# Category-Specific Object Reconstruction from a Single Image

Abhishek Kar*, Shubham Tulsiani*, João Carreira, Jitendra Malik
University of California, Berkeley

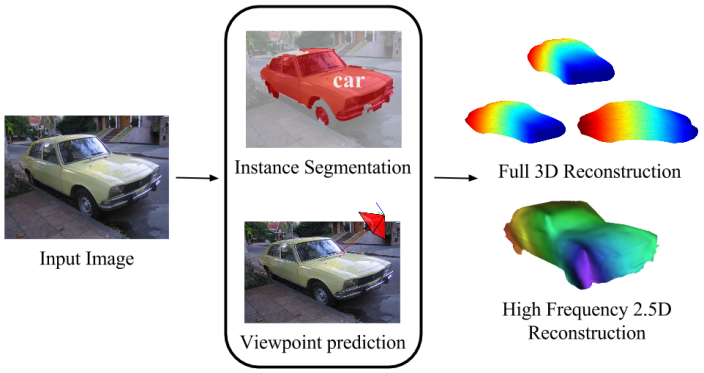


Figure 1: Automatic object reconstruction from a single image. Our method leverages estimated instance segmentations and predicted viewpoints to generate a full 3D mesh and high frequency 2.5D depth maps.

Object detection and segmentation have witnessed rapid progress owing mostly to convolutional neural networks trained on large datasets. We show that building upon these detectors, 3D object reconstruction from a single image can be readily addressed and give an existence proof in Fig. 1.

At the core of our approach are novel deformable 3D models that can be learned from 2D annotations available in existing object detection datasets and can be driven by noisy automatic object segmentations. These allow us to overcome the two main challenges to object reconstruction in the wild: 1) detection datasets typically have many classes and each may encompass wildly different shapes, making 3D model acquisition expensive. 2) 3D shape inference should be robust to any small imperfections that detectors produce, yet be expressive enough to represent rich shapes.We use the learned deformable 3D models to infer the shape for each detection in a novel input image and further complement it with a bottom-up module for recovering high-frequency shape details. The learning and inference steps are described below.

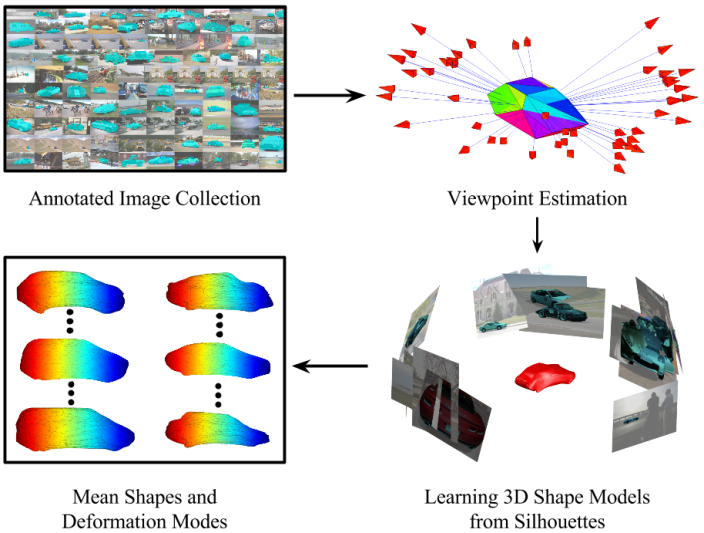


Figure 2: Overview of our training pipeline.

**Deformable Model Learning**: Figure 2 illustrates our pipeline for learning shape models from just 2D training images, aided by ground truth segmentations and a few keypoints. We first use the NRSfM [4] framework, extended to incorporate silhouette information, to jointly estimate the camera viewpoints (rotation, translation and scale) for all training instances in each class.

We then learn the category level shape models by optimizing over a deformation basis of representative 3D shapes that best explain all silhouettes, conditioned on the camera viewpoints. We model our category shapes as deformable point clouds - one for each subcategory of the class. Our shape model $M = (\bar{S}, V)$ comprises of a mean shape $\bar{S}$ (learnt mean shapes for several classes are shown in Figure 3) and linear deformation bases $V = \{V_1, .., V_K\}$. We formulate our energy primarily based on image silhouettes and priors on natural shapes. These energies enforce that the shape for an instance is consistent with its silhouette $(E_s, E_c)$, shapes are locally consistent $(E_l)$, normals vary smoothly $(E_n)$ and the deformation parameters are small $(\|\alpha_{ik} V_k\|_F^2)$. Finally, we solve the constrained optimization in equation 1 using block-coordinate descent to learn the category level shape model. We refer the reader to the main text for details regarding our optimization and formulations of our shape energies.

$$\begin{aligned} \min_{\bar{S},V,\alpha} \quad & E_l(\bar{S},V) + \sum_i (E_s^i + E_{kp}^i + E_c^i + E_n^i + \sum_k (\|\alpha_{ik} V_k\|_F^2)) \\ \text{subject to:} \quad & S^i = \bar{S} + \sum_k \alpha_{ik} V_k \end{aligned} \tag{1}$$

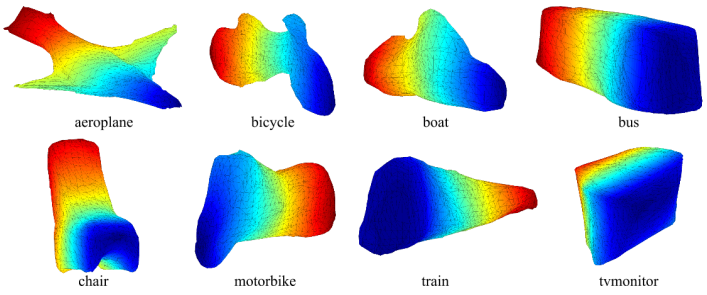


Figure 3: Mean shapes learnt for rigid classes in PASCAL VOC using our formulation. Color encodes depth when viewed frontally.

**Reconstruction in the Wild**: We approach object reconstruction from the big picture downward - like a sculptor first hammering out the big chunks and then chiseling out the details. After objects are detected and approximately segmented [3] and coarse poses predicted [5], we use these in fitting our top down deformable shape models to the noisy silhouettes. Finally, we recover high frequency shape details from shading cues by conditioning an intrinsic image algorithm [1] with our inferred coarse 3D shape. We present the first fully automatic reconstructions on PASCAL VOC[2] and state of the art object reconstruction as benchmarked on PASCAL3D+[6].

[1] Jonathan T Barron and Jitendra Malik. Shape, illumination, and reflectance from shading. Technical Report UCB/EECS-2013-117, EECS, UC Berkeley, May 2013.

[2] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html.

[3] Bharath Hariharan, Pablo Arbeláez, Ross Girshick, and Jitendra Malik. Simultaneous detection and segmentation. In *European Conference on Computer Vision (ECCV)*, 2014.

[4] L. Torresani, A. Hertzmann, and C. Bregler. Non-rigid structure-from-motion: Estimating shape and motion with hierarchical priors. *TPAMI*, 2008.

[5] Shubham Tulsiani and Jitendra Malik. Viewpoints and keypoints. In *CVPR*. 2015.

[6] Yu Xiang, Roozbeh Mottaghi, and Silvio Savarese. Beyond pascal: A benchmark for 3d object detection in the wild. In *WACV*, 2014.

* Authors contributed equally

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.