.

With our CorrAE, one can establish correspondence across frames and then optimize the human representation for human tracking. One straightforward option is to optimize the latent code $\mathbf{z}_i^h$ of each frame with temporal smoothness loss. However, the latent space is not interpretable, without decoupling of the human pose and shape. Instead, we leverage the SMPL representation [44] which has disentangled pose and shape parameter space.

Specifically, we first use the smplx library [56] to obtain each frame's SMPL parameters $\boldsymbol{\theta}_i, \boldsymbol{\beta}_i$ from our CorrAE predictions $\hat{\mathbf{P}}_i^h$. Note that our CorrAE is the key here as without correspondence, it is impossible to obtain accurate SMPL parameters [4, 5]. We then compute a mean body shape $\bar{\boldsymbol{\beta}} = \frac{1}{T}\sum_i \boldsymbol{\beta}_i$ as the shape parameter for the full sequence and optimize only the per-frame translation $\mathbf{t}_i$, scale s_i and local poses $\boldsymbol{\theta}_i$. Let $\mathbf{H}(\boldsymbol{\theta}, \boldsymbol{\beta})$ denote the SMPL model that outputs SMPL vertices given pose and shape parameters, we compute the human points in camera space by: $\mathbf{H}_i = s_i \mathbf{H}(\boldsymbol{\theta}_i, \bar{\boldsymbol{\beta}}) + \mathbf{t}_i$. We optimize the per-frame parameters $\boldsymbol{\theta}_i, \mathbf{t}_i, s_i$ to fit into original HDM human reconstructions $\mathbf{P}_i^h$ with temporal smoothness and pose regularization:

$$\mathcal{L}_{\text{hum}} = \sum_i^T \lambda_{\text{cd}}^h d_{\text{cd}}(\mathbf{P}_i^h, \mathbf{H}_i) + \lambda_p L_{\text{pr}}(\boldsymbol{\theta}_i) + \lambda_a^h L_{\text{acc}}(\mathcal{H}) \quad (2)$$

here L_{pr} is the pose prior loss [4, 24, 65], and $L_{\text{acc}}(\mathcal{H}) = \sum_{i=2}^T ||\mathbf{H}_i - 2\mathbf{H}_{i-1} + \mathbf{H}_{i-2}||_2^2$ is a smoothness loss based on acceleration. We show in Tab. 5 that optimizing via SMPL model is better than optimizing the latent vectors directly.

3.3. Rigid object shape reconstruction

Similar to the human points, the object points $\mathbf{P}_i^o$ predicted by HDM [79] are also unordered and lack temporal consistency. Furthermore, the model generates different 3D shapes when the object is occluded in the image, making tracking more challenging. In a video, the object can be decomposed into one global shape in canonical space and per-frame relative transformations. Estimating object pose without shape is however non-trivial as the object can be fully occluded during interaction. Methods like COLMAP [59] and OnePose [63] fail to track the object due to limited object features. To this end, we propose TOPNet,

a transformer-based network that leverages the temporal information to predict object rotations in a video. Our idea is to use transformer [66] to exchange features across frames and output smooth object poses.

Specifically, we first use DINOv2 encoder [53] followed with two additional convolution layers [76] to compress each input image $\mathbf{I}_i$ into a feature vector $\mathbf{F}_i^I$. We concatenate the image feature with human pose information $\mathbf{F}_i^h$ and object visibility ratio $v_i \in [0, 1]$ (0-fully occluded and 1-fully visible). The human feature $\mathbf{F}_i^h = (\boldsymbol{\theta}_i, \mathbf{J}_i, \mathbf{J}_i')$ consists of SMPL pose $\boldsymbol{\theta}_i$, 3D joint location $\mathbf{J}_i$ and joint velocity $\mathbf{J}_i'$. Hence, the feature vector for each image is $\mathbf{F}_i = (\mathbf{F}_i^I, \mathbf{F}_i^h, v_i)$. We then stack features from W consecutive frames as a feature matrix $\mathcal{F} = (\mathbf{F}_1, ..., \mathbf{F}_W)$ and use transformer encoder [66] to exchange temporal information. The attention layer in transformer outputs more temporally consistent features which are sent to MLP to predict object rotation, represented via the 6D representation [105]. We train our TOPNet with L1 distance between the predicted rotations $\hat{\mathcal{R}} = \{\hat{\mathbf{R}}_1, ..., \hat{\mathbf{R}}_W\}$ and ground truth $\mathbf{R}_i$, and acceleration loss L_{acc} (Eq. (2)): $\mathcal{L}_{\text{rot}} = \sum_i ||\hat{\mathbf{R}}_i - \mathbf{R}_i||_1 + \lambda_a^p L_{\text{acc}}(\hat{\mathcal{R}})$. Our TOPNet is trained with W=16 consecutive frames to learn temporal correlations due to limited data IO speed. At test time, we find it better to use a longer window W=64 and average predictions of each frame in different sliding windows. See Supp. for analysis.

Training this model requires ground truth object pose in a video during realistic interaction. We propose a method to synthesize such video dataset in Sec. 3.5. At training time, we use ground truth human pose and object visibility. At test time, we use PARE [35] to estimate the human pose. For the object, we render the HDM object predictions twice, with and without reconstructed human points, and then compute the mask area ratio as the visibility ratio.

With the predicted rotations, we can then decompose the object tracking into global shape and per-frame object pose optimization. Specifically, we optimize an object shape $\mathbf{O} \in \mathbb{R}^{N\times 3}$ in canonical space and its corresponding rotation $\mathbf{R}_i$, translation $\mathbf{t}_i^o$ and scale s_i^o at frame i. We transform the object from canonical space to frame i via: $\mathbf{O}_i' = s_i^o \mathbf{O}\mathbf{R}_i + \mathbf{t}_i$. Note that the translation $\mathbf{t}_i$ is initialized from the mean of HDM object points hence our TOPNet only needs to predict rotation. We optimize the shape and pose to fit into 2D object masks M_i and HDM object reconstructions:

$$\mathcal{L}_{\text{obj}} = \sum_i^T \lambda_{\text{cd}}^o d_{\text{cd}}(\mathbf{O}_i', \mathbf{P}_i^o) + \lambda_{\text{occ}} L_{\text{occ-sil}}(\pi(\mathbf{O}_i'), M_i) + \lambda_a^o L_{\text{acc}}(\mathcal{O}') + \lambda_a^r L_{\text{acc}}(\mathcal{R}) + \lambda_a^t L_{\text{acc}}(\mathcal{T}) + \lambda_a^s L_{\text{acc}}(\mathbf{s}) \quad (3)$$

where $\pi(\cdot)$ denotes differentiable rendering and $L_{\text{occ-sil}}$ is the occlusion aware silhouette loss [97]. L_{acc} is temporal smoothness (Eq. (2)) for a sequence of object points: $\mathcal{O}' = \{\mathbf{O}_1', ..., \mathbf{O}_T'\}$, rotations $\mathcal{R}$, translations $\mathcal{T}$ and scales $\mathbf{s}$.

We initialize the canonical object shape $\mathbf{O}$ from HDM reconstruction of one random frame where the object visibility is higher than a threshold σ. For more details about loss weights λ_*, see Supp.

3.4. Joint human object optimization

The shape optimization discussed in Sec. 3.2 and Sec. 3.3 deals with human and object separately which can lead to unrealistic interaction. Hence, we propose to further jointly optimize human and object together to satisfy contacts predicted by the initial HDM [79] reconstructions.

HDM reconstructs separate human $\mathbf{P}_i^h$ and object $\mathbf{P}_i^o$ point clouds, which allows computing the contact points to pull our optimized human and object together when there are contacts. Specifically, we identify the human points that are in contact as the points whose distance to the object is smaller than a threshold: $\tilde{\mathbf{P}}_i^h = \{\mathbf{p}_{i,j}^h | \min_k |\mathbf{p}_{i,j}^h - \mathbf{p}_{i,k}^o|_2 < \delta, \forall j, k \in \{1, ..., N\}\}$. The corresponding object points $\mathbf{p}_{i,k}^o$ are identified as the object contacts $\tilde{\mathbf{P}}_i^o$. We then transfer the contact points to the optimized human $\mathbf{H}_i$ and object points $\mathbf{O}_i'$ by finding their closest points in $\tilde{\mathbf{P}}_i^h, \tilde{\mathbf{P}}_i^o$. Denoting the transferred contact points as $\tilde{\mathbf{H}}_i, \tilde{\mathbf{O}}_i$, the joint optimization objective combines contact distance loss and separate human and object losses:

$$\mathcal{L}_{\text{joint}} = \mathcal{L}_{\text{hum}} + \mathcal{L}_{\text{obj}} + \lambda_c \sum_i ||\tilde{\mathbf{H}}_i - \tilde{\mathbf{O}}_i||_2^2 \quad (4)$$

where L_{hum} and L_{obj} are defined in Eq. (2) and Eq. (3) respectively. See Supp. for more loss weight details.

3.5. ProciGen-V: Synthetic Interaction Videos

Training our video-based TOPNet requires video datasets of human object interaction. Previous methods[77, 78] were trained on BEHAVE[6] and InterCap [27] which cannot generalize to other objects as shown in [79]. Capturing more interaction is not scalable due to the compositionality of human object interaction. Recent work ProciGen [79] can generate interaction with new objects but is limited to static frames only. To this end, we introduce ProciGen-Video, a method to synthesize videos for interaction.

We start with ProciGen [79], which generates new interactions by replacing the real captured object with new shapes from the same category. To generate an interaction video, we sample a chunk of an interaction sequence with human and object poses from real data. We then randomly select one shape of the same category from shape databases [8, 13] and replace the original object with the new shape. We initialize the object pose for the new shape via dense correspondence [103]. This initialization can have interpenetration as the new shape is different from the original, hence, we further fine-tune the human and object pose

to satisfy contacts and temporal smoothness. We then render the human with SMPL-D texture [3] and object with original textures to obtain images in blender. Please see Supp. for optimization details and example videos.

We apply our method to generate interaction videos of 10 different object categories with interaction motions sampled from BEHAVE and InterCap training set, and object shapes from Shapenet [8], Objaverse [13]. This leads to around 10 hours of video with 8337 sequences and 2M images. We call this dataset ProciGen-Video, or ProciGen-V for short. Since all the objects used in our dataset are aligned in the canonical space, models trained on our data have category-level pose estimation ability and they generalize well to real data (Fig. 1, Tab. 1, Tab. 3).

4. Experiments

In this section, we first compare our method against prior works on reconstructing human object interaction from images or videos and then evaluate our ProciGen-V dataset. We further ablate the design choices of our human, object and joint optimization modules. Implementation details, method limitation and failure case are discussed in Supp.

4.1. Comparison with prior methods

Baselines. We compare our method against image-based methods CHORE [77], HDM [79] and video based method VisTracker [78]. CHORE relies on known object templates while HDM is template-free. VisTracker also requires known object templates but it improves CHORE on video by leveraging image aligned SMPL and temporal information. HOLD [15] is a template-free method for video based hand-object tracking that is relevant to us. However, it relies on hand-specific inverse skinning and hand-object contact annotations which makes it not directly applicable to our setup. Hence we do not consider HOLD as a baseline.
Datasets. We conduct experiments on the BEHAVE [6], InterCap [27], synthetic ProciGen [79] and our ProciGen-V datasets. BEHAVE and InterCap capture realistic humans interacting with 20 and 10 objects, respectively. ProciGen extends these dataset by replacing the objects with shapes from ShapeNet [8], Objaverse [13], and ABO [10] datasets and synthesizing new interaction images. Following HDM [79], we first train HDM and our CorrAE on the ProciGen dataset and then fine tune HDM on BEHAVE and InterCap training sets. Our TOPNet is also first trained on ProciGen-V and then tuned on BEHAVE and InterCap. We test the methods on sequences from the 10 overlapping categories between ProciGen-V and BEHAVE as well as InterCap. In total there are 66 test sequences (51k images) in BEHAVE and 15 sequences (2k images) in InterCap.
Evaluation Metrics. Following [64, 79], we evaluate the shape accuracy using F-score with threshold 0.01m. We report the F-score for human, object and combined shape.

	Method	Human↑	Object↑	Comb.↑
BEHAVE	HDM synth. only	0.3574	0.4118	0.4118
	Ours synth. only	0.3780	0.4658	0.4488
	HDM	0.3909	0.5111	0.4622
	Ours	**0.4113**	**0.5849**	**0.5169**
InterCap	HDM synth. only	0.3570	0.4342	0.4080
	Ours synth. only	0.3715	0.5141	0.4646
	HDM	0.4325	0.6267	0.5362
	Ours	**0.4463**	**0.6329**	**0.5555**

Table 1. **Reconstruction results** (F-sc.@0.01m) on BEHAVE [6] and InterCap [27] of *template-free* methods. Synth. only denotes model trained *only* on synthetic ProciGen[79] and our ProciGen-V. Our method consistently outperforms HDM [79] in all settings.

We first compare with template-free method HDM [79] in Tab. 1. HDM is image-based method that cannot reason temporal consistency and fails when the object is occluded. Following HDM [79], we also report the results when both methods are trained only on the synthetic ProciGen and ProciGen-V datasets (synth. only). It can be seen that our method consistently improves over HDM in both without and with training on real data.

We then compare with CHORE [77] and VisTracker [78] that require object templates. For fair comparison, we adapt our method to use the same object template as the shape $\mathbf{O}$ and optimize only the transformation parameters. We report the results in Tab. 2 following the same Procrustes alignment used by CHORE [77] to avoid depth-scale ambiguity. CHORE predicts noisy object poses when they are occluded, leading to inconsistent tracking. VisTracker improves CHORE via ad-hoc pose smoothing and infilling which is still suboptimal. Our method takes temporal information into account in the first place and produces more stable tracking. Notably, our method without template (Tab. 1) also archives better performance compared to template-based VisTracker and CHORE. We show some qualitative comparisons of our method without template and baselines in Fig. 3. It can be seen that our method is more robust to occlusions and produces coherent object shapes.

4.2. Evaluating the ProciGen-Video dataset

Our ProciGen-V dataset allows us to train a video-based pose estimator that generalizes to other instances of the same category. To evaluate this, we train our object pose TOPNet on ProciGen-V, BEHAVE, and a combination of both. For other parts of our method, we use the same HDM trained on ProciGen [79] only and perform the same optimization process. Due to compute limitation, we test only on 14 BEHAVE sequences (9.1k images). We report the object pose error, the chamfer distance between predicted and GT rotation applied to GT meshes, and the final tracking results in Tab. 3. The performance gap between models trained on synthetic and real data is small and pre-training

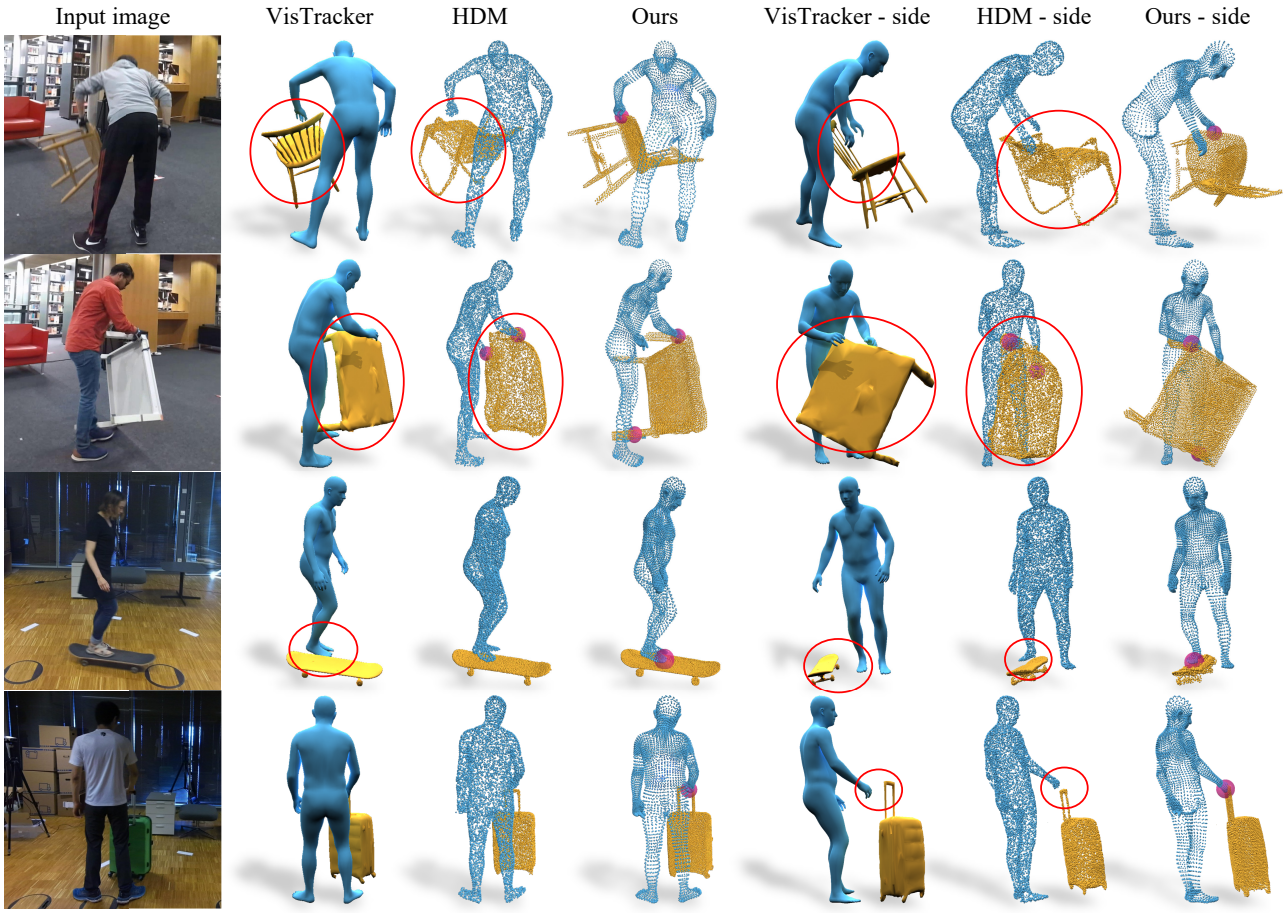

Figure 3. Comparing our method against VisTracker [78] and HDM [79] on BAHEVE (row1-2) [6] and InterCap (row3-4) [27]. VisTracker relies on post-hoc processing to refine object pose which is inaccurate and HDM reconstructs inconsistent object shapes (row 1-2) or interactions (row 3-4). Our temporal based pose estimation and optimization leads to consistent shape and interaction.

model on our ProciGen-V can boost the performance.

	Method	Human↑	Object↑	Comb.↑
BEHAVE	CHORE [77]	0.3807	0.4423	0.4224
	VisTracker [78]	0.4189	0.5607	0.5012
	Ours + template	**0.4972**	**0.5854**	**0.5525**
InterCap	CHORE [77]	0.4065	0.4972	0.4601
	VisTracker [78]	0.4293	0.5316	0.4902
	Ours + template	**0.4826**	**0.5349**	**0.5349**

Table 2. **Reconstruction results** (F-sc.@0.01m) on BEHAVE [6] and InterCap [27] for *template based* methods. Our method obtains more accurate tracking in both datasets.

4.3. Evaluating object pose estimators

We propose TOPNet, a video based method to predict category level object rotations conditioned on human pose. We compare our TOPNet against a prior method CenterPose [41] for category-level pose estimation. Note that we exclude methods [36, 39] that require additional optimization as it is too slow to process video data or methods that require object templates [51, 69, 98] or depth [7, 70, 72]. We train CenterPose, our TOPNet with and without human

TOPNet Training Data	Obj. CD ↓	Hum.↑	Obj.↑	Comb.↑
a. ProciGen-V only	4.859	**0.3392**	0.4662	0.4327
b. BEHAVE only	3.175	0.3376	0.4815	0.4385
c. ProciGen-V+BEHAVE	**2.922**	**0.3392**	**0.4980**	**0.4468**

Table 3. **Training object pose TOPNet on different datasets**. Our TOPNet trained only on our synthetic ProciGen-V generalizes to unseen instances in BEHAVE [6]. Further fine-tuning the model on real data archives the best performance.

Method	Obj. CD↓	Hum.↑	Obj.↑	Comb.↑
a. CenterPose + Our opt.	7.889	0.3332	0.3272	0.3641
b. Ours w/o hum. cond.	4.910	0.3380	0.4610	0.4286
c. Ours full model	**4.859**	**0.3392**	**0.4662**	**0.4327**

Table 4. **Object pose error** (Chamfer Distance in cm) and joint tracking results using different object pose estimator. Our temporal based TOPNet predicts more accurate object poses than CenterPose [41] and human condition improves the performance.

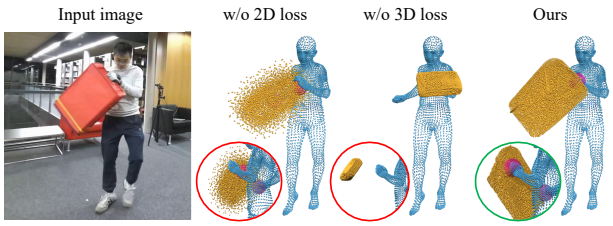

Figure 4. **The effects of 2D and 3D losses** for object optimization. Without the 2D mask loss, the object shape is very noisy and without 3D chamfer loss the relative object position is incorrect.

pose condition on our synthetic ProciGen-V and test it on 14 BEHAVE sequences (same as Sec. 4.2 Tab. 3a). Results are shown in Tab. 4. CenterPose predicts noisy pose under occlusions while our model with human conditioning achieves the best overall results.

4.4. Evaluating human reconstruction module

We propose a simple yet efficient network CorrAE to obtain correspondence for a sequence of unordered human points. Keeping other modules the same as Tab. 3a, we compare results of using our CorrAE or human registration methods ArtEq [16], NICP [46] for the human reconstruction part in Tab. 5. It can be seen that our method achieves similar performance but is much faster than NICP. We also report the results of optimizing the latent code of CorrAE instead of optimizing SMPL parameters in Tab. 5c. The latent space entangles human pose and shape which is less controllable than SMPL model [44]. Optimizing the latent code also leads to unsmooth surface points, please see supplementary for examples and more analysis. Optimizing via the SMPL layer leads to better human reconstruction (Tab. 5d).

Method	Hum.↑	Obj.↑	Comb.↑	Corr. Time↓
a. ArtEq + our opt.	0.3386	0.3386	0.4230	0.467
b. NICP + our opt.	0.3380	**0.4669**	0.4325	72.322
c. Ours opt. latent	0.3281	0.4631	0.4283	**0.001**
d. Ours	**0.3392**	0.4662	**0.4327**	2.553

Table 5. **Different methods for human optimization** and the runtime (seconds/image) to obtain correspondences. ArtEq [16] (a) is fast but performance is suboptimal. NICP [46] (b) obtains similar results yet the runtime is significantly longer than our autoencoder (c, d). Optimizing through the SMPL body model achieves a good balance between runtime and accuracy.

4.5. Importance of different losses

We ablate the effect of the 2D mask loss ($L_{\text{occ-sil}}$) and the 3D chamfer loss (d_{cd}) for object optimization (Eq. (3)) in Fig. 4. It can be seen that omitting both will lead to a low-fidelity object shape. We also ablate our contact-based pose refinement in Tab. 6 and Fig. 5. Quantitatively, the contact refinement has tiny difference yet qualitatively it significantly improves the physical plausibility of the reconstructed interaction (Fig. 5). Without this refinement, the object will be floating in the air which is not physically plausible.

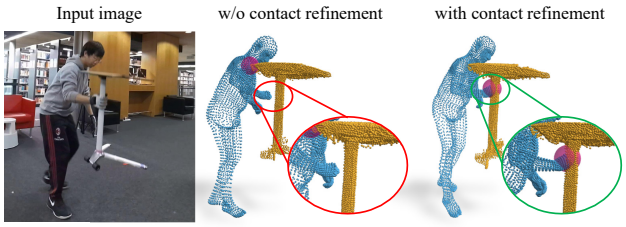

Figure 5. **The effect of the contact-based refinement.** Without contact, the hand and object can be far apart, leading to implausible interaction.

Method	Hum.	Obj.	Comb.
a. w/o contact refinement	**0.3430**	0.4653	**0.4338**
b. Ours full model	0.3392	**0.4662**	0.4327

Table 6. **Contact based refinement** leads to tiny quantitative difference but more plausible interaction qualitatively, see Fig. 5.

5. Conclusion

In this paper, we present InterTrack, the first approach to track full-body interaction with dynamic object from monocular RGB video without pre-scanned object templates. Our first contribution is a simple yet efficient autoencoder CorrAE to directly predict SMPL vertices from unordered human reconstructions. Our second contribution is a video-based pose estimator TOPNet that leverages temporal information to predict smooth rotation under occlusions. We also introduce a method to generate synthetic videos for interaction and we synthesize 10-hour videos of ∼8.5k sequences for training video based methods.

Our experiments on BEHAVE and InterCap show that our method significantly outperforms prior template-free and template-based approaches. Ablation studies also show that our method trained on our synthetic video dataset generalizes to real data. We also show that our CorrAE is much more efficient than SoTA human registration method with similar performance and our TOPNet is more accurate than another category-level pose estimator method. Our code, models and dataset are publicly released.

Acknowledgements. We thank RVH group members [1] for their helpful discussions. This work is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - 409792180 (Emmy Noether Programme, project: Real Virtual Humans), and German Federal Ministry of Education and Research (BMBF): Tübingen AI Center, FKZ: 01IS18039A, and Amazon-MPI science hub. Gerard Pons-Moll is a Professor at the University of Tübingen endowed by the Carl Zeiss Foundation, at the Department of Computer Science and a member of the Machine Learning Cluster of Excellence, EXC number 2064/1 – Project number 390727645.

References

[1] http://virtualhumans.mpi-inf.mpg.de/people.html. 8

[2] Chen Bao, Helin Xu, Yuzhe Qin, and Xiaolong Wang. Dexart: Benchmarking generalizable dexterous manipulation with articulated objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21190–21200, 2023. 2

[3] Bharat Lal Bhatnagar, Garvita Tiwari, Christian Theobalt, and Gerard Pons-Moll. Multi-garment net: Learning to dress 3d people from images. In *IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2019. 6, 2

[4] Bharat Lal Bhatnagar, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Combining implicit function learning and parametric models for 3d human reconstruction. In *European Conference on Computer Vision (ECCV)*. Springer, 2020. 3, 4

[5] Bharat Lal Bhatnagar, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Loopreg: Self-supervised learning of implicit surface correspondences, pose and shape for 3d human mesh registration. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 3, 4

[6] Bharat Lal Bhatnagar, Xianghui Xie, Ilya Petrov, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Behave: Dataset and method for tracking human object interactions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2, 3, 5, 6, 7

[7] Junhao Cai, Yisheng He, Weihao Yuan, Siyu Zhu, Zilong Dong, Liefeng Bo, and Qifeng Chen. Ov9d: Open-vocabulary category-level 9d object pose and size estimation. *arXiv preprint arXiv:2403.12396*, 2024. 7

[8] Angel X. Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, Li Yi, and Fisher Yu. ShapeNet: An Information-Rich 3D Model Repository. Technical Report arXiv:1512.03012 [cs.GR], Stanford University — Princeton University — Toyota Technological Institute at Chicago, 2015. 5, 6, 2

[9] Yuhang Chen and Chenxing Wang. Kinematics-based 3D Human-Object Interaction Reconstruction from Single View, 2024. 3

[10] Jasmine Collins, Shubham Goel, Kenan Deng, Achleshwar Luthra, Leon Xu, Erhan Gundogdu, Xi Zhang, Tomas F Yago Vicente, Thomas Dideriksen, Himanshu Arora, Matthieu Guillaumin, and Jitendra Malik. Abo: Dataset and benchmarks for real-world 3d object understanding. *CVPR*, 2022. 6

[11] Enric Corona, Gerard Pons-Moll, Guillem Alenya, and Francesc Moreno-Noguer. Learned vertex descent: A new direction for 3d human model fitting. In *European Conference on Computer Vision (ECCV)*. Springer, 2022. 3

[12] Devikalyan Das, Christopher Wewer, Raza Yunus, Eddy Ilg, and Jan Eric Lenssen. Neural parametric gaussians for monocular non-rigid object reconstruction. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2, 3, 4

[13] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. *arXiv preprint arXiv:2212.08051*, 2022. 5, 6, 2

[14] Yan Di, Fabian Manhardt, Gu Wang, Xiangyang Ji, Nassir Navab, and Federico Tombari. SO-Pose: Exploiting Self-Occlusion for Direct 6D Pose Estimation. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12376–12385, Montreal, QC, Canada, 2021. IEEE. 4

[15] Zicong Fan, Maria Parelli, Maria Eleni Kadoglou, Muhammed Kocabas, Xu Chen, Michael J Black, and Otmar Hilliges. HOLD: Category-agnostic 3d reconstruction of interacting hands and objects from video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 494–504, 2024. 3, 6

[16] Haiwen Feng, Peter Kulits, Shichen Liu, Michael J. Black, and Victoria Abrevaya. Generalizing neural human fitting to unseen poses with articulated se(3) equivariance, 2023. 3, 8

[17] Thibault Groueix, Matthew Fisher, Vladimir G. Kim, Bryan Russell, and Mathieu Aubry. 3d-coded : 3d correspondences by deep deformation. In *ECCV*, 2018. 3

[18] Vladimir Guzov, Aymen Mir, Torsten Sattler, and Gerard Pons-Moll. Human poseitioning system (hps): 3d human pose estimation and self-localization in large scenes from body-mounted sensors. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2021. 3

[19] Vladimir Guzov, Julian Chibane, Riccardo Marin, Yannan He, Yunus Saracoglu, Torsten Sattler, and Gerard Pons-Moll. Interaction replica: Tracking human–object interaction and scene changes from human motion. In *International Conference on 3D Vision (3DV)*, 2024. 3

[20] Oshri Halimi, Or Litany, Emanuele Rodola, Alex M Bronstein, and Ron Kimmel. Unsupervised learning of dense shape correspondence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4370–4379, 2019. 3

[21] Shreyas Hampali, Tomas Hodan, Luan Tran, Lingni Ma, Cem Keskin, and Vincent Lepetit. In-hand 3d object scanning from an rgb sequence. *CVPR*, 2023. 3

[22] Yana Hasson, Gül Varol, Dimitrios Tzionas, Igor Kalevatykh, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning joint reconstruction of hands and manipulated objects. In *CVPR*, 2019. 2

[23] Yana Hasson, Bugra Tekin, Federica Bogo, Ivan Laptev, Marc Pollefeys, and Cordelia Schmid. Leveraging photometric consistency over time for sparsely supervised hand-object reconstruction. In *CVPR*, 2020. 3

[24] Yannan He, Garvita Tiwari, Tolga Birdal, Jan Eric Lenssen, and Gerard Pons-Moll. Nrdf: Neural riemannian distance fields for learning articulated pose priors. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 4

[25] Chun-Hao P. Huang, Hongwei Yi, Markus Höschle, Matvey Safroshkin, Tsvetelina Alexiadis, Senya Polikovsky, Daniel Scharstein, and Michael J. Black. Capturing and inferring dense full-body human-scene contact. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 13274–13285, 2022. 3

[26] Di Huang, Xiaopeng Ji, Xingyi He, Jiaming Sun, Tong He, Qing Shuai, Wanli Ouyang, and Xiaowei Zhou. Reconstructing hand-held objects from monocular video. In *SIGGRAPH Asia Conference Proceedings*, 2022. 3

[27] Yinghao Huang, Omid Taheri, Michael J. Black, and Dimitrios Tzionas. InterCap: Joint markerless 3D tracking of humans and objects in interaction. In *German Conference on Pattern Recognition (GCPR)*, pages 281–299. Springer, 2022. 2, 3, 5, 6, 7

[28] Chaofan Huo, Ye Shi, Yuexin Ma, Lan Xu, Jingyi Yu, and Jingya Wang. StackFLOW: Monocular Human-Object Reconstruction by Stacked Normalizing Flow with Offset. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pages 902–910, Macau, SAR China, 2023. International Joint Conferences on Artificial Intelligence Organization. 3

[29] Yuheng Jiang, Suyi Jiang, Guoxing Sun, Zhuo Su, Kaiwen Guo, Minye Wu, Jingyi Yu, and Lan Xu. Neuralhofusion: Neural volumetric rendering under human-object interactions. *arXiv preprint arXiv:2202.12825*, 2022. 3

[30] Yuheng Jiang, Kaixin Yao, Zhuo Su, Zhehao Shen, Haimin Luo, and Lan Xu. Instant-nvr: Instant neural volumetric rendering for human-object interactions from monocular rgbd stream, 2023. 3

[31] Zeren Jiang, Chen Guo, Manuel Kaufmann, Tianjian Jiang, Julien Valentin, Otmar Hilliges, and Jie Song. Multiply: Reconstruction of multiple people from monocular video in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2

[32] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4), 2023. 4

[33] Hyeonwoo Kim, Sookwan Han, Patrick Kwon, and Hanbyul Joo. Beyond the contact: Discovering comprehensive affordance for 3d objects from pre-trained 2d diffusion models, 2024. 4

[34] Muhammed Kocabas, Nikos Athanasiou, and Michael J. Black. VIBE: Video inference for human body pose and shape estimation. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 5252–5262. IEEE, 2020. 2

[35] Muhammed Kocabas, Chun-Hao P. Huang, Otmar Hilliges, and Michael J. Black. PARE: Part attention regressor for 3D human body estimation. In *Proceedings International Conference on Computer Vision (ICCV)*, pages 11127–11137. IEEE, 2021. 5

[36] Taeyeop Lee, Jonathan Tremblay, Valts Blukis, Bowen Wen, Byeong-Uk Lee, Inkyu Shin, Stan Birchfield, In So Kweon, and Kuk-Jin Yoon. TTA-COPE: Test-Time Adaptation for Category-Level Object Pose Estimation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21285–21295, Vancouver, BC, Canada, 2023. IEEE. 7

[37] Jiaman Li, Jiajun Wu, and C Karen Liu. Object motion guided human motion synthesis. *ACM Trans. Graph.*, 42 (6), 2023. 2

[38] Lei Li and Angela Dai. GenZI: Zero-shot 3D human-scene interaction generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2

[39] Yi Li, Gu Wang, Xiangyang Ji, Yu Xiang, and Dieter Fox. DeepIM: Deep Iterative Matching for 6D Pose Estimation. In *European Conference on Computer Vision (ECCV)*, page 16, 2018. 7

[40] Zizhang Li, Dor Litvak, Ruining Li, Yunzhi Zhang, Tomas Jakab, Christian Rupprecht, Shangzhe Wu, Andrea Vedaldi, and Jiajun Wu. Learning the 3d fauna of the web. *arXiv preprint arXiv:2401.02400*, 2024. 2

[41] Yunzhi Lin, Jonathan Tremblay, Stephen Tyree, Patricio A. Vela, and Stan Birchfield. Single-stage keypoint-based category-level object pose estimation from an RGB image. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2022. 2, 7

[42] Yunze Liu, Yun Liu, Che Jiang, Kangbo Lyu, Weikang Wan, Hao Shen, Boqiang Liang, Zhoujie Fu, He Wang, and Li Yi. Hoi4d: A 4d egocentric dataset for category-level human-object interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21013–21022, 2022. 2

[43] Zhijian Liu, Haotian Tang, Yujun Lin, and Song Han. Point-voxel cnn for efficient 3d deep learning. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2019. 3, 1

[44] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. In *ACM Transactions on Graphics*. ACM, 2015. 2, 3, 4, 8, 1

[45] Riccardo Marin, Simone Melzi, Emanuele Rodolà, and Umberto Castellani. Farm: Functional automatic registration method for 3d human bodies, 2018. 3

[46] Riccardo Marin, Enric Corona, and Gerard Pons-Moll. Nicp: Neural icp for 3d human registration at scale. In *European Conference on Computer Vision*, 2024. 2, 3, 4, 8

[47] Luke Melas-Kyriazi, Christian Rupprecht, and Andrea Vedaldi. Pc2: Projection-conditioned point cloud diffusion for single-image 3d reconstruction. In *CVPR*, 2023. 1

[48] Hyeongjin Nam, Daniel Sungho Jung, Gyeongsik Moon, and Kyoung Mu Lee. Joint reconstruction of 3d human and object via contact-based refinement transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 3

[49] Richard A Newcombe, Shahram Izadi, Otmar Hilliges, David Molyneaux, David Kim, Andrew J Davison, Pushmeet Kohi, Jamie Shotton, Steve Hodges, and Andrew Fitzgibbon. Kinectfusion: Real-time dense surface mapping and tracking. In *IEEE International Symposium on Mixed and Augmented Reality*, pages 127–136, 2011. 2

[50] Richard A Newcombe, Dieter Fox, and Steven M Seitz. Dynamicfusion: Reconstruction and tracking of non-rigid scenes in real-time. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 343–352, 2015. 2

[51] Van Nguyen Nguyen, Thibault Groueix, Mathieu Salzmann, and Vincent Lepetit. GigaPose: Fast and Robust

Novel Object Pose Estimation via One Correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 7

[52] David Novotny, Ignacio Rocco, Samarth Sinha, Alexandre Carlier, Gael Kerchenbaum, Roman Shapovalov, Nikita Smetanin, Natalia Neverova, Benjamin Graham, and Andrea Vedaldi. Keytr: Keypoint transporter for 3d reconstruction of deformable objects in videos. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5585–5594, 2022. 2

[53] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2024. 5, 1

[54] Litany Or, Remez Tal, Rodolà Emanuele, Bronstein Alex M., and Bronstein Michael M. Deep functional maps: Structured prediction for dense shape correspondence. In *International Conference on Computer Vision (ICCV)*, 2017. 3

[55] Maks Ovsjanikov, Mirela Ben-Chen, Justin Solomon, Adrian Butscher, and Leonidas Guibas. Functional maps: A flexible representation of maps between shapes. *ACM Transactions on Graphics - TOG*, 31, 2012. 3

[56] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019. 4

[57] Ilya A Petrov, Riccardo Marin, Julian Chibane, and Gerard Pons-Moll. Object pop-up: Can we infer 3d objects and their poses from human interactions alone? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 2

[58] Haozhe Qi, Chen Zhao, Mathieu Salzmann, and Alexander Mathis. HOISDF: Constraining 3D Hand-Object Pose Estimation with Global Signed Distance Fields. *CVPR*, 2024. 3

[59] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 4

[60] Philipp Schröppel, Christopher Wewer, Jan Eric Lenssen, Eddy Ilg, and Thomas Brox. Neural point cloud diffusion for disentangled 3d shape and appearance generation. In *CVPR*, 2024. 3

[61] Haixin Shi, Yinlin Hu, Daniel Koguciuk, Juan-Ting Lin, Mathieu Salzmann, and David Ferstl. Free-Moving Object Reconstruction and Pose Estimation with Virtual Camera, 2024. 3

[62] Zhuo Su, Lan Xu, Dawei Zhong, Zhong Li, Fan Deng, Shuxue Quan, and Lu Fang. Robustfusion: Robust volumetric performance reconstruction under human-object interactions from monocular RGBD stream. *CoRR*, abs/2104.14837, 2021. 3

[63] Jiaming Sun, Zihao Wang, Siyu Zhang, Xingyi He, Hongcheng Zhao, Guofeng Zhang, and Xiaowei Zhou. OnePose: One-shot object pose estimation without CAD models. *CVPR*, 2022. 4

[64] Maxim Tatarchenko*, Stephan R. Richter*, René Ranftl, Zhuwen Li, Vladlen Koltun, and Thomas Brox. What do single-view 3d reconstruction networks learn? In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 6

[65] Garvita Tiwari, Dimitrije Antic, Jan Eric Lenssen, Nikolaos Sarafianos, Tony Tung, and Gerard Pons-Moll. Pose-ndf: Modeling human pose manifolds with neural distance fields. In *European Conference on Computer Vision (ECCV)*. Springer, 2022. 4

[66] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2017. 5, 1

[67] Gu Wang, Fabian Manhardt, Federico Tombari, and Xiangyang Ji. GDR-Net: Geometry-guided direct regression network for monocular 6d object pose estimation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16611–16621, 2021. 4

[68] Xi Wang, Gen Li, Yen-Ling Kuo, Muhammed Kocabas, Emre Aksan, and Otmar Hilliges. Reconstructing action-conditioned human-object interactions using commonsense knowledge priors. In *International Conference on 3D Vision (3DV)*, 2022. 3

[69] Jiaxin Wei, Xibin Song, Weizhe Liu, Laurent Kneip, Hongdong Li, and Pan Ji. RGB-based Category-level Object Pose Estimation via Decoupled Metric Scale Recovery, 2023. 7

[70] B Wen and Kostas E Bekris. Bundletrack: 6d pose tracking for novel objects without instance or category-level 3d models. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2021. 7

[71] Bowen Wen, Jonathan Tremblay, Valts Blukis, Stephen Tyree, Thomas Muller, Alex Evans, Dieter Fox, Jan Kautz, and Stan Birchfield. Bundlesdf: Neural 6-dof tracking and 3d reconstruction of unknown objects. *CVPR*, 2023. 2

[72] Bowen Wen, Wei Yang, Jan Kautz, and Stan Birchfield. FoundationPose: Unified 6D Pose Estimation and Tracking of Novel Objects. In *CVPR*, 2024. 7, 4

[73] Chung-Yi Weng, Brian Curless, Pratul P. Srinivasan, Jonathan T. Barron, and Ira Kemelmacher-Shlizerman. HumanNeRF: Free-viewpoint rendering of moving people from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16210–16220, 2022. 2

[74] Zhenzhen Weng and Serena Yeung. Holistic 3d human and scene mesh estimation from single view images. *arXiv preprint arXiv:2012.01591*, 2020. 3

[75] Christopher Wewer, Eddy Ilg, Bernt Schiele, and Jan Eric Lenssen. Simnp: Learning self-similarity priors between

neural points. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 3

[76] Shangzhe Wu, Ruining Li, Tomas Jakab, Christian Rupprecht, and Andrea Vedaldi. MagicPony: Learning articulated 3d animals in the wild. In *CVPR*, 2023. 2, 5, 1

[77] Xianghui Xie, Bharat Lal Bhatnagar, and Gerard Pons-Moll. Chore: Contact, human and object reconstruction from a single rgb image. In *European Conference on Computer Vision (ECCV)*. Springer, 2022. 2, 3, 5, 6, 7

[78] Xianghui Xie, Bharat Lal Bhatnagar, and Gerard Pons-Moll. Visibility aware human-object interaction tracking from single rgb camera. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2, 3, 5, 6, 7

[79] Xianghui Xie, Bharat Lal Bhatnagar, Jan Eric Lenssen, and Gerard Pons-Moll. Template free reconstruction of human-object interaction with procedural interaction generation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2, 3, 4, 5, 6, 7, 1

[80] Xianghui Xie, Xi Wang, Nikos Athanasiou, Bharat Lal Bhatnagar, Chun-Hao P. Huang, Kaichun Mo, Hao Chen, Xia Jia, Zerui Zhang, Liangxian Cui, Xiao Lin, Bingqiao Qian, Jie Xiao, Wenfei Yang, Hyeongjin Nam, Daniel Sungho Jung, Kihoon Kim, Kyoung Mu Lee, Otmar Hilliges, and Gerard Pons-Moll. RHOBIN Challenge: Reconstruction of human object interaction. *arXiv preprint arXiv:2401.04143*, 2024. 2

[81] Xianghui Xie, Xi Wang, Nikos Athanasiou, Bharat Lal Bhatnagar, Chun-Hao P. Huang, Kaichun Mo, Hao Chen, Xia Jia, Zerui Zhang, Liangxian Cui, Xiao Lin, Bingqiao Qian, Jie Xiao, Wenfei Yang, Hyeongjin Nam, Daniel Sungho Jung, Kihoon Kim, Kyoung Mu Lee, Otmar Hilliges, and Gerard Pons-Moll. Rhobin challenge: Reconstruction of human object interaction, 2024. 2

[82] Xiang Xu, Hanbyul Joo, Greg Mori, and Manolis Savva. D3d-hoi: Dynamic 3d human-object interactions from videos. *arXiv preprint arXiv:2108.08420*, 2021. 3

[83] Yuxuan Xue, Bharat Lal Bhatnagar, Riccardo Marin, Nikolaos Sarafianos, Yuanlu Xu, Gerard Pons-Moll, and Tony Tung. Nsf: Neural surface fields for human modeling from monocular depth. In *ICCV*, 2023. 2

[84] Yuxuan Xue, Xianghui Xie, Riccardo Marin, and Gerard Pons-Moll. Gen-3diffusion: Realistic image-to-3d generation via 2d & 3d diffusion synergy, 2024. 2

[85] Yuxuan Xue, Xianghui Xie, Riccardo Marin, and Gerard Pons-Moll. Human 3diffusion: Realistic avatar creation via explicit 3d consistent diffusion models. In *Arxiv*, 2024. 2

[86] Gengshan Yang, Deqing Sun, Varun Jampani, Daniel Vlasic, Forrester Cole, Huiwen Chang, Deva Ramanan, William T Freeman, and Ce Liu. Lasr: Learning articulated shape reconstruction from a monocular video. In *CVPR*, 2021. 2

[87] Gengshan Yang, Deqing Sun, Varun Jampani, Daniel Vlasic, Forrester Cole, Ce Liu, and Deva Ramanan. Viser: Video-specific surface embeddings for articulated 3d shape reconstruction. In *NeurIPS*, 2021. 2

[88] Gengshan Yang, Minh Vo, Natalia Neverova, Deva Ramanan, Andrea Vedaldi, and Hanbyul Joo. Banmo: Building animatable 3d neural models from many casual videos. In *CVPR*, 2022. 2

[89] Gengshan Yang, Chaoyang Wang, N. Dinesh Reddy, and Deva Ramanan. Reconstructing animatable categories from videos. In *CVPR*, 2023. 2

[90] Yufei Ye, Abhinav Gupta, and Shubham Tulsiani. What's in your hands? 3d reconstruction of generic objects in hands. In *CVPR*, 2022. 3

[91] Yufei Ye, Poorvi Hebbar, Abhinav Gupta, and Shubham Tulsiani. Diffusion-guided reconstruction of everyday hand-object interaction clips. In *ICCV*, 2023. 3

[92] Yufei Ye, Xueting Li, Abhinav Gupta, Shalini De Mello, Stan Birchfield, Jiaming Song, Shubham Tulsiani, and Sifei Liu. Affordance diffusion: Synthesizing hand-object interactions. In *CVPR*, 2023. 3

[93] Yufei Ye, Abhinav Gupta, Kris Kitani, and Shubham Tulsiani. G-hop: Generative hand-object prior for interaction reconstruction and grasp synthesis. In *CVPR*, 2024. 3

[94] Raza Yunus, Jan Eric Lenssen, Michael Niemeyer, Yiyi Liao, Christian Rupprecht, Christian Theobalt, Gerard Pons-Moll, Jia-Bin Huang, Vladislav Golyanik, and Eddy Ilg. Recent Trends in 3D Reconstruction of General Non-Rigid Scenes. *Computer Graphics Forum*, 2024. 2

[95] 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. In *CVPR*, 2023. 3

[96] Juze Zhang, Jingyan Zhang, Zining Song, Zhanhe Shi, Chengfeng Zhao, Ye Shi, Jingyi Yu, Lan Xu, and Jingya Wang. Hoi-m3: Capture multiple humans and objects interaction within contextual environment. In *CVPR*, 2024. 2

[97] 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 *European Conference on Computer Vision (ECCV)*, 2020. 3, 5

[98] Kaifeng Zhang, Yang Fu, Shubhankar Borse, Hong Cai, Fatih Porikli, and Xiaolong Wang. Self-Supervised Geometric Correspondence for Category-Level 6D Object Pose Estimation in the Wild, 2023. 7

[99] Mengqi Zhang, Yang Fu, Zheng Ding, Sifei Liu, Zhuowen Tu, and Xiaolong Wang. Hoidiffusion: Generating realistic 3d hand-object interaction data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8521–8531, 2024. 2

[100] Xiaohan Zhang, Bharat Lal Bhatnagar, Sebastian Starke, Vladimir Guzov, and Gerard Pons-Moll. Couch: Towards controllable human-chair interactions. In *European Conference on Computer Vision (ECCV)*. Springer, 2022. 2

[101] Zerong Zheng, Tao Yu, Qionghai Dai, and Yebin Liu. Deep implicit templates for 3d shape representation. In *CVPR*, 2021. 3

[102] Keyang Zhou, Bharat Lal Bhatnagar, Jan Eric Lenssen, and Gerard Pons-Moll. Toch: Spatio-temporal object correspondence to hand for motion refinement. In *ECCV*. Springer, 2022. 2

[103] Keyang Zhou, Bharat Lal Bhatnagar, Bernt Schiele, and Gerard Pons-Moll. Adjoint rigid transform network: Task-conditioned alignment of 3d shapes. In *2022 International Conference on 3D Vision (3DV)*. IEEE, 2022. 3, 5, 1

[104] Keyang Zhou, Bharat Lal Bhatnagar, Jan Eric Lenssen, and Gerard Pons-Moll. Gears: Local geometry-aware hand-object interaction synthesis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2

[105] Yi Zhou, Connelly Barnes, Lu Jingwan, Yang Jimei, and Li Hao. On the continuity of rotation representations in neural networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 5, 1