

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 an automated pipeline with pixels as inputs and 3D surfaces of various rigid categories as outputs in images of realistic scenes. At the core of our approach are 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

Consider the car in Figure 1. As humans, not only can we infer at a glance that the image contains a car, we also construct a rich internal representation of it such as its location and 3D pose. Moreover, we have a guess of its 3D shape, even though we might never have seen this particular car. We can do this because we don’t experience the image of this car *tabula rasa*, but in the context of our “remembrance of things past”. Previously seen cars enable us to develop a notion of the 3D shape of cars, which we can project to this particular instance. We also specialize our representation to this particular instance (e.g. any custom decorations it might have), signalling that both top-down and bottom-up cues influence our percept [26].

A key component in such a process would be a mechanism to build 3D shape models from past visual experiences. We have developed an algorithm that can build category-specific shape models from just images with 2D annotations (segmentation masks and a small set of key-points) present in modern computer vision datasets (e.g.

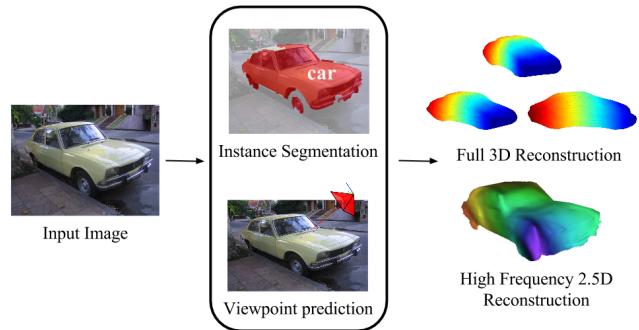


Figure 1: Automatic object reconstruction from a single image obtained by our system. Our method leverages estimated instance segmentations and predicted viewpoints to generate a full 3D mesh and high frequency 2.5D depth maps.

PASCAL VOC [15]). These models are then used to guide the top down 3D shape reconstruction of novel 2D car images. We complement our top-down shape inference algorithm with a bottom-up module that further refines our shape estimate for a particular instance. Finally, building upon the rapid recent progress in recognition modules [2, 11, 17, 20, 34] (object detection, segmentation and pose estimation), we demonstrate that our learnt models are robust when applied “in the wild” enabling fully automatic reconstructions with just images as inputs.

The recent method of Vicente *et al.* [36] reconstructs 3D models from similar annotations as we do but it has a different focus: it aims to reconstruct a fully 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 can work in such settings, partly because it uses explicit 3D shape models. Our work also has connections to that of Kemelmacher-Shlizerman *et al.* [23, 32] 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.

In the history of computer vision, model-based object

* Authors contributed equally

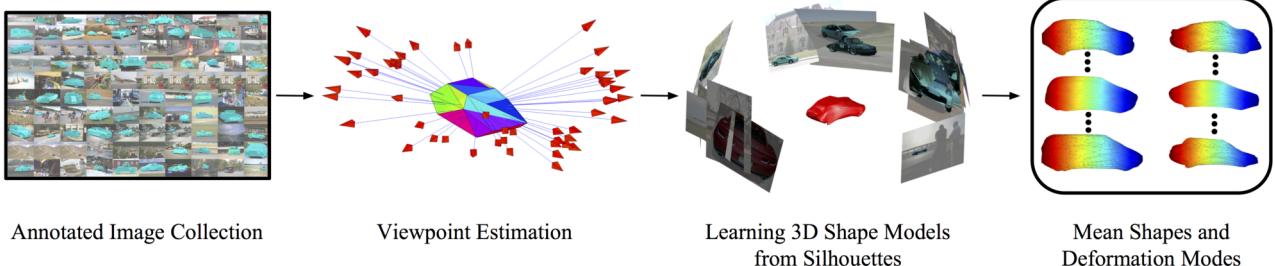


Figure 2: Overview of our training pipeline. We use an annotated image collection to estimate camera viewpoints which we then use alongwith object silhouettes to learn 3D shape models. Our learnt shape models, as illustrated in the rightmost figure are capable of deforming to capture intra-class shape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

reconstruction from a single image has reflected varying preferences on model representations. Generalized cylinders [27] resulted in very compact descriptions for certain classes of shapes, and can be used for category level descriptions, but the fitting problem for general shapes is challenging. Polyhedral models [18, 40], which trace back to the early work of Roberts [29], and CAD models [25, 31] provide crude approximations of shape and given a set of point correspondences can be quite effective for determining instance viewpoints. Here we pursue more expressive basis shape models [1, 7, 42] which establish a balance between the two extremes as they can deform but only along class-specific modes of variation. In contrast to previous work (e.g. [42]), we fit them to automatic figure-ground 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 [36]. Our approach operates 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 describe these two stages of model learning in the following subsections. Figure 2 illustrates this training pipeline of ours.

ure 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. Originally proposed for recovering shape and deformations from video [6, 33, 16, 10], NRSfM is a natural choice for viewpoint estimation from sparse correspondences 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 [41] or flower petal and clown fish [28]. Closer to our work, Hejrati and Ramanan [21] 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.* [33] and 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 [37] and with rigid structure-from-motion [36] but, to the best of our knowledge, not in conjunction with NRSfM.

NRSfM Model. Given K keypoint correspondences per instance $n \in \{1, \dots, N\}$, our adaptation of the NRSfM algorithm in [33] corresponds to maximizing the likelihood of the following model:

$$\begin{aligned} P_n &= (I_K \otimes c_n R_n) S_n + T_n + N_n \\ S_n &= \bar{S} + V z_n \\ z_n &\sim \mathcal{N}(0, I), \quad N_n \sim \mathcal{N}(0, \sigma^2 I) \end{aligned} \tag{1}$$

$$\begin{aligned} \text{subject to: } R_n R_n^T &= I_2 \\ \sum_{k=1}^K C_n^{mask}(p_{k,n}) &= 0, \quad \forall n \in \{1, \dots, N\} \end{aligned} \tag{2}$$

Here, P_n is the 2D projection of the 3D shape S_n with white noise N_n 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 $[V_1, V_2, \dots, V_m] = V$ and latent deformation parameters z_n . Our key modification is constraint (2) 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. Missing data (occluded keypoints) is dealt with by “filling-in” the values using the forward equations after the E-step. The algorithm computes shape parameters $\{\bar{S}, V\}$, 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.

2.2. 3D Basis Shape Model Learning

Equipped with camera projection parameters and keypoint correspondences (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* [24] and variational methods derived from *snakes* e.g [14, 30] which deform a surface mesh iteratively until convergence. Some interesting recent papers have extended variational

approaches to handle categories [12, 13] but typically require some form of 3D annotations to bootstrap models. A recently proposed visual-hull based approach [36] 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 sail boats and cruise liners. Such models typically require knowledge of object parts, their spatial arrangements etc. [22] 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 enables us to learn deformable models without relying on point correspondences between scanned 3D exemplars [8].

Our shape model $M = (\bar{S}, V)$ comprises of a mean shape \bar{S} and deformation bases $V = \{V_1, \dots, V_K\}$ 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} refers to the Chamfer distance field of the binary mask of silhouette O and $\Delta^k(p; Q)$ is defined as the squared 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 Δ notation it can be written as follows:

$$E_s(S, O, P) = \sum_{C^{\text{mask}}(p) > 0} \Delta^1(p; O) \quad (3)$$

Silhouette Coverage. Using silhouette consistency alone

would just drive points projected outside in towards the silhouette. This wouldn't ensure though that the object silhouette is "filled" - i.e. there might be overcarving. We deal with it by having an energy term that encourages points on the silhouette to pull nearby projected points towards them. Formally, this can be expressed as:

$$E_c(S, O, P) = \sum_{p \in O} \Delta^m(p; P(S)) \quad (4)$$

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 corresponding energy term penalizes deviation of the shape from the 3D keypoints KP for each instance. Specifically, this can be written as:

$$E_{kp}(S, O, P) = \sum_{\kappa \in KP} \Delta^m(\kappa; S) \quad (5)$$

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 δ parameter represents the mean squared displacement between neighboring points and it encourages all faces to have similar size. Here V_{ki} is the i^{th} point in the k^{th} basis.

$$E_l(\bar{S}, V) = \sum_i \sum_{j \in N(i)} ((\|\bar{S}_i - \bar{S}_j\| - \delta)^2 + \sum_k \|V_{ki} - V_{kj}\|^2) \quad (6)$$

Normal 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

$$E_n(S) = \sum_i \sum_{j \in N(i)} (1 - \vec{\mathcal{N}}_i \cdot \vec{\mathcal{N}}_j) \quad (7)$$

Here, $\vec{\mathcal{N}}_i$ 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. Note 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 [4].

Our total energy is given in equation 8. In addition to the above smoothness priors we also penalize the L_2 norm of

the deformation parameters α_i to prevent unnaturally large deformations.

$$E_{tot}(\bar{S}, V, \alpha) = E_l(\bar{S}, V) + \sum_i (E_s^i + E_{kp}^i + E_c^i + E_n^i + \sum_k (\|\alpha_{ik} V_k\|_F^2)) \quad (8)$$

Learning. We solve the optimization problem in equation 9 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 α 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 α in our formulation. In order to deal with imperfect segmentations and wrongly estimated keypoints, we use truncated versions of the above energies that reduce the impact of 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 learns finer structures like chair legs and bicycle handles, which become more prominent with deformations.

$$\begin{aligned} \min_{\bar{S}, V, \alpha} & E_{tot}(\bar{S}, V, \alpha) \\ \text{subject to: } & S^i = \bar{S} + \sum_k \alpha_{ik} V_k \end{aligned} \quad (9)$$

Our training objective is highly non-convex and non-smooth and is susceptible to initialization. We follow the suggestion of [14] and initialize our mean shape with a soft visual hull computed using all training instances. The deformation bases and deformation weights are initialized randomly.

3. 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 detecting and segmenting objects in the scene, we infer their coarse 3D poses and use them to fit our top-down shape models to the noisy segmentation masks. Finally, we recover high frequency shape details from shading cues. We will now explain these components one at a time.

Initialization. During inference, we first detect and segment the object in the image [20] and then predict viewpoint (rotation matrix) and subcategory for the object using a CNN based system similar to [34] (augmented to predict subcategories). Our learnt models are at a canonical bounding box scale - all objects are first resized to a particular width during training. Given the predicted bounding box, we scale the learnt mean shape of the predicted subcategory

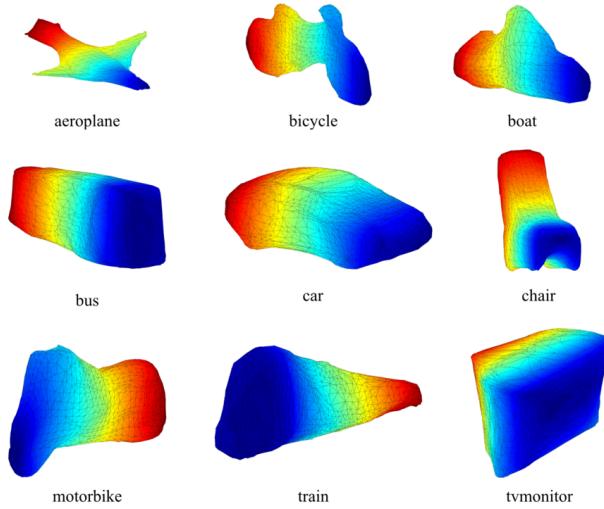


Figure 4: Mean shapes learnt for rigid classes in PASCAL VOC obtained using our basis shape formulation. Color encodes depth when viewed frontally.

accordingly. Finally, the mean shape is rotated as per the predicted viewpoint and translated to the center of the predicted bounding box.

Shape Inference. After initialization, we solve for the deformation weights α (initialized to 0) as well as all the camera projection parameters (scale, translation and rotation) by optimizing equation (9) for fixed \bar{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 [5, 4], SIRFS, which exploits statistical regularities between shapes, reflectance and illumination. Formally, SIRFS is formulated as the following optimization problem:

$$\underset{Z, L}{\text{minimize}} \quad g(I - S(Z, L)) + f(Z) + h(L)$$

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 ϵ 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 a few minutes to learn a shape model for a given object category.

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 consider images from the challenging PASCAL VOC 2012 dataset [15] which contain objects from the 10 rigid object categories (as listed in Table 1). We use the publicly available ground truth class-specific keypoints [9] and object segmentations [19]. Since ground truth 3D shapes are unavailable for PASCAL VOC and most other detection datasets, we evaluated the expressiveness of our learned 3D models on the next best thing we managed to obtain: the PASCAL3D+ dataset [39] 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

| | 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[36] | 5.07 | 6.03 | 8.80 | 8.76 | 4.38 | 5.74 | 4.86 | 6.49 | 17.52 | 8.37 | 7.60 |
| | Puffball[35] | 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[36] | 9.39 | 7.24 | 11.43 | 18.42 | 6.86 | 7.39 | 8.06 | 12.21 | 29.57 | 5.75 | 11.63 |
| | SIRFS[4] | 12.98 | 12.31 | 16.03 | 29.21 | 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 [36, 35] using ground truth keypoints and masks on PASCAL VOC. Note that [36] 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 | 5.34 | 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 | 8.24 |
| | PP+SDS | 6.58 | 14.02 | 14.43 | 6.65 | 7.96 | 7.47 | 7.57 | 15.21 | 15.23 | 13.24 | 10.84 |
| | Puffball[35](SDS) | 9.68 | 10.23 | 11.80 | 15.95 | 12.42 | 8.28 | 9.45 | 9.60 | 23.38 | 9.26 | 12.00 |
| Depth | KP+Mask | 9.02 | 7.26 | 13.51 | 12.10 | 8.04 | 8.02 | 10.00 | 23.05 | 25.57 | 7.48 | 12.41 |
| | KP+SDS | 9.07 | 7.98 | 13.57 | 9.90 | 7.98 | 7.96 | 9.99 | 22.57 | 23.59 | 7.64 | 12.03 |
| | PP+SDS | 10.94 | 11.64 | 12.26 | 15.95 | 13.17 | 10.06 | 12.55 | 21.19 | 36.37 | 8.98 | 15.31 |
| | SIRFS[4] | 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.

- 1) the Hausdorff distance normalized by the 3D bounding box size of the ground truth model [3] and 2) a depth map error to evaluate the quality of the reconstructed visible object surface, measured as the mean absolute distance between reconstructed and ground truth depth:

$$Z\text{-MAE}(\hat{Z}, Z^*) = \frac{1}{n \cdot \gamma} \min_{\beta} \sum_{x,y} |\hat{Z}_{x,y} - Z_{x,y}^* - \beta| \quad (10)$$

where \hat{Z} and Z^* represent predicted and ground truth depth maps respectively. Analytically, β can be computed as the median of $\hat{Z} - Z^*$ and γ 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 and scale invariant.

4.1. Expressiveness of Learned 3D Models

We learn and fit our 3D models on the same whole dataset (no train/test split), following the setup of Vicente et al [36]. 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[35] and SIRFS[4] – 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 [34] and augment it to predict subtypes at test time. This is achieved by training the system as described in [34] with additional subcategory labels 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 [20].

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. 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 errors can be high as the metrics are sensitive to bad alignment. The data sparsity issue is especially visible in the case of sofas where in a train/test setting in Table 2 the numbers drop significantly with less training data (only 34 instances). Note we do not evaluate our bottom-up component as the PASCAL 3D+ meshes provided do not share the same high frequency shape details as the instance. We will show qualitative results in the next subsection.

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 [20] in Figure 5. We can see 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.

5. Conclusion

We have proposed what may be the first approach to perform fully automatic object reconstruction from a single image on a large and realistic dataset. 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 [39] 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-linear shape models [38]

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 João 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] 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
- [2] P. Arbeláez, J. Pont-Tuset, J. T. Barron, F. Marques, and J. Malik. Multiscale combinatorial grouping. *CVPR*, 2014. 1
- [3] N. Aspert, D. Santa-Cruz, and T. Ebrahimi. Mesh: Measuring errors between surfaces using the hausdorff distance. In *ICME*, 2002. 6
- [4] J. T. Barron and J. Malik. Color constancy, intrinsic images, and shape estimation. *ECCV*, 2012. 4, 5, 6
- [5] J. T. Barron and J. Malik. Shape, illumination, and reflectance from shading. Technical Report UCB/EECS-2013-117, EECS, UC Berkeley, May 2013. 5
- [6] A. Bartoli, V. Gay-Bellile, U. Castellani, J. Peyras, S. Olsen, and P. 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
- [7] 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. 2
- [8] V. Blanz and T. Vetter. Face recognition based on fitting a 3d morphable model. *TPAMI*, 25(9):1063–1074, 2003. 3
- [9] L. Bourdev, S. Maji, T. Brox, and J. Malik. Detecting people using mutually consistent poselet activations. In *European Conference on Computer Vision (ECCV)*, 2010. 5
- [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 and C. Sminchisescu. Constrained parametric min-cuts for automatic object segmentation. In *CVPR*, 2010. 1
- [12] 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
- [13] 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

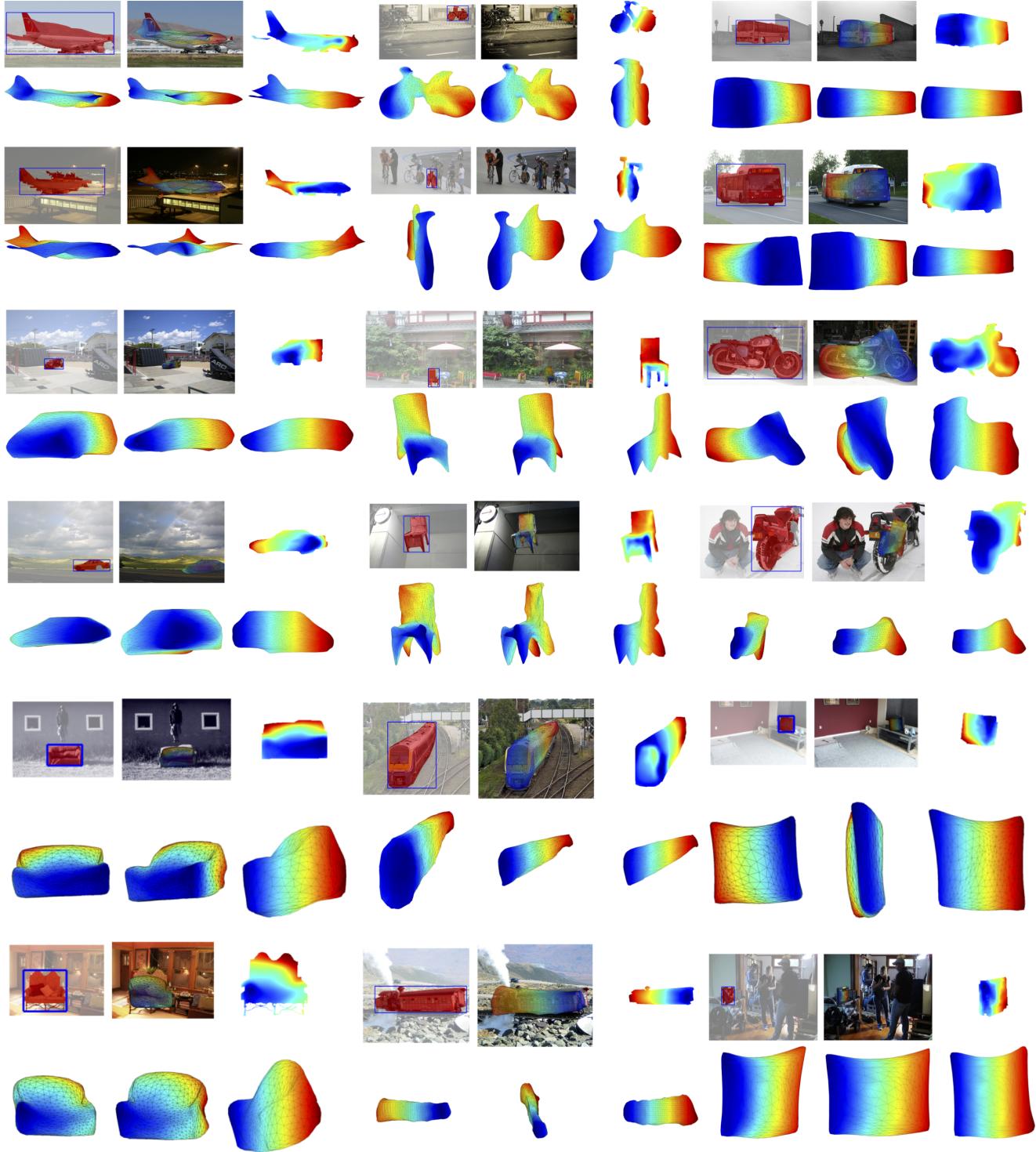


Figure 5: Fully automatic reconstructions on detected instances (0.5 IoU with ground truth) using our models on rigid categories in PASCAL VOC. We show our instance segmentation input, the inferred shape overlaid on the image, a 2.5D depth map (after the bottom-up refinement stage), the mesh in the image 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/1mALxQ>.

- [14] 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, 4
- [15] 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>. 1, 5
- [16] R. Garg, A. Roussos, and L. Agapito. Dense variational reconstruction of non-rigid surfaces from monocular video. In *CVPR*, June 2013. 2
- [17] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014. 1
- [18] 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
- [19] B. Hariharan, P. Arbelaez, L. Bourdev, S. Maji, and J. Malik. Semantic contours from inverse detectors. In *ICCV*, 2011. 5
- [20] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik. Simultaneous detection and segmentation. In *European Conference on Computer Vision (ECCV)*, 2014. 1, 4, 6, 7
- [21] M. Hejrati and D. Ramanan. Analyzing 3d objects in cluttered images. In *NIPS*, pages 602–610, 2012. 2
- [22] 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
- [23] I. Kemelmacher-Shlizerman. Internet based morphable model. In *International Conference on Computer Vision (ICCV)*, 2011. 1
- [24] 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
- [25] J. J. Lim, H. Pirsiavash, and A. Torralba. Parsing ikea objects: Fine pose estimation. In *ICCV*, 2013. 2
- [26] C. Nandakumar, A. Torralba, and J. Malik. How little do we need for 3-d shape perception? *Perception-London*, 40(3):257, 2011. 1
- [27] R. Nevatia and T. O. Binford. Description and recognition of curved objects. *Artificial Intelligence*, 8(1):77–98, 1977. 2
- [28] M. Prasad, A. Fitzgibbon, A. Zisserman, and L. Van Gool. Finding nemo: Deformable object class modelling using curve matching. In *CVPR*, 2010. 2
- [29] L. G. Roberts. *Machine Perception of Three-Dimensional Solids*. PhD thesis, Massachusetts Institute of Technology, 1963. 2
- [30] Y. Sahillioglu 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
- [31] S. Satkin, M. Rashid, J. Lin, and M. Hebert. 3dnn: 3d nearest neighbor. *International Journal of Computer Vision*, pages 1–29, 2014. 2
- [32] 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. 1
- [33] L. Torresani, A. Hertzmann, and C. Bregler. Non-rigid structure-from-motion: Estimating shape and motion with hierarchical priors. *TPAMI*, 2008. 2
- [34] S. Tulsiani and J. Malik. Viewpoints and keypoints. In *CVPR*, 2015. 1, 4, 6
- [35] N. R. Twarog, M. F. Tappen, and E. H. Adelson. Playing with puffedball: simple scale-invariant inflation for use in vision and graphics. In *ACM Symp. on Applied Perception*, 2012. 6
- [36] S. Vicente, J. Carreira, L. Agapito, and J. Batista. Reconstructing pascal voc. *CVPR 2014*, 2014. 1, 2, 3, 6
- [37] S. Vicente and L. de Agapito. Balloon shapes: Reconstructing and deforming objects with volume from images. In *3DV*, pages 223–230. IEEE, 2013. 2
- [38] Z. Wu, S. Song, A. Khosla, F. Yu, L. Zhang, X. Tang, and J. Xiao. 3d shapenets: A deep representation for volumetric shape modeling. In *CVPR*, 2015. 7
- [39] Y. Xiang, R. Mottaghi, and S. Savarese. Beyond pascal: A benchmark for 3d object detection in the wild. In *WACV*, 2014. 5, 7
- [40] 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
- [41] S. Zhu, L. Zhang, and B. Smith. Model evolution: An incremental approach to non-rigid structure from motion. In *CVPR*, 2010. 2
- [42] M. Z. 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