Learning Category-Specific Deformable 3D Models for Object Reconstruction

Shubham Tulsiani*, Abhishek Kar*, João Carreira and Jitendra Malik, *Fellow, IEEE*

Abstract—We address the problem of fully automatic object localization and reconstruction from a single image. This is both a very challenging and very important problem which has, until recently, received limited attention due to difficulties in segmenting objects and predicting their poses. Here we leverage recent advances in learning convolutional networks for object detection and segmentation and introduce a complementary network for the task of camera viewpoint prediction. These predictors are very powerful, but still not perfect given the stringent requirements of shape reconstruction. Our main contribution is a new class of deformable 3D models that can be robustly fitted to images based on noisy pose and silhouette estimates computed upstream and that can be learned directly from 2D annotations available in object detection datasets. Our models capture top-down information about the main global modes of shape variation within a class providing a "low-frequency" shape. In order to capture fine instance-specific shape details, we fuse it with a high-frequency component recovered from shading cues. A comprehensive quantitative analysis and ablation study on the PASCAL 3D+ dataset validates the approach as we show fully automatic reconstructions on PASCAL VOC as well as large improvements on the task of viewpoint prediction.

Index Terms—Object Reconstruction, 3D Shape Modeling, Viewpoint Estimation, Scene Understanding

1 Introduction

CONSIDER the chairs in Figure 1. As humans, not only can we infer at a glance that the image contains three chairs, we also construct a rich internal representation of each of them such as their locations and 3D poses. Moreover, we have a guess of their 3D shapes, even though we might never have seen these particular chairs. We can do this because we do not experience this image *tabula rasa*, but in the context of our "remembrance of things past". Previously seen chairs enable us to develop a notion of the 3D shape of chairs, which we can project to the instances in this particular image. We also specialize our representation to these particular instances (e.g. any custom decorations they might have), signalling that both top-down and bottom-up cues influence our percept [1]. In this work, we incorporate these principles in a computational approach for reconstructing objects given a single image.

The task of reconstructing objects from a single image is a challenging one – a typical image depicts many objects, each possibly belonging to a different object category; an object category, in turn, comprises instances of varying shapes, textures, size *etc.* and any particular instance may be viewed from a different viewpoint. Previous approaches to this problem can be broadly grouped into two paradigms. The paradigm of model-based object reconstruction has reflected varying preferences on model representations. Generalized cylinders [2] 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 [3], [4], which trace back to the early work of Roberts [5], and CAD models [6], [7], [8], cannot perfectly deform into shapes even slightly different from those in training data, but given a set of point correspondences can be quite effective for determining approximate instance viewpoints. Some recent methods have proposed using similar instances from a collection of CAD models [9], [10] for non-parametric reconstruction but their applications have been restricted to pre-segmented online product images or recovering 3D from 2.5D object scans [11]. Here we pursue more expressive basis shape models [12], [13], [14] which establish a balance between the two extremes as they can deform but only along class-specific modes of variation.

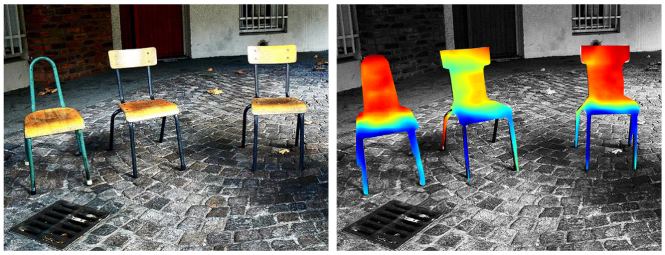

Fig. 1: Example outputs of our system, given a single image of a scene having chairs, a class that the system was exposed to during training. The coloring on the right image signals object-centric depth (we do not aim for globally consistent depths across multiple objects). Blue means close to the camera, red means far from the camera.

The alternate paradigm comprises of approaches that target the problem of object reconstruction in a class or object agnostic manner, either implicitly or explicitly using generic learned 3D shape cues [15], [16], or bottom-up cues

- *S.Tulsiani, A.Kar and J. Malik are with the Department of Electrical Engineering and Computer Science, University of California at Berkeley. E-mail: {shubhtuls, akar, malik}@eecs.berkeley.edu*
- *J. Carreira is with Google DeepMind and was with the Department of Electrical Engineering and Computer Science, University of California at Berkeley during the majority of this work. E-mail: joaoluis@google.com*

**Authors contributed equally.*

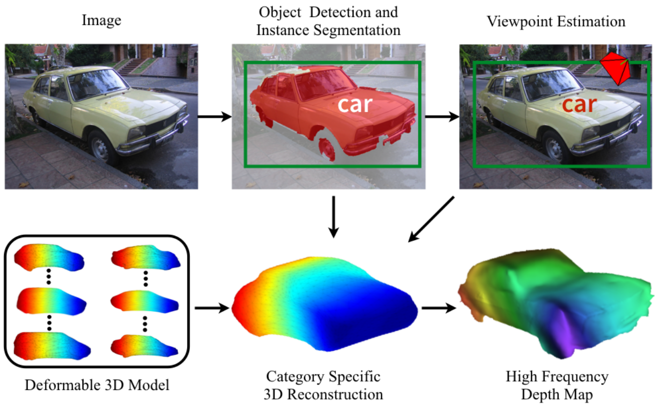
Image
Object Detection and Instance Segmentation
Viewpoint Estimation
car
car
Deformable 3D Model
Category Specific 3D Reconstruction
High Frequency Depth Map

Fig. 2: Overview of our full reconstruction method. We leverage estimated instance segmentations and predicted viewpoints to generate a full 3D mesh and a high frequency 2.5D depth map for each object in the image.

and the physics of image formation [17], [18] building upon the long tradition of shape-from-X, which traces back to seminal work by Horn [19]. These methods, while quite general, have not yet been demonstrated for 3D reconstruction – as opposed to 2.5D – and typically assume known object segmentation [18]. Some recent approaches have demonstrated the use of supervised learning techniques to implcitly learn generic cues to predict depth maps [20] and surface normals [21], [22] but these have primarily focused on inferring scene-level information which differs from our goal of perceiving the shape of objects.

In this work, we combine both these reconstruction paradigms - we obtain top-down shape information from our model-based reconstruction approach and complement it with bottom-up shape information obtained via an intrinsic image decomposition method. Crucially, in contrast to previous work (e.g. [18], [23], [24]), we do not require perfect knowledge of object localization and pose as our reconstruction is driven by automatic figure-ground object segmentations and viewpoint estimations.

The framework we propose to reconstruct the objects present in an image is outlined in Figure 2. As a first step, we leverage the recent progress made by the computer vision community in object detection [25] and instance segmentation [26], [27] to identify and localize objects in the image. For each object, we also predict a viewpoint in the form of three euler angles. We then use our learned deformable 3D shape models in conjunction with the viewpoint and localization information to produce a "top-down" 3D reconstruction for the object guided primarily by category level cues. Finally, we infuse our 3D shape with high frequency local shape cues to obtain our end result - a rich 3D reconstruction of the object. We briefly outline each of the components required for the above proposed framework.

Learning Deformable 3D Models. As noted earlier, previously seen objects allow us to develop a notion of 3D shape which informs inference for new instances. We present an algorithm that can build category-specific deformable shape models from just images with 2D annotations (segmentation masks and a small set of keypoints) present in modern computer vision datasets (e.g. PASCAL VOC [28]). These learnt shape models and deformations allow us to robustly infer shape while capturing intra-class shape variation.

Learning to Estimate Viewpoint. The first step towards being able to represent objects in 3D is to predict their viewpoint. This intermediate representation provides coarse information about the shape and its inference is a well studied problem in computer vision [29], [30], [31], [32], [33], [34], [35]. We train a Convolutional Neural Network (CNN) [36], [37] based architecture which can implicitly capture and aggregate local evidence to obtain a viewpoint estimate and demonstrate improvements over the state-of-the-art for this task.

Object Shape Recovery. Given an object's category, approximate localization and viewpoint, we obtain a 3D reconstruction for the corresponding object using the learned category-specific deformable shape model. We complement the top-down shape inferred via this inference with a bottom-up module that further refines our shape estimate for a particular instance. This framework allows us to capture the coarse as well as fine level shape details for objects from a single image.

Our paper is organized as follows: in Section 2 we describe our model learning pipeline where we estimate camera parameters for all training objects (Section 2.1) followed by our shape model formulation (Section 2.2) to learn 3D models. We then present our viewpoint estimation method in Section 3 and Section 4 describes our testing pipeline where we leverage our learnt models to reconstruct novel instances without assuming any annotations. We evaluate the various components of our approach in Section 5 and provide sample reconstructions in the wild.

This journal paper extends our earlier work [38] by providing a detailed exposition of our viewpoint prediction system and its systematic evaluation previously presented in [39]. We also report updated experiments with a slightly modified mesh metric and using improved versions of our pose prediction [39] and instance segmentation [27] systems.

2 Learning Deformable 3D Models

We are interested in learning 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 onto the image. We want to learn such models from just 2D training images, aided by ground truth segmentations and a few keypoints, similar to [23]. Our approach operates by first estimating the projection parameters (camera) for 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 estimated cameras. We describe these two stages of model learning in the following subsections. Figure 3 illustrates this training pipeline of ours.

2.1 Camera Estimation

We use the framework of NRSfM [40] to jointly estimate the projection parameters (rotation, translation and scale)

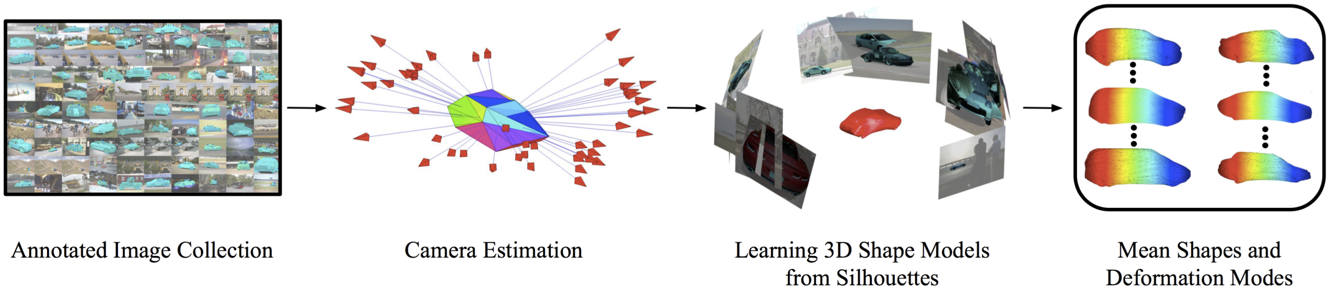
Annotated Image Collection
Camera Estimation
Learning 3D Shape Models from Silhouettes
Mean Shapes and Deformation Modes

Fig. 3: Overview of our training pipeline. We use an annotated image collection to estimate camera projection parameters 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.

for all training instances in each class. Originally proposed for recovering shape and deformations from video [40], [41], [42], [43], NRSfM is a natural choice for camera 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 [44] or flower petal and clown fish [45]. Closer to our work, Hejrati and Ramanan [46] 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.* [42] 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 [47] and rigid structure-from-motion [23] literature but, to the best of our knowledge, not in conjunction with NRSfM.

NRSfM Model Formulation. We are provided with an annotated training set $T : \{(O_n, P_n)\}_{n=1}^{N}$, where O_n is the instance silhouette and $P_n \in \mathbb{R}^{2\times K}$ denotes the annotated keypoint coordinates, possibly with missing entries (occluded/truncated keypoints). The annotated keypoints P_n are projections of the underlying 3D points $W_n \in \mathbb{R}^{3\times K}$ via the projection function π_n. In the NRSfM model, the space of 3D keypoint locations W_n is parametrized linearly and the projection function is assumed to be weakly orthographic *i.e.* $\pi_n \equiv (c_n, R_n, T_n)$, where c_n represents scale, $R_n \in \mathbb{R}^{2\times 3}$ denotes rotation and $T_n \in \mathbb{R}^{1\times 2}$ corresponds to 2D translation. Our goal is to infer the camera parameters (c_n, R_n, T_n) as well as 3D keypoint locations W_n for all instances in the annotated training set.

Formally, our adaptation of the NRSfM algorithm in [42] corresponds to maximizing the likelihood of the following model:

$$\begin{aligned} P_n &= c_n R_n W_n + 1^T T_n + N_n \\ W_n &= \bar{W} + \sum_{k=1}^{B} U_b z_{nb} \qquad (1) \\ z_n &\sim \mathcal{N}(0, I), \quad N_n^k \sim \mathcal{N}(0, \sigma^2 I) \end{aligned}$$

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

Here, the (partially) observed keypoint locations P_n are assumed to be the projection under $\pi_n \equiv (c_n, R_n, T_n)$ of the 3D shape W_n with white noise N_n. The shape is parameterized as a factored Gaussian with a mean shape $\bar{W}$, B basis vectors $[U_1, U_2, \cdots, U_B] = U$ and latent deformation parameters z_n. Our key modification is constraint in Eq. 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 cameras 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 $\{\bar{W}, U\}$, 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 4, NRSfM allows us to reliably predict cameras while being robust to intraclass variations.

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 the same 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* [48] and variational methods derived from *snakes* e.g [49], [50] which deform a surface mesh iteratively until convergence. Some interesting recent papers have extended variational approaches to handle categories [24], [51] but typically require some form of 3D annotations to bootstrap models. A recently proposed visual-hull based approach [23] requires

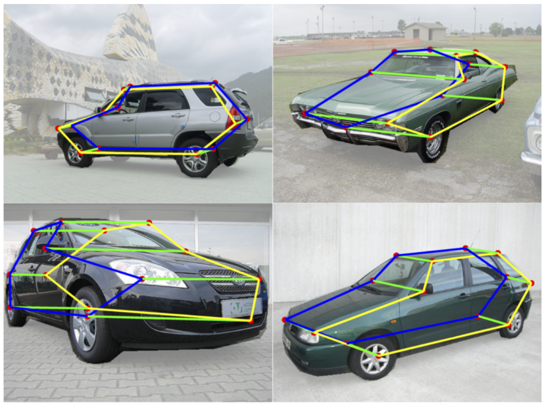

Fig. 4: NRSfM camera estimation: Estimated cameras visualized using a 3D car wireframe.

only 2D annotations as we do for class-based reconstruction and it was successfully demonstrated on PASCAL VOC but does not serve our purpose as it makes strong assumptions about the accuracy of the segmentation and will in fact fill entirely any segmentation with a voxel layer. In contrast, we build parametric shape models for categories that compactly capture intra class shape variations. The benefits of having a model of 3D shape are manifold: 1) we are more robust to noisy inputs (silhouettes and pose) allowing us to pursue reconstruction in a fully automatic setting and 2) we can potentially sample novel shapes from an object category.

Shape Model Formulation. We model our category shapes as a deformable point cloud[1]. As in the NRSfM model, we use a linear combination of basis vectors 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 [52].

The annotated training set $T : \{(O_n, P_n)\}_{n=1}^{N}$, where O_n is the instance silhouette and $P_n \in \mathbb{R}^{2\times K}$ denotes the annotated keypoint coordinates, is augmented after NRSfM to contain π_n (the projection function from world to image coordinates) and W_n (3D coordinates for a small set of keypoints). Our shape model $M = (\overline{S}, V)$ comprises of a mean shape $\overline{S}$ and deformation bases $V = \{V_1, ., V_K\}$ learnt from the augmented training set $T : \{(O_n, \pi_n, W_n)\}_{n=1}^{N}$. Note that the π_i we obtain using NRSfM corresponds to orthographic projection but our algorithm could handle perspective projection as well.

In addition to the above, we use the following notations – $\pi(S)$ 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.

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:

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 (*i.e.* squared distance to the closest silhouette point). In our Δ notation it can be written as follows:

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

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:

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

Keypoint Consistency. Our NRSfM algorithm provides us with sparse 3D keypoints along with camera projection parameters. 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 W for each instance. Specifically, this can be written as:

$$E_{kp}(S, W) = \sum_{\kappa\in W} \Delta^m(\kappa; S) \tag{5}$$

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.

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

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

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

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. Note that this, in conjunction with

[1] Differently from our earlier work [38] which learned a deformable model for each manually annotated subcategory, we simply learn one deformable shape model per object class.

our previous energies automatically enforces the commonly used prior that normals should be perpendicular to the viewing direction at the occluding contour [53].

Our total energy is given in equation Eq. 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 Eq. 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 5. 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.

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

Our training objective is highly non-convex and non-smooth and is susceptible to initialization. We follow the suggestion of [49] and initialize our mean shape with a soft visual hull computed using all training instances. The deformation bases and deformation weights are initialized randomly.

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', 'Keypoint Consistency' gradients and leverage Chamfer distance fields for obtaining 'Silhouette Consistency' gradients. Our overall computation takes only about 15 min to learn a deformable shape model for an object category with about 500 annotated examples.

3 Learning to Predict Viewpoint

In our proposed framework, viewpoint prediction is an important component towards reconstructing the objects present in an image. We are interested in a system that is accurate across all instances of a category as well as robust to localization errors in object detection. We present a CNN based system for the viewpoint prediction task and demonstrate that it leads to significant improvements over previous approaches.

Related Work. Recently, CNNs have been shown to outperform Deformable Part Model (DPM) [54] based methods for recognition tasks [25], [55]. Whereas DPMs explicitly model part appearances and their deformations, the CNN architecture allows such relations to be captured implicitly using a hierarchical convolutional structure. Girshick *et al.* [56] argued that DPMs could also be thought as a specific instantiation of CNNs and therefore training an end-to-end CNN for the corresponding task should outperform a method which instead explicitly models part appearances and relations.

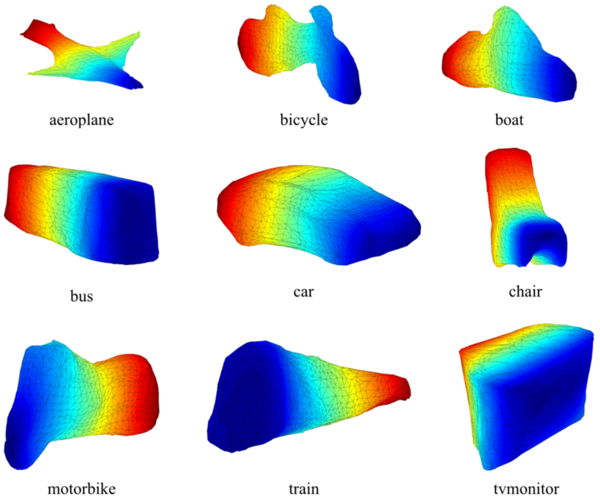

Fig. 5: Mean shapes learnt for rigid classes in PASCAL VOC obtained using our basis shape formulation. Color encodes depth when viewed frontally.

This result is particularly applicable to viewpoint estimation where the prominent approaches, from the initial instance based methods [29] to current state-of-the-art [57], [58] explicitly model local appearances and aggregate evidence to infer viewpoint. Pepik *et al.* [58] extend DPMs to 3D to model part appearances and rely on these to infer pose and Xiang *et al.* [57] introduce a separate DPM component corresponding to each viewpoint. Ghodrati *et al.* [59] differ from the explicit part-based methodology, using a fixed global descriptor to estimate viewpoint. We build on both these approaches by using a method which, while using a global descriptor, can implicitly capture part appearances.

Formulation. We formulate the global pose estimation for rigid categories as predicting the viewpoint wrt to a canonical pose. This is equivalent to determining the three euler angles corresponding to azimuth (ϕ), elevation(φ) and cyclo-rotation(ψ). We frame the task of predicting the euler angles as a classification problem where the classes $\{1, \ldots N_\theta\}$ correspond to N_θ disjoint angular bins. We note that the euler angles, and therefore every viewpoint, can be equivalently described by a rotation matrix. We will use the notion of viewpoints, euler angles and rotation matrices interchangeably.

Learning. Viewpoint is manifested in a 2D image by the spatial relationships among the different features of the object. CNN based methods which can implicitly capture and hierarchically build on such relations are therefore suitable candidates for viewpoint prediction. Let N_c be the number of object classes, N_a be number of angles to be predicted per instance. The number of output units per class is $N_a \times N_\theta$ resulting in a total of $N_c \times N_a \times N_\theta$ outputs. We adopt an approach similar to Girshick *et al.* [25] and finetune a CNN model whose weights are initialized from a model pretrained on the Imagenet [60] classification task. We experimented with the architectures from Krizhevsky *et al.* [55] (denoted as TNet) and Simonyan *et al.* [61] (denoted as ONet). The architecture of our network is the same as the corresponding pre-trained network with an additional fully-connected layer having $N_c \times N_a \times N_\theta$ output units.

Instead of training a separate CNN for each class, we implement a loss layer that selectively considers the $N_a \times N_\theta$ outputs corresponding the class of the training instance and computes a logistic loss for each of the angle predictions. This allows us to train a CNN which can jointly predict viewpoint for all classes, thus enabling learning a shared feature representation across all categories. We use the Caffe framework [62] to train and extract features from the CNN described above. We also use data-augmentation by jittering ground-truth bounding boxes and generating additional training examples by using boxes that overlap with the annotated bounding box with IoU > 0.7.

4 Reconstruction in the Wild

Given an image, our goal is to reconstruct the depicted objects. As the initial step, we use existing state-of-the-art systems [26] to detect and segment the objects present in the image. We then proceed to individually reconstruct each of the detected objects. We approach the problem of reconstructing these objects from the big picture downward - like a sculptor first hammering out the big chunks and then chiseling out the details. We infer their coarse 3D poses and use these along with the predicted instance segmentations to fit our top-down shape models to obtain a coarse top-down shape (Section 4.1). Finally, we recover high frequency shape details from shading cues present in the image (Section 4.2).

4.1 Category Specific Shape Inference

We have at our disposal category-level deformable shape models which can be driven by data-specific and shape-prior based energy terms to infer an object's shape. Recall that the proposed energy terms (Section 2.2), in particular 'Silhouette Consistency' ($E_s(S, O, \pi)$) and 'Silhouette Coverage' ($E_c(S, O, \pi)$) depend on a known object silhouette O and camera projection π. We first describe how we estimate O, π and then formulate an optimization problem to infer object shape S.

Initialization. Given an object detection along with its predicted instance segmentation, we use the largest connected component in the predicted segmentation to obtain the object silhouette O. We use the viewpoint prediction system described in Section 3 to predict the viewpoint for the detected object, thereby obtaining the camera rotation R. 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 accordingly and obtain camera scale c. The translation T is initialized to be the center of the predicted bounding box. These provide us an initial estimate of the camera parameters $\pi_0 \equiv (c, R, T)$.

Formulation. We want to infer a shape that best explains the observed object silhouette, respects generic shape priors (smoothness, continuity) and lies on the linear manifold of category-level shapes. Note that, unlike model learning phase, we do not have access to annotated keypoint locations and thus do not enforce the reconstruction to explain any keypoint locations. These observations are incorporated by the reconstruction energy defined in (using E_s, E_c, E_n defined in Section 2.2).

$$E_r = E_s + E_c + E_n \tag{10}$$

In addition to inferring the instance shape, we also observe that the initial camera estimate π_0 is only approximate as the R is predicted upto a dicretization and c, T are initialized coarsely. To alleviate this, we treat the camera parameters π as optimization variables. We further add regularizers to enforce the prior that shape deformation should be small and the the estimated camera should not deviate significantly from the initial camera estimate π_0 . Our final optimization for inferring the object reconstruction is given in Eq. 11.

$$\begin{aligned} \min_{\alpha,\pi} \quad & E_r(S, \pi) + \delta(\pi, \pi_0) + \sum_k (\|\alpha_k V_k\|_F^2)) \\ \text{subject to:} \quad & S = \bar{S} + \sum_k \alpha_k V_k \end{aligned} \tag{11}$$

Inference. In the above optimization, we first set the optimization variables α, π to $0, \pi_0$ respectively. We then solve the above minimization for the deformation weights α as well as all the camera projection parameters π (scale, translation and rotation) by optimizing Eq. 9 using block-coordinate descent (alternately optimizing π and α). The resulting output from the minimization provides us the projection parameters π as well as the inferred 3D shape $S = \bar{S} + \sum_k \alpha_k V_k$. We use the efficient implementations of energy gradients described earlier and consequently, our overall computation takes only about 2 sec to reconstruct a novel instance using a single CPU core.

4.2 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 [18], [53], 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

		aero	bike	boat	bus	car	chair	mbike	sofa	train	tv	mean
Mesh	Ours	**1.72**	**1.78**	3.01	**1.90**	1.77	**2.18**	1.88	**2.13**	**2.39**	**3.28**	**2.20**
	Carvi [23]	1.87	1.87	**2.51**	2.36	**1.41**	2.42	**1.82**	2.31	3.10	3.39	2.31
	Puff [63]	3.30	2.52	2.90	3.32	2.82	3.09	2.58	2.53	3.92	3.31	3.03
Depth	Ours	**9.51**	**9.27**	17.20	**12.71**	9.94	**7.78**	9.61	**13.70**	31.58	8.78	**13.01**
	Carvi [23]	10.05	9.28	**15.06**	18.51	**8.14**	7.98	**9.38**	13.71	**31.25**	**8.33**	13.17
	SIRFS [18]	13.52	13.79	20.78	29.93	22.48	18.59	16.80	18.28	40.56	20.18	21.49

TABLE 1: Studying the quality of our learnt 3D models: comparison between our method and [23], [63] using ground truth keypoints and masks on PASCAL VOC.

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.

5 Experiments

We have presented several contributions towards the goal of object reconstruction from a single image – Section 2 proposed a method to learn deformable 3D models from an annotated image set, Section 3 introduced a CNN based system to predict viewpoints and Section 4 put forward a framework for reconstructing objects from a single image. Our goal in the experiments was to empirically evaluate and qualitatively demonstrate the efficacy of each of these contributions.

We first examine the quality and expressiveness of our learned 3D models by evaluating how well they matched the underlying 3D shapes of the training data (Section 5.1). We also evaluate the accuracy of our viewpoint prediction system (Section 5.2). We then study their sensitivity of obtained reconstructions when fit to images using noisy automatic segmentations and pose predictions (Section 5.3) and finally present qualitative results for reconstructions from a single image (Section 5.4).

5.1 Quality of Learned 3D Models

The first question we address is whether the category-specific shape models we learn for each object class (Section 2) using an annotated image collection correctly explain the underlying 3D object shape for these annotated instances. Note that while it is not our final goal, this is itself a very challenging task - we have to obtain a dense 3D reconstruction for annotated images using just silhouettes and sparse keypoint correspondences. Recent work by Vicente *et al.* [23] addressed this task of 'lifting' an annotated image collection to 3D and we compare the performance of our model learning stage against their approach. We also incorporate category-agnostic shape inflation [63] and intrinsic image [53] methods as baselines. The evaluation metrics, dataset and results are described below.

Dataset. We consider images from the challenging PASCAL VOC 2012 dataset [28] which contain objects from the 10 rigid object categories (as listed in Table 1). We use the publicly available ground truth class-specific keypoints [64] and object segmentations [65] to learn category-specific shape models for each class. We learn and fit our 3D models on the whole dataset (no train/test split), following the setup of Vicente *et al.* [23].

Since ground truth 3D shapes are unavailable for PASCAL VOC and most other detection datasets, we evaluated the quality of our learned 3D models on the next best thing we managed to obtain: the PASCAL3D+ dataset [57] 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 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".

Metrics. We quantify the quality of our 3D models by comparing against the PASCAL 3D+ models using two metrics - 1) a mesh error metric computed as the Hausdorff distance between the ground truth and predicted mesh after translating both to the origin and normalizing by the diagonal of the tighest 3D bounding box of the ground truth mesh [66] 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:

$$Z\text{-MAE}(\hat{Z}, Z^*) = \frac{1}{n \cdot \gamma} \min_{\beta} \sum_{x,y} |\hat{Z}_{x,y} - Z^*_{x,y} - \beta| \qquad (12)$$

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.

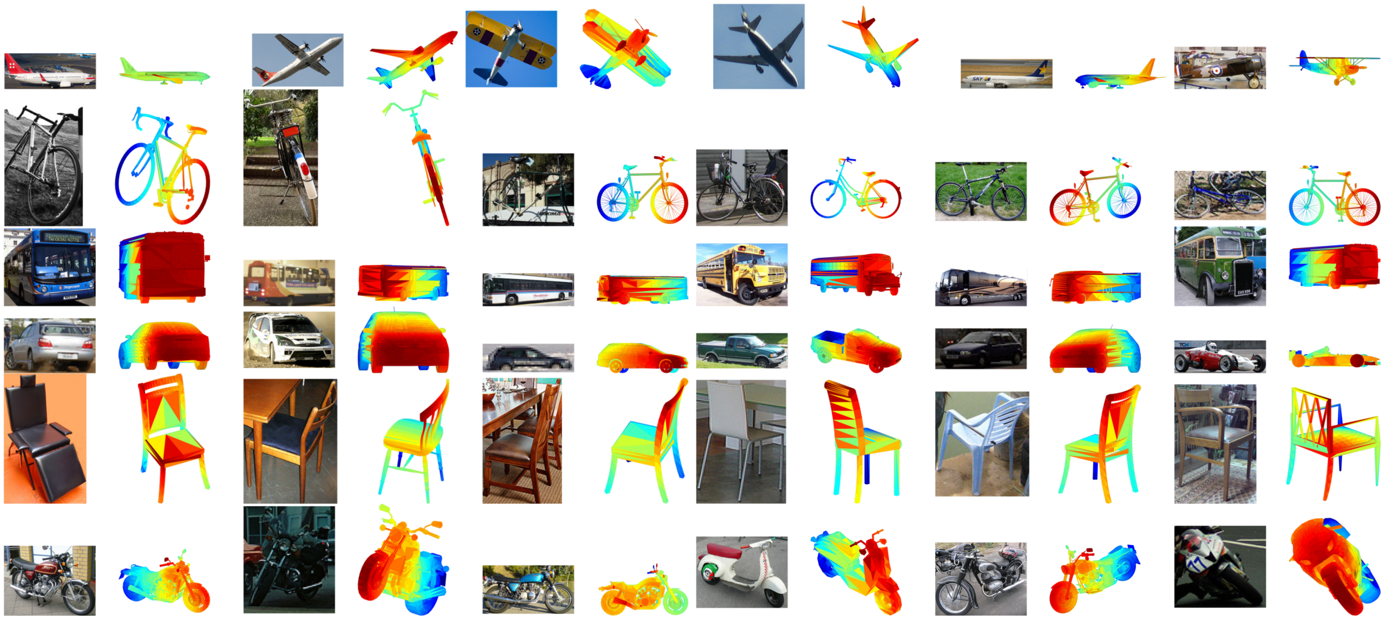

Fig. 6: Viewpoint predictions for unoccluded groundtruth instances using our algorithm. The columns show 15th, 30th, 45th, 60th, 75th and 90th percentile instances respectively in terms of the error. We visualize the predictions by rendering a 3D model using our predicted viewpoint.

Results. We report the performance of our model learning approach in Table 1. Here, 'SIRFS' denotes a state-of-the art intrinsic image decomposition method and 'Puffball'l [63] denotes a shape-inflation method for reconstruction. 'Carvi' denotes the recent method by Vicente *et al.* [23] which is specifically designed for the task of reconstructing an annotated image collection as their visual hull based reconstruction technique makes strong assumptions regarding the accuracy of the object mask and predicted viewpoint.

We observe that category-agnostic methods – Puffball [63] and SIRFS [18], [53] – consistently perform worse on the benchmark by themselves as they use generic priors to reconstruct each image individually and cannot reason over the image collection jointly. Our model learning performs comparably to the specialized approach of Vicente *et al.*– we demonstrate competitive, if not better, performance on both benchmarks with our models showing greater robustnes to perspective foreshortening effects on "trains" and "buses". Certain classes like "boat" and "sofa" are especially hard because of large intra-class variance and data sparsity respectively.

5.2 Accuracy of Viewpoint Estimation

An important component of the proposed reconstruction framework is the viewpoint estimation system Section 3 which allows us to fit learned models to objects in new images. We evaluate this component under two settings – viewpoint prediction accuracy when the object localization is known and a detection setting with unkown localization. We observe that our proposed approach significantly improves the state-of-the-art for viewpoint estimation in both these settings.

Dataset. Xiang *et al.* [57] provide annotations for (ϕ, φ, ψ) corresponding to all the instances in the PASCAL VOC 2012 detection train, validation set as well as for ImageNet images. We use the PASCAL train set and the ImageNet annotations to train the CNN described in Section 3 and use the PASCAL VOC 2012 validation set annotations to evaluate our performance.

Viewpoint Estimation with Ground Truth box. To analyze the performance of our viewpoint estimation method independent of factors like mis-localization, we first tackle the task of estimating the viewpoint of an object with known bounds. Let $\Delta(R_1, R_2) = \frac{\|log(R_1^T R_2)\|_F}{\sqrt{2}}$ denote the geodesic distance function over the manifold of rotation matrices. $\Delta(R_{gt}, R_{pred})$ captures the difference between ground truth viewpoint R_{gt} and predicted viewpoint R_{pred}. We use two complementary metrics for evaluation -

- **Median Error :** The common confusions for the task of viewpoint estimation often are predictions which are far apart (eg. left facing vs right facing car) and the median error ($MedErr$) is a widely use metric that is robust to these if a significant fraction of the estimates are accurate.
- **Accuracy at θ :** A small median error does not necessarily imply accurate estimates for all instances, a complementary performance measure is the fraction of instances whose predicted viewpoint is within a fixed threshold of the target viewpoint. We denote this metric by Acc_θ where θ is the threshold. We use $\theta = \frac{\pi}{6}$.

Recently, Ghodrati *et al.* [59] achieved results comparable to state-of-the art by using a linear classifier over layer 5 features of TNet. We denote this method as 'Pool5-TNet' and implement it as a baseline. To study the effect of end-to-end training of the CNN architecture, we use a linear classifier on top of the fc7 layer of TNet as another baseline (denoted

	aero	bike	boat	bottle	bus	car	chair	table	mbike	sofa	train	tv	mean
$Acc_{\frac{\pi}{6}}$ (Pool5-TNet)	0.27	0.18	0.36	0.81	0.71	0.36	0.52	0.52	0.38	0.67	0.70	0.71	0.52
$Acc_{\frac{\pi}{6}}$ (fc7-TNet)	0.50	0.44	0.39	0.88	0.81	0.70	0.39	0.38	0.48	0.44	0.78	0.65	0.57
$Acc_{\frac{\pi}{6}}$ (ours-TNet)	0.78	0.74	0.49	**0.93**	0.94	**0.90**	0.65	**0.67**	0.83	0.67	0.79	0.76	0.76
$Acc_{\frac{\pi}{6}}$ (ours-ONet)	**0.81**	**0.77**	**0.59**	**0.93**	**0.98**	0.89	**0.80**	0.62	**0.88**	**0.82**	**0.80**	**0.80**	**0.81**
$MedErr$ (Pool5-TNet)	42.6	52.3	46.3	18.5	17.5	45.6	28.6	27.7	37.0	25.9	20.6	21.5	32.0
$MedErr$(fc7-TNet)	29.8	40.3	49.5	13.5	7.6	13.6	45.5	38.7	31.4	38.5	9.9	22.6	28.4
$MedErr$(ours-TNet)	14.7	18.6	31.2	13.5	6.3	**8.8**	17.7	17.4	17.6	15.1	8.9	17.8	15.6
$MedErr$(ours-ONet)	**13.8**	**17.7**	**21.3**	**12.9**	**5.8**	9.1	**14.8**	**15.2**	**14.7**	**13.7**	**8.7**	**15.4**	**13.6**

TABLE 2: Viewpoint Estimation with Ground Truth box

as 'fc7-TNet'). With the aim of analyzing viewpoint estimation independently, the evaluations were restricted only to objects marked as non-occluded and non-truncated. The performance of our method and comparisons to the baseline are shown in Table 2. The results clearly demonstrate that end-to-end training improves results and that our method with the TNet architecture performs significantly better than the 'Pool5-TNet' method used in [59]. We also observe a significant improvement by using the ONet architecture and only use this architecture for further experiments/analysis. In Figure 6, we show our predictions sorted in terms of the error and it can be seen that the predictions for most categories are reliable even at the 90th percentile.

	AVP				$AVP_{\frac{\pi}{6}}$	$ARP_{\frac{\pi}{6}}$
Number of bins	**4**	**8**	**16**	**24**	**-**	**-**
Xiang *et al.* [57]	19.5	18.7	15.6	12.1	-	-
Pepik *et al.* [58]	23.8	21.5	17.3	13.6	-	-
Ghodrati *et al.* [59]	24.1	22.3	17.3	13.7	-	-
ours	**49.1**	**44.5**	**36.0**	**31.1**	50.7	46.5

TABLE 3: Mean performance of our approach for various metrics. The detailed results for individual classes can be found at the PASCAL3D leaderboard (http://cvgl.stanford.edu/projects/pascal3d.html).

Viewpoint Estimation with Detection. Xiang *et al.* [57] introduced the AVP metric to measure advances in the task of viewpoint estimation in the setting where localizations are not known a priori. The metric is similar to the AP criterion used for PASCAL VOC detection except that each detection candidate has an associated viewpoint and the detection is labeled correct if it has a correct predicted viewpoint bin as well as a correct localization (bounding box IoU $>$ 0.5). Xiang *et al.* [57] also compared to Pepik *et al.* [58] on the AVP metric using various viewpoint bin sizes and Ghodrati *et al.* [59] also showed comparable results on the metric. To evaluate our method, we obtain detections from RCNN [25] using MCG [67] object proposals and augment them with a pose predicted using the corresponding detection's bounding box.

We note that there are two issues with the AVP metric - it only evaluates the prediction for the azimuth (ϕ) angle and discretizes viewpoint instead of treating it continuously. Therefore, we also introduce two additional evaluation metrics which follow the IoU $>$ 0.5 criteria for localization but modify the criteria for assigning a viewpoint prediction to be correct as follows -

- $AVP_\theta : \delta(\phi_{gt}, \phi_{pred}) < \theta$
- $ARP_\theta : \Delta(R_{gt}, R_{pred}) < \theta$

Note that ARP_θ requires the prediction of all euler angles instead of just ϕ and therefore, is a stricter metric.

The performance of our CNN based approach for viewpoint prediction in the detection setting is shown in Table 3 and it can be seen that we significantly outperform the state-of-the-art methods across all categories. While it is not possible to compare our pose estimation performance independent of detection with DPM based methods like [57], [58], an indirect comparison results from the analysis using ground truth boxes where we demonstrate that our pose estimation approach is an improvement over [59] which in turn performs similar to [57], [58] while using similar detectors.

5.3 Sensitivity Analysis for Recognition based Reconstruction

Our primary goal is to reconstruct objects in an image automatically. Towards this goal, we study the performance of our system when relaxing the availability of various expensive annotations of the form of keypoint correspondences or instance segmentations.

Dataset and Metrics. The reconstruction error metrics for measuring mesh and depth error are the same as described previously (Section 5.1). The segmentation, keypoint annotations for learning and the mesh annotations for evaluation are also similarly obtained. However, for the sensitivity analysis, we introduce a train/test split since the recognition components used for instance segmentation and viewpoint estimation are trained on the PASCAL VOC train set. We therefore train our category-shape models on only the subset of the data corresponding to PASCAL VOC train set. We then reconstruct the held out objects in the PASCAL validation set and report performance for these test objects.

Results. In order to analyze sensitivity of our models to noisy inputs we reconstructed held-out test in-

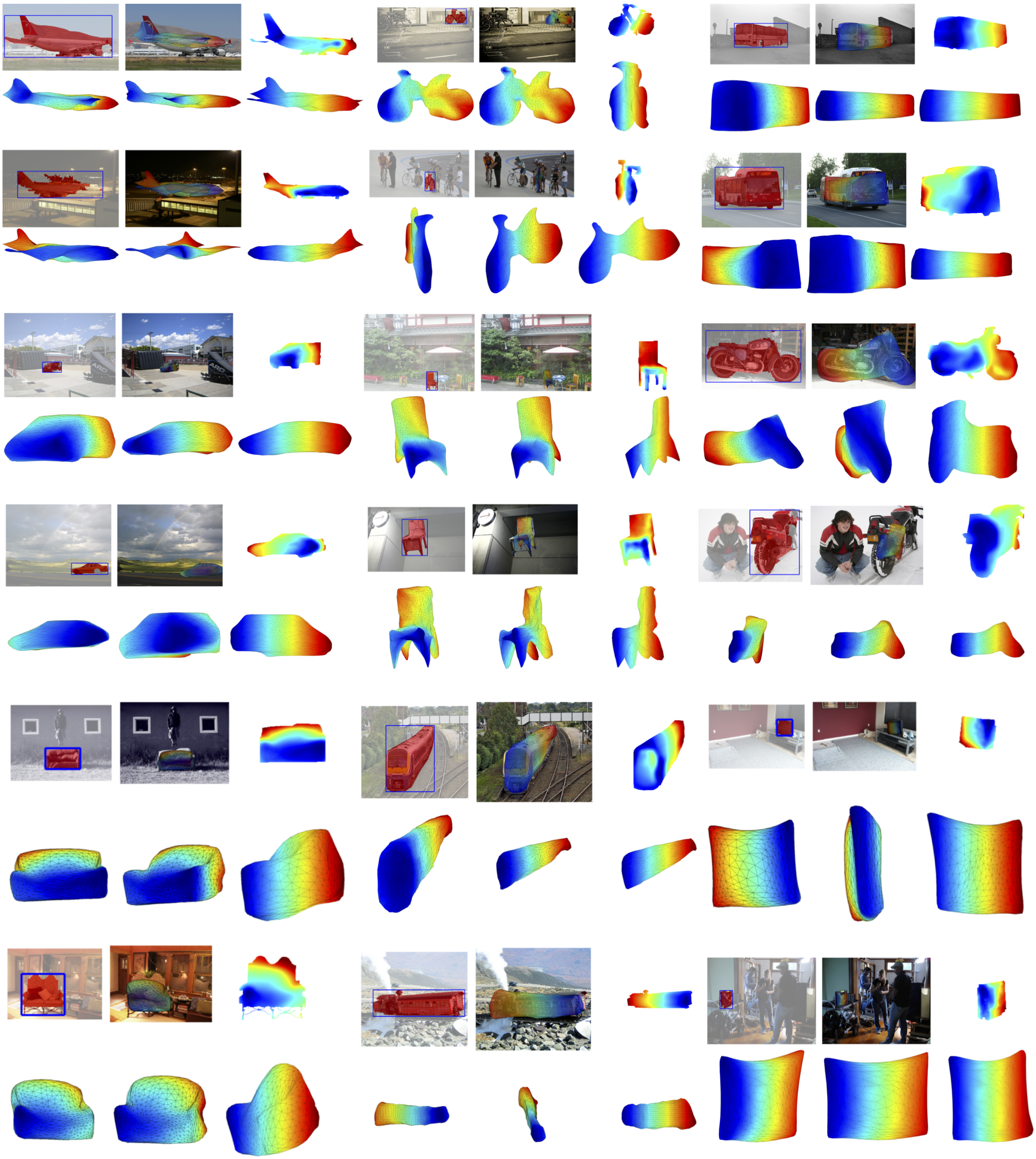

Fig. 7: Fully automatic reconstructions on detected instances (0.5 IoU with ground truth) using our models on rigid categories in PASCAL VOC. We show our instance segmentation input, the inferred shape overlaid on the image, a 2.5D depth map (after the bottom-up refinement stage), the mesh in the image 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 coordinate frame (blue is closer). More results can be found at https://goo.gl/MgVQzZ.

		aero	bike	boat	bus	car	chair	mbike	sofa	train	tv	mean
Mesh	KP+Mask	1.77	1.85	3.68	1.90	1.80	2.26	1.83	6.86	2.69	3.40	2.80
	KP+SDS	1.75	1.89	3.71	1.87	1.75	2.27	1.84	6.56	2.76	3.39	2.78
	PP+SDS	1.84	2.02	4.59	1.86	1.88	2.41	2.01	7.30	2.74	3.27	2.99
	Puff [63]	3.31	2.49	2.95	3.40	2.87	3.09	2.65	2.73	3.91	3.33	3.07
Depth	KP+Mask	9.83	9.95	21.07	12.80	10.07	9.10	9.98	29.39	25.70	9.85	14.77
	KP+SDS	9.95	10.35	20.11	13.06	10.49	9.24	10.61	27.94	26.13	10.10	14.80
	PP+SDS	11.42	11.25	21.93	22.04	13.69	10.27	11.71	26.76	34.92	9.88	17.39
	SIRFS [18]	13.58	14.48	19.64	30.14	22.60	20.12	16.81	21.54	41.40	23.67	22.40

TABLE 4: 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.

stances 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 described in Section 3. 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 [26].

Table 4 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 4 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.

5.4 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 [26] in Figure 7. 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.

6 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). Another important component of our framework is a convolutional neural network based viewpoint prediction system which we have shown to be considerably more accurate than previous approaches – in context of objects with known localization as well as automatically detected objects. We report an extensive evaluation of the quality of the learned 3D models on a recent 3D benchmarking dataset for PASCAL VOC [57] showing competitive results with models that specialize in shape reconstruction using ground truth annotations as 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 [68] may prove helpful, as well as a tighter integration between segmentation and reconstruction.

Acknowledgement

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] C. Nandakumar, A. Torralba, and J. Malik, “How little do we need for 3-d shape perception?” *Perception-London*, 2011. 1

[2] R. Nevatia and T. O. Binford, “Description and recognition of curved objects,” *Artificial Intelligence*, 1977. 1

[3] A. Gupta, A. A. Efros, and M. Hebert, “Blocks world revisited: Image understanding using qualitative geometry and mechanics,” in *European Conference on Computer Vision*, 2010. 1

[4] J. Xiao, B. Russell, and A. Torralba, "Localizing 3d cuboids in single-view images," in *Advances in Neural Information Processing Systems*, 2012. 1

[5] L. G. Roberts, "Machine perception of three-dimensional solids," Ph.D. dissertation, Massachusetts Institute of Technology, 1963. 1

[6] J. J. Lim, H. Pirsiavash, and A. Torralba, "Parsing ikea objects: Fine pose estimation," in *IEEE International Conference on Computer Vision*, 2013. 1

[7] S. Satkin, M. Rashid, J. Lin, and M. Hebert, "3dnn: 3d nearest neighbor," *International Journal of Computer Vision*, 2014. 1

[8] B. Pepik, M. Stark, P. Gehler, T. Ritschel, and B. Schiele, "3d object class detection in the wild," in *Workshop on 3D from a Single Image (3DSI) (in conjunction with CVPR'15)*, 2015. 1

[9] H. Su, Q. Huang, N. J. Mitra, Y. Li, and L. Guibas, "Estimating image depth using shape collections," *ACM Transactions on Graphics (TOG)*, vol. 33, 2014. 1

[10] Q. Huang, H. Wang, and V. Koltun, "Single-view reconstruction via joint analysis of image and shape collections," *ACM Transactions on Graphics (TOG)*, vol. 34, 2015. 1

[11] M. Sung, V. G. Kim, R. Angst, and L. Guibas, "Data-driven structural priors for shape completion," *ACM Transactions on Graphics (TOG)*, 2015. 1

[12] D. Anguelov, P. Srinivasan, D. Koller, S. Thrun, J. Rodgers, and J. Davis, "Scape: Shape completion and animation of people," in *ACM Transactions on Graphics (TOG)*, 2005. 1

[13] V. Blanz and T. Vetter, "A morphable model for the synthesis of 3d faces," in *Computer Graphics and Interactive Techniques*, 1999. 1

[14] M. Z. Zia, M. Stark, B. Schiele, and K. Schindler, "Detailed 3d representations for object recognition and modeling," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2013. 1

[15] D. Hoiem, A. A. Efros, and M. Hebert, "Automatic photo pop-up," *ACM Transactions on Graphics (TOG)*, 2005. 1

[16] A. Saxena, M. Sun, and A. Y. Ng, "Make3d: Learning 3d scene structure from a single still image," *PAMI*, 2009. 1

[17] K. Karsch, Z. Liao, J. Rock, J. T. Barron, and D. Hoiem, "Boundary cues for 3d object shape recovery," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013. 2

[18] J. T. Barron and J. Malik, "Shape, illumination, and reflectance from shading," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2015. 2, 6, 7, 8, 11

[19] B. Horn, "Shape from shading: A method for obtaining the shape of a smooth opaque object from one view," PhD thesis, Massachusetts Inst. of Technology, 1970. 2

[20] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," in *Advances in Neural Information Processing Systems*, 2014. 2

[21] D. Eigen and R. Fergus, "Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture," in *IEEE International Conference on Computer Vision*, 2015. 2

[22] X. Wang, D. Fouhey, and A. Gupta, "Designing deep networks for surface normal estimation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015. 2

[23] S. Vicente, J. Carreira, L. Agapito, and J. Batista, "Reconstructing pascal voc," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014. 2, 3, 7, 8

[24] T. Cashman and A. Fitzgibbon, "What shape are dolphins? building 3d morphable models from 2d images," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2013. 2, 3

[25] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014. 2, 5, 6, 9

[26] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik, "Simultaneous detection and segmentation," in *European Conference on Computer Vision*, 2014. 2, 6, 11

[27] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik, "Hypercolumns for object segmentation and fine-grained localization," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015. 2

[28] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International Journal of Computer Vision*, 2010. 2, 7

[29] D. P. Huttenlocher and S. Ullman, "Recognizing solid objects by alignment with an image," *International Journal of Computer Vision*, 1990. 2, 5

[30] F. Rothganger, S. Lazebnik, C. Schmid, and J. Ponce, "3d object modeling and recognition using local affine-invariant image descriptors and multi-view spatial constraints," *International Journal of Computer Vision*, pp. 231–259, 2006. 2

[31] I. Gordon and D. G. Lowe, "What and where: 3d object recognition with accurate pose," in *Toward category-level object recognition*. Springer, 2006, pp. 67–82. 2

[32] S. Savarese and L. Fei-Fei, "View synthesis for recognizing unseen poses of object classes," in *European Conference on Computer Vision*. Springer, 2008. 2

[33] J. Xiao, J. Chen, D.-Y. Yeung, and L. Quan, "Structuring visual words in 3d for arbitrary-view object localization," in *European Conference on Computer Vision*. Springer, 2008. 2

[34] C. Gu and X. Ren, "Discriminative mixture-of-templates for viewpoint classification," in *European Conference on Computer Vision*. Springer, 2010. 2

[35] M. Ozuysal, V. Lepetit, and P. Fua, "Pose estimation for category specific multiview object localization," in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2009. 2

[36] K. Fukushima, "Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position," *Biological Cybernetics*, pp. 193–202, 1980. 2

[37] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation applied to handwritten zip code recognition," *Neural Comput.*, pp. 541–551, 1989. 2

[38] A. Kar, S. Tulsiani, J. Carreira, and J. Malik, "Category-specific object reconstruction from a single image," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015. 2, 4

[39] S. Tulsiani and J. Malik, "Viewpoints and keypoints," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015. 2

[40] C. Bregler, A. Hertzmann, and H. Biermann, "Recovering non-rigid 3d shape from image streams," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2000. 2, 3

[41] A. Bartoli, V. Gay-Bellile, U. Castellani, J. Peyras, S. Olsen, and P. Sayd, "Coarse-to-fine low-rank structure-from-motion," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2008. 3

[42] L. Torresani, A. Hertzmann, and C. Bregler., "Non-rigid structure-from-motion: Estimating shape and motion with hierarchical priors." *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2008. 3

[43] R. Garg, A. Roussos, and L. Agapito, "Dense variational reconstruction of non-rigid surfaces from monocular video," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013. 3

[44] S. Zhu, L. Zhang, and B. Smith, "Model evolution: An incremental approach to non-rigid structure from motion." in *IEEE Conference on Computer Vision and Pattern Recognition*, 2010. 3

[45] M. Prasad, A. Fitzgibbon, A. Zisserman, and L. Van Gool, "Finding nemo: Deformable object class modelling using curve matching," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2010. 3

[46] M. Hejrati and D. Ramanan, "Analyzing 3d objects in cluttered images," in *Advances in Neural Information Processing Systems*, 2012. 3

[47] S. Vicente and L. de Agapito, "Balloon shapes: Reconstructing and deforming objects with volume from images," in *3DV*, 2013. 3

[48] A. Laurentini, "The visual hull concept for silhouette-based image understanding," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 1994. 3

[49] C. H. Esteban and F. Schmitt, "Silhouette and stereo fusion for 3d object modeling," *Computer Vision and Image Understanding*, 2004. 3, 5

[50] Y. Sahilliolu and Y. Yemez, "A surface deformation framework for 3d shape recovery," in *Multimedia Content Representation, Classification and Security*, 2006. 3

[51] Y. Chen, T.-K. Kim, and R. Cipolla, "Inferring 3d shapes and deformations from single views," in *European Conference on Computer Vision*, 2010. 3

[52] V. Blanz and T. Vetter, "Face recognition based on fitting a 3d morphable model," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2003. 4

[53] J. T. Barron and J. Malik, "Color constancy, intrinsic images, and shape estimation," *European Conference on Computer Vision*, 2012. 5, 6, 7, 8

[54] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2010. 5

[55] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in *Advances in Neural Information Processing Systems*, 2012. 5, 6

[56] R. Girshick, F. Iandola, T. Darrell, and J. Malik, “Deformable part models are convolutional neural networks,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015. 5

[57] Y. Xiang, R. Mottaghi, and S. Savarese, “Beyond pascal: A benchmark for 3d object detection in the wild,” in *WACV*, 2014. 5, 7, 8, 9, 11

[58] B. Pepik, M. Stark, P. Gehler, and B. Schiele, “Teaching 3d geometry to deformable part models,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012. 5, 9

[59] A. Ghodrati, M. Pedersoli, and T. Tuytelaars, “Is 2d information enough for viewpoint estimation?” in *BMVC*, 2014. 5, 8, 9

[60] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009. 6

[61] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *CoRR*, vol. abs/1409.1556, 2014. 6

[62] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, “Caffe: Convolutional architecture for fast feature embedding,” *arXiv preprint arXiv:1408.5093*, 2014. 6

[63] 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 Symposium on Applied Perception*, 2012. 7, 8, 11

[64] L. Bourdev, S. Maji, T. Brox, and J. Malik, “Detecting people using mutually consistent poselet activations,” in *European Conference on Computer Vision*, 2010. 7

[65] B. Hariharan, P. Arbelaez, L. Bourdev, S. Maji, and J. Malik, “Semantic contours from inverse detectors,” in *IEEE International Conference on Computer Vision*, 2011. 7

[66] N. Aspert, D. Santa-Cruz, and T. Ebrahimi, “Mesh: Measuring errors between surfaces using the hausdorff distance,” in *ICME*, 2002. 7

[67] P. Arbeláez, J. Pont-Tuset, J. Barron, F. Marques, and J. Malik, “Multiscale combinatorial grouping,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014. 9

[68] 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 *IEEE Conference on Computer Vision and Pattern Recognition*, 2015. 11

Shubham Tulsiani received the B.Tech degree in Computer Science and Engineering from Indian Institute of Technology, Kanpur in 2013. He is currently a graduate student with the EECS Department, University of California, Berkeley. His research interests lie at the intersection of recognition, pose estimation and reconstruction from a single image. He is a recipient of the Berkeley Graduate Fellowship, CVPR Best Student Paper Award and International Physics Olympiad Gold medal.

Abhishek Kar is a fourth year PhD student at UC Berkeley advised by Prof. Jitendra Malik. He obtained his B.Tech degree in Computer Science and Engineering from Indian Institute of Technology, Kanpur in 2012 where he worked on problems in human computer interaction and intelligent tutoring systems and received the Honda Young Engineer and Scientist (YES) Award. His primary research interest lies in 3D reconstruction of objects and scenes in the wild by leveraging advances in object/scene recognition. He is a recipient of the CVPR Best Student Paper Award and the Outstanding GSI Award at UC Berkeley.

João Carreira received his doctorate from the University of Bonn, Germany. His thesis focused on sampling class-independent object segmentation proposals using the CPMC algorithm, and on applying them in object recognition and localization. Systems authored by him and colleagues were winners of all four PASCAL VOC Segmentation challenges, 2009-2012. He did post-doctoral work at the Institute of Systems and Robotics in Coimbra, Portugal and is currently with the EECS department, at the University of California in Berkeley, USA. His research interests lie at the intersection of recognition, segmentation, pose estimation and shape reconstruction of objects from a single image.

Jitendra Malik is Arthur J. Chick Professor in the Department of Electrical Engineering and Computer Science at the University of California at Berkeley, where he also holds appointments in vision science, cognitive science and Bioengineering. He received the PhD degree in Computer Science from Stanford University in 1985 following which he joined UC Berkeley as a faculty member. He served as Chair of the Computer Science Division during 2002-2006, and of the Department of EECS during 2004-2006. Jitendra's group has worked on computer vision, computational modeling of biological vision, computer graphics and machine learning. Several well-known concepts and algorithms arose in this work, such as anisotropic diffusion, normalized cuts, high dynamic range imaging and shape contexts. He was awarded the Longuet-Higgins Award for A Contribution that has Stood the Test of Time twice, in 2007 and 2008, received the PAMI Distinguished Researcher Award in computer vision in 2013 and the K.S. Fu prize in 2014. Jitendra Malik is a Fellow of the IEEE, ACM, and the American Academy of Arts and Sciences, and a member of the National Academy of Sciences and the National Academy of Engineering.