Shape2Pose: Human-Centric Shape Analysis

Vladimir G. Kim
Stanford University

Siddhartha Chaudhuri
Princeton University

Leonidas Guibas
Stanford University

Thomas Funkhouser
Princeton University

Abstract

As 3D acquisition devices and modeling tools become widely available there is a growing need for automatic algorithms that analyze the semantics and functionality of digitized shapes. Most recent research has focused on analyzing geometric structures of shapes. Our work is motivated by the observation that a majority of man-made shapes are designed to be used by people. Thus, in order to fully understand their semantics, one needs to answer a fundamental question: "how do people interact with these objects?" As an initial step towards this goal, we offer a novel algorithm for automatically predicting a static pose that a person would need to adopt in order to use an object. Specifically, given an input 3D shape, the goal of our analysis is to predict a corresponding human pose, including contact points and kinematic parameters. This is especially challenging for man-made objects that commonly exhibit a lot of variance in their geometric structure. We address this challenge by observing that contact points usually share consistent local geometric features related to the anthropometric properties of corresponding parts and that human body is subject to kinematic constraints and priors. Accordingly, our method effectively combines local region classification and global kinematically-constrained search to successfully predict poses for various objects. We also evaluate our algorithm on six diverse collections of 3D polygonal models (chairs, gym equipment, cockpits, carts, bicycles, and bipedal devices) containing a total of 147 models. Finally, we demonstrate that the poses predicted by our algorithm can be used in several shape analysis problems, such as establishing correspondences between objects, detecting salient regions, finding informative viewpoints, and retrieving functionally-similar shapes.

CR Categories: I.3.5 [Computer Graphics]: Computational Geometry and Object Modeling—Geometric algorithms, languages, and systems;

Keywords: shape analysis, affordance analysis

Links: DL PDF WEB DATA CODE

1 Introduction

With the increasing availability of digitized 3D models, there is a growing need for automatic algorithms that can assist in semantic parsing of geometric shapes. To meet this need, a variety of shape analysis algorithms have been proposed in recent years, including methods for saliency estimation, shape segmentation, feature detection, symmetry analysis, and surface correspondence; and, many tools have used these algorithms for analyzing, searching, organizing, designing, and editing 3D shapes [Mitra et al. 2013].

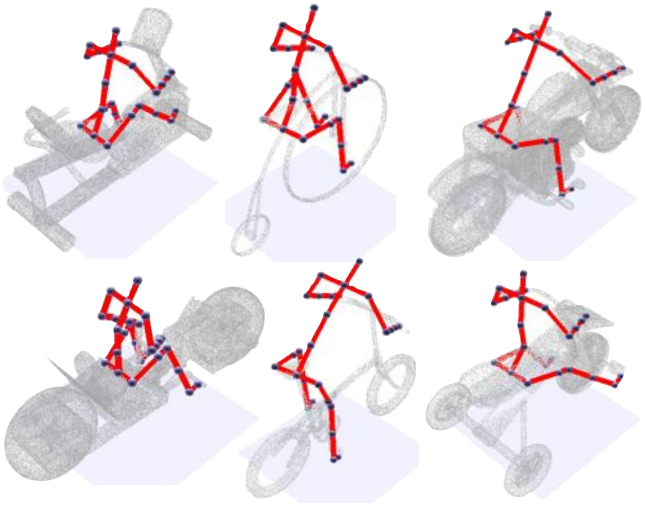


Figure 1: *Given an input shape, our algorithm predicts a human pose using a trained affordance model. The predicted joint angles and surface contact points can be used to detect functional similarities between the shapes, establish a set of key point correspondences, and mark semantically salient surface regions.*

Most previous algorithms have focused on geometric analysis, mainly using techniques that compute global shape properties, extract part structures, and/or detect local shape features. While these algorithms have advanced greatly within recent years, they usually can provide little information about the semantics or function of an object, and they often struggle to provide any information for classes of objects with high intra-class shape diversity.

In this paper, we propose a new shape analysis tool based on *object affordance* – a quality of an object that allows someone to perform an action [Gibson 1977]. We observe that knowing how a human interacts with an object provides useful information about its semantics and function, even when the shape of the object is difficult to parse.

As a demonstration of this idea, consider the six shapes depicted in Figure 1. Although the shapes are quite different from one another globally, and they share limited similarity in part shapes and arrangements (one does not even have wheels), it is easy to tell that they are all some form of bipedal device based on the riding pose taken by the person shown in red. Moreover, by simply observing the person's pose, we can immediately tell the relevant symmetries of the objects, the functions of some parts (e.g., the handlebar is for grasping by a hand), semantic correspondences between points on different surfaces (e.g., all points contacted by the right foot are in functional correspondence), and important geometric properties to preserve during shape editing (e.g., maintain the spatial relationships between pedals, seats, and handlebars).

The main challenge in leveraging affordance for shape analysis is to automatically predict the pose that a human will take when using a given object. We need an algorithm that can produce a semantically appropriate human pose for any given 3D model. We address this

challenge with an algorithm that leverages consistency of anthropometric features across different shapes and poses. Our approach relies upon the following observations: (i) local geometric features strongly correlate with geometry of corresponding body parts (e.g. we sit on relatively flat areas and grab cylinder-like regions), (ii) human poses are subject to kinematic constraints and priors (e.g. a knee cannot bend backwards), (iii) humans exhibit bilateral symmetry, and (iv) shape surfaces cannot be penetrated by a person.

Although local geometric features provide very strong cues for human-object interaction [Norman 1988], they alone are not sufficient for predicting contact points in most cases (e.g. grasping a bike's frame is as easy as holding its handles). This motivates a joint bottom-up and top-down search that leverages both local features (i) and global constraints (ii, iii, iv). For the bottom-up search we use a classifier to predict candidate contact points on the surface, and our top-down optimization searches for the most plausible pose that allows reaching the high-probability contacts. Our pipeline relies on training data to learn pose priors and geometric features of contact points for different body parts. The main technical contribution of this work is a polynomial-time optimization algorithm that finds an approximate minimum in the combinatorial space of body-to-contact assignments. Our algorithm stems from the insight that after learning a pose prior one can sample a large number of poses from the prior and pre-compute probability distributions for positions of body parts. This allows our method to quickly find combinations of high-probability contact points that can be reached with plausible poses.

To summarize, the main contributions of this paper are: (i) we introduce *Shape2Pose*, a novel affordance-inspired shape analysis tool that predicts human pose parameters and surface contact points, (ii) we develop an efficient polynomial-time algorithm for exploring the combinatorial space of contact assignments by pre-computing a probability distribution for body parts, and (iii) we create a dataset of 147 models from diverse object classes with annotated ground truth poses. Finally, we thoroughly evaluate our method and demonstrate favorable performance in comparison to direction extensions of related methods.

2 Related Work

This work describes a geometry analysis tool for semantic understanding of shapes from an affordance perspective. In this section we review the current research on shape analysis and affordance analysis.

Shape Analysis. The availability of 3D shape repositories and the rising number of applications that leverage geometric data demand algorithms for structural analysis and structure-aware manipulation of shapes [Mitra et al. 2013]. Previous work concentrates on detecting symmetries [Mitra et al. 2012], upright orientation [Fu et al. 2008], geometric variations [Ovsjanikov et al. 2011; Kim et al. 2013], consistent segmentations [Golovinskiy and Funkhouser 2009; Huang et al. 2011; Sidi et al. 2011], region classification [Kalogerakis et al. 2010], and correspondences [Huang et al. 2012; Kim et al. 2012] in collections and in individual shapes. These analyses can facilitate synthesis [Kalogerakis et al. 2012], exploration [Ovsjanikov et al. 2011; Kim et al. 2012], and interactive modeling [Gal et al. 2009; Chaudhuri et al. 2011]. In this work we extend this toolkit with affordance analysis that predicts contact points and pose parameters, essentially detecting invariant structural relations that can serve to facilitate applications and to relate diverse shapes via common use patterns.

Affordance Analysis. Recent work in computer vision demonstrates that observing how people interact with shapes can help in shape recognition [Delaitre et al. 2012; Fouhey et al. 2012; Wei et al. 2013]. But even without observing the actual interaction, one can predict a list of semantic tags to represent the types of actions an object affords [Fritz et al. 2006; Hermans et al. 2011], note that this analysis is also commonly performed jointly with object classification [Sun et al. 2009; Stark et al. 2008]. A more detailed affordance representations also demonstrated to be fruitful, for example predicting an approximate alignment of a human model to a shape [Grabner et al. 2011; Gupta et al. 2011]. A recent approach by [Jiang et al. 2013] uses a probabilistic framework to select a pose from a list of six most common poses and rigidly align it to models in a 3D scene. They also demonstrate applications in object labeling and automatic object placement [Jiang et al. 2012; Jiang and Saxena 2012; Jiang and Saxena 2013]. The main advantage of our work is that it does not assume that there is a small set of discrete poses; instead, we search a continuous pose parameter space. This finer representation enables higher accuracy and applications beyond shape classification.

Grasp Prediction. Modeling and predicting grasping interactions received special attention in robotics [Bohg et al. 2013]. Data-driven techniques often rely on machine learning and train on annotated example shapes to predict graspable regions based on their geometric features [Saxena et al. 2006; Saxena 2009; Lenz et al. 2013]. Alternatively, one can use shape retrieval to find a similar object in an annotated database and transfer a grasping pose [Goldfeder and Allen 2011]. For more complex geometries the query shape can also be decomposed into graspable parts [Bard and Troccaz 1990; Przybylski et al. 2012]. Finally, applications that require a robot to be able to manipulate an object often rely on a physical simulation [Rosales et al. 2011; Feix et al. 2013]. Previous work in graphics also leverages kinematic models to create realistically looking grasping actions that are similar to example captured data [Pollard and Zordan 2005; Ying et al. 2007; Zhao et al. 2013].

These previous approaches focus on robot and virtual hand interaction, while we focus on shape analysis. Also note that a grasping interaction has a simple functional purpose (to be able to pick something up), while in our work we face the challenge of producing semantic poses without hardcoding into the algorithm the functional context of the pose.

3 Human-Centric Shape Analysis

The key idea behind our work is to leverage the prediction of object affordances during shape analysis. Following the insight that the pose adopted by a human body when interacting with a shape provides strong and persistent cues about its semantics and function [Gibson 1977], we propose a system that integrates affordance analysis into semantic parsing of 3D shapes. To investigate this idea, we have developed an algorithm called *Shape2Pose* that simultaneously predicts kinematic parameters for a static human pose and points of contact between a human body and the shape's surface, and then we use the algorithm to assist classical problems in 3D shape analysis, including salience estimation, surface correspondence, viewpoint selection, and shape retrieval.

The main difference from previous work on shape parsing in computer graphics [Mitra et al. 2013] is that we supply extra critical information to our analysis algorithm: *the shape and deformation modes of a human body.* This extra input allows our algorithm to search for specific geometric structures and spatial arrangements of relevance to a person using the object. That is, rather than asking the system: "what significant patterns you can find in this shape?"

we ask it: "how can this human body be posed to interact with this shape?" The second question is far more specific, easier to solve, and more likely to reveal semantic information relevant to the function of an object.

The main difference from recent work on affordance analysis in computer vision [Jiang et al. 2013; Grabner et al. 2011] is that we establish human-object contact points by searching a continuous space of human body deformations (rather than a small, discrete number of fixed poses). This difference is important for providing precise contact point labels for shape parsing applications and for analyzing object classes with large intra-class pose variations (e.g., gym equipment). It is akin to docking flexible molecules to minimize an energy function based on explicit atomic bonds rather than aligning molecules to fit roughly with rigid transformations.

The advantages of our approach are highlighted in Figure 2. In the three examples shown, each of the input shapes has global shape that is quite different from any of the others, multiple partial symmetries, many local geometries that could be labeled as salient, many similarly-shaped surfaces that could serve as human contact points, and multiple places that could support a typical human pose. However, by fitting a human pose into the shape to optimize both the plausibility of joint angles and the precise locations of contact points (e.g., hands on the steering wheel, feet on the foot pedals, etc.), the functionally relevant human pose and contact points can be identified automatically (our result is shown in red). Moreover, from the predicted human pose, important shape features can be detected (handles for grasping), semantic similarities between shapes can be found (used in a sitting position), and functional annotations can be inferred (the pose suggests steering wheels are turned by a person).

4 System Overview

Our system consists of two stages: training and prediction.

For the training stage, the input is a collection of shapes with manually prescribed contact points and poses represented by joint angles, and the output is an affordance model represented by an objective function that measures the quality of a pose for any novel shape. The affordance model incorporates terms learned from examples to model the local geometry of contact points and the joint angles for human interaction poses, and it includes penalty terms for deviations from reflectional symmetry and intersections with the shape.

For the prediction stage, the input is a novel shape, and the output is a set of joint angles and contact points for a likely human interaction pose. The key algorithm in this stage searches the combinatorial space of human poses to find the one with small energy according to the affordance model. It interleaves sampling from the distribution of likely contact points, constraining the search to consider likely solutions in the joint distribution. In order to align these two distributions that are defined over different domains, we sample a large number of poses from the joint angle distributions, and for each region in space store a probability of it being reached by a body part. Since distribution of body parts and potential contact points are now defined in Euclidean space, they can be aligned with a rigid transformation. The peaks in joint probability under such an alignment define a candidate pose. For each of the best candidate poses, we run an inverse kinematic optimization to evaluate the exact value of the affordance model, returning the best found as the final solution. For example, in Figure 3, given a bicycle (a), our system predicts high probability contacts (b), and computes distribution of body parts (c). Optimizing the joint probability leads to the final kinematically plausible pose (d).

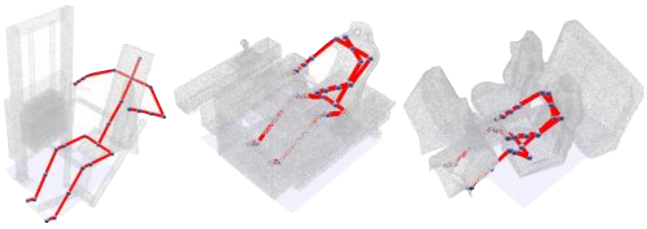


Figure 2: *Predicted poses for some partially symmetric shapes from the cockpit and gym equipment datasets.*

The following two sections describe the algorithmic components of these stages in more detail.

4.1 Learning an Affordance Model

In the training stage of our process, we learn an affordance model for a class of shapes. Our goal in this stage is to build an energy function that can be used to evaluate the interaction between a shape S and a human pose represented by a rigid transformation T, joint angles $\theta = \{\theta_1, ..., \theta_n\}$, key body parts P (*back*, *pelvis*, *palms*, and *toes*), and contact point assignments $m : P \rightarrow S \cup \{ground, unassigned\}$.[1]

Our affordance model is defined as the minimizer of an energy function over the space of possible poses for the shape:

$$E(T,\theta,m,S) = w_{\text{dist}}E_{\text{dist}}(T,\theta,m,S) + w_{\text{feat}}E_{\text{feat}}(m,S)$$
$$+w_{\text{pose}}E_{\text{pose}}(\theta) + w_{\text{sym}}E_{\text{sym}}(T,m,S) + w_{\text{isect}}E_{\text{isect}}(T,\theta,S) \qquad (1)$$

The first two terms on the right are local penalties defined for salient parts on the body assigned to contact points on the shape ($p \not\rightarrow$ *unassigned*): E_{dist} penalizes parts that do not touch the target point (Section 4.1.1), and E_{feat} penalizes contact assignments if the local surface geometry is ill-suited for placement of the corresponding part (Section 4.1.2). The remaining terms define global pose constraints: E_{pose} penalizes implausible poses (Section 4.1.3), E_{sym} penalizes misalignment of object and human symmetries (Section 4.1.4), and E_{isect} penalizes surface intersections (Section 4.1.5). We set weights $w_{\text{dist}} = 1000$, $w_{\text{feat}} = 10$, $w_{\text{pose}} = 0.3$, $w_{\text{sym}} = 1$ and $w_{\text{isect}} = 0.05$ for all experiments presented in this paper. We now describe the energy terms in more detail.

4.1.1 Contact Distance. If a body part is assigned to a surface point, we want to ensure they actually make contact. Hence, we penalize large separations between them. This can be viewed as a hard constraint, since it is enforced with very high weight.

$$E_{\text{dist}} = \sum_{p\in P,\ m(p)\neq unassigned} \|T\mathbf{p}_\theta - m(p)\|^2 \qquad (2)$$

where $\mathbf{p}_\theta$ is the position of p given joint angles θ. If a part is assigned to the ground, we measure the separation in height.

4.1.2 Feature Compatibility. Feature compatibility measures how likely it is for a surface point to be in contact with a particular body part. For example, hands should be placed on graspable points, and the pelvis on areas that resemble seats. Given training shapes $S_1, S_2, \ldots, S_M$ with annotated ground truth contacts $m_i : P \rightarrow S_i$, we learn a regression model $V_p : S \rightarrow [0,1]$ for each body part $p \in P$ which estimates the probability that it will be placed on a point on a query surface S. The model relies on local shape features to predict which regions are compatible with the body part: for instance, large flat areas afford sitting. The same predictor is used for symmetric body parts, such as hands.

[1] Some parts may be unassigned and rest in free space: $p \rightarrow$ *unassigned*, or may be placed on the ground plane: $p \rightarrow$ *ground*.

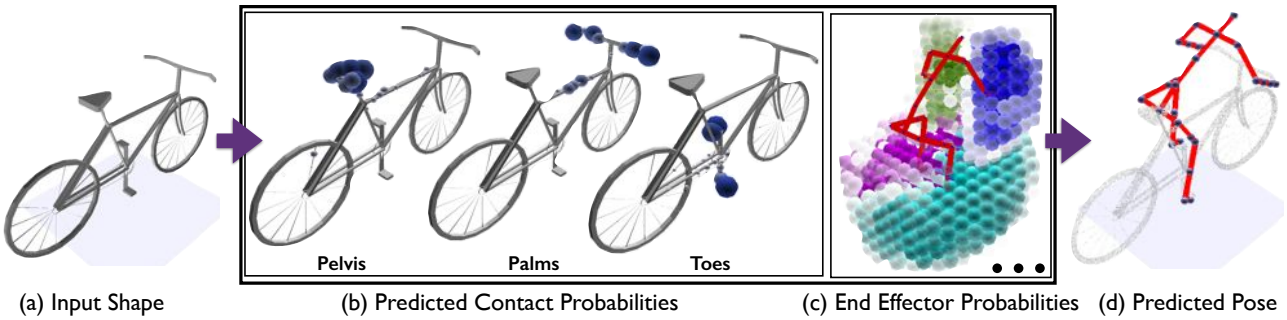


Figure 3: *Example execution of our pose prediction pipeline: (a) starting with the input shape, we (b) classify surface regions as potential contact points (darker color indicates higher confidence), (c) find end effector probability distributions, and (d) find the best pose, minimizing terms from (b) and (c).*

We sample $1000 \cdot A$ points $C_{S_i} = \{c_1, c_2, .., c_K\}$ with the iterative farthest-point algorithm on each shape S_i, where A is the shape's surface area in square meters. Geometric features are computed at these points. The features include principal component analysis on local neighborhoods, local symmetry axes, height above the ground plane, curvature, shape diameter function, and a histogram of distances to points in a human-sized neighborhood (please refer to the appendix for details).

Next, we produce a training signal V_p^i for each body part p and training shape S_i, which has value 1 at the ground truth contact point $m_i(p)$ and falls off smoothly to zero in a geodesic neighborhood. Specifically, the signal $V_p^i(c_j)$ at sample point c_j is $\exp\left(\frac{-g(c_j,\, m_i(p))^2}{\tau^2}\right)$, where $g(\cdot,\cdot)$ is the geodesic distance, and τ is chosen so that the signal at 20cm from the contact point is 0.4. Figure 4(a) shows several training signals overlaid on shapes.

Finally, for each body part p, we train a random regression forest [Breiman 2001] with 30 trees to estimate the function V_p. During the prediction stage, the regression forest is used to predict feature compatibility at each candidate contact point assigned to a body part. The overall compatibility energy E_{feat} is given by the sum:

$$E_{\text{feat}} = \sum_{p \in P} -\log V_p(m(p)) \tag{3}$$

For parts mapped to the ground plane or unassigned, compatibility is estimated from training data statistics. Specifically, $V_p(ground) = M_p^{ground}/M$, where M_p^{ground} is the number of times part p was placed on the ground (or similarly, unassigned). To avoid infinite energy we set a lower bound of 0.1 on the unassigned compatibility.

Figure 4(b) shows high-probability contacts predicted by the model on a scanned query shape (for this result we omitted features that require a polygonal surface, such as curvature and shape diameter). Note that while locally supported features are useful for identifying potential contact regions, especially in partially occluded shapes such as this scan, they usually provide only a weak and diffuse indication of actual contact points. For example, there is confusion between hands and backs since the training examples have both armrests and backs with flat, unbroken vertical structure. Only the global pose prior can resolve this ambiguity, as shown in Figure 4(c).

4.1.3 Pose Prior. The pose prior distinguishes plausible poses from implausible ones. We learn the pose prior from training examples and represent it as mixtures of Gaussians over joint angles. We use the same skeletal model, constructed from linked hinge and ball-and-socket joints, in all examples (see Figure 5).

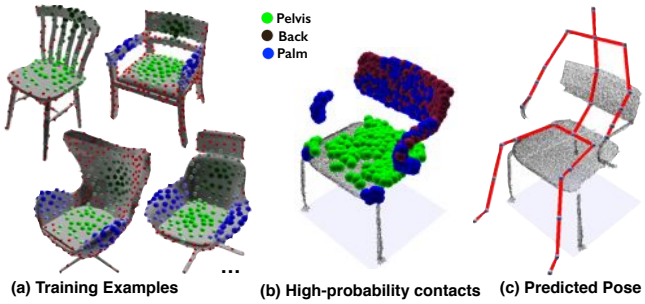


Figure 4: *Example predictions on a 3D Kinect scan of a chair. (a) We train affordance priors on the chairs dataset, then (b) predict high-probability contact points, and (c) predict the final pose. Note that while predictions based on local geometric features are robust to occlusions, they still suffer from large variance in the geometry of the chairs. Thus, the pose prior provides an essential cue for final positioning of contact points.*

Each pose is represented as a concatenated 40-dimensional vector of joint angles θ. We use mean-shift clustering to group the input poses into L clusters (in our experiments we obtain $L \leq 2$ for most datasets, except for Gym Equipment where $L \leq 5$). Within each cluster l, the variation of the i^{th} joint angle θ_i is represented by a Gaussian with learned mean μ_i^l and standard deviation σ_i^l. We also observe that natural poses are often symmetric, so for each pair of symmetric joints $(\theta_i, \theta_i^{\text{sym}})$ (e.g. left and right knees) we learn a Gaussian representing its deviation from perfect left-right symmetry: $|\theta_i - \theta_i^{\text{sym}}| \sim \mathcal{N}(\mu_i^{\text{sym}}, \sigma_i^{\text{sym}})$. For the angle θ_i controlling left-right bending of the spine, we set $\theta_i^{\text{sym}} = 0$ to discourage tilting. Our final pose prior energy is:

$$E_{\text{pose}} = \min_{l \in L} \sum_{i}^{40} \frac{|\theta_i - \mu_i^l|^2}{(\sigma_i^l)^2} + \frac{\left(|\theta_i - \theta_i^{\text{sym}}| - \mu_i^{\text{sym}}\right)^2}{(\sigma_i^{\text{sym}})^2} \tag{4}$$

Figure 5 shows the distribution of end effectors generated by sampling pose priors in different datasets. Note that we chose to model the distribution of each joint angle independently, to allow a wide range of poses of varying plausibility. Even with this fairly general prior, we are able to achieve good results.

4.1.4 Symmetry. We observe that the parts of shapes with which humans interact typically have local bilateral symmetry, presumably to reflect the symmetry of the human body. Hence, we penalize poses which are not symmetric with respect to a local symmetry plane, if one can be detected. We run the algorithm of Podolak et al. [2006] to find all prominent symmetry planes supported by approximately symmetric neighborhoods of diameter 1m. These planes define mirror maps over Euclidean space: $SP = \{sp : \mathbb{R}^3 \rightarrow \mathbb{R}^3\}$. Given a candidate pose, we penalize devia-

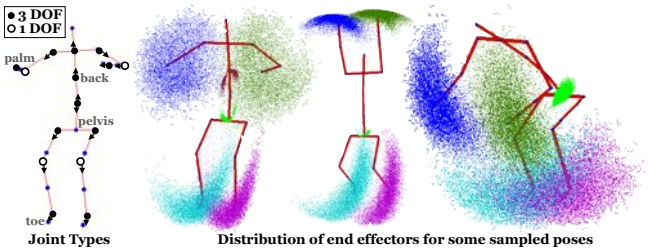


Figure 5: *Example pose distributions in the gym dataset. For the skeletal structure on the left, we show the distribution in positions of body parts. Each point corresponds to a relative body part position in a different pose.*

tions from symmetry in the most prominent plane sp^*, if any, within 2 meters. We also penalize improbable separations between the center of local surface symmetry $\mathbf{sp}_c^*$ and the center of the human $\mathbf{h}_c$. The separation $d = T\mathbf{h}_c - \mathbf{sp}_c^*$ is assumed to follow a Gaussian $\mathcal{N}(\mu_c, \sigma_c)$ learned from training data. The final symmetry energy is:

$$E_{\text{sym}} = w_{\text{plane}} \sum_{p \in P} \|m(p^{\text{sym}}) - sp^*(m(p))\| + w_{\text{center}} \frac{|d - \mu_c|^2}{\sigma_c^2} \quad (5)$$

where p^{sym} is the body part symmetric to p. $w_{\text{plane}} = 3$ and $w_{\text{center}} = 0.5$ in all experiments.

4.1.5 Surface Intersection. The final energy term helps avoid intersections between the shape and the human. For simplicity, we assume the body is represented as a skeleton, with linear bones $B = \{b_1, b_2, \ldots, b_K\}$ connecting joints. For each link b_i we compute its intersections $I_S(b_i)$ with the input shape S, ignoring intersections within 5cm of body parts assigned to contact points. We ascribe higher penalty if the bone intersects the surface orthogonally. The intersection energy is the sum of maximal per-link penalties:

$$E_{\text{isect}} = \sum_{b_i \in B} \max_{q \in I_S(b_i)} |\text{normal}(q) \cdot \text{direction}(b_i)| \quad (6)$$

4.2 Predicting a Human Pose

In the prediction stage of our system, we use a learned affordance model to predict an interaction pose (T, θ, m) for a novel shape S. The key challenge is to sample the space of human-shape interactions efficiently. To address this challenge, we interleave sampling from contact point assignments and joint angles.

Note that our pose representation is overspecified: if contact points m are assigned, one can solve for (T, θ) using inverse kinematics. Similarly, given T and θ, one can infer contact points m via nearest neighbors (or keep them unassigned), since distance to assigned contact points E_{dist} dominates other energy terms. However, parameterizing the problem by just (T, θ), or by just m, makes it difficult to efficiently explore minima of the overall energy function $E(T, \theta, m, S)$, since energy terms would then be expressed in terms of the complex nonlinearities of a nearest neighbor search or an IK solve. The key insight behind our optimization procedure is that it is possible to sample high-probability contact assignments m and high-probability poses θ *independently*, since they contribute to different major energy terms in Equation 1, $E_{\text{feat}}(m)$ and $E_{\text{pose}}(\theta)$ respectively.

We sample high probability assignments of contact point $m(p)$ for each body part p independently, by picking candidate points on the shape whose compatibility energy with p is lower than the cost of leaving them unassigned. The toy situation in Figure 6(a) shows points with high feature compatibility for two different types of body parts (green and blue).

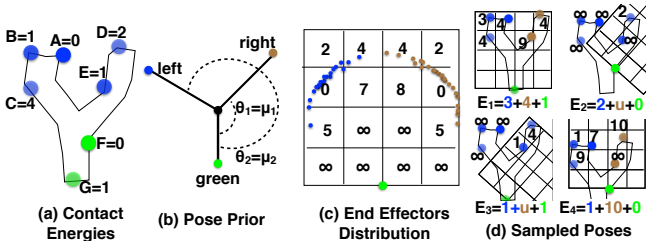


Figure 6: *We demonstrate our optimization procedure for a toy example: (a) the classified surface, where each point has a certain energy penalty, (b) a pose prior for a two-angle skeleton, (c) distribution of end effectors (left and right) with respect to the green reference point, and (d) some sampled poses with the final energies, where* u *is the penalty for keeping an end effector unassigned.*

We sample plausible poses with low energy E_{pose} by directly sampling the joint angle Gaussians from the pose prior. In our experiments we sample 50,000 poses: the whole process takes only a fraction of a second. If we fix an anchor point and assume T is the identity transform, we can co-align the sampled poses and produce a distribution of salient body part positions relative to the anchor. Figure 6(c) shows the distribution of left and right end-effectors with respect to the green point for a toy two-angle skeleton. Now, given an initial assignment of a body part $m(p_0)$ and a transformation T, we can treat p_0 as the anchored point and align the distributions over the surface to look for contact points that can be reached with a high-probability pose.

In practice, we discretize the space into a grid of 10cm^3 voxels, each storing a portion of the energy of the most plausible sampled pose that places a given body part in this voxel. The stored partial energy is the sum of pose prior penalties (Equation 5) due to all joints on the path to the body part from the anchor. Thus, a lower bound on the overall energy of a pose can be estimated by simply adding up the entries of the voxels containing the individual body parts. Note that joints might contribute to more than one partial energy if paths to different body parts overlap: if so, we average their contributions over overlapping paths. Because of these simplifications, as well as the finite resolution of the grid, the grid penalty is only an approximate lower bound on E_{pose}.

Next, we try every sampled contact point as an anchor for the corresponding body part, and consider 32 rotations around the up axis. The anchor and the rotation define the transform T, which aligns part distribution grids to the surface (Figure 6(d) shows four example alignments). Given the aligned grid, we estimate a lower bound on the feature, pose and symmetry energy, as well as the corresponding pose, by greedily assigning body parts to contact points. Each successive assignment $m(p_i)$ is chosen to be the one that least increases $E_{\text{feat}} + E_{\text{pose}} + E_{\text{sym}}$, where symmetry is measured w.r.t. the previously assigned $i-1$ points, and the pose prior is bounded from below by the entry in the aligned grid cell containing the assigned contact point. In our toy example, for each alignment, we show penalties due to each contact point, and the final energy produced by selecting the best contacts (for simplicity we omit the symmetry term). Note that although contact points F, A have the best feature compatibility, and G, C, E would be the least distorted pose, the optimal pose combining both features and pose prior is defined by a different set of points: G, B, D.

Finally, in order to produce the overall best pose, we must compute the full energy, which requires knowledge of the exact joint angles θ. We sort all candidate poses in order of increasing estimated lower bound on energy, and for each pose we minimize $E_{\text{dist}} + E_{\text{pose}}$ with respect to θ via inverse kinematics. We solve for θ using a variant of the Jacobian inverse technique (similar to the method

described by Buss [2005]). In brief, given target positions $\{m(p_i)\}$ and current body part positions $\{\mathbf{p}_{\theta,i}\}$, we find the contact distance energies $i = 1, 2 \ldots k$: $e_i = \|m(p) - \mathbf{p}_{\theta,i}\|^2$, and pose energy due to each angle $j = 1, 2 \ldots n$: $e_{k+j} = E_{\text{pose}}(\theta_j)$. The Jacobian matrix J has k rows with derivatives for contact distance $\partial E_{\text{dist}}(p_i)/\partial\theta_j$, and n rows for pose priors $\partial E_{\text{pose}}(\theta_j)/\partial\theta_j$. We solve the damped least squares problem: $\Delta\theta = (J^T J + \lambda^2 I)^{-1} J^T e$, and iteratively update θ until the energy $E_{\text{dist}} + E_{\text{pose}}$ stops improving. As a further refinement, we perform fine-scale greedy optimization of contact assignments up to a search radius of 1.5 times the grid cell size, to fix potential misalignments due to grid discretization. We iterate through the poses until the lower bound on energy of all remaining poses is higher than the current best pose, following which we terminate and return the best pose.

5 Results

In this section, we present results of experiments with our affordance analysis technique. The goals of these experiments are to: 1) test whether our algorithm can correctly predict human poses for diverse classes of objects, 2) test whether the algorithm works for shapes where people use them in unusual poses, 3) compare the results of our algorithm to alternative approaches, and 4) evaluate the impact of different aspects of our algorithm on the final results.

For these experiments, we created a benchmark of 6 data sets comprising a total of 147 polygonal models extracted from [van Kaick et al. 2013; Kim et al. 2013; Trimble 2013]. Each data set contains a collection of shapes from one object class with great shape diversity and interesting human-object interactions (gym equipment, cockpits, bicycles, carts, chairs, and bipedal devices). All shapes are represented as polygon soups without color or structure and have consistent up-right orientation and scale.

To generate ground-truth affordances for these data sets, we asked a student volunteer to manually provide ground truth poses for typical human interactions in a two stage procedure. First, the volunteer prescribed contact points for salient body regions (pelvis, back, arms and toes), with the possibility of assigning a point to a ground or leaving it unassigned. Then we ran an inverse kinematics procedure to produce a pose that allowed reaching the contact points. In the second stage the volunteer fixed the implausible poses by directly editing the joint angles. This process resulted in a single ground-truth pose for each shape.

Pose Prediction. To test our algorithms for predicting poses for new shapes, we ran a leave-one-out experiment for each dataset (i.e. we train on all except for one model and then predict the pose for the omitted model).

Figures 1, 2, and 7 show representative results. Qualitative inspection of these results suggests that our method successfully predicts correct poses among a diverse set of objects and for a wide range of poses. It also appears that the poses predicted by our algorithm yield significant semantic information about an object. For example, the relationships between poses predicted for the shopping cart and wheel barrow in the bottom left of 7 reveal the similarities in their functions (they both are pushed by people), provide cues about structural relationships between parts (e.g., the handles on both are for grasping, the two handles of the wheelbarrow have symmetric functions, etc.), and suggest ways that the shape could be optimized to improve human affordance (e.g., the handles of the wheelbarrow could be spaced further apart).

In order to quantitatively evaluate the correctness of the poses predicted by our algorithm, we measure the distances between all predicted and ground-truth joint positions for each model. Figure 9 show plots of the errors, where each point on a curve represents the fraction of joints whose error is less than the distance threshold listed on the horizontal axis (ranging from 0 to 25 centimeters). From these plots, we see that our algorithm (the red curve) predicts positions within 20cm of their manually prescribed locations for 70%-85% of joints in all datasets, except the gym equipment which is still over 50%.

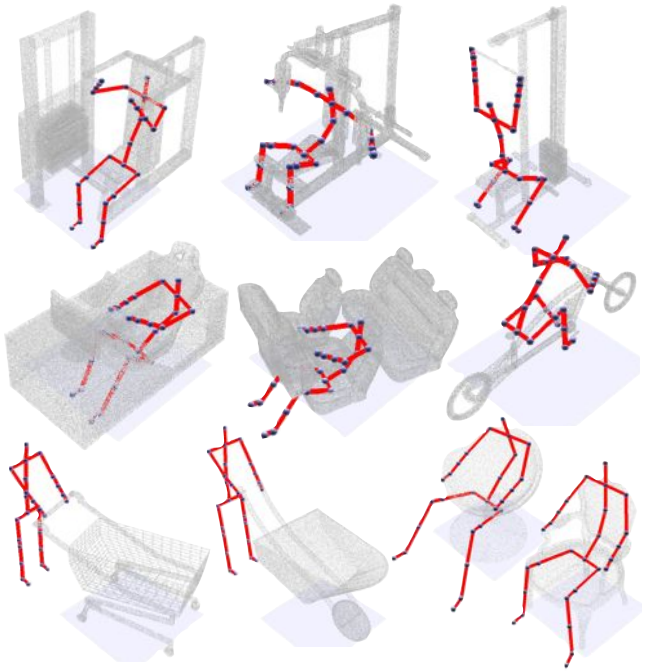


Figure 7: *Example predicted poses for different classes of shapes.*

Comparison. In order to evaluate the benefits of our particular implementation of affordance analysis, we compare to potential alternative techniques. Note that none of the existing methods address the problem we are trying to solve in this paper (predicting kinematic parameters for the interaction pose), and so we must compare to extensions of existing techniques:

- **Shape Matching** - we investigate whether global shape similarity provides enough information to predict a pose. In particular, we use the method of [Huang et al. 2013] to co-align all shapes in the entire dataset. Then, given a query shape, we find the most similar model and transfer contact points from that model via closest co-aligned points. Then we run our inverse kinematics optimization to produce a valid pose.

- **Rigid Pose** - we compare to affordance analysis techniques that model the affordance with a small set of rigid poses. In particular, we execute our algorithm, but disallow any deviations from a best mean pose in a mixture of Gaussians.

- **Local** - we compare our method to surface classification methods. In particular, we pick the best contact point to every body part and then run the inverse kinematics optimization to produce a valid pose.

Figure 8 shows visualizations of results predicted with these methods, and Figure 9 shows plots of their average prediction errors for each data set in comparsion to our method. These results suggest that our algorithm provides the most accurate predictions for most data sets. Looking in more detail, we find that the global shape matching approach (blue curve) fails if the training dataset does

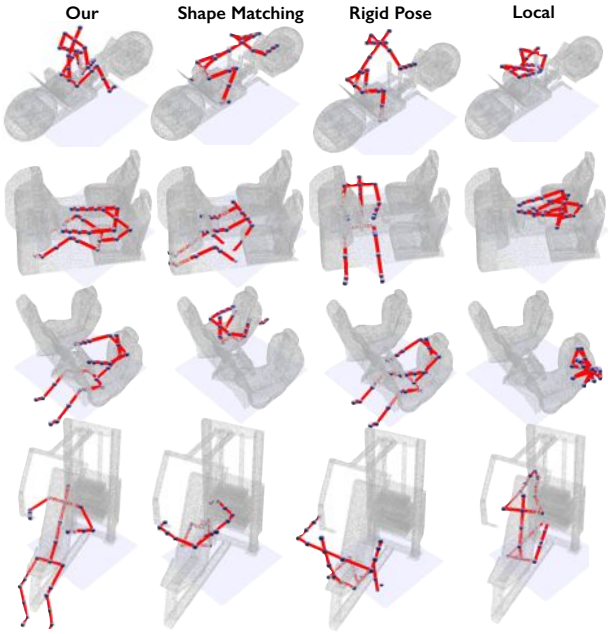


Figure 8: *A comparison of our method to possible alternatives: shape matching, computing affordance with a rigid pose, and using only local region classifier.*

not have a globally similar shape (e.g. the cockpits fail, since other models in the training data only include one-seat or five-seat examples). Also, since shape matching is not aware of local geometry of functional parts it often misaligns contact points (e.g. the handles are transferred to the front wheel of the bipedal vehicle). That said, the technique performs well on object classes that have similar global structure, such as chairs. Matching a rigid pose (green curve) performs well only on datasets where the variance in poses is relatively small (e.g. cockpits), and degrades quickly as relative arrangement of functional parts becomes more diverse (all other datasets). Classification of local geometry features (purple curve) provide very weak cues about shape affordance without the additional global pose constraints. This last result confirms our expectation that human-centric shape analysis can provide information beyond that of local shape analysis.

Training Data Size. We test how the size of training data affects the performance of our method. Specifically, in this leave-one-out experiment we use random training sets of different sizes for each test shape. Figure 10 shows changes in accuracy for different classes of shapes as we vary the number of training examples. Note that there is a general upward trend in accuracy as we use more training models, however the curves are not strictly increasing because training subsets are chosen at random independently for each experiment. We also observe that the first few training examples significantly boost the accuracy, but the curves remain relatively flat after about 15 training models for most classes of shapes. If shapes and poses exhibit higher variance, such as in the Gym Equipment dataset, more training examples are necessary, and an overall lower accuracy is achieved.

Mixed Classes. Although our method is designed for analyzing relatively homogeneous datasets, we perform a stress-test experiment where both training and test data are heterogenous. Specifically, we pick three very dissimilar classes of shapes and execute our algorithm as-is on datasets obtained by mixing pairs of classes, without any class-specific annotations. Figure 11 shows the accuracy curves for pose predictors trained on each class separately

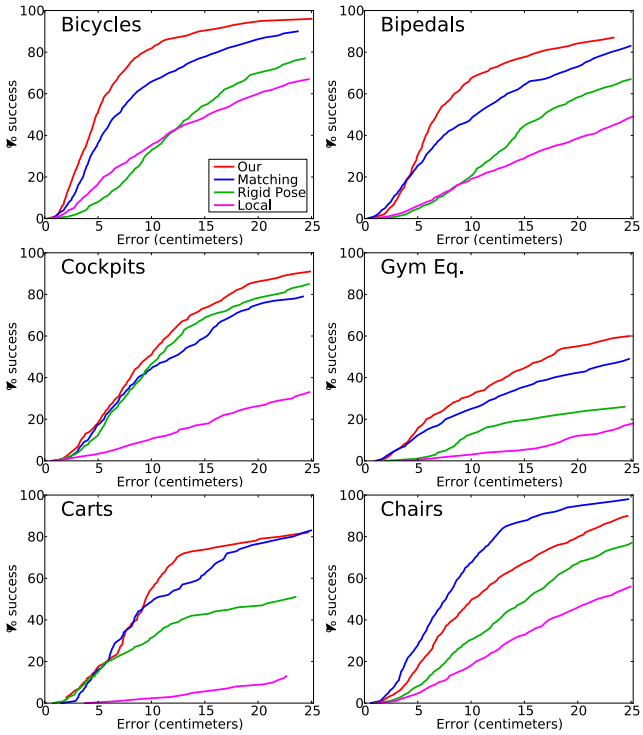


Figure 9: *We quantitatively compare our technique to potential alternatives, such as predicting poses using global shape matching, using a small set of rigid poses, or using only local features. This plot depicts fraction of correct joint positions (y-axis) for a given distance threshold (x-axis, given in centimeters).*

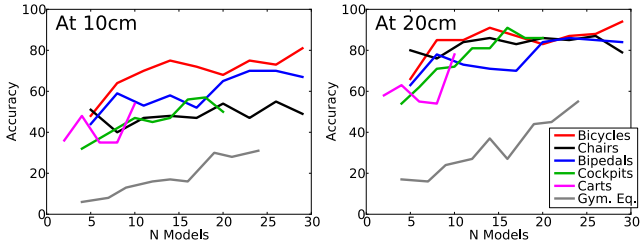


Figure 10: *Effect of size of training set on accuracy of final poses. In this figure each curve shows the accuracy of predicted poses (y-axis) relative to the number of training examples (x-axis), for different threshold distances (*10*cm and* 20*cm).*

(dashed lines) and trained on the mixed datasets (solid lines). Note that for related shapes (e.g. chairs and cockpits) the drop in performance is small, and it is somewhat more significant for less similar classes of shapes (e.g. bicycles and either chairs or cockpits). In the future, it would be interesting to combine geometry-based classification algorithms with our pose prediction method to facilitate both recognition of objects as well as the quality of predicted poses in heterogenous model collections.

In the remainder of the results section we evaluate contribution of different aspects of our algorithm to the final results.

Contact Point Classification. First, we evaluate how well our surface classification step can use local geometric features to predict contact points. We evaluate each contact type for each dataset separately in our leave-one-out experimental setup. In particular, given a prediction function and a threshold τ_{val} we find a set of positively-labeled points $C(\tau_{\text{val}}) = \{c_j\}$, s.t. $V_i(c_j) > \tau_{\text{val}}$. Then,

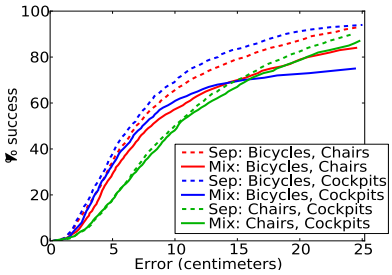


Figure 11: *Effect of highly heterogenous datasets. We plot the accuracy of our method using a pose predictor trained separately on different classes of shapes (dashed lines) and trained on mixed datasets (solid lines).*

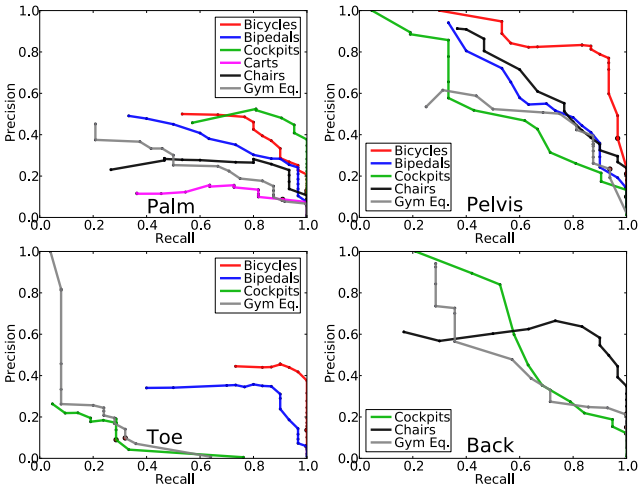


Figure 12: *Precision-recall curves for different types of contact points; higher recall means that the ground truth contact was chosen as a candidate, and higher precision means that candidate contacts are near a true contact. Clearly, local features alone are not sufficient to detect contact points.*

we compute precision for the set $C(\tau_{\text{val}})$ assuming that the point is true positive if it is within $\tau_{\text{gt}} = 10$cm of a ground truth contact. We similarly compute recall for the set of ground truth contacts assuming that a true contact was detected if there is a positive prediction within τ_{gt}. See Figure 12 for precision and recall for values of $\tau_{\text{val}} \in [0, 1]$. The curves for predictions of palm and toe placements are rather low, further confirming that local geometry alone is not sufficient to robustly predict the best human pose.

Energy Terms. Next, we investigate the importance of other energy terms by excluding one term at a time. Figure 14 shows some typical failure examples and Figure 13 quantitatively evaluates overall quality of resulting poses using our leave-one-out experimental setup with per-joint error metric. We found that our energy terms play different roles depending on the dataset. For example, performance on cockpits significantly degrades if pose prior is omitted, since these models typically have more candidate contacts. The symmetry penalty plays an important role for chairs since it provides additional cues when larger areas (e.g. seat and back) are detected as candidate contacts. Finally, we found that intersection plays a relatively minor role with respect to our evaluation metric. However, it eliminates some visual artifacts, such as a person's legs penetrating through the seat.

Computational Complexity and Timing. Finally, we discuss the computational complexity of our method. Note that our optimization procedure starts by trying different rotations around every candidate contact point, so the outer loop executes $N_{\text{cand}}N_{\text{rot}}$ times. For each iteration, one can find the minimal energy contact assignment by visiting every candidate contact and lookup its value in end effector probability distribution grid. Thus, the final complexity of our algorithm is $O(N_{\text{cand}}^2)$, if we assume that number of rotations is constant.

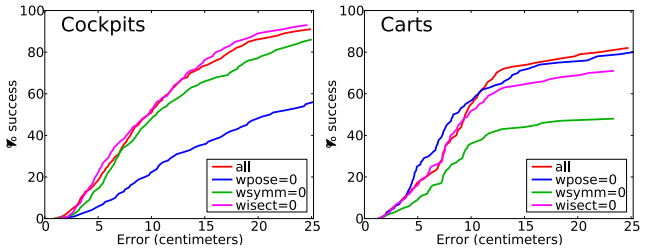


Figure 13: *We evaluate how much different energy terms affect the quality of the final results. These plots show the fraction of correctly mapped joints (y-axis) for different error thresholds (x-axis), and each curve corresponds to an energy weight assigned to zero. Results for other datasets are available in the supplemental material.*

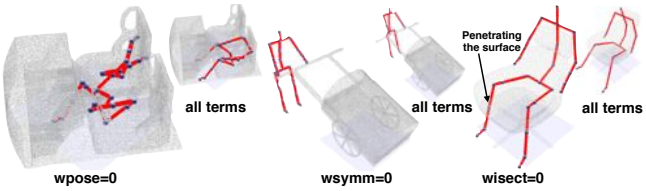


Figure 14: *Typical failures due to the absence of different energy terms.*

Data	N	Prep	Train	Opt
Bicycles	30	80s	115s	130s
Bipedals	30	225s	200s	590s
Cockpits	21	1150s	550s	970s
Carts	11	235s	25s	15s
Chairs	30	50s	60s	80s
Gym Equipment	25	345s	270s	500s

Table 1: *Some average (per model) execution times on different datasets. Preparation time is spent on point and feature computation. Training involves point feature classifiers and pose priors (where the former takes the majority of the time), note that once the model is trained local feature classification takes only a fraction of a second. The last column records the timing for our optimization procedure.*

We report the average running time our algorithm on 2.6 GHz Intel processor at different stages in Table 1. The optimization time ranges from 2s (for some carts) to as much as an hour for one of the car models with large surface area. Evidently from the complexity analysis the running time is mostly affected by the total number of surface points that were classified as potential contacts, and thus, suffers if there are many false positives during the classification stage. Note that poses are detected for a vast majority of models in the order of 10 minutes. While training a surface classifier can take up to a few minutes, the classification (i.e. evaluating the random regression forest) takes only a fraction of a second. On average, the most time-consuming step at the prediction stage is the data preparation (sampling surface points, computing geodesic distances, and computing local geometric features) which is highly non-optimized and can take up to 30 minutes per model.

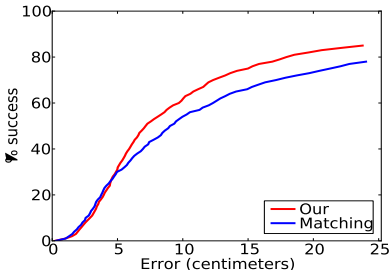


Figure 15: *We quantitatively evaluate the quality of sparse correspondences produced by our method vs shape matching. This plot shows the fraction of correctly predicted contact points (y-axis) that are within a distance threshold ground truth contact point (x-axis, in centimeters).*

6 Applications

In this section, we investigate applications of our affordance prediction algorithm for shape analysis in computer graphics. We conjecture that predicting how people use an object provides opportunities for novel "human-centric" algorithms to analyze and process shapes. The following paragraphs describe a few early investigations of this idea.

Sparse Correspondence. Establishing semantically similar points between related 3D models is an important problem in geometry analysis with applications in morphing, property transfer, and similarity measurement. A common scenario is that one has a pre-processed collection of 3D models and wants to find correspondences from key points on shapes in the collection to the surface of a new model.

Since human contact points provide a strong and persistent feature among a large variety of man-made objects, and since their global arrangements and local geometric features are constrained by human biomechanics, we conjecture that the contact points computed by our affordance algorithm provide good predictors for sparse correspondences between shapes (i.e., all points predicted in contact with the left-palm are marked in correspondence).

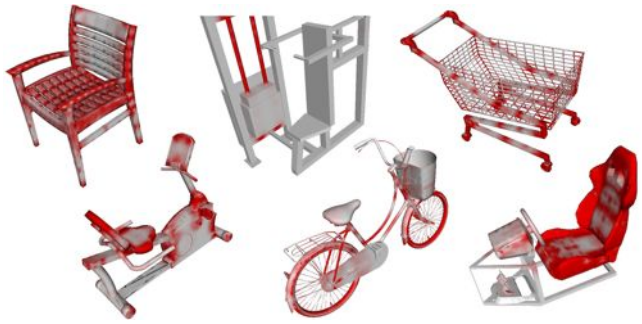
To test our methods in this setting, we train the affordance model on a dataset with annotated contacts, and then use it to predict contacts on a new input surface. Figure 15 shows a plot measuring the accuracy of the predicted surface contact points averaged over all datasets. Note that about 80% of contact points are predicted within 20cm distance of their true location by our method (the red curve).

To evaluate the quality of our results with respect to the state-of-the-art in surface correspondence, we compared to the Global Matching method described in the previous section [Huang et al. 2013] (the blue curve in Figure 15). Note that our method provides better contact point correspondences on average. More details can be found in the supplemental results.

Salience Estimation. Predicting the functional "importance" of a surface patch or object part is a critical problem in 3D shape analysis, with applications in feature detection, mesh processing, etc. Previous methods have considered salience measures based on analysis of curvatures [Gal and Cohen-Or 2006; Lee et al. 2005], shape similarities [Shilane and Funkhouser 2007], and other properties [Chen et al. 2012].

We propose a new human-centric measure of salience. Intuitively speaking, we try to predict the importance of a point on an object to a person using the object. To do so, we execute the proposed

(a) Human-centric saliency [this paper]



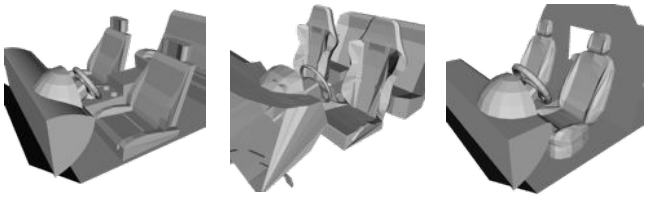
(b) Mesh saliency [Lee et al., 2005]

Figure 16: *Comparison of surface salience estimates using a) our predicted human poses versus b) traditional methods based on mesh curvatures. Surface regions shown with more red are stronger predictions.*

affordance analysis algorithm and then analyze the relationships between points on the object surface and predicted human poses, promoting positions within reach of a human, nearby human contact points, and along direct visibility sightlines. Specifically, for any point q and pose (T, θ, C) with rigid transform T, joint angles θ, and contact points C, our human-centric salience measure is the sum of three terms: $S(q, T, \theta, C) = S_P(q, C) + S_C(q, C) + S_V(q, T, \theta)$, where $S_P(q, C)$ measures the proximity of q to the centroid of C, $S_C(q, C)$ measures the proximity of q to the closest contact point in C, and $S_V(q, T, \theta)$ measures the visibility of q to a person in pose (T, θ). Details on the computation of this function can be found in the appendix.

We have computed this new salience measure using the predicted poses for all 147 meshes in the test data sets. Figure 16 shows visualizations of the results (top row) in direct comparison with a more traditional mesh salience method based on surface curvatures (bottom row) [Lee et al. 2005]. These comparisons suggest that our salience estimator captures a different notion of surface importance than previous work – it models relevance to a person using the object. See the supplemental materials for more examples and comparisons.

View Selection. Automatic placement of a virtual camera for display of a 3D mesh is a classical problem in computer graphics. Traditionally, research on this problem has focused on selection of "zoomed-out" views, which show the entire mesh in frame, usually with the centroid of the object at center of the frame and the orientation chosen to maximize visible surface area, mesh salience, or some other property of the projected view [Secord et al. 2011]. A much less-studied problem is how to automatically select views for "zoomed-in" images, where details of an object are shown via close-ups, for example.



(a) Human-centric viewpoint selection [this paper]

(b) Mesh-centric viewpoint selection

Figure 17: *Comparison of zoomed viewpoints selected using a) estimated human poses versus b) a typical method based on mesh centeredness and salience.*

We propose that affordance analysis can be useful to guide automatic selection of views for zoomed-in images. Intuitively, if we are going to show only a fraction of an object, it makes sense to focus on the part of relevance to a person. To investigate this idea, we implemented an automatic view selection algorithm that places the centroid of a predicted human pose at the center of the frame and then selects the view direction that maximizes the sum of human-centric salience over the visible surface area.

Figure 17 shows results of our algorithm (top row) on several cockpit examples, where the zoom factor is approximately 2X and the view direction is constrained to be looking down at a 30 degree angle. Note that our method automatically selects views that focus on the driver's seat and dashboard in a way that matches ones commonly found in car catalogs. The lower images show a comparison to results that would be achieved by centering the image on the centroid of the object and then rotating to maximize the sum of mesh salience over the visible surface area, a logical alternative motivated by [Lee et al. 2005]. More comparisons can be found in the supplemental material.

Shape Retrieval. Similarity-based retrieval of shapes is common problem in graphics, vision, and robotics. Given a query object, the goal is to find similar objects in a database, where similarity is usually defined by matching global shapes, structural properties, and/or local shape features. In this section, we consider a new way to measure the similarity of two shapes – based on the similarity of the poses people are in when they use them. Intuitively, human poses are similar for related actions (sitting, riding, pushing, etc.) and thus similarities between poses can reveal functional similarities between objects.

To investigate this idea, we implemented a "pose-based" retrieval system that uses the average distance between predicted joint locations (after optimal rigid pose alignment) as a similarity metric for shape retrieval. Figure 18 shows representative results for this system, where the left-most image is a query followed by sorted list of the most similar models. Note how the top ranked matches with pose-based retrieval reveal similarities in human use rather than overall shape. For example, a person might be looking for a bicycle that is ridden with an up-right posture (Figure 18a), a front-leaning speedy bicycle (Figure 18b). We find that affordance is one of the most persistent cues for some classes of objects, such as benchpress gym equipment (Figure 18c), and thus could be used to enhance retrieval results for certain types of retrieval tasks.

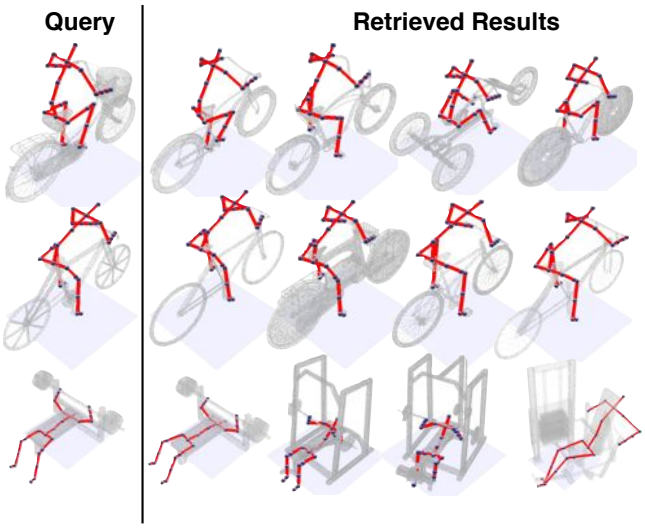


Figure 18: *Some pose-based shape retrieval results, where in each row the query shape is on the left. Our automatically predicted poses are rendered for reference.*

7 Conclusion and Future Work

In this paper, we propose that predicting shape affordances is useful for computer analysis of 3D models. We investigate this idea by implementing a novel algorithm for generating the static pose that a person would most likely adopt when interacting with an object. Our algorithm uses a combination of local anthropometric classifiers as well as global biomechanics constraints to search for a plausible pose. To test this algorithm, we created human pose annotations for a dataset of diverse shapes from various object classes and developed experiments to measure the accuracy of algorithmic predictions. We find that our algorithm produces results within a 20cm tolerance in the vast majority of cases and makes predictions better than any tested extension of existing techniques. Finally, we investigate how the algorithm can be used in novel way for human-centric shape analysis by describing novel methods for coarse surface correspondence, salience estimation, viewpoint selection, and shape retrieval.

This work has several limitations that suggest topics for future work. First, our algorithm performs only static analysis of an object's shape, which is not sufficient to understand all functional interactions with a human. A higher level reasoning is necessary for some scenarios, e.g. in Figure 19 one needs to understand the purpose of weightlifting (semantics), how parts of an elliptical device and a treadmill move over time (dynamics), and that pushing wheelbarrow might be easier using the handles (biomechanics). Future work could incorporate biomechanical models, contact dynamics, and other physical simulations to produce more accurate and general affordance models at higher computational cost. It also would be interesting to consider objects suitable only for some body shapes (e.g. children vs adult bicycles), objects that are used without body contact (e.g. a TV set), and objects that can be used in multiple poses (e.g. one can sit or slouch on a chair).

Second, we investigate only a few low-level shape analysis applications in this paper. Those applications were chosen because they provide fundamental building blocks for other applications, but further work is required to investigate the applicability of our method in a broader sense. For example, we conjecture that affordance analysis might: 1) help autonomous agents to interact appropriately with unlabeled objects in virtual worlds (e.g., sit here, grab it like this, etc.); 2) help automatic segmentation and recognition algorithms

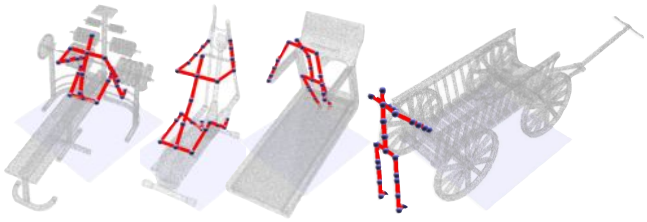


Figure 19: *A few failure examples for our method. Note that in some cases understanding the functional use of the shape is essential for estimating the pose (e.g. lifting weights is not about sitting comfortably). Similarly one would need to model the dynamics of an object as it is being used to understand that it will be difficult to use the elliptical machine while sitting, or walk on a treadmill in the opposite direction. Finally, there could be multiple ways to plausibly interact with a shape, such as pushing the wagon from the front. For these cases a more accurate biomechanical model might help suggest the most efficient interaction given the dynamics and semantics.*

to assign functional labels to parts of 3D models (e.g., this looks like a handle, but it's probably not one because a person can't grab it easily); 3) provide a way to align collections of shapes consistently for database exploration (e.g., rotate them all so that a person using the object is looking left); 4) guide optimization of shapes to better fit people with specific body shapes (e.g., adjust this to better fit a child); or 5) help makers of instructional tools to provide visualizations of how people typically use an object (e.g., here's how you should sit on this gym equipment). Exploring these and other applications is beyond the scope of this paper, but provides interesting topics for future work.

In the end, we believe that this work is just a first step towards the challenging goal of understanding the semantics of human-object interaction. Capturing such interactions can provide us with insights on what are the important structural properties of shapes, how objects relate to one another, and how objects can be optimized to facilitate certain interactions. Leveraging pose prediction for these kinds of applications outlines novel and interesting research directions in geometry analysis.

Acknowledgements. We acknowledge Qi-Xing Huang for helping with the comparisons. We also thank Ashutosh Saxena, Peter Minary, and anonymous reviewers for their comments and suggestions. This project was supported by NSF grants DMS 1228304, CCF 1161480, IIS-1251217, CNS-0831374, AFOSR grant FA9550-12-1-0372, ONR MURI N00014-13-1-0341, Intel, Adobe, and a Google research award.

References

BARD, C., AND TROCCAZ, J. 1990. Automatic preshaping for a dextrous hand from a simple description of objects. *Intelligent Robots and Systems*, 865–872.

BOHG, J., MORALES, A., ASFOUR, T., AND KRAGIC, D. 2013. Data driven grasp synthesis - a survey. *IEEE Transactions on Robotics*.

BREIMAN, L. 2001. Random forests. *Mach. Learning 45*, 1, 5–32.

BUSS, S. R. 2005. Introduction to inverse kinematics with jacobian transpose, pseudoinverse and damped least squares methods. *Unpublished survey*.

CHAUDHURI, S., KALOGERAKIS, E., GUIBAS, L., AND KOLTUN, V. 2011. Probabilistic reasoning for assembly-based 3D modeling. *SIGGRAPH*, 35:1–35:10.

CHEN, X., SAPAROV, A., PANG, B., AND FUNKHOUSER, T. 2012. Schelling points on 3D surface meshes. *SIGGRAPH 31*, 4.

DELAITRE, V., FOUHEY, D., LAPTEV, I., SIVIC, J., GUPTA, A., AND EFROS, A. 2012. Scene semantics from long-term observation of people. In *ECCV*.

FEIX, T., ROMERO, J., EK, C., SCHMIEDMAYER, H., AND KRAGIC, D. 2013. A metric for comparing the anthropomorphic motion capability of artificial hands. *IEEE Transactions on Robotics 29*, 1, 82–93.

FOUHEY, D. F., DELAITRE, V., GUPTA, A., EFROS, A. A., LAPTEV, I., AND SIVIC, J. 2012. People watching: Human actions as a cue for single-view geometry. In *ECCV*.

FRITZ, G., PALETTA, L., BREITHAUPT, R., AND ROME, E. 2006. Learning predictive features in affordance based robotic perception systems. In *Intelligent Robots and Systems*, 3642–3647.

FU, H., COHEN-OR, D., DROR, G., AND SHEFFER, A. 2008. Upright orientation of man-made objects. *SIGGRAPH*.

GAL, R., AND COHEN-OR, D. 2006. Salient geometric features for partial shape matching and similarity. *ACM Trans. Graph.*.

GAL, R., SORKINE, O., MITRA, N. J., AND COHEN-OR, D. 2009. iWIRES: An analyze-and-edit approach to shape manipulation. *SIGGRAPH 28*, 3, #33, 1–10.

GIBSON, J. J. 1977. The theory of affordances. *Lawrence Erlbaum*.

GOLDFEDER, C., AND ALLEN, P. K. 2011. Data-driven grasping. *Auton. Robots 31*, 1, 1–20.

GOLOVINSKIY, A., AND FUNKHOUSER, T. 2009. Consistent segmentation of 3D models. *Proc. SMI 33*, 3, 262–269.

GRABNER, H., GALL, J., , AND VAN GOOL, L. 2011. What makes a chair a chair? *CVPR*.

GUPTA, A., SATKIN, S., EFROS, A. A., AND HEBERT, M. 2011. From 3D scene geometry to human workspace. In *IEEE CVPR*.

HERMANS, T., REHG, J. M., AND BOBICK, A. 2011. Affordance prediction via learned object attributes. *ICRA*.

HUANG, Q., KOLTUN, V., AND GUIBAS, L. 2011. Joint shape segmentation with linear programming. In *SIGGRAPH Asia*.

HUANG, Q.-X., ZHANG, G.-X., GAO, L., HU, S.-M., BUTSCHER, A., AND GUIBAS, L. 2012. An optimization approach for extracting and encoding consistent maps. *SIGGRAPH Asia*.

HUANG, Q., SU, H., AND GUIBAS, L. 2013. Fine-grained semi-supervised labeling of large shape collections. *SIGGRAPH Asia*.

JIANG, Y., AND SAXENA, A. 2012. Hallucinating humans for learning robotic placement of objects. *ISER*.

JIANG, Y., AND SAXENA, A. 2013. Infinite latent conditional random fields for modeling environments through humans. *RSS*.

JIANG, Y., LIM, M., AND SAXENA, A. 2012. Learning object arrangements in 3D scenes using human context. *ICML*.

JIANG, Y., KOPPULA, H. S., AND SAXENA, A. 2013. Hallucinated humans as the hidden context for labeling 3D scenes. *CVPR*.

KALOGERAKIS, E., HERTZMANN, A., AND SINGH, K. 2010. Learning 3D mesh segmentation and labeling. In *SIGGRAPH*.

KALOGERAKIS, E., CHAUDHURI, S., KOLLER, D., AND KOLTUN, V. 2012. A probabilistic model for component-based shape synthesis. *SIGGRAPH*.

KIM, V. G., LI, W., MITRA, N., DIVERDI, S., AND FUNKHOUSER, T. 2012. Exploring collections of 3D models using fuzzy correspondences. *SIGGRAPH*.

KIM, V. G., LI, W., MITRA, N. J., CHAUDHURI, S., DIVERDI, S., AND FUNKHOUSER, T. 2013. Learning part-based templates from large collections of 3D shapes. *SIGGRAPH*.

LEE, C., VARSHNEY, A., AND JACOBS, D. 2005. Mesh saliency. *SIGGRAPH*.

LENZ, I., LEE, H., AND SAXENA, A. 2013. Deep learning for detecting robotic grasps. In *RSS*.

MITRA, N. J., PAULY, M., WAND, M., AND CEYLAN, D. 2012. Symmetry in 3D geometry: Extraction and applications. In *EUROGRAPHICS State-of-the-art Report*.

MITRA, N. J., WAND, M., ZHANG, H., COHEN-OR, D., KIM, V., AND HUANG, Q.-X. 2013. Structure-aware shape processing. In *Courses Siggraph Asia*.

NORMAN, D. 1988. *The Psychology of Everyday Things*. Basic Books.

OVSJANIKOV, M., LI, W., GUIBAS, L., AND MITRA, N. J. 2011. Exploration of continuous variability in collections of 3D shapes. *SIGGRAPH 30*, 4, 33:1–33:10.

PODOLAK, J., SHILANE, P., GOLOVINSKIY, A., RUSINKIEWICZ, S., AND FUNKHOUSER, T. 2006. A planar-reflective symmetry transform for 3D shapes. *ACM Trans. Graph. 25*, 3.

POLLARD, N. S., AND ZORDAN, V. B. 2005. Physically based grasping control from example. *SCA*.

PRZYBYLSKI, M., WACHTER, M., ASFOUR, T., AND DILLMANN, R. 2012. A skeleton-based approach to grasp known objects with a humanoid robot. *Humanoid Robots*.

ROSALES, C., PORTA, J., AND ROS, L. 2011. Global optimization of robotic grasps. *RSS*.

SAXENA, A., DRIEMEYER, J., KEARNS, J., AND NG, A. 2006. Robotic grasping of novel objects. In *NIPS*.

SAXENA, A. 2009. *Monocular depth perception and robotic grasping of novel objects*. PhD thesis, Stanford University.

SECORD, A., LU, C., FINKELSTEIN, A., SINGH, M., AND NEALEN, A. 2011. Perceptual models of viewpoint preference. *ACM Trans. Graph. 50*, 5.

SHAPIRA, L., SHAMIR, A., AND COHEN-OR, D. 2008. Consistent mesh partitioning and skeletonisation using the shape diameter function. *Vis. Comput. 24*, 4, 249–259.

SHILANE, P., AND FUNKHOUSER, T. 2007. Distinctive regions of 3d surfaces. *ACM Trans. Graph. 26*, 2 (June).

SIDI, O., VAN KAICK, O., KLEIMAN, Y., ZHANG, H., AND COHEN-OR, D. 2011. Unsupervised co-segmentation of a set of shapes via descriptor-space spectral clustering. *SIGGRAPH Asia 30*, 6, 126:1–126:9.

STARK, M., LIES, P., ZILLICH, M., WYATT, J., AND SCHIELE, B. 2008. Functional object class detection based on learned affordance cues. *Computer Vision Systems*.

SUN, J., MOORE, J. L., BOBICK, A., AND REHG, J. M. 2009. Learning visual object categories for robot affordance prediction. *The International Journal of Robotics Research*.

TRIMBLE, 2013. Trimble 3D warehouse, http://sketchup.google.com/3dwarehouse/.

VAN KAICK, O., XU, K., ZHANG, H., WANG, Y., SUN, S., SHAMIR, A., AND COHEN-OR, D. 2013. Co-hierarchical analysis of shape structures. *SIGGRAPH 32*, 4, 69:1–69:10.

WEI, P., ZHAO, Y. B., ZHENG, N., AND ZHU, S. 2013. Modeling 4D human-object interactions for event and object recognition. *ICCV*.

YING, L., FU, J. L., AND POLLARD, N. S. 2007. Data-driven grasp synthesis using shape matching and task-based pruning. *Transactions on Visualization and Computer Graphics 13*, 4, 732–747.

ZHAO, W., ZHANG, J., MIN, J., AND CHAI, J. 2013. Robust realtime physically based motion control for human grasping. In *SIGGRAPH Asia*.

Appendix

This section provides implementation details omitted from previous sections to improve clarity of exposition.

Point Features. We estimate geometric features at a sparse set of candidate contact points on an input shape S. First, we densely sample $100000 \cdot A$ points P_{dense} from the surface, where A is the surface area in square meters. We next compute approximate geodesic distances between the points by connecting each point to its 10 nearest neighbors and running the Dijkstra algorithm on the resulting graph. Our features are as follows: for a point $c \in S$ we take its geodesic neighborhood in P_{dense} and compute eigenvalues λ and eigenvectors v of the covariance matrix. We next define the following features: λ_1/λ_0, λ_2/λ_0, $v_0 \cdot \text{up}$, $v_2 \cdot \text{up}$, and the variance in height of the neighborhood. These features are estimated for geodesic neighborhoods of radii 6, 12, 18, 24, and 30 cm. Additionally, we compute (absolute) curvature, SDF (shape diameter function) [Shapira et al. 2008], average curvature and SDF over a geodesic neighborhood of radius 18cm, the distance to the best local reflection plane [Podolak et al. 2006], the point height, and an 8-bin histogram of distances to other points on the shape up to a representative human armspan of 1.8m.

Salience Estimation. The human-centric salience estimator $S(q,T,\theta,C)$ for any point q, and pose (T,θ,C) with rigid transform T, joint angles θ and contact points C is computed as follows:

$$S(q,T,\theta,C) = S_P(q,C) + S_C(q,C) + S_V(q,T,\theta)$$

where $S_P(q,C)$ is a Gaussian function of the distance from q to the centroid $\hat{C}$ of C ($S_P(q,C) = \lambda_P \exp\left(-\|q-\hat{C}\|^2/2\sigma_P^2\right), \lambda_P = 4, \sigma_P = 1\text{m}$). $S_C(q,C)$ is a Gaussian function of the distance from q to the closest point c^* in C ($S_P(q,C) = \lambda_C \exp\left(-\|q-c^*\|^2/2\sigma_C^2\right), \lambda_C = 1, \sigma_C = 10\text{cm}$). $S_V(q,T,\theta)$ is a function that estimates the visibility of q to a person in pose (T,θ) with four factors accounting for occlusion, depth, foveation, and surface normal orientation: $S_V(q,T,\theta) = S_{VO}(q,e) \cdot S_{VD}(q,e,d) \cdot S_{VF}(q,e,d) \cdot S_{VN}(q,e)$, where e and v are the estimated eye position and view direction of (T,θ), respectively, n is the normal to the surface at q, and the four factors are computed as follows. $S_{VO}(q,e)$ is 1 if q is visible to e, and 0 otherwise; $S_{VD}(q,e,d) = \exp\left(-((q-e)\cdot d)^2/2\sigma_{VD}^2\right)$ ($\sigma_{VD} = 1\text{m}$); $S_{VF}(q,e,d) = (((q-e)\cdot d)/\|q-e\|)^{\alpha}$ ($\alpha = 8$); and, $S_{VN}(q,e) = (((q-e)\cdot n)/\|q-e\|)^{\beta}$ ($\beta = 8$).