Category-Specific Object Reconstruction from a Single Image

Abhishek Kar*, Shubham Tulsiani*, João Carreira and Jitendra Malik
University of California, Berkeley - Berkeley, CA 94720
{akar,shubhtuls,carreira,malik}@eecs.berkeley.edu

Abstract

Object reconstruction from a single image – in the wild – is a problem where we can make progress and get meaningful results today. This is the main message of this paper, which introduces the first fully automatic pipeline having pixels as inputs and dense 3D surfaces of various rigid categories as outputs in images of realistic scenes. At the core of our approach are novel deformable 3D models that can be learned from 2D annotations available in existing object detection datasets, that can be driven by noisy automatic object segmentations and which we complement with a bottom-up module for recovering high-frequency shape details. We perform a comprehensive quantitative analysis and ablation study of our approach using the recently introduced PASCAL 3D+ dataset and show very encouraging automatic reconstructions on PASCAL VOC.

1. Introduction

We consider the problem of fully automatic object reconstruction from a single image, which has received little attention in recent decades largely because objects could not be localized, much less reconstructed. Now times have changed. The realization that convolutional networks [31] trained on huge datasets worked [29], together with the development of methods for generating object segmentation proposals [12, 1, 15, 43, 3, 28] has resulted in much sudden progress in object detection and segmentation [20, 23, 11]. We will show that building upon these detectors, object reconstruction from a single image is ready to be attacked and give an existence proof in fig. 1.

Any attempt at object reconstruction in the wild faces two main challenges: 1) detection datasets typically have many classes and each may encompass wildly different shapes, making 3D model acquisition expensive; it would be ideal to learn 3D models directly from their training data. 2) No matter how much progress occurs in detection, it is unlikely that it will become completely flawless. This means that fitting the learned 3D models should be robust to any small imperfections that detectors produce, yet be expressive enough to represent rich shapes.

Here we introduce the first approach that meets both requirements. Our proposed method can learn dense deformable 3D basis shape models [8] directly from ground truth silhouettes and a small set of keypoints that are available in modern detection datasets, such as PASCAL VOC [18] or COCO [33], and we demonstrate that these models degrade gracefully when fit to images in the wild, leveraging current state-of-the-art detectors.

The recent method of Vicente et al [44] reconstructs 3D models from similar annotations as we do but it has a different focus: it aims to reconstruct a fully

* Authors contributed equally
Figure 2: Overview of our training pipeline. We use an annotated image collection to estimate camera viewpoints which we then use along with object silhouettes to learn 3D shape models. Our learnt shape models, as illustrated in the rightmost figure are capable of deforming to capture intra-class shape variation.

annotated image set while making strong assumptions about the quality of the segmentations it fits to and is hence inappropriate for reconstruction in an unconstrained setting. Our approach embodies the top-down tradition and is designed from scratch for operating in noisy conditions by resorting to explicit 3D shape models. Our work is also close to that of Kemelmacher-Shlizerman et al. [27, 40] which aims to learn morphable models for faces from 2D images, but we focus on richer shapes in unconstrained settings, at the expense of lower resolution reconstructions.

The evolution of ideas on object reconstruction from a single image over time has reflected different preferences along the shape representation bias-variance trade-off [25]. Expressive models such as generalized cylinders [34] could represent a very rich family of shapes but were overly sensitive to noise in lower-level inputs. Other representations such as solid models [21, 48], which trace back to the early work of Roberts [36], and CAD models [32, 38], can represent a limited set of shapes but are robust to noise in lower level inputs, which are used mostly for determining viewpoint. Here we pursue more expressive basis shape models [8, 2, 50] which establish a balance between the two extremes as they can deform but only along class-specific modes of variation. Differently from previous work (e.g. [50]), we fit them to automatic figure-ground object segmentations.

Our paper is organized as follows: in section 2 we describe our model learning pipeline where we estimate camera viewpoints for all training objects (section 2.1) followed by our shape model formulation (section 2.2) to learn 3D models. Section 3 describes our testing pipeline where we use our learnt models to reconstruct novel instances without assuming any annotations. We evaluate our reconstructions under various settings in section 4 and provide sample reconstructions in the wild.

2. Learning Deformable 3D Models

We are interested in 3D shape models that can be robustly aligned to noisy object segmentations by incorporating top-down class-specific knowledge of how shapes from the class typically project into the image. We want to learn such models from just 2D training images, aided by ground truth segmentations and a few keypoints, similar to [44]. We will do this by first estimating the viewpoints of all objects in a class using a structure-from-motion approach, followed by optimizing over a deformation basis of representative 3D shapes that best explain all silhouettes, conditioned on the viewpoints. We will describe these two stages of model learning in the following subsections. Figure 2 illustrates this training pipeline of ours.

2.1. Viewpoint Estimation

We use the framework of NRSfM [10] to jointly estimate the camera viewpoints (rotation, translation and scale) for all training instances in each class. NRSfM, which was originally proposed for recovering shape and deformations from video [10, 41, 7, 19], is a natural choice for viewpoint estimation from sparse class keypoints as intra-class variation may become a confounding factor if not modeled explicitly. However, the performance of such algorithms has only been explored on simple categories, such as SUV’s [49] or flower petal and clown fish [35]. Closer to our work, Hejrati and Ramanan [24] used NRSfM on a larger class (cars) but need a predictive detector to fill-in missing data (occluded keypoints) which we do not assume to have here.

We closely follow the EM-PPCA formulation of Torresani et al. [41], which uses a Gaussian factor analyzer as a shape prior and employs an expectation-maximization algorithm to minimize re-projection error while regularizing the shape deformation parameters. We propose a simple extension to the algorithm that incorporates silhouette information in addition to
keypoint correspondences to robustly recover cameras and shape bases. Energies similar to ours have been proposed in the shape-from-silhouette literature [45] and with rigid structure-from-motion [44] but, to the best of our knowledge, not in conjunction with NRSfM.

**NRSfM Model.** Given $K$ keypoint correspondences per image and instances $n = \{1, \cdots, N\}$, our adaptation of the NRSfM algorithm in [41] corresponds to maximizing the likelihood of the following factor analysis model:

$$
P_n = (I_K \otimes c_n R_n)S_n + T_n + N_n$$
$$S_n = S + VZ_n$$
$$z_n \sim \mathcal{N}(0, I), \quad N_n \sim \mathcal{N}(0, \sigma^2 I)$$

subject to:

$$R_nR_n^T = I_2$$
$$\sum_{k=1}^{K} C_n^{mask}(p_k, n) = 0, \quad \forall n \in \{1, \cdots, N\}$$

Here, $P_n$ is the 2D projection of the 3D shape $S_n$ with white noise and the rigid transformation given by the orthographic projection matrix $R_n$, scale $c_n$ and 2D translation $T_n$. The shape is parameterized as a factored Gaussian with a mean shape $\bar{S}$, $m$ basis vectors $[V_1, V_2, \cdots, V_m] = V$ and latent deformation parameters $z_n$. Our key modification is constraint (3) where $C_n^{mask}$ denotes the Chamfer distance field of the $n$th instance’s binary mask and says that all keypoints $p_k, n$ of instance $n$ should lie inside its binary mask. We observed that this results in more accurate viewpoints as well as more meaningful shape bases learnt from the data.

**Learning.** The likelihood of the above model is maximized using the EM algorithm with an axis-angle parameterization for rotation matrices. Missing data (occluded keypoints) is dealt with by “filling-in” the values by running the forward equations after the E-step. The algorithm computes shape parameters $\{S, V\}$, rigid body transformations $\{c_n, R_n, T_n\}$ as well as the deformation parameters $\{z_n\}$ for each training instance $n$. In practice, we augment the data using horizontally mirrored images to exploit bilateral symmetry in the object classes considered. We also precompute the Chamfer distance fields for the whole set to speed up computation. As shown in figure 3, NRSfM allows us to reliably predict viewpoint while being robust to intraclass variations.

![Figure 3: NRSfM viewpoint estimation: Estimated viewpoints visualized using a 3D car wireframe.](image)

**2.2. 3D Basis Shape Model Learning**

Equipped with camera projection parameters and keypoint correspondences (now lifted to 3D by NRSfM) on the whole training set, we proceed to build deformable 3D shape models from object silhouettes within a class. 3D shape reconstruction from multiple silhouettes projected from a single object in calibrated settings has been widely studied. Two prominent approaches are visual hulls [30] and variational methods derived from snakes e.g. [16, 37] which deform a surface mesh iteratively until convergence. Some interesting recent papers have extended variational approaches to handle categories [13, 14] but typically require some form of 3D annotations to bootstrap models. A recently proposed visual-hull based approach [44] requires only 2D annotations as we do for class-based reconstruction and it was successfully demonstrated on PASCAL VOC but does not serve our purposes as it makes strong assumptions about the accuracy of the segmentation and will in fact fill entirely any segmentation with a voxel layer.

**Shape Model Formulation.** We model our category shapes as deformable point clouds – one for each subcategory of the class. The underlying intuition is the following: some types of shape variation may be well explained by a parametric model e.g. a Toyota sedan and a Lexus sedan, but it is unreasonable to expect them to model the variations between sailboats and cruise liners. Such models typically require knowledge of object parts, their spatial arrangements etc. [26] and involve complicated formulations that are difficult to optimize. We instead train separate linear shape models for different subcategories of a class. As in the NRSfM model, we use a linear combination of bases to model these deformations. Note that we learn such models from silhouettes and this is what allows...
to do without a large number of point correspondences between scanned 3D training exemplars necessary for building morphable models [9].

Our shape model $M = (\bar{S}, V)$ comprises of a mean shape $\bar{S}$ and deformation bases $V = \{V_1, .., V_K\}$ learnt from a training set $T : \{(O^i, P^i)\}_{i=1}^N$, where $O^i$ is the instance silhouette and $P^i$ is the projection function from world to image coordinates. Note that the $P^i$ we obtain using NRSfM corresponds to orthographic projection but our algorithm could handle perspective projection as well.

**Energy Formulation.** We formulate our objective function primarily based on image silhouettes. For example, the shape for an instance should always project within its silhouette and should agree with the keypoints (lifted to 3D by NRSfM). We capture these by defining corresponding energy terms as follows: (here $P(S)$ corresponds to the 2D projection of shape $S$, $C_{mask}$ corresponds to the Chamfer distance field of the binary mask corresponding to silhouette $O$, $\Delta(Q, p, k)$ corresponds to the average distance of point $p$ to its $k$ nearest neighbors in set $Q$)

**Silhouette Consistency.** Silhouette consistency simply enforces the predicted shape for an instance to project inside its silhouette. This can be achieved by penalizing the points projected outside the instance mask by their distance from the silhouette. In our $\Delta$ notation it can be written as follows:

$$E_s(S, O, P) = \sum_{C_{mask} > 0} \Delta(O, p, 1)^2$$

**Silhouette Coverage.** Using silhouette consistency alone would just drive points in towards the silhouette. This wouldn’t ensure that the object silhouette is “filled” - i.e. there might be overcarving. We deal with it by having an energy term that encues points on the silhouette to pull nearby projected points towards them. Formally, this can be expressed as:

$$E_c(S, O, P) = \sum_{p \in O} \Delta(P(S), p, m)^2$$

**Keypoint Consistency.** Our NRSfM algorithm provides us with sparse 3D keypoints along with camera viewpoints. We use these sparse correspondences on the training set to deform the shape to explain these 3D points. The energy term corresponding to it penalizes deviation of the shape from the 3D keypoints for each instance. Specifically, this can be written as:

$$E_{kp}(S, O, P) = \sum_{k \in KP} \Delta(P(S), k, m)^2$$

**Local Consistency.** In addition to the above data terms, we use a simple shape regularizer to restrict arbitrary deformations by imposing a quadratic deformation penalty between every point and its neighbors. We also impose a similar penalty on deformations to ensure local smoothness. The $\delta$ parameter represents the mean squared displacement between neighboring points and it encourages all faces to have similar size.

$$E_l(S, V) = \sum_i \sum_{j \in N(i)} ((\bar{S}_i - \bar{S}_j)^2 - \delta)^2 + \sum_k \|V_{ki} - V_{kj}\|^2$$

**Normals Smoothness.** Shapes occurring in the natural world tend to be locally smooth. We capture this prior on shapes by placing a cost on the variation of normal directions in a local neighborhood in the shape. Our normal smoothness energy is formulated as

$$E_n(S) = \sum_i \sum_{j \in N(i)} (1 - \vec{N}_i \cdot \vec{N}_j)$$

Here, $\vec{N}_i$ represents the normal for the $i^{th}$ point in shape $S$ which is computed by fitting planes to local point neighborhoods. Our prior essentially states that local point neighborhoods should be flat. Not that this, in conjunction with our previous energies automatically enforces the commonly used prior that normals should be perpendicular to the viewing direction at the occluding contour [5].

Our total energy is given in equation 9. In addition to the above smoothness priors we also penalize the $L2$ norm of the deformation parameters $\alpha_i$ to prevent unnaturally large deformations.

$$E_{tot}(S, V, \alpha) = E_l(S, V) + \sum_i (E_s^i + E_c^i + E_n^i + \sum_k (\|\alpha_{ik}V_k\|_F^2))$$

**Learning.** We solve the optimization problem in equation 10 to obtain our shape model $M = (\bar{S}, V)$. The mean shape and deformation basis are inferred via block-coordinate descent on $(\bar{S}, V)$ and $\alpha$ using sub-gradient computations over the training set. We restrict $\|V_k\|_F$ to be a constant to address the scale ambiguity between $V$ and $\alpha$ in our formulation. In order to deal with imperfect segmentations and wrongly estimated keypoints, we use truncated versions of the above energies that avoid fitting to outliers. The mean shapes learnt using our algorithm for 9 rigid categories in PASCAL VOC are shown in figure 4. Note that in
addition to representing the coarse shape details of a category, the model also ends up learning finer structures like chair legs and bicycle handles, which become more prominent with deformations.

$$\min_{\overline{S}, V, \alpha} E_{\text{tot}}(\overline{S}, V, \alpha)$$

subject to: $$S^i = \overline{S} + \sum_k \alpha ik V_k$$

Our training objective is highly non-convex and non-smooth and is susceptible to initialization. We follow the suggestion of [16] and initialize our mean shape with a soft visual hull computed using all training instances. The deformation bases and deformation weights are initialized randomly.

Figure 4: Mean shapes learnt for rigid classes in PASCAL VOC obtained using our basis shape formulation. Color encodes depth when viewed frontally.

3. Reconstruction in the Wild

We approach object reconstruction from the big picture downward, rather like a sculptor first hammering out the big chunks and then chiseling out the details. After objects in a scene are detected and approximately segmented, coarse poses are predicted and we use these in fitting our top-down shape models to the noisy silhouettes. Finally, we recover high frequency shape details from shading cues. We will now explain these components one at a time.

Initialization. Unlike during 3D model learning when projection parameters (rotation, translation and scale) were provided by structure from motion on annotated keypoints, at test time all we have are detector outputs.

While rotation is predicted, scale and translation of a detected object are initialized using its bounding box dimensions. We also predict fine subcategories and use these to select the correct mean shape and associated deformation bases ($\overline{S}, V$). We initialize the shape of the instance to the subcategory mean shape and the deformation weights $\alpha$ to zero.

Shape Inference. After initialization, we solve for the deformation weights $\alpha$ by optimizing equation (9) for fixed $\overline{S}, V$. Note that we do not have access to annotated keypoint locations at test time, the ‘Keypoint Consistency’ energy $E_{kp}$ is ignored during the optimization.

Bottom-up Shape Refinement. The above optimization results in a top-down 3D reconstruction based on the category-level models, inferred object silhouette, viewpoint and our shape priors. We propose an additional processing step to recover high frequency shape information by adapting the intrinsic images algorithm of Barron and Malik [6, 5], SIRFS, which exploits statistical regularities between shapes, reflectance and illumination. Formally, SIRFS is formulated as the following optimization problem:

$$\min_{Z, L} g(I - S(Z, L)) + f(Z) + h(L)$$

where $R = I - S(Z, L)$ is a log-reflectance image, $Z$ is a depth map and $L$ is a spherical-harmonic model of illumination. $S(Z, L)$ is a rendering engine which produces a log shading image with the illumination $L$. $g$, $f$ and $h$ are the loss functions corresponding to reflectance, shape and illumination respectively.

We incorporate our current coarse estimate of shape into SIRFS through an additional loss term:

$$f_o(Z, Z') = \sum_i ((Z_i - Z'_i)^2 + \epsilon^2)^{\gamma_o}$$

where $Z'$ is the initial coarse shape and $\epsilon$ a parameter added to make the loss differentiable everywhere. We obtain $Z'$ for an object by rendering a depth map of our fitted 3D shape model which guides the optimization of this highly non-convex cost function. The outputs from this bottom-up refinement are reflectance, shape and illumination maps of which we retain the shape.

Implementation Details. The gradients involved in our optimization for shape and projection parameters are extremely efficient to compute. We use approximate nearest neighbors computed using k-d tree to implement the ‘Silhouette Coverage’ gradients and leverage Chamfer distance fields for obtaining ‘Silhouette Consistency’ gradients. Our overall computation takes
only about 2 sec to reconstruct a novel instance using a single CPU core. Our training pipeline is also equally efficient - taking only about 15 minutes to learn a shape model (for a category with 500 images) for a given object category. We plan to make our implementations and learned shape models publicly available.

4. Experiments

Experiments were performed to assess two things: 1) how expressive our learned 3D models are by evaluating how well they matched the underlying 3D shapes of the training data 2) study their sensitivity when fit to images using noisy automatic segmentations and pose predictions.

Datasets. For all our experiments, we considered those images having objects from 10 rigid categories of the challenging PASCAL VOC 2012 [17] dataset and used ground truth class-specific keypoints and object segmentations available online [22]. Since ground truth 3D shapes are unavailable for PASCAL VOC and the other high profile detections datasets, we evaluated the expressiveness of our learned 3D models on the next best thing we managed to obtain: the PASCAL3D+ dataset [47] which has up to 10 3D CAD models for the rigid categories in PASCAL VOC. PASCAL3D+ provides between 4 different models for “tvmonitor” and “train” and 10 for “car” and “chair”. The different meshes primarily distinguish between subcategories but may also be redundant (e.g. there are more than 3 meshes for sedans in “car”). We obtain our subcategory labels on the training data by merging some of these cases, which also helps us in tackling data sparsity for some subcategories. The subset of PASCAL we considered after filtering occluded instances, which we do not tackle in this paper, had between 70 images for “sofa” and 500 images for classes “aeroplanes” and “cars”. We will make all our image sets available along with our implementation.

Metrics. We quantify the quality of our 3D models by comparing against the PASCAL 3D+ models using two metrics - 1) the Hausdorff distance normalized by the 3D bounding box size of the ground truth model [4] and 2) a depth map error to evaluate the quality of the reconstructed visible object surface, hereby defined as the mean absolute distance between reconstructed and ground truth depth:

\[
Z\text{-MAE}(\hat{Z}, Z^*) = \frac{1}{n \cdot \gamma} \min_{\beta} \sum_{x,y} |\hat{Z}_{x,y} - Z^*_{x,y} - \beta| \quad (11)
\]

where \(\hat{Z}\) and \(Z^*\) represent predicted and ground truth depth maps respectively. Analytically, \(\beta\) can be computed as the median of \(\hat{Z} - Z^*\) and \(\gamma\) is a normalization factor to account for absolute object size for which we use the bounding box diagonal. Note that our depth map error is translation invariant but not scale invariant.

4.1. Expressiveness of Learned 3D Models

We learned and then fit our 3D models on the same whole dataset (no train/test split), following the setup of Vicente et al [44]. Table 1 compares our reconstructions on PASCAL VOC with those of this recently proposed method which is specialized for this task (e.g. it is not designed for fitting to noisy data), as well as to a state of the art class-agnostic shape inflation method that reconstructs also from a single silhouette. We demonstrate competitive performance on both benchmarks with our models showing greater robustness to perspective foreshortening effects on “trains” and “buses”. Category-agnostic methods – Puffball [42] and SIRFS [5] – consistently perform worse on the benchmark by themselves. Certain classes like “boat” and “tvmonitor” are especially hard because of large intraclass variance and data sparsity respectively.

4.2. Sensitivity Analysis

In order to analyze sensitivity of our models to noisy inputs we reconstructed held-out test instances using our models given just ground truth bounding boxes. We compare various versions of our method using ground truth (Mask)/imperfect segmentations (SDS) and keypoints (KP)/our pose predictor (PP) for viewpoint estimation respectively. For pose prediction, we use the CNN-based system of [39] and augment it to predict subtypes at test time. This is achieved by training the system as described in [39] after adding an extra label - the subtypes obtained from PASCAL 3D+ as described above. To obtain an approximate segmentation from the bounding box, we use the refinement stage of the state-of-the-art joint detection and segmentation system proposed in [23].

Here, we use a train/test setting where our models are trained on only a subset of the data and used to reconstruct the held out data from bounding boxes. Table 2 shows that our results degrade gracefully from the fully annotated to the fully automatic setting as expected. Our method is robust to some mis-segmentation owing to our shape model that prevents shapes from bending unnaturally to explain noisy silhouettes. Our reconstructions degrade slightly with imperfect pose initializations even though our projection parameter optimization deals with it to some extent. With predicted poses, we observe that sometimes even when our reconstructions look plausible, the er-
### Classes

|         | aero | bike | boat | bus | car | chair | mbike | sofa | train | tv | mean |
|---------|------|------|------|-----|-----|-------|-------|------|-------|----|------|
| **Mesh** |      |      |      |     |     |       |       |      |       |    |      |
| KP+Mask | 5.00 | 6.27 | 9.94 | 6.22 | 5.18 | 5.20  | 4.98  | 6.58 | 12.60 | 9.64 | 7.16 |
| Carvi[44] | 5.07 | 6.03 | 8.80 | 8.76 | 4.38 | 5.74  | 4.86  | 6.49 | 17.52 | 8.37 | 7.60 |
| Puffball[42] | 9.73 | 10.39 | 11.68 | 15.40 | 11.77 | 8.58  | 8.99  | 8.62 | 23.68 | 9.45 | 11.83 |
| **Depth** |      |      |      |     |     |       |       |      |       |    |      |
| KP+Mask | 9.25 | 7.87 | 12.36 | 11.77 | 7.22 | 7.51  | 8.97  | 9.70 | 30.91 | 6.84 | 11.24 |
| Carvi[44] | 9.39 | 7.24 | 11.43 | 18.42 | 7.96 | 7.47  | 7.57  | 15.21 | 15.63 | 12.44 | 11.63 |
| SIRFS[5] | 12.98 | 12.31 | 16.03 | 29.31 | 21.58 | 15.53 | 16.30 | 18.08 | 38.54 | 21.36 | 20.19 |

Table 1: Studying the expressiveness of our learnt 3D models: comparison between our method and [44, 42] using ground truth keypoints and masks on PASCAL VOC. Note that [44] operates with ground truth annotations and reconstructs an image corpus and our method is used here on the same task for a fair comparison. Please see text for more details.

### Classes

|         | aero | bike | boat | bus | car | chair | mbike | sofa | train | tv | mean |
|---------|------|------|------|-----|-----|-------|-------|------|-------|----|------|
| **Mesh** |      |      |      |     |     |       |       |      |       |    |      |
| KP+Mask | 5.13 | 6.46 | 10.46 | 5.89 | 5.07 | 5.15  | 15.07 | 12.16 | 11.69 | 8.24 |
| KP+SDS  | 4.96 | 6.58 | 10.58 | 4.67 | 4.97 | 5.40  | 5.21  | 15.08 | 12.78 | 12.18 |
| PP+SDS  | 6.58 | 14.02 | 14.43 | 6.65 | 7.96 | 7.47  | 7.57  | 15.21 | 15.63 | 10.84 |
| Puffball[42] (SDS) | 9.68 | 10.23 | 11.80 | 15.95 | 12.42 | 8.28  | 9.45  | 9.60  | 23.38 | 9.26  |
| **Depth** |      |      |      |     |     |       |       |      |       |    |      |
| KP+Mask | 9.02 | 7.26 | 13.51 | 12.10 | 8.04 | 8.02  | 10.00 | 23.65 | 25.37 | 7.48  |
| KP+SDS  | 9.07 | 7.98 | 13.57 | 9.90 | 7.98 | 7.56  | 9.99  | 22.57 | 23.59 | 7.64  |
| PP+SDS  | 10.94 | 11.64 | 12.26 | 15.95 | 13.17 | 10.06 | 12.55 | 21.19 | 36.37 | 8.98  |
| SIRFS[5] | 11.80 | 11.83 | 15.98 | 29.15 | 21.64 | 15.58 | 16.91 | 19.64 | 37.58 | 23.01 | 20.31 |

Table 2: Ablation study for our method assuming/relaxing various annotations at test time on objects in PASCAL VOC. As can be seen, our method degrades gracefully with relaxed annotations. Note that these experiments are in a train/test setting and numbers will differ from table 1. Please see text for more details.

4.3. Fully Automatic Reconstruction

We qualitatively demonstrate reconstructions on automatically detected and segmented instances with 0.5 IoU overlap with the ground truth in whole images in PASCAL VOC using [23] in figure 5. It can be seen that our method is able to deal with some degree of mis-segmentation. Some of our major failure modes include not being able to capture the correct scale and pose of the object and thus badly fitting to the silhouette in some cases. Our subtype prediction also fails on some instances (e.g. CRT vs flat screen “tvmonitors”) leading to incorrect reconstructions. We include more such images in the supplementary material for the reader to peruse. More results can be found at [http://goo.gl/lmALxQ](http://goo.gl/lmALxQ).

### 5. Conclusion

We have proposed what may be the first approach to perform fully automatic object reconstruction from a single image on large, realistic datasets. Critically, our deformable 3D shape model can be bootstrapped from easily acquired ground-truth 2D annotations thereby bypassing the need for a-priori manual mesh design or 3D scanning and making it possible for convenient use of these types of models on large real-world datasets (e.g. PASCAL VOC). We report an extensive evaluation of the quality of the learned 3D models on a recent 3D benchmarking dataset for PASCAL VOC [47] showing competitive results with models that specialize in shape reconstruction from ground truth segmentations inputs while demonstrating that our method is equally capable in the wild, on top of automatic object detectors.

Much research lies ahead, both in terms of improving the quality and the robustness of reconstruction at test time (both bottom-up and top-down components), developing benchmarks for joint recognition and reconstruction and relaxing the need for annotations during training: all of these constitute interesting and important directions for future work. More expressive non-
Figure 5: Fully automatic reconstructions on detected instances (0.5 IoU with ground truth) using our models on rigid categories in PASCAL VOC. We show our instance segmentation input, the inferred shape overlaid on the image, a 2.5D depth map (after the bottom-up refinement stage), the mesh in the image viewpoint and two other views. It can be seen that our method produces plausible reconstructions which is a remarkable achievement given just a single image and noisy instance segmentations. Color encodes depth in the image co-ordinate frame (blue is closer). More results can be found at http://goo.gl/lmALxQ
linear shape models [46] may prove helpful, as well as a tighter integration between segmentation and reconstruction.

Acknowledgements

This work was supported in part by NSF Award IIS-1212798 and ONR MURI-N00014-10-1-0933. Shubham Tulsiani was supported by the Berkeley fellowship and Joao Carreira was supported by the Portuguese Science Foundation, FCT, under grant SFRH/BPD/84194/2012. We gratefully acknowledge NVIDIA corporation for the donation of Tesla GPUs for this research.

References

[1] B. Alexe, T. Deselaers, and V. Ferrari. What is an object? In CVPR, 2010. 1
[2] D. Anguelov, P. Srinivasan, D. Koller, S. Thrun, J. Rodgers, and J. Davis. Scape: shape completion and animation of people. In ACM Trans. Graph., 2005. 2
[3] P. Arbeláez, J. Pont-Tuset, J. T. Barron, F. Marques, and J. Malik. Multiscale combinatorial grouping. In CVPR, 2014. 1
[4] N. Aspert, D. Santa-Cruz, and T. Ebrahimi. Mesh: Measuring errors between surfaces using the Hausdorff distance. In ICME, 2002. 6
[5] J. T. Barron and J. Malik. Color constancy, intrinsic images, and shape estimation. ECCV, 2012. 4, 5, 6, 7
[6] J. T. Barron and J. Malik. Shape, illumination, and reflectance from shading. Technical Report UCB/EECS-2013-117, EECS, UC Berkeley, May 2013. 5
[7] A. Bartoli, V. Gay-Bellile, U. Castellani, J. Peyras, S. Olsen, and F. Sayd. Coarse-to-fine low-rank structure-from-motion. In Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on, pages 1–8, June 2008. 2
[8] V. Blanz and T. Vetter. A morphable model for the synthesis of 3d faces. In Proceedings of the 26th annual conference on Computer graphics and interactive techniques, 1999. 1, 2
[9] V. Blanz and T. Vetter. Face recognition based on fitting a 3d morphable model. TPAMI, 25(9):1063–1074, 2003. 4
[10] C. Bregler, A. Hertzmann, and H. Biermann. Recovering non-rigid 3d shape from image streams. In Computer Vision and Pattern Recognition, 2000. Proceedings. IEEE Conference on, volume 2, pages 690–696 vol.2, 2000. 2
[11] J. Carreira, R. Caseiro, J. Batista, and C. Sminchisescu. Semantic segmentation with second-order pooling. In ECCV, 2012. 1
[12] J. Carreira and C. Sminchisescu. Constrained parametric min-cuts for automatic object segmentation. In CVPR, 2010. 1
[13] T. Cashman and A. Fitzgibbon. What shape are dolphins? building 3d morphable models from 2d images. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 35(1):232–244, Jan 2013. 3
[14] Y. Chen, T.-K. Kim, and R. Cipolla. Inferring 3d shapes and deformations from single views. In Proceedings of the 11th European Conference on Computer Vision Conference on Computer Vision: Part III, ECCV’10, pages 300–313, Berlin, Heidelberg, 2010. Springer-Verlag. 3
[15] I. Endres and A. Hoiem. Category independent object proposals. In ECCV, 2010. 1
[16] C. H. Esteban and F. Schmitt. Silhouette and stereo fusion for 3d object modeling. Comput. Vis. Image Underst., 96(3):367–392, Dec. 2004. 3, 5
[17] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012). http://www.pascal-network.org/challenges/VOC/voc2012/workshop/. 6
[18] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. IJCV, 2010. 1
[19] R. Garg, A. Roussos, and L. Agapito. Dense variational reconstruction of non-rigid surfaces from monocular video. In CVPR, June 2013. 2
[20] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In CVPR, 2014. 1
[21] A. Gupta, A. A. Efros, and M. Hebert. Blocks world revisited: Image understanding using qualitative geometry and mechanics. In Computer Vision–ECCV 2010, pages 482–496. Springer, 2010. 2
[22] B. Hariharan, P. Arbelaez, L. Bourdev, S. Maji, and J. Malik. Semantic contours from inverse detectors. In ICCV, 2011. 6
[23] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik. Simultaneous detection and segmentation. In European Conference on Computer Vision (ECCV), 2014. 1, 6, 7
[24] M. Hejrati and D. Ramanan. Analyzing 3d objects in cluttered images. In NIPS, pages 602–610, 2012. 2
[25] D. Hoiem and S. Savarese. Representations and techniques for 3d object recognition and scene interpretation, volume 15. Morgan & Claypool Publishers, 2011. 2
[26] E. Kalogerakis, S. Chaudhuri, D. Koller, and V. Koltun. A Probabilistic Model of Component-Based Shape Synthesis. ACM Transactions on Graphics, 31(4), 2012. 3
[27] I. Kemelmacher-Shlizerman. Internet based morphable model. In International Conference on Computer Vision (ICCV), 2011. 2
[28] P. Krähenbühl and V. Koltun. Geodesic object proposals. In Computer Vision–ECCV 2014, pages 725–739. Springer International Publishing, 2014. 1
[29] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012. 1
[30] A. Laurentini. The visual hull concept for silhouette-based image understanding. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 16(2):150–162, Feb 1994. 3
[31] Y. LeCun, B. Boser, J. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Backpropagation applied to hand-written zip code recognition. In *Neural Computation*, 1989. 1
[32] J. J. Lim, H. Pirsiavash, and A. Torralba. Parsing ikea objects: Fine pose estimation. In *ICCV*, 2013. 2
[33] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. *arXiv preprint arXiv:1405.0312*, 2014. 1
[34] R. Nevatia and T. O. Binford. Description and recognition of curved objects. *Artificial Intelligence*, 8(1):77–98, 1977. 2
[35] M. Prasad, A. Fitzgibbon, A. Zisserman, and L. Van Gool. Finding nemo: Deformable object class modelling using curve matching. In *CVPR*, 2010. 2
[36] L. G. Roberts. *Machine Perception of Three-Dimensional Solids*. PhD thesis, Massachusetts Institute of Technology, 1963. 2
[37] Y. Sahillioli and Y. Yemez. A surface deformation framework for 3d shape recovery. In *Multimedia Content Representation, Classification and Security*, volume 4105 of *Lecture Notes in Computer Science*, pages 570–577. Springer Berlin Heidelberg, 2006. 3
[38] S. Satkin, M. Rashid, J. Lin, and M. Hebert. 3dnn: 3d nearest neighbor. *International Journal of Computer Vision*, pages 1–29, 2014. 2
[39] Submitted. Viewpoints and keypoints. In *Attached In Supplementary*. 6
[40] S. Suwajanakorn, I. Kemelmacher-Shlizerman, and S. Seitz. Total moving face reconstruction. In D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, editors, *Computer Vision ECCV 2014*, volume 8692 of *Lecture Notes in Computer Science*, pages 796–812. Springer International Publishing, 2014. 2
[41] L. Torresani, A. Hertzmann, and C. Bregler. Non-rigid structure-from-motion: Estimating shape and motion with hierarchical priors. *TPAMI*, 2008. 2, 3
[42] N. R. Twarog, M. F. Tappen, and E. H. Adelson. Playing with puffball: simple scale-invariant inflation for use in vision and graphics. In *ACM Symp. on Applied Perception*, 2012. 2, 3
[43] K. E. Van de Sande, J. R. Uijlings, T. Gevers, and A. W. M. Smeulders. Segmentation as selective search for object recognition. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 1879–1886. IEEE, 2011. 1
[44] S. Vicente, J. Carreira, L. Agapito, and J. Batista. Reconstructing pascal voc. *CVPR 2014*, 2014. 1, 2, 3, 6, 7
[45] S. Vicente and L. de Agapito. Balloon shapes: Reconstructing and deforming objects with volume from images. In *3DV*, pages 223–230. IEEE, 2013. 3
[46] Z. Wu, S. Song, A. Khosla, X. Tang, and J. Xiao. 3d shapenets for 2.5 d object recognition and next-best-view prediction. *arXiv preprint arXiv:1406.5670*, 2014. 9
[47] Y. Xiang, R. Mottaghi, and S. Savarese. Beyond pascal: A benchmark for 3d object detection in the wild. In *WACV*, 2014. 6, 7
[48] J. Xiao, B. Russell, and A. Torralba. Localizing 3d cuboids in single-view images. In *Advances in Neural Information Processing Systems*, pages 746–754, 2012. 2
[49] S. Zhu, L. Zhang, and B. Smith. Model evolution: An incremental approach to non-rigid structure from motion. In *CVPR*, 2010. 2
[50] M. Zia, M. Stark, B. Schiele, and K. Schindler. Detailed 3d representations for object recognition and modeling. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(11):2608–2623, 2013. 2
1. Additional Results

We present additional reconstruction results on PASCAL VOC using both the fully automatic and fully annotated settings. As described in the main manuscript, the fully automatic setting involves reconstructing from automatically detected and segmented instances given the full image (using the technique of [1] in our case). In the fully annotated setting, we present results for the methodology described in section 4.1 where we inspect the expressiveness of our models by reconstructing the whole annotated dataset.

Figure 1 shows results using the fully annotated setting. Note that the reconstructions qualitatively look better owing to poses obtained from ground truth keypoints and clean segmentations. The reconstructed models also tend to fit the silhouettes better. This setting is directly comparable to the class-based reconstruction algorithm presented in [2]. The category “boat” is almost always bad because the large intra-class variation results in very noisy estimated cameras in the first place that leads to bad basis shape models being learnt. In some cases like “trains”, the visual hull initialization of our basis shape models (section 2.2) results in an open front as a result of the isosurface extraction failing. Given this bad initial topology, our algorithm does the best it can to explain the silhouettes in the dataset. Figure 2 presents more example reconstructions in the fully automatic setting.

Videos showing more of our results can be found at http://goo.gl/1mALxQ. The videos show full 3D reconstructions on PASCAL VOC using both settings above and also texture-mapped depth maps with upto 30° out of plane rotations.

References

[1] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik. Simultaneous detection and segmentation. In European Conference on Computer Vision (ECCV), 2014.

[2] S. Vicente, J. Carreira, L. Agapito, and J. Batista. Reconstructing pascal voc. CVPR 2014, 2014.
Figure 1: Reconstructions obtained using our algorithm on PASCAL VOC using the fully annotated setting (please see text for details). We show the clean segmentation input, the inferred shape overlaid on the image, a 2.5D depth map (after the bottom-up refinement stage), the mesh in the image viewpoint and two other views. Note that here we use ground truth masks and obtain pose using the ground truth keypoints. Color encodes depth in the image co-ordinate frame (blue is closer). More results can be found at http://goo.gl/lmALxQ
Figure 2: Fully automatic reconstructions on detected instances (0.5 IoU with ground truth) using our models on rigid categories in PASCAL VOC. We show our instance segmentation input, the inferred shape overlaid on the image, a 2.5D depth map (after the bottom-up refinement stage), the mesh in the image viewpoint and two other views. It can be seen that our method produces plausible reconstructions which is a remarkable achievement given just a single image and noisy instance segmentations. Color encodes depth in the image co-ordinate frame (blue is closer). More results can be found at http://goo.gl/lmALxQ