Perceiving 3D Human-Object Spatial Arrangements from a Single Image in the Wild

Jason Y. Zhang[1]*, Sam Pepose[2]*, Hanbyul Joo[2],
Deva Ramanan[1,3], Jitendra Malik[2,4], and Angjoo Kanazawa[4]

[1] Carnegie Mellon University
[2] Facebook AI Research
[3] Argo AI
[4] UC Berkeley

Abstract. We present a method that infers spatial arrangements and shapes of humans and objects in a globally consistent 3D scene, all from a single image in-the-wild captured in an uncontrolled environment. Notably, our method runs on datasets without any scene- or object-level 3D supervision. Our key insight is that considering humans and objects jointly gives rise to "3D common sense" constraints that can be used to resolve ambiguity. In particular, we introduce a scale loss that learns the distribution of object size from data; an occlusion-aware silhouette re-projection loss to optimize object pose; and a human-object interaction loss to capture the spatial layout of objects with which humans interact. We empirically validate that our constraints dramatically reduce the space of likely 3D spatial configurations. We demonstrate our approach on challenging, in-the-wild images of humans interacting with large objects (such as bicycles, motorcycles, and surfboards) and handheld objects (such as laptops, tennis rackets, and skateboards). We quantify the ability of our approach to recover human-object arrangements and outline remaining challenges in this relatively unexplored domain. The project webpage can be found at https://jasonyzhang.com/phosa.

1 Introduction

Tremendous strides have been made in estimating the 2D structure of in-the-wild scenes in terms of their constituent objects. While recent work has also demonstrated impressive results in estimating 3D structures, particularly human bodies, the focus is often on bodies [29,36] and objects [10,16] imaged in isolation or in controlled lab conditions [22,53]. To enable true 3D in-the-wild scene understanding, we argue that one must look at the *holistic* 3D scene, where objects and bodies can provide contextual cues for each other so as to correct local ambiguities. Consider the task of understanding the image in Figure 1. Independently estimated 3D poses of humans and objects are not necessarily consistent in the spatial arrangement of the 3D world of the scene (top row). When processed holistically, one can produce far more plausible 3D arrangements

* denotes equal contribution. Please direct correspondence to jasonyzhang@cmu.edu.

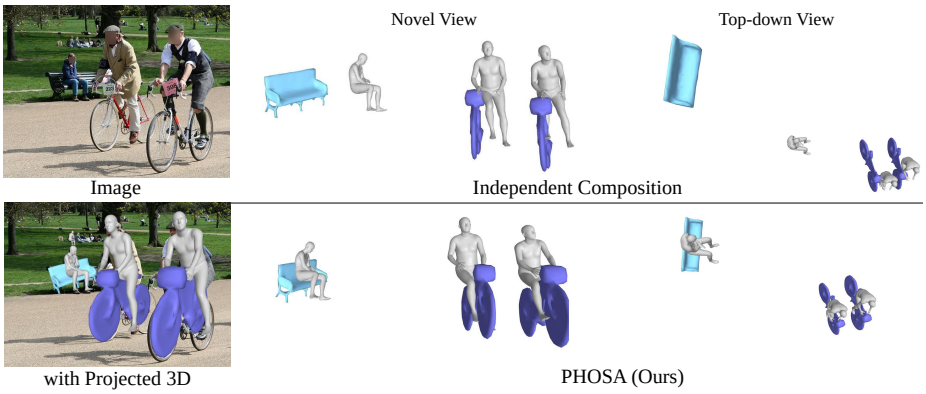

Fig. 1: We present PHOSA, Perceiving Human-Object Spatial Arrangements, an approach that recovers the spatial arrangements of humans and objects in 3D space from a single image by reasoning about their intrinsic scale, human-object interaction, and depth ordering. Given the input image (top left), we show two possible interpretations of the 3D scene that have similar 2D projections (bottom left). Using priors of humans, objects, and their interactions, our approach is able to recover the more reasonable interpretation (bottom row).

by exploiting contextual cues, such as the fact that humans tend to sit on park benches and ride bicycles rather than float mid-air.

In this paper, we present a method that can similarly recover the 3D spatial arrangement and shape of humans and objects in the scene from a single image. We demonstrate our approach on challenging, in-the-wild images containing multiple and diverse human-object interactions. We propose an optimization framework that relies on automatically predicted 2D segmentation masks to recover the 3D pose, shape, and location of humans along with the 6-DoF pose and intrinsic scale of key objects in the scene. Per-instance intrinsic scale allows one to convert each instance's local 3D coordinate system to coherent world coordinates, imbuing people and objects with a consistent notion of metric size.

There are three significant challenges to address. First is that the problem is inherently ill-posed as multiple 3D configurations can result in the same 2D projection. It is attractive to make use of data-driven priors to resolve such ambiguities. But we immediately run into the second challenge: obtaining training data with 3D supervision is notoriously challenging, particularly for entire 3D scenes captured in-the-wild. Our key insight is that considering humans and objects jointly gives rise to 3D scene constraints that reduce ambiguity. We make use of physical 3D constraints including a prior on the typical size of objects within a category. We also incorporate spatial constraints that encode typical modes of interactions with humans (e.g. humans typically interact with a bicycle by grabbing its handlebars). Our final challenge is that while there exists numerous mature technologies supporting 3D understanding of humans (including shape models and keypoint detectors), the same tools do not exist for the collective space of all objects. In this paper, we take the first step toward

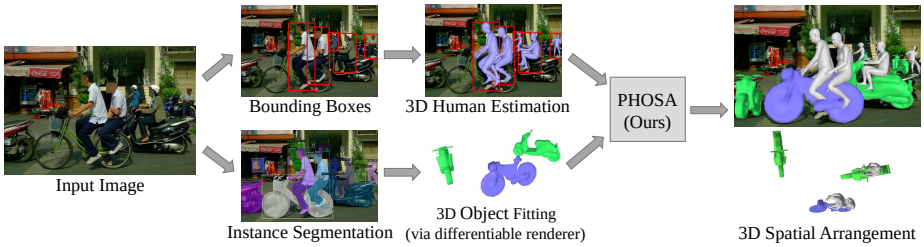

Fig. 2: **Overview of our method, PHOSA.** Given an image, we first detect instances of humans and objects [24]. We predict the 3D pose and shape of each person [27] and optimize for the 3D pose of each object by fitting to a segmentation mask [31]. Then, we convert each 3D instance in its own local coordinate frame into world coordinates using an intrinsic scale. Using our Human-Object Spatial Arrangement optimization, we produce a globally consistent output, as shown here. Our framework produces plausible reconstructions that capture realistic human-object interaction, preserve depth ordering, and obey physical constraints.

building such tools by learning the natural size distributions of object categories without any supervision. Our underlying thesis, bourne out by experiment, is that contextual cues arising from holistic processing of human-object arrangements can still provide enough information to understand objects in 3D.

We design an optimization-based framework, where we first reconstruct the humans and objects *locally* in each detected bounding box. For humans, we make use of state-of-the-art 3D human reconstruction output [27]. For objects, we solve for the 6-DoF parameters of a category-specific 3D shape exemplar that fits the local 2D object instance segmentation mask [31]. We then use a per-instance intrinsic scale to convert each local 3D prediction into a world coordinate frame by endowing metric size to each object and define a *global* objective function that scores different 3D object layouts, orientations, and shape exemplars. We operationalize constraints through loss terms in this objective. We make use of gradient-based solvers to optimize for the globally consistent layout. Although no ground truth is available for this task, we evaluate our approach qualitatively and quantitatively on the COCO-2017 dataset [40], which contains challenging images of humans interacting with everyday objects obtained in uncontrolled settings. We demonstrate the genericity of our approach by evaluating on objects from 8 categories of varying size and interaction types: baseball bats, bicycles, laptops, motorcycles, park benches, skateboards, surfboards, and tennis rackets.

2 Related Work

3D human pose and shape from a single image. Recovering the 3D pose and shape of a person from a single image is a fundamentally ambiguous task. As such, most methods employ statistical 3D body models with strong priors on shape learned from large-scale 3D scans and with known kinematic structure to model the articulation [4,65,41,47,28]. Seminal works in this area [17,54,3] fit the

parameters of a 3D body model to manual annotation such as silhouettes and keypoints obtained from users interaction [54,19,65]. Taking advantage of the progress in 2D pose estimation, [7] proposes a fully automatic approach where the parameters of the SMPL body model [41] are fit to automatically detected 2D joint locations in combination with shape and pose priors. More recent approaches employ expressive human models with faces and fingers [61,48]. Another line of work develops a learning based framework, using a feed-forward model to directly predict the parameters of the body model from a single image [57,51,29,45,59,42]. More recent approaches combine human detection with 3D human pose and shape prediction [20]. Much of the focus in such approaches is training on in-the-wild images of humans without any paired 3D supervision. [29] employ an adversarial prior on pose, [50] explore using ordinal supervision, and [49] use texture consistency. More recently, [36,27] have proposed hybrid approaches that combine feed-forward networks with optimization-based methods to improve 2D keypoint fit, achieving state-of-the-art results. In this work, we use the 3D regression network from [27] to recover the 3d pose and shape of humans.

Note that all of these approaches consider the human in isolation. More related to our work are methods that recover 3D pose and shape of multiple people [63,64,26]. These approaches use collision constraints to avoid intersection, and use bottom-up grouping or ordinal-depth constraints to resolve ambiguities. We take inspiration from these works for the collision loss and the depth ordering loss from [26], but our focus is on humans and objects in this work.

3D objects from a single image. There has also been significant literature in single-view 3D object reconstruction. Earlier methods optimize a deformable shape model to image silhouettes [39,32,5,30,25]. Recent approaches train a deep network to predict the 3D shape from an image [10,16,12,18,46,43,9]. Most of these approaches require 3D supervision or multi-view cues and are trained on synthetic datasets such as [60,55]. Many of these approaches reason about 3D object shape in isolation, while in this work we focus on their spatial arrangements. There are several works [56,38,15] that recover the 3D shape of multiple objects but still reason about them independently. More recently, [37] proposes a graph neural network to reason about the relationships between object to infer their layout, trained on a synthetic dataset with no humans. In this work we explore 3D spatial arrangements of humans and objects in the wild. As there is no 3D supervision for these images, we take the traditional category-based model-fitting approach to get the initial 6-DoF pose of the objects and refine their spatial arrangements in relation to humans and other objects in the scene.

3D human-to-object interaction. Related to our work are earlier approaches that infer about the 3D geometry and affordances from observing people interacting with scenes over time [11,13,21]. These approaches are similar to our work in spirit in that they use the ability to perceive humans to understand the 3D properties of the scene. The majority of the recent works rely on a pre-captured 3D scene to reason about 3D human-object interaction. [53] use RGB-D sensors to capture videos of people interacting with indoor scenes and use this data to

learn a probabilistic model that reasons about how humans interact with its environment. Having access to 3D scenes provides scene constraints that improve 3D human pose perception [62,35,52]. The recent PROX system [22] demonstrates this idea through an optimization-based approach to improve 3D human pose estimation conditioned on a known 3D scene captured by RGB-D sensors. While we draw inspiration from their contact terms to model human object interaction, we critically do not assume that 3D scenes are available. We experiment on single images captured in an uncontrolled *in-the-wild* environment, often outdoors.

More related to our approach are those that operate on images. There are several hand-object papers that recover both 3D object and 3D hand configurations [23]. In this work we focus on 3D spatial arrangements of humans and objects. Imapper [44] uses priors built from RGB-D data [53] to recover a plausible global 3D human motion and a global scene layout from an indoor video. Most related to our work is [8], who develop an approach that recovers a parse graph that represents the 3D human pose, 3D object, and scene layout from a single image. They similarly recover the spatial arrangements of humans and objects, but rely on synthetic 3D data to (a) learn priors over human-object interaction and (b) train 3D bounding box detectors that initialize the object locations. In this work we focus on recovering 3D spatial arrangements of humans and objects in the wild where no 3D supervision is available for 3D objects and humans and their layout. Due to the reliance on 3D scene capture and/or 3D synthetic data, many previous work on 3D human object interaction focus on indoor office scenes. By stepping out of this supervised realm, we are able to explore and analyze how 3D humans and objects interact in the wild.

3 Method

Our method takes a single RGB image as input and outputs humans and various categories of objects in a common 3D coordinate system. We begin by separately estimating 3D humans and 3D objects in each predicted bounding box provided by an object detector [24]. We use a state-of-the art 3D human pose estimator [27] to obtain 3D humans in the form of a parametric 3D human model (SMPL [41]) (in Sec. 3.1), and use a differentiable renderer to obtain 3D object pose (6-DoF translation and orientation) by fitting 3D mesh object models to predicted 2D segmentation masks [34] (in Sec. 3.2). The core idea of our method is to exploit the interaction between humans and objects to spatially arrange them in a common 3D coordinate system by optimizing for the per-instance *intrinsic scale*, which specifies their metric size (in Sec. 3.3). In particular, our method can also improve the performance of 3D pose estimation for objects by exploiting cues from the estimated 3D human pose. See Fig. 2 for the overview of our method.

3.1 Estimating 3D Humans

Given a bounding box for a human provided by a detection algorithm [24], we estimate the 3D shape and pose parameters of SMPL [41] using [27]. The 3D

Fig. 3: **Ambiguity in scale.** Recovering a 3D scene from a 2D image is fundamentally ambiguous because multiple 3D interpretations can have the same 2D projection. Consider the photo of the surfer on the left. A large surfboard far away (c) and a small one closer (a) have the same 2D projection (second panel) as the correct interpretation (b). In this work, we aim to resolve this scale ambiguity by exploiting cues from human-object interaction, essentially using the human as a "ruler."

human is parameterized by pose $\theta \in \mathbb{R}^{72}$ and shape $\beta \in \mathbb{R}^{10}$, as well as as a weak-perspective camera $\Pi = [\sigma, t_x, t_y] \in \mathbb{R}^3$ to project the mesh into image coordinates. To position the humans in the 3D space, we convert the weak-perspective camera to the perspective camera projection by assuming a fixed focal length f for all images, where the distance of the person is determined by the reciprocal of the camera scale parameter σ. Thus, the 3D vertices of the SMPL model for the i-th human is represented as,

$$V_h^i = \mathcal{M}(\beta^i, \theta^i) + \left[t_x^i \; t_y^i \; f/\sigma^i\right] \tag{1}$$

where $\mathcal{M}$ is the differentiable SMPL mapping from pose and shape to a human mesh with 6890 vertices in meters. The SMPL shape parameter β controls the height and size of the person. In practice, this is difficult to reliably estimate from an image since a tall, far-away person and a short, closeby person may project to similar image regions (Fig. 1 and Fig. 6). To address this ambiguity, we fix the estimated SMPL pose and shape and introduce an additional per-human intrinsic scale parameter $s_i \in \mathbb{R}$ that changes the size and thus depth of the human in world coordinates: $V_h^{i*} = s_i V_h^i$. While the shape parameter β also captures size, we opt for this parameterization as it can also be applied to objects (described below) and thus optimized to yield a globally consistent layout of the scene.

3.2 Estimating 3D Objects

We consider each object as a rigid body mesh model and estimate the 3D location $\mathbf{t} \in \mathbb{R}^3$, 3D orientation $\mathcal{R} \in SO(3)$, and an intrinsic scale $s \in \mathbb{R}$. The intrinsic scale converts the local coordinate frame of the template 3D mesh to the world frame. We consider single or multiple exemplar mesh models for each object category, pre-selected based on the shape variation within each category. For example, we use a single mesh for skateboards but four meshes for motorcycle. The mesh models are obtained from [1,2,38] and are pre-processed to have fewer faces (about 1000) to make optimization more efficient. See Fig. 5 for some

examples of mesh models and the supplementary for a full list. The 3D state of the jth object is represented as,

$$V_o^j = s^j \mathcal{R}^j \mathcal{O}(c^j, k^j) + \mathbf{t}^j, \tag{2}$$

where $\mathcal{O}(c^j, k^j)$ specifies the k^j-th exemplar mesh for category c^j. Note that the object category c^j is provided by the object detection algorithm [24], and k^j is automatically determined in our optimization framework (by selecting the exemplar that minimizes reprojection error).

Our first goal is to estimate the 3D pose of each object independently. However, estimating 3D object pose in the wild is challenging because (1) there are no existing parametric 3D models for target objects; (2) 2D keypoint annotations or 3D pose annotations for objects in the wild images are rare; and (3) occlusions are common in cluttered scenes, particularly those with humans. We propose an optimization-based approach using a differentiable renderer [31] to fit the 3D object to instance masks from [24] in a manner that is robust to partial occlusions. We began with an pixel-wise L2 loss over rendered silhouettes S versus predicted masks M, but found that it ignored boundary details that were important for reliable pose estimation. We added a symmetric chamfer loss [14] which focuses on boundary alignment, but found it computationally prohibitive since it required recomputing a distance transform of S at each gradient iteration. We found good results with L2 mask loss augmented with a *one-way* chamfer loss that computes the distance of each silhouette boundary pixel to the nearest mask boundary pixel, which requires computing a *single* distance transform once for the mask M. Given an no-occlusion indicator I (0 if pixel only corresponds to mask of different instance, 1 else), we write our loss as follows:

$$L_{\text{occ-sil}} = \sum (I \circ S - M)^2 + \sum_{p \in E(I \circ S)} \min_{\hat{p} \in E(M)} \|p - \hat{p}_2\| \tag{3}$$

where $E(M)$ computes the edge map of mask M. Note that this formulation can handle partial occlusions by object categories for which we do not have 3D models, as illustrated in Fig. 4. We also add an offscreen penalty to avoid degenerate solutions when minimizing the chamfer loss. To estimate the 3D object pose, we minimize the occlusion-aware silhouette loss:

$$\{\mathcal{R}^j, \mathbf{t}^j\}^* = \underset{\mathcal{R}, \mathbf{t}}{\operatorname{argmin}} L_{\text{occ-sil}} \left(\Pi_{\text{sil}}(V_o^j), M^j \right), \tag{4}$$

where Π_{sil} is the silhouette rendering of a 3D mesh model via a perspective camera with a fixed focal length (same as f in (1)) and M^j is a 2D instance mask for the j-th object. We use PointRend [34] to compute the instance masks. See Fig. 4 for a visualization and the supplementary for more implementation details. While this per-instance optimization provides a reasonable 3D pose estimate, the mask-based 3D object pose estimation is insufficient since there remains a fundamental ambiguity in determining the global location relative to other objects or people, as shown in Fig. 3. In other words, reasoning about instances in isolation cannot resolve ambiguity in the intrinsic scale of the object.

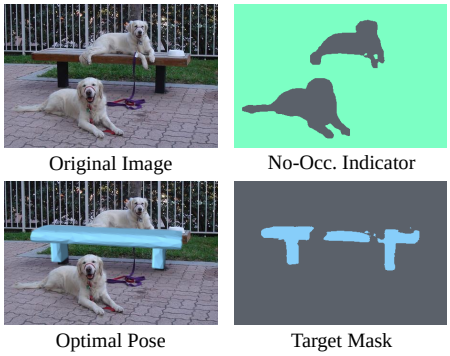

Fig. 4: **Occlusion-Aware Silhouette Loss for optimizing object pose.** Given an image, a 3D mesh model, and instance masks, our occlusion-aware silhouette loss finds the 6-DoF pose that most closely matches the target mask (bottom right). To be more robust to partial occlusions, we use a no-occlusion indicator (top right) to ignore regions that correspond to other object instances, including those for which we do not have 3D mesh models (e.g. dogs).

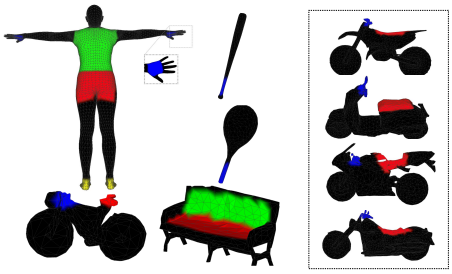

Fig. 5: **Part labels for fine-grained interaction.** To model human-object interactions, we label each object mesh with interaction regions corresponding to parts of the human body. Each color-coded region on the person (top left) interacts with the matching colored region for each object. The interaction loss pulls pairs of corresponding parts closer together. To better capture variation in shape, we can use multiple mesh instances for the same category (e.g. the motorcycles shown on the right). See the supplementary for all mesh models and interaction correspondences.

3.3 Modeling Human-Object Interaction for 3D Spatial Arrangement

Reasoning about the 3D poses of humans and objects independently may produce inconsistent 3D scene arrangements. In particular, objects suffer from a fundamental depth ambiguity: a large object further away can project to the same image coordinates as a small object closer to the camera (see Fig. 3). As such, the absolute 3D depth cannot be estimated. The interactions between humans and objects can provide crucial cues to reason about the relative spatial arrangement among them. For example, knowing that two people are riding on the same bike suggests that they should have similar depths in Fig 2. This pair-wise interaction cue can be propagated to determine the spatial arrangement of multiple humans and objects together. Furthermore, given the fact that a wide range of 2D and 3D supervision exists for human pose, we can leverage 3D human pose estimation to further adjust the orientation of the 3D object. For instance, knowing that a person is sitting down on the bench can provide a strong prior to determine the 3D orientation of the bench. Leveraging this requires two important steps: (1) identifying a human and an object that are interacting and (2) defining an objective function to correctly adjust their spatial arrangements.

Identifying human-object interaction. We hypothesize that an interacting person and object must be nearby in the world coordinates. In our formulation, we solve for the 6-DoF object pose and the intrinsic scale parameter s^j which

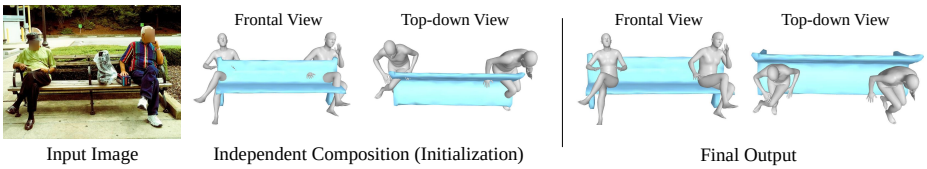

Fig. 6: **Recovering realistic human-object spatial arrangements by reasoning about depth ordering and interpenetration.** Given the image on the left as input, we first initialize the spatial arrangement by independently estimating the 3D human and object poses (Independent Composition). By incorporating physical priors such as avoiding mesh inter-penetration as well as preserving the depth ordering inferred from the predicted segmentation mask, we can produce the much more plausible spatial arrangement shown on the right.

places the object into world coordinates and imbues the objects with metric size. We use 3D bounding box overlap between the person and object to determine whether the object is interacting with a person. The size of the per-category 3D bounding box in world coordinates is set larger for larger object categories. See supplementary for a full list of 3D box sizes. Endowing a reasonable initial scale is important for identifying human-object interaction because if the object is scaled to be too large or too small in size, it will not be nearby the person. We first initialize the scale using common sense reasoning, via an internet search to find the average size of objects (e.g. baseball bats and bicycles are $\sim$0.9 meters $\sim$2 meters long respectively). Through our proposed method, the per-instance intrinsic scales change during optimization. From the final distribution of scales obtained over the test set, we compute the empirical mean scale and repeat this process using this as the new initialization (Fig. 7).

Objective function to optimize 3D spatial arrangements. Our objective includes multiple terms to provide constraints for interacting humans and objects:

$$L = \lambda_1 L_{\text{occ-sil}} + \lambda_2 L_{\text{interaction}} + \lambda_3 L_{\text{depth}} + \lambda_4 L_{\text{collision}} + \lambda_5 L_{\text{scale}}. \quad (5)$$

We optimize (5) using a gradient-based optimizer [33] w.r.t. intrinsic scale $s^i \in \mathbb{R}$ for the i-th human and intrinsic scale $s^j \in \mathbb{R}$, rotation $\mathcal{R}^j \in SO(3)$, and translation $\mathbf{t}^j \in \mathbb{R}^3$ for the j-th object instance jointly. The object poses are initialized from Sec. 3.2. $L_{\text{occ-sil}}$ is the same as (4) except without the chamfer loss which didn't help during joint optimization. We define the other terms below.

Interaction loss: We first introduce a coarse, instance-level interaction loss to pull the interacting object and person close together:

$$L_{\text{coarse inter}} = \sum_{h \in \mathcal{H}, o \in \mathcal{O}} \mathbb{1}(h, o) \|C(h) - C(o)\|_2, \quad (6)$$

where $\mathbb{1}(h, o)$ identifies whether human h and object o are interacting according to the 3D bounding box overlap criteria described before.

Humans generally interact with objects in specific ways. For example, humans hold tennis rackets by the handle. This can be used as a strong prior for human-object interaction and adjust their spatial arrangement. To do this, we annotate surface regions on the SMPL mesh and on our 3D object meshes where there is likely to be interaction, similar to PROX [22]. These include the hands, feet, and back of a person or the handlebars and seat of a bicycle, as shown in Fig. 5. To encode spatial priors about human-object interaction (e.g. people grab bicycle handlebars by the hand and sit on the seat), we enumerate pairs of object and human part regions that interact (see supplementary for a full list). We incorporate a fine-grained, parts-level interaction loss by using the part-labels (Fig. 5) to pull the interaction regions closer to achieve better alignment:

$$L_{\text{fine inter}} = \sum_{h \in \mathcal{H}, o \in \mathcal{O}} \sum_{\substack{\mathcal{P}_h, \mathcal{P}_o \in \\ \mathcal{P}(h,o)}} \mathbb{1}(\mathcal{P}_h, \mathcal{P}_o) \|C(\mathcal{P}_h) - C(\mathcal{P}_o)\|_2, \tag{7}$$

where $\mathcal{P}_h$ and $\mathcal{P}_o$ are the interaction regions on the person and object respectively. Note that we define the parts interaction indicator $\mathbb{1}(\mathcal{P}_h, \mathcal{P}_o)$ using the same criteria as instances, i.e. 3D bounding box overlap. The interactions are recomputed at each iteration. Finally, $L_{\text{interaction}} = L_{\text{coarse inter}} + L_{\text{fine inter}}$.

Scale loss: We observe that there is a limit to the variation in size within a category. Thus, we incorporate a Gaussian prior on the intrinsic scales of instances in the same category using a category-specific mean scale:

$$L_{\text{scale}} = \sum_{c} \sum_{j \in [|\mathcal{O}_c|]} \|s_j - \bar{s}_c\|_2. \tag{8}$$

We initialize the intrinsic scale of all objects in category c to $\bar{s}_c$. The mean object scale $\bar{s}_c$ is initially set using common sense estimates of object size. In Fig. 7, we visualize the final distribution of object sizes learned for the COCO-2017 [40] test set after optimizing for human interaction. We then repeat the process with the empirical mean as a better initialization for $\bar{s}_c$. We also incorporate the scale loss for the human scales s_i with a mean of 1 (original size) and a small variance.

Ordinal Depth loss: The depth ordering inferred from the 3D placement should match that of the image. While the correct depth ordering of people and objects would also minimize the occlusion-aware silhouette loss, we posit that the ordinal depth loss introduced in Jiang et al [26] can help recover more accurate depth orderings from the modal masks. Using an ordinal depth can give smoother gradients to both the occluder and occluded object. Formally, for each pair of instances, we compare the pixels at the intersections of the silhouettes with the segmentation mask. If at pixel p, instance i is closer than instance j but the segmentation masks at p show j and not i, then we apply a ranking loss on the depths of both instances at pixel p:

$$L_{\text{depth}} = \sum_{o_i \in \mathcal{H} \cup \mathcal{O}} \sum_{o_j \in \mathcal{H} \cup \mathcal{O}} \sum_{\substack{p \in \text{Sil}(o_i) \\ \cap \text{Sil}(o_j)}} \mathbb{1}(p, o_i, o_j) \log\left(1 + \exp(D_{o_j}(p) - D_{o_i}(p))\right), \tag{9}$$

where $\text{Sil}(o)$ is the rendered silhouette of instance o, $D_o(p)$ is the depth of instance o at pixel p, and $\mathbb{1}(p, o_i, o_j)$ is 1 if the segmentation label at pixel p is o_j but $D_{o_i}(p) < D_{o_j}(p)$. See [26] for more details.

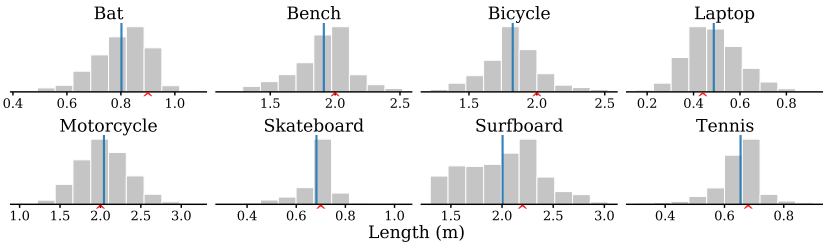

Fig. 7: **Learned size distribution from human interaction:** Here, we visualize the distribution of object sizes across the COCO-2017 [40] test set at the end of optimization. The red caret denotes the size resulting from the hand-picked scale used for initialization. The blue line denotes the size produced by the empirical mean scale of all category instances at the end of optimization. We then use the empirical mean as the new initialization for intrinsic object scale.

Collision loss: Promoting proximity between people and objects can exacerbate the problem of instances occupying the same 3D space. To address this, we penalize poses that would human and/or object interpenetration using the collision loss $L_{\text{collision}}$ introduced in [6,58]. We use a GPU implementation based on [47] which detects colliding mesh triangles, computes a 3D distance field, and penalizes based on the depth of the penetration. See [47] for more details.

4 Evaluation

In this section, we provide quantitative and qualitative analysis on the performance of our method on the COCO-2017 [40] dataset. We focus our evaluation on 8 categories: baseball bats, benches, bicycles, laptops, motorcycles, skateboards, surfboards, and tennis rackets. These categories cover a significant variety in size, shape, and types of interaction with humans.

4.1 Quantitative Analysis

Since 3D ground truth annotations for both humans and objects do not exist for in-the-wild images, we used a forced-choice evaluation procedure on COCO-2017 [40] images. To test the contribution of the holistic processing of human and object instances, we evaluate our approach against an "independent composition," which uses our approach from Sec. 3.1 and Sec. 3.2 to independently estimate the human and object poses. To make the independent composition competitive, we set the intrinsic scale to be the empirical mean per-category scale learned over the test set by our proposed method in Sec. 3.3. This actually gives the independent composition global information through human-object interaction. This is the best that can be done without considering all instances holistically.

We collected a random subset of images from the COCO 2017 test set in which at least one person and object overlap in 2D. For each of our object categories, we randomly sample 50 images with at least one instance of that category. For each

Fig. 8: **Qualitative results of our method on test images from COCO 2017.** Our method, PHOSA, recovers plausible human-object spatial arrangements by explicitly reasoning about human interaction. We evaluate the importance of modeling interaction by comparing with independently estimated human and object poses also using our method (Independent Composition). The intrinsic scale for the independent composition is set to the per-category empirical mean scale learned by our method.

Ours vs.	Bat	Bench	Bike	Laptop	Motor.	Skate.	Surf.	Tennis
Indep. Comp.	83	74	73	61	74	71	82	82
No $L_{\text{occ-sil}}$	79	74	87	76	70	96	80	77
No $L_{\text{interaction}}$	82	59	57	46	71	71	76	68
No L_{scale}	77	49	54	51	54	55	55	56
No L_{depth}	50	55	55	55	52	50	51	50
No $L_{\text{collision}}$	52	40	51	51	50	52	50	50

Table 1: **Percentage of images for which our proposed method performs better on a subset of COCO 2017 test set.** In this table, we evaluate our approach against an independent composition and ablations of our method. The independent composition estimates human and object pose independently using Sec. 3.1 and Sec. 3.2 and sets the intrinsic scale to the empirical mean category scale learned by our method. The ablations each drop one loss term from our proposed method. In each row, we compute the average percentage of images for which our method performs better across a random sample of COCO 2017 test set images [40]. A number greater than 50 implies that our proposed method performs better than the independent composition or that the ablated loss term is beneficial for that category.

image, annotators see the independent composition and the result of our proposed method in random order, marking whether our result looks better than, equal to, or worse than the independent composition (see supplementary for screenshots of our annotating tool). We compute the average percentage of images for which our method performs better (treating equal as 50) in Tab. 1. Overall, we find that our method safely outperforms the independent composition.

To evaluate the importance of the individual loss terms, we run an ablative study. We run the same forced choice test for the full proposed method compared with dropping a single loss term in Tab. 1. We find that omitting the occlusion-aware silhouette loss and the interaction loss has the most significant effect. Using the silhouette loss during global optimization ensures that the object poses continue to respect image evidence, and the interaction loss encodes the spatial arrangement of the object relative to the person. We did observe that the interaction loss occasionally pulls the object too aggressively toward the person for the laptop category. The scale loss appears to have a positive effect for most categories. Note that because we initialized the scale to the empirical mean, the effects of the scale loss are not as pronounced as they would be if initialized to something else. The depth ordering loss gave a moderate boost while the collision loss had a less pronounced effect. We attribute this to the collision loss operating only on the surface triangles and thus prone to getting stuck in local minima in which objects get embedded inside the person (especially for benches).

4.2 Qualitative Analysis

In Fig. 8, we demonstrate the generality of our approach on the COCO 2017 dataset by reconstructing the spatial arrangement of multiple people and objects

engaged in a wide range of 3D human-object interactions. We find that our method works on a variety of everyday objects with which people interact, ranging from handheld objects (baseball bats, tennis rackets, laptops) to full-sized objects (skateboards, bicycles, motorcycles) to large objects (surfboards, benches). In the middle column of Fig. 8, we visualize the spatial arrangement produced by the independent composition introduced in Sec. 4. We find that independently estimating human and object poses is often insufficient for resolving fundamental ambiguities in scale. Explicitly reasoning about human-object interaction produces more realistic spatial arrangements. Please refer to the Supplementary Materials for discussion of failure modes and significantly more qualitative results.

5 Discussion

In summary, we have found that 2D and 3D technologies for understanding objects and humans have advanced considerably. Armed with these advances, we believe the time is right to start tackling broader questions of holistic 3D scene-understanding—and moving such questions from the lab to uncontrolled in-the-wild imagery! Our qualitative analysis suggests that 3D human understanding has particularly matured, even for intricate interactions with objects in the wild. Detailed recovery of 3D object shape is still a challenge, as illustrated by our rather impoverished but surprisingly effective exemplar-based shape model. It would be transformative to learn statistical models (a "SMPL-for-objects"), and we take the first step by learning the intrinsic scale distribution from data.

A number of conclusions somewhat surprised us. First, even though object shape understanding lacks some tools compared to its human counterpart (such as statistical 3D shape models and keypoint detectors), 2D object instance masks combined with a differentiable renderer and a 3D shape library proves to be a rather effective initialization for 3D object understanding. Perhaps even more remarkable is the scalability of such an approach. Adding new objects and defining their modes of interactions is relatively straightforward, because it is far easier to "paint" annotations on 3D models than annotate individual image instances. Hence 3D shapes provide a convenient coordinate frame for *meta*-level supervision. This is dramatically different from the typical supervised pipeline, in which adding a new object category is typically quite involved.

While the ontology of objects will likely be diverse and continue to grow and evolve over time, humans will likely remain a consistent area of intense focus and targeted annotation. Because of this, we believe it will continue to be fruitful to pursue approaches that leverage contextual constraints from humans that act as "rulers" to help reason about objects. In some sense, this philosophy harkens back to Protagoras's quote from Ancient Greece—"man is the measure of all things"!

Acknowledgements: We thank Georgia Gkioxari and Shubham Tulsiani for insightful discussion and Victoria Dean and Gengshan Yang for useful feedback. We also thank Senthil Purushwalkam for deadline reminders. This work was funded in part by the CMU Argo AI Center for Autonomous Vehicle Research.

References

1. 3d Warehouse, https://3dwarehouse.sketchup.com
2. Free3d, https://free3d.com
3. Agarwal, A., Triggs, B.: Recovering 3d human pose from monocular images. TPAMI (2006)
4. Anguelov, D., Srinivasan, P., Koller, D., Thrun, S., Rodgers, J., Davis, J.: SCAPE: Shape Completion and Animation of PEople. SIGGRAPH (2005)
5. Aubry, M., Maturana, D., Efros, A.A., Russell, B.C., Sivic, J.: Seeing 3d chairs: exemplar part-based 2d-3d alignment using a large dataset of cad models. In: CVPR (2014)
6. Ballan, L., Taneja, A., Gall, J., Van Gool, L., Pollefeys, M.: Motion capture of hands in action using discriminative salient points. In: ECCV (2012)
7. Bogo, F., Kanazawa, A., Lassner, C., Gehler, P., Romero, J., Black, M.J.: Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In: ECCV (2016)
8. Chen, Y., Huang, S., Yuan, T., Qi, S., Zhu, Y., Zhu, S.C.: Holistic++ scene understanding: Single-view 3d holistic scene parsing and human pose estimation with human-object interaction and physical commonsense. In: CVPR (2019)
9. Chen, Z., Zhang, H.: Learning implicit fields for generative shape modeling. In: CVPR (2019)
10. Choy, C.B., Xu, D., Gwak, J., Chen, K., Savarese, S.: 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In: ECCV (2016)
11. Delaitre, V., Fouhey, D.F., Laptev, I., Sivic, J., Gupta, A., Efros, A.A.: Scene semantics from long-term observation of people. In: ECCV (2012)
12. Fan, H., Su, H., Guibas, L.J.: A point set generation network for 3d object reconstruction from a single image. In: CVPR (2017)
13. Fouhey, D.F., Delaitre, V., Gupta, A., Efros, A.A., Laptev, I., Sivic, J.: People watching: Human actions as a cue for single view geometry. IJCV (2014)
14. Gavrila, D.M.: Pedestrian detection from a moving vehicle. In: ECCV. pp. 37–49. Springer (2000)
15. Georgia Gkioxari, Jitendra Malik, J.J.: Mesh r-cnn. ICCV (2019)
16. Girdhar, R., Fouhey, D., Rodriguez, M., Gupta, A.: Learning a predictable and generative vector representation for objects. In: ECCV (2016)
17. Grauman, K., Shakhnarovich, G., Darrell, T.: Inferring 3d structure with a statistical image-based shape model. In: ICCV (2003)
18. Groueix, T., Fisher, M., Kim, V.G., Russell, B.C., Aubry, M.: A papier-mâché approach to learning 3d surface generation. In: CVPR (2018)
19. Guan, P., Weiss, A., Balan, A.O., Black, M.J.: Estimating human shape and pose from a single image. In: ICCV (2009)
20. Guler, R.A., Kokkinos, I.: Holopose: Holistic 3d human reconstruction in-the-wild. In: CVPR (2019)
21. Gupta, A., Satkin, S., Efros, A.A., Hebert, M.: From 3d scene geometry to human workspace. In: CVPR (2011)
22. Hassan, M., Choutas, V., Tzionas, D., Black, M.J.: Resolving 3d human pose ambiguities with 3d scene constraints. In: ICCV (2019)
23. Hasson, Y., Varol, G., Tzionas, D., Kalevatykh, I., Black, M.J., Laptev, I., Schmid, C.: Learning joint reconstruction of hands and manipulated objects. In: CVPR (2019)
24. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: ICCV (2017)

25. Izadinia, H., Shan, Q., Seitz, S.M.: Im2cad. In: CVPR (2017)
26. Jiang, W., Kolotouros, N., Pavlakos, G., Zhou, X., Daniilidis, K.: Coherent reconstruction of multiple humans from a single image. In: CVPR. pp. 5579–5588 (2020)
27. Joo, H., Neverova, N., Vedaldi, A.: Exemplar fine-tuning for 3d human pose fitting towards in-the-wild 3d human pose estimation. arXiv preprint arXiv:2004.03686 (2020)
28. Joo, H., Simon, T., Sheikh, Y.: Total capture: A 3d deformation model for tracking faces, hands, and bodies. In: CVPR (2018)
29. Kanazawa, A., Black, M.J., Jacobs, D.W., Malik, J.: End-to-end recovery of human shape and pose. In: CVPR (2018)
30. Kar, A., Tulsiani, S., Carreira, J., Malik, J.: Category-specific object reconstruction from a single image. In: CVPR (2015)
31. Kato, H., Ushiku, Y., Harada, T.: Neural 3d mesh renderer. In: CVPR (2018)
32. Kholgade, N., Simon, T., Efros, A., Sheikh, Y.: 3d object manipulation in a single photograph using stock 3d models. ACM Transactions on Graphics (TOG) (2014)
33. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
34. Kirillov, A., Wu, Y., He, K., Girshick, R.: Pointrend: Image segmentation as rendering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9799–9808 (2020)
35. Kjellström, H., Kragić, D., Black, M.J.: Tracking people interacting with objects. In: CVPR (2010)
36. Kolotouros, N., Pavlakos, G., Black, M.J., Daniilidis, K.: Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In: ICCV (2019)
37. Kulkarni, N., Misra, I., Tulsiani, S., Gupta, A.: 3d-relnet: Joint object and relational network for 3d prediction. In: CVPR (2019)
38. Kundu, A., Li, Y., Rehg, J.M.: 3d-rcnn: Instance-level 3d object reconstruction via render-and-compare. In: CVPR (2018)
39. Lim, J.J., Pirsiavash, H., Torralba, A.: Parsing ikea objects: Fine pose estimation. In: ICCV (2013)
40. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: ECCV (2014)
41. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: SMPL: A skinned multi-person linear model. SIGGRAPH Asia (2015)
42. Mehta, D., Sridhar, S., Sotnychenko, O., Rhodin, H., Shafiei, M., Seidel, H.P., Xu, W., Casas, D., Theobalt, C.: Vnect: Real-time 3d human pose estimation with a single rgb camera. In: SIGGRAPH (2017)
43. Mescheder, L., Oechsle, M., Niemeyer, M., Nowozin, S., Geiger, A.: Occupancy networks: Learning 3d reconstruction in function space. In: CVPR (2019)
44. Monszpart, A., Guerrero, P., Ceylan, D., Yumer, E., Mitra, N.J.: imapper: interaction-guided scene mapping from monocular videos. ACM Transactions on Graphics (TOG) (2019)
45. Omran, M., Lassner, C., Pons-Moll, G., Gehler, P.V., Schiele, B.: Neural body fitting: Unifying deep learning and model-based human pose and shape estimation. In: 3DV (2018)
46. Park, J.J., Florence, P., Straub, J., Newcombe, R., Lovegrove, S.: Deepsdf: Learning continuous signed distance functions for shape representation. In: CVPR (2019)
47. Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A.A.A., Tzionas, D., Black, M.J.: Expressive body capture: 3d hands, face, and body from a single image. In: CVPR (2019)

48. Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A.A.A., Tzionas, D., Black, M.J.: Expressive body capture: 3d hands, face, and body from a single image. In: CVPR (2019)
49. Pavlakos, G., Kolotouros, N., Daniilidis, K.: TexturePose: Supervising human mesh estimation with texture consistency. In: ICCV (2019)
50. Pavlakos, G., Zhou, X., Daniilidis, K.: Ordinal depth supervision for 3D human pose estimation. In: CVPR (2018)
51. Pavlakos, G., Zhu, L., Zhou, X., Daniilidis, K.: Learning to estimate 3D human pose and shape from a single color image. In: CVPR (2018)
52. Rosenhahn, B., Schmaltz, C., Brox, T., Weickert, J., Cremers, D., Seidel, H.P.: Markerless motion capture of man-machine interaction. In: CVPR (2008)
53. Savva, M., Chang, A.X., Hanrahan, P., Fisher, M., Nießner, M.: Pigraphs: learning interaction snapshots from observations. ACM Transactions on Graphics (TOG) (2016)
54. Sigal, L., Balan, A., Black, M.J.: Combined discriminative and generative articulated pose and non-rigid shape estimation. In: NeurIPS (2008)
55. Song, S., Yu, F., Zeng, A., Chang, A.X., Savva, M., Funkhouser, T.: Semantic scene completion from a single depth image. In: CVPR (2017)
56. Tulsiani, S., Gupta, S., Fouhey, D., Efros, A.A., Malik, J.: Factoring shape, pose, and layout from the 2d image of a 3d scene. In: CVPR (2018)
57. Tung, H.Y.F., Tung, H.W., Yumer, E., Fragkiadaki, K.: Self-supervised learning of motion capture. In: NeurIPS (2017)
58. Tzionas, D., Ballan, L., Srikantha, A., Aponte, P., Pollefeys, M., Gall, J.: Capturing hands in action using discriminative salient points and physics simulation. IJCV (2016)
59. Varol, G., Ceylan, D., Russell, B., Yang, J., Yumer, E., Laptev, I., Schmid, C.: Bodynet: Volumetric inference of 3d human body shapes. In: ECCV (2018)
60. Wu, Z., Song, S., Khosla, A., Yu, F., Zhang, L., Tang, X., Xiao, J.: 3d shapenets: A deep representation for volumetric shapes. In: CVPR (2015)
61. Xiang, D., Joo, H., Sheikh, Y.: Monocular total capture: Posing face, body, and hands in the wild. In: CVPR (2019)
62. Yamamoto, M., Yagishita, K.: Scene constraints-aided tracking of human body. In: CVPR (2000)
63. Zanfir, A., Marinoiu, E., Sminchisescu, C.: Monocular 3D pose and shape estimation of multiple people in natural scenes — the importance of multiple scene constraints. In: CVPR (2018)
64. Zanfir, A., Marinoiu, E., Zanfir, M., Popa, A.I., Sminchisescu, C.: Deep network for the integrated 3D sensing of multiple people in natural images. In: NeurIPS (2018)
65. Zhou, S., Fu, H., Liu, L., Cohen-Or, D., Han, X.: Parametric reshaping of human bodies in images. In: SIGGRAPH (2010)