Dense Object Reconstruction with Semantic Priors

Sid Yingze Bao[†] Manmohan Chandraker[*] Yuanqing Lin[*] Silvio Savarese[†]

[*]NEC Labs America, Cupertino, CA, USA
[†]University of Michigan at Ann Arbor, MI, USA

Abstract

We present a dense reconstruction approach that overcomes the drawbacks of traditional multiview stereo by incorporating semantic information in the form of learned category-level shape priors and object detection. Given training data comprised of 3D scans and images of objects from various viewpoints, we learn a prior comprised of a mean shape and a set of weighted anchor points. The former captures the commonality of shapes across the category, while the latter encodes similarities between instances in the form of appearance and spatial consistency. We propose robust algorithms to match anchor points across instances that enable learning a mean shape for the category, even with large shape variations across instances. We model the shape of an object instance as a warped version of the category mean, along with instance-specific details. Given multiple images of an unseen instance, we collate information from 2D object detectors to align the structure from motion point cloud with the mean shape, which is subsequently warped and refined to approach the actual shape. Extensive experiments demonstrate that our model is general enough to learn semantic priors for different object categories, yet powerful enough to reconstruct individual shapes with large variations. Qualitative and quantitative evaluations show that our framework can produce more accurate reconstructions than alternative state-of-the-art multiview stereo systems.

1. Introduction

Recent years have seen rapid strides in dense 3D shape recovery, with multiview stereo (MVS) systems capable of reconstructing entire monuments [14, 17]. Despite this progress, MVS has remained largely applicable only in favorable imaging conditions. Lack of texture leads to extended troughs in photoconsistency-based cost functions, while specularities violate inherent Lambertian assumptions. Diffuse photoconsistency is not a reliable metric with wide baselines in scenarios with few images, leading to sparse, noisy MVS outputs. Under these circumstances, MVS reconstructions often display holes or artifacts (see Figure 1 dashed box).

On the other hand, there have been crucial developments in two seemingly disjoint areas of computer vision. With the advent of cheap commerical scanners and depth sensors, it is now possible to easily acquire 3D shapes. Concurrently, the performance of modern object detection algorithms [9, 11, 22, 32] has rapidly improved to allow inference of reliable bounding boxes in the presence of clutter, especially when information is shared across multiple views. This paper presents a framework for dense 3D reconstruction that overcomes the drawbacks of traditional MVS by leveraging semantic information in the form of object detection and shape priors learned from a database of training images and 3D shapes.

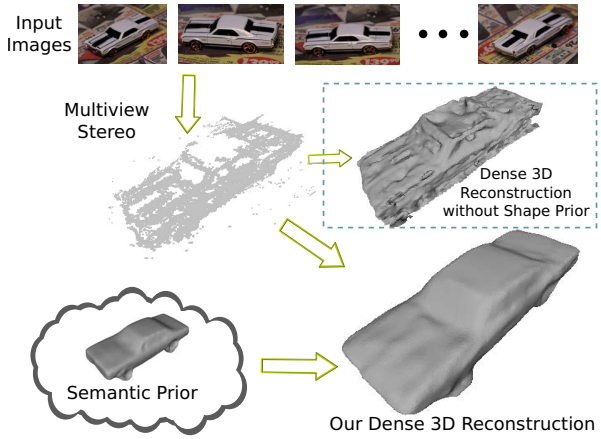


Figure 1. Traditional multiview stereo faces challenges due to lack of texture, wide baselines or specularities. We propose a framework for semantic dense reconstruction that learns a category-level shape prior, which is used with weighted warping and refinement mechanisms to reconstruct regularized, high-quality 3D shapes.

The aforementioned drawbacks of MVS have been widely recognized and several prior works share our philosophy of augmenting reconstruction with prior knowledge. For instance, Furukawa et al. [13] reconstruct indoor environments by incorporating Manhattan priors, Gallup et al. [16] recover urban façades with a piecewise planar assumption and Wu et al. [31] recover building models with a prior derived from architectural schematic curves. All the above approaches use application-specific information to provide the shape priors.

In contrast, our priors are far more general – they are category-level and learned from training data. An overview of our reconstruction framework is shown in Figure 2. We postulate in Section 3 that while object instances within a category might have very different shapes and appearances, they share certain similarities at a semantic level. For example, both sedans and sports cars have bonnets and wheels. We model semantic similarity as a shape prior, which consists of a set of automatically learned anchor points across several instances, along with a learned mean shape that captures the

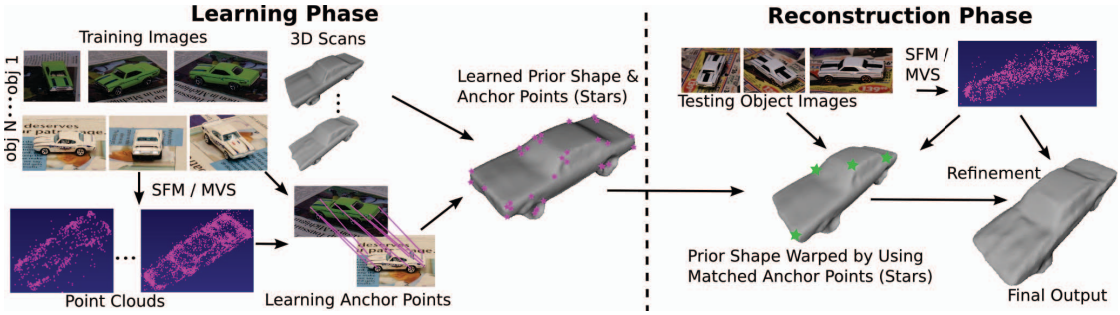


Figure 2. Outline of our semantic dense reconstruction framework. Please see Section 1 for an overview.

shared commonality of the entire category. Our experiments demonstrate that this novel representation can successfully achieve the balance between capturing semantic similarities and shape variation across instances.

In the learning phase (Section 4), the anchor points encode attributes such as frequency, appearance and location similarity of features across instances. The associated weights aid in discarding spurious texture matches, while determining a weighted regularization for both mean shape learning and reconstruction. Based on matched anchor points, the shape prior for a category is determined by a series of weighted thin-plate spline (TPS) warps over the scans of training objects.

Our reconstruction phase (Section 5) starts with a point cloud obtained by applying a structure-from-motion (SFM) or MVS system to images of an unseen instance (with a shape different from training objects). Bounding boxes from object detection in individual images are collated using the SFM camera poses and used to localize and orient the object in the point cloud. This guides the process of matching anchor points – shown by green stars in right panel in Figure 2 – between the learned prior and the test object's SFM point cloud, followed by a warping of the prior shape in order to closely resemble the true shape. Finer details not captured by the shape prior may be recovered by a refinement step, using guidance from SFM or MVS output. The refinement combines confidence scores from anchor points and photoconsistency in order to produce a regularized, high quality output shape. Not only are our reconstructions visually pleasant, they are also quantitatively closer to the ground truth than other baselines (Section 6).

2. Relation to Prior Work

Our comprehensive reconstruction pipeline relates to several areas of computer vision, as briefly explored in this section.

Multiview Stereo. This paper provides a framework to augment traditional multiview stereo (MVS) reconstruction methods with semantic information. Broadly, MVS approaches in computer vision may be categorized as patch-growing, depth-map based and volumetric methods. The former uses a locally planar patch model to perform a succession of expansion steps to maximize photoconsistency and filtering steps to remove inaccurate patches [15]. Depth map-based methods seek a labeling from the space of pixels to a set of discrete depth labels [21]. Volumetric methods, on the other hand, seek a binary partitioning of 3D space into object and non-object [18, 30]. We choose the patch-based system [15] for demonstration, but our framework can be generalized to other approaches too.

Example-Based Reconstruction. A set of example shapes is used by active shape models (ASM) to encode patterns of variability, thereby ensuring a fitted shape consistent with deformations observed in training [8]. However, it requires heavy manual annotation and only models linear variations. While reasonable in 2D, it is arguably not well-suited for the far higher shape and appearance variations in general 3D scenes. Subsequent works on statistical shape analysis [10] allow non-rigid TPS warps between shapes [5], but often require landmark identification and initial rigid alignment based on point distributions, which is not feasible for general scenes [24]. We use semantic information, namely object detection for localization and anchor point matching, to overcome those drawbacks. Learned anchor points yield confidence scores, which guide our deformation process through a weighted TPS [26].

Morphable models in 3D demonstrate realistic shape recovery, but are limited to categories like faces with low shape variation that can be accurately modeled with a linear PCA basis [4]. Pauly et al. propose a framework for example-based 3D scan completion, but require dense 3D scans [25]. By exploiting semantics in the form of object detection and anchor point matching, we handle both greater shape variation and noisy, incomplete, image-based MVS inputs.

Shape Matching. Determining correspondence across instances with varying shape is a key step in shape matching. Belongie et al. pose correspondence search as a bipartite matching problem with shape context descriptors [2], Berg at al. find points with similar geometric blur descriptors by solving an integer quadratic program[3], while Chui and Rangarajan's TPS-RPM determines matches with a soft assign [6]. A 3D CAD model is aligned to images in [23], but the model and features are manually defined. The demands on correspondences for 3D reconstruction are far higher than 2D shape matching – competing factors like high localization accuracy, stringent outlier rejection and good density are all crucial to obtaining a high quality dense reconstruction. Algorithms 1 and 2 are designed to robustly meet these challenges.

Object Detection and 3D Information. The mutual benefit of combining object detection and SFM is demonstrated in [1]. The flexibility of implicit shape-based detection frameworks

[22] is used to transfer depth annotations from training images to test objects in [29, 28]. TPS-RPM is combined with Hough voting to localize object boundaries in [12]. Object recognition is improved in [19] by computing deformation priors directly in transformation space. However, the complexity of 3D shapes and the accuracy demands of 3D reconstruction necessitate far greater control over the deformation process, so we consider it advantageous to compute priors in the mesh space.

3. Our Model

We assume that for each object category, there exists a prior that consists of a 3D *mean shape* $\mathbf{S}^*$ that captures the commonality of shapes across all instances and a set of *anchor points* $\mathbf{A}$ that captures similarities between subsets of instances. The shape of any particular object $\mathbf{S}^i$ is a transformation of $\mathbf{S}^*$, plus specific details Δ^i not shared by other instances:

$$\mathbf{S}^i = T(\{\mathbf{S}^*, \mathbf{A}\}, \theta^i) + \Delta^i, \tag{1}$$

where T is a warping (transformation) function and θ^i is the warping parameter that is unique to each object instance. In the following, we briefly explain the various aspects of our model.

Anchor points. The key to reconstructing an object instance is to estimate the warping parameters θ^i. We leverage on certain reliable features associated with the shape prior, which we call anchor points. Anchor points form the backbone of our framework, since they are representative of object shape and the relative importance of different object structures. Anchor points with high weights, ω, are considered stable in terms of location and appearance, and thus, more representative of object shape across instances. They guide the learning of the mean shape for a category, as well as the deformation processes during actual 3D reconstruction. In Section 4.1, we detail the mechanism of learning anchor points from training data.

Warping function. We assume that the functional form of T is known. In particular, prior work on shape matching [2, 19] has demonstrated inspiring results using regularized thin-plate spline (TPS) transformations [5] to capture deformations. Let $\{\mathbf{x}_i\}$ and $\{\mathbf{x'}_i\}$, $i = 1, \cdots, n$, be two sets of anchor points for object instances O and O'. The TPS mapping T is given by

$$T(\mathbf{x}, \{\boldsymbol{\alpha}, \boldsymbol{\beta}\}) = \sum_{j=0}^{3} \alpha_j \phi_j(\mathbf{x}) + \sum_{i=1}^{n} \beta_i U(\mathbf{x}, \mathbf{x}_i) \tag{2}$$

where $\phi_0(\mathbf{x}) = 1$, $\phi_j(\mathbf{x}) = x_j$ and $U(\mathbf{x}, \mathbf{x}_i) = \|\mathbf{x} - \mathbf{x}_i\|$. Note that our TPS representation is in 3D, instead of the more common 2D representation in traditional shape matching. The solution for the parameters $\theta = \{\boldsymbol{\alpha}, \boldsymbol{\beta}\}$ in a regularized framework is given by the system of equations:

$$(\mathbf{K} + n\lambda\mathbf{I})\boldsymbol{\beta} + \boldsymbol{\Phi}\boldsymbol{\alpha} = \mathbf{x}', \quad \boldsymbol{\Phi}^\top\boldsymbol{\beta} = \mathbf{0}, \tag{3}$$

where $K_{ij} = U(\mathbf{x}_i, \mathbf{x}_j)$, $\boldsymbol{\Phi}_{ij} = \phi_j(\mathbf{x}_i)$ and λ is a regularization parameter. Regularized TPS yields a solution that interpolates between two point sets and is sufficiently smooth. However, greater control is required for 3D reconstruction applications, since the extent of deformations must be determined by the local level of detail. Semantic information of this nature is determined automatically in our framework by the anchor point learning mechanism. To incorporate semantic information from anchor points, in the form of a weight matrix $\mathbf{W} = \text{diag}(\omega_1, \cdots, \omega_n)$, we use an extension of TPS [26]:

$$(\mathbf{K} + n\lambda\mathbf{W}^{-1})\boldsymbol{\beta} + \boldsymbol{\Phi}\boldsymbol{\alpha} = \mathbf{x}', \quad \boldsymbol{\Phi}^\top\boldsymbol{\beta} = \mathbf{0}, \tag{4}$$

which is again solvable analytically like regularized TPS.

Unique Details. Details specific to each object that are not captured in the shape prior are recovered by a refinement step. This refinement is used in both mean shape learning and during reconstruction of a particular test object.

To refine a shape $\mathbf{S}^i$ (a mesh) towards shape $\mathbf{S}^j$, we compute displacements for vertices in $\mathbf{S}^i$. For a vertex $\mathbf{p}^i_k$ in $\mathbf{S}^i$, we estimate the surface normal $\mathbf{n}^i_k$ by a local tangent space computation. The vertex $\mathbf{p}^i_k$ is matched to $\mathbf{p}^j_k$ in $\mathbf{S}^j$ if $\|\mathbf{p}^j_k - \mathbf{p}^i_k\| < \tau_1$ and $|(\mathbf{p}^j_k - \mathbf{p}^i_k)^\top \mathbf{n}^i_k| < 1 - \tau_2$, where τ_1, τ_2 are predefined thresholds. Let $\mathcal{P}^i$ be the set of vertices in $\mathbf{S}^i$ that can be matched as above to the set $\mathcal{P}^j$ in $\mathbf{S}^j$ and $\mathcal{N}^i_k$ be the set of 1-nearest neighbors of $\mathbf{p}^i_k$ in $\mathcal{P}^i$. Then, the set of displacements, $\Delta^i = \{\mathbf{d}^i_k\}$, for $1 \le k \le |\mathcal{P}^i|$, are computed by minimizing:

$$\sum_{\mathbf{p}^i_k \in \mathcal{P}^i} \epsilon^i_k (\mathbf{d}^i_k - (\mathbf{p}^j_k - \mathbf{p}^i_k))^2 + \mu \sum_{\mathbf{p}^i_k \in \mathbf{S}^i} \sum_{\mathbf{p}^i_l \in \mathcal{N}^i_k} (\mathbf{d}^i_k - \mathbf{d}^i_l)^2, \tag{5}$$

where ϵ^i_k is a weight factor. The above cost function encourages the refined shape to lie closer to $\mathbf{S}^j$, while minimizing the local distortion induced by such displacement. The parameter μ is empirically determined for the training set. Note that (5) represents an extremely sparse linear system that can be solved efficiently. The vertices of the refined shape are obtained as $\mathbf{p}^i_k + \mathbf{d}^i_k$ and it inherits the connectivity of $\mathbf{S}^i$.

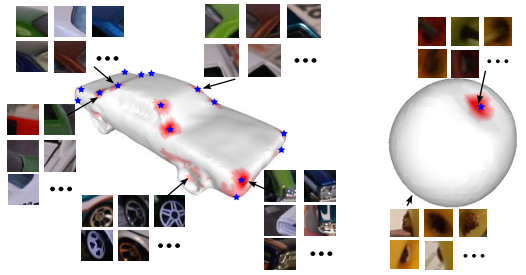
In the above, we are purposefully vague on the representation for the shape $\mathbf{S}^j$. This is because the above mechanism can be used, with minor changes, for both mean shape learning with the shape $\mathbf{S}^j$ being a mesh and for reconstruction with $\mathbf{S}^j$ being the oriented point cloud output of MVS, as elaborated in Sections 4.2 and 5.2, respectively.

4. Learning Reconstruction Priors

For each object category, we use a set of object instances $\{O^n\}$ to learn a mean shape $\mathbf{S}^*$ and a set of anchor points $\mathbf{A}$. For each object instance O^i in this training set, we capture a set of images $\mathbf{I}^i$ and use a 3D scanner to obtain a detailed 3D shape $\mathbf{S}^i_{\text{scan}}$. Given $\mathbf{I}^i$, we use a standard SFM pipeline to reconstruct a point cloud $\mathbf{S}^i_{\text{sfm}} = \{\mathbf{p}^i_j\}$, where $\mathbf{p}^i_j$ is a 3D point. We manually label a small number of SFM points, $\mathbf{M}^i = \{\mathbf{p}^i_1, \mathbf{p}^i_2, \cdots, \mathbf{p}^i_m\}$ (see the stars in Figure 3 and 4). The labelled points $\mathbf{M}$ are used to align the scanned shapes $\{\mathbf{S}^i_{\text{scan}}\}$ and their reconstructed point clouds $\{\mathbf{S}^i_{\text{sfm}}\}$ in our training dataset. They also serve as the initialization for the anchor point learning, as described in the following.

4.1. Learning Anchor Points

An anchor point, $A = \{\boldsymbol{\Gamma}, \boldsymbol{\chi}, \omega\}$, consists of a feature vector $\boldsymbol{\Gamma}$ that describes appearance, the 3D location $\boldsymbol{\chi}$ with respect to the mean shape and a scalar weight ω. $\boldsymbol{\Gamma}$ is the aggregation of HOG features [9] in all images where A is visible and of



(a) Car (b) Fruit

Figure 3. Learned mean shape and anchor points density. Darker red indicates greater density of anchor points. For cars, most anchor points are located around wheels and body corners since those parts are shared across instances. For fruits, anchor points are distributed around the stem and bottom. Blue stars show initially labelled points and the rest are learned by the proposed method. We also show image patches associated with the features of a few example anchor points.

every object where A exists. For an anchor point A, if $\mathcal{V}$ are the indices of objects across which the corresponding SFM points are matched and Ω^i are the indices of images of O^i where A is visible, the corresponding feature vector is:

$$\mathbf{\Gamma} = \{\{\mathbf{f}^i_{k^i}\}_{k^i \in \Omega^i}\}_{i \in \mathcal{V}}. \tag{6}$$

where $\mathbf{f}^i_{k^i}$ is the HOG feature of the image point associated with A in image $I^i_{k^i}$. Let $\mathbf{p}^i_j$ be the locations of the corresponding 3D points, normalized with respect to object centroid and scale. Then, the location for the anchor point is

$$\boldsymbol{\chi}_j = \frac{1}{|\mathcal{V}|} \sum_{i \in \mathcal{V}} \mathbf{p}^i_j. \tag{7}$$

The weight ω reflects "importance" of an anchor point. We consider an anchor point important if it appears across many instances, with low position and appearance variance. That is,

$$\omega = w_x \, w_a \, w_f \tag{8}$$

where $w_x = \exp(-\frac{\sum_{i \neq k} \|\mathbf{p}^i - \mathbf{p}^k\|}{\sigma_x N_2})$, $w_a = \exp(-\frac{\sum_{i \neq k} d^{i,k}}{\sigma_a N_2})$ and $w_f = \log|\mathcal{V}|$ encode location stability, appearance similarity and instance frequency, respectively. N_2 is the number of combinations. The coefficients σ_a and σ_x are determined empirically from training data for each category. In the above,

$$d^{i,k} = \min_{l^i \in \Omega^i, l^k \in \Omega^k} (|\mathbf{f}^i_{l^i} - \mathbf{f}^k_{l^k}|), \quad \text{for } i \neq k, \tag{9}$$

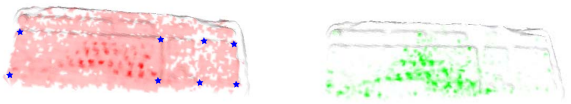
where Ω^i is the set of images of O^i where the point is visible.

In contrast to applications like shape matching, the quality of dense reconstruction is greatly affected by the order and extent of deformations. Thus, the learned anchor point weights ω are crucial to the success of dense reconstruction. Note that while ASM frameworks also associate a weight with landmark points, they are computed solely based on location uncertainty. By encoding appearance similarity and instance frequency, we impart greater semantic knowledge to our reconstruction stage.

The key precursor to learning anchor points is matching 3D points across instances, which is far from trivial. Besides within-class variation, another challenge is the fact that most SFM points correspond to texture. Such points usually dominate an SFM point cloud, but do not generalize across instances

Algorithm 1 Learning anchor points

Set Parameters δ_f, δ_p.
For objects O^i, $i \in [1, N]$, label m points to get $\mathbf{M}^i$.
Use $\mathbf{M}^i$ to align $\mathbf{S}^i_{\text{sfm}}$ with $\mathbf{S}^i_{\text{scan}}$.
$\forall \mathbf{p}^i_j \subset \mathbf{M}^i$, find $A_j = \{\mathbf{\Gamma}_j, \boldsymbol{\chi}_j, \omega_j\}$ using (6), (7), (8).
Initialize $\mathbf{A} = \{A_j\}$, $j = 1, \cdots, m$.
while anchor point set $\mathbf{A}$ is updated **do**
  for $i = 1 : N$ **do**
    Solve $\theta = \arg\min \sum_k \|T(\mathbf{p}^i_k, \theta) - \boldsymbol{\chi}_k\|$.
    Warp SFM point cloud $\mathbf{S}^i_{\text{sfm}} \leftarrow T(\mathbf{S}^i_{\text{sfm}}, \theta)$.
  end for
  for all $\mathbf{p}^i_k \in \mathbf{S}^i_{\text{sfm}}$ **do**
    for all $\mathbf{p}^j_l \in \mathbf{S}^j_{\text{sfm}}$, where $j \neq i$ **do**
      if $d(\mathbf{f}^i_k, \mathbf{f}^j_l) < \delta_f$ and $\|\mathbf{p}^i_k - \mathbf{p}^j_l\| < \delta_p$ **then**
        Match $\mathbf{p}^i_k$ to $\mathbf{p}^j_l$.
      end if
    end for
  end for
  Filter conflicting matches.
  Identify sets of matched SFM points $\mathbf{B}_h$, $h \in [1, H]$.
  for $h = 1 : H$ **do**
    Find $A_h = \{\mathbf{\Gamma}_h, \boldsymbol{\chi}_h, \omega_h\}$ using (6), (7), (8).
  end for
  Update $\mathbf{A} = \mathbf{A} \cup \{A_h\}$, for $h = 1, \cdots, H$.
end while
Output: denser anchor point set $\mathbf{A}$.



(a) Density (b) Weights

Figure 4. Learned shape prior and anchor points for keyboard category. (a) Density of anchor point distribution. Blue stars show the initially labelled anchor points. (b) Learned weights of anchor points. Deeper color means higher weight.

since they do not correspond to the object shape, thus, may not be anchor point candidates. Moreover, the density of anchor points cannot be too low, since they guide the deformation process that computes the mean shape and fits it to the 3D point cloud. To ensure the robustness of anchor point matching and good density, we propose an iterative algorithm, detailed in Algorithm 1. The distribution and weights of the learned anchor points are visualized in Figure 3 and 4.

4.2. Mean Shape Construction

The learned anchor points are used to compute a mean shape for an object category. Recall that we have a mapping from the set of anchor points to each instance in the training set. Thus, we can warp successive shapes closer to a mean shape using the anchor points. The mean shape is constructed by combining these aligned and warped shapes of different instances. Since there are multiple shape instances, the order of combining them is a critical design issue, because improperly combining dissimilar shapes may introduce severe artifacts. To determine the order for combining shapes, we first measure the pairwise similarity between all pairs of training instances. In our experiments, we use the weighted number of commonly matched anchor points as the similarity cue. Given the pairwise similarities, we use hierarchical clustering to group the shapes. The similarity relationships can be represented as a

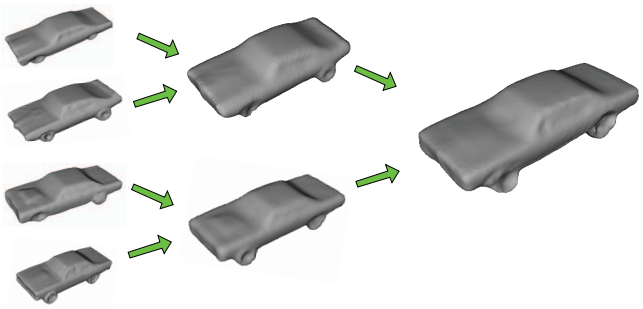


Figure 5. The mean shape computation proceeds by systematic combination of training instances, based on a binary tree traversal. The leaf nodes of the tree are the individual training instances, with assignments based on a pairwise shape similarity computation followed by hierarchical clustering. Note that unique details are lost, while features representative of the entire class are preserved.

binary tree where each leaf node is an object. We combine the warped shapes $T(\mathbf{S}^i_{\text{scan}})$ following the order of merging successive branches, to eventually obtain a single shape $\mathbf{S}^*$, which represents the commonality of all training instances. We use $\mathbf{S}^*$ as the mean shape. The mean shape learning procedure is shown for a subset of the car dataset in Fig. 5. Note that $\mathbf{S}^*$ is computed by using the warped training examples, where the warping maps the 3D locations of learned anchor points. Thus, the prior shape is always aligned with the anchor points.

In the above, the warp $T(\mathbf{S}^i_{\text{scan}}) \rightarrow \mathbf{S}^j_{\text{scan}}$, with $i < j$ according to the above defined ordering, is computed as the weighted thin plate spline transformation given by (4). Two shapes aligned by anchor points are eventually combined into a single one using displacement vectors computed by minimizing (5). The learned mean models for car, fruit and keyboard categories are shown in Figs. 3 and 4.

5. Semantic Reconstruction with Shape Priors

Given a number of images of an object O, we can reconstruct its 3D shape by warping the learned prior shape $\mathbf{S}^*$ based on the estimated θ and by recovering Δ in (1) subsequently. The reconstruction consists of three steps: matching anchor points, warping by anchor points, and refinement. Accurately recovering warp parameters θ requires accurate matches between anchor points in $\mathbf{S}^*$ and SFM points in $\mathbf{S}_{\text{sfm}}$. This is facilitated by an initial coarse alignment between $\mathbf{S}^*$ and $\mathbf{S}_{\text{sfm}}$.

5.1. Initial Alignment

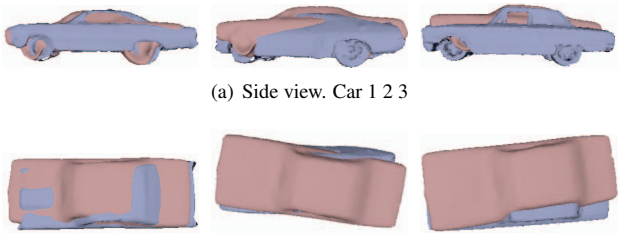
It is conventional in shape modeling literature to compute shape alignments using Procrustes analysis or ICP [8]. However, reconstructed SFM point clouds are typically sparse, contain several outliers and the point set of the object of interest might be dominated by background clutter. The second semantic component of our framework, object detection, is used to alleviate these issues for initial alignment.

State-of-the-art object detectors like [11] can detect objects in an image with cluttered background, with reasonably accurate estimates of object pose. Further, as demonstrated by [1], multiple images can significantly improve detection accuracy in both image and 3D space. In image I_j, the detector returns the confidence value $p_i(\mathbf{u}, s, \pi)$ of a detection hypothesis which appears in image location $\mathbf{u}$, with scale (height and width) s and pose π. Given the estimated camera poses, a hypothesized 3D object O can be projected to each image I_j at location $\mathbf{u}_j$, scale s_j and pose π_j. Thereby, the object O in 3D space may be estimated as

$$O = \arg\max_O \sum p_j(\mathbf{u}_j, s_j, \pi_j). \tag{10}$$

Please refer to [1] for details. This allows approximate estimation of the centroid, 3D pose and scale of an object. Since we also know those for the shape prior, we can use a rigid transformation to coarsely align the prior shape and its anchor points to fit the SFM point cloud of the object. The initial alignment for a car reconstruction is shown in Figure 6.

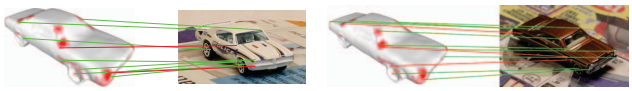
Note that unlike Procrustes alignment, this detection-based alignment does not rely on any SFM points (only camera poses), thus, it is robust to the sparsity and noise that pervade SFM point clouds obtained from few images.



(a) Side view. Car 1 2 3

(b) Top view. Car 1 2 3

Figure 6. Initial alignment using object detection. Blue shows ground truth position of the object to be reconstructed. Red shows object position and orientation estimated from detection [11] across 15 views.



(a) Car 1 (b) Car 2

Figure 7. Matching anchor points from leaned model (left) to new object (right). We show the high confidence matches visible under the displayed viewpoint. The green/red lines show the good/bad matches.

5.2. Reconstruction

Given a set of images $\mathbf{I}$ of an object with unknown shape $\mathbf{S}$, we use standard SFM to recover the 3D point cloud $\mathbf{S}_{\text{sfm}}$. Our goal is to use the mean shape $\mathbf{S}^*$ to produce a dense reconstruction that closely resembles $\mathbf{S}$.

Matching Anchor Points. Since the initial alignment uses the object's location, pose and scale, anchor points are likely to be aligned to 3D locations in the vicinity of their true matches. Thus, the burden of identifying the point in $\mathbf{S}_{\text{sfm}}$ that corresponds to an anchor point in $\mathbf{S}^*$ is reduced to a local search. We use HOG features to match anchor points to SFM points. To further improve the robustness, Algorithm 2 proposes an iterative matching scheme. Examples of robust anchor point matches from our algorithm are shown in Figure 7.

Algorithm 2 Matching anchor points

Set parameters δ_1 δ_2 η.
for $k = 1 : K$ (total number of iterations) **do**
  Initialize match set $\mathbf{B}_k = \{\}$.
  for all $A_i = \{\mathbf{\Gamma}_i, \boldsymbol{\chi}_i, \omega_i\} \in \{\mathbf{A}\}$ **do**
    Define $P = \{\mathbf{p}_k \in \mathbf{S}_{\text{sfm}} : \|\mathbf{p}_k - \boldsymbol{\chi}_i\| < \delta_1\}$.
    Find $\mathbf{p}_j \in \mathbf{S}_{\text{sfm}}$ s.t. $\mathbf{p}_j = \arg\min_P d^{i,j}$ (Eq. 9)
    If $d(\mathbf{f}_j, \mathbf{f}_i) < \delta_2$, match $(A_i, \mathbf{p}_j)$, $\mathbf{B}_k = \mathbf{B}_k \cup \{\mathbf{p}_j\}$.
    Record 3D distance $r_i = \|\boldsymbol{\chi}_i - \mathbf{p}_j\|$.
  end for
  Solve $\theta'_k = \arg\min \|T(\mathbf{A}, \theta) - \mathbf{B}_k\|$.
  for all $A_i \in \mathbf{A}$ **do**
    if $\|T(\boldsymbol{\chi}_i, \theta'_k) - \mathbf{b}_i\| > r_i$ **then**
      Discard match $(A_i, \mathbf{b}_i)$, $\mathbf{B}_k = \mathbf{B}_k \backslash \{\mathbf{b}_i\}$.
    end if
  end for
  Solve $\theta_k = \arg\min \|T(\mathbf{A}, \theta) - \mathbf{B}_k\|$.
  $\forall A_i \in \mathbf{A}$, $\boldsymbol{\chi}_i \leftarrow T(\boldsymbol{\chi}_i)$.
  $\delta_1 \leftarrow \eta\delta_1$.
end for
Output: the set of matches $\mathbf{B}_K$.

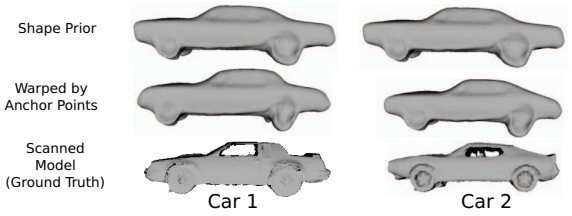


Figure 8. Warping of the shape prior with the learned anchor points matched to SFM points using Algorithm 2. Note that while the shape prior represents the commonality of all instances, anchor point-based warping recovers coarse aspects of instance-specific shape, such as the back geometry of Car 2.

Warping Based on Anchor Points. Assume $\mathbf{S}^*$ is the shape prior after the initial alignment of Section 5.1. We use the above matches between anchor points in $\mathbf{S}^*$ and SFM points in $\mathbf{S}_{\text{sfm}}$ to estimate parameters θ for the weighted TPS warping (4) and obtain $\mathbf{S}' = T(\mathbf{S}^*, \theta)$ that further approaches the actual shape. Notice that, this warping not only reduces the alignment error from the initial detection-based alignment, it also deforms the prior to fit the actual shape of the object. See Figure 8.

Refinement. The final step in the reconstruction process is to recover the unique details of the object. These unique details cannot be learned a priori, so they may not be captured by the warped shape $\mathbf{S}'$. We use the output of an MVS algorithm [15], $\mathbf{S}_{\text{mvs}}$, to supply these details. While MVS may have several missing regions and outliers for the object we consider, it may reconstruct accurate oriented patches in textured or Lambertian regions where diffuse photoconsistency is a reliable metric. Using the refinement process governed by (5), we move the vertices of $\mathbf{S}'$ closer to $\mathbf{S}_{\text{mvs}}$. The weights ϵ_k now incorporate the confidence in the corresponding matched MVS point, which is encoded by the normalized cross-correlation photoconsistency.

The effect of refinement is shown in Figure 9. Note that not only are the holes and outliers of traditional MVS eliminated in our reconstruction, but fine details that are missing in the warped prior shape are also recovered by refinement – see the front bonnet and rear spoiler of Car 1, or the inset rear window edges and the protruding trunk of Car 2. This refined shape is the final output of our dense reconstruction framework.

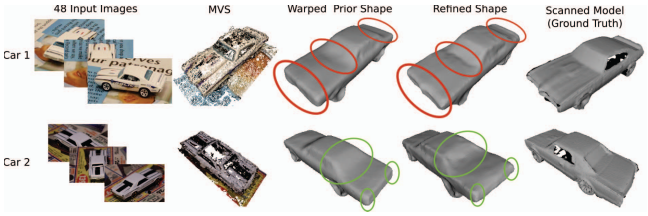


Figure 9. Refinement recovers unique details of an instance that are lost during mean shape learning. Examples such as the rear spoiler of Car 1 and the inset rear window of Car 2 are highlighted.

6. Experiments

We evaluate our method on three categories: car, fruit and keyboard. We use a structured light 3D scanner to acquire ground truth shapes for learning and evaluation. Our testing is leave-one-out, that is, to reconstruct one instance, we train our model on all the rest. The model parameters are obtained by cross-validation in the training set. We compare against state-of-the-art MVS methods, show reconstruction results in Figure 11 and report quantitative evaluation results for the car dataset in Tables 1 and 2. Example results from individual stages of our framework are also depicted in Figures 3–9.

The car dataset comprises ten instances with lengths between 65 – 73mm. Using the detection-based initial alignment (Section 5.1), the estimated centroids of test objects are localized within 20% of object length and the orientation estimation error is within 10°, as shown in Figure 6. The fruit dataset consists of life-size models for twelve fruits of varying shapes and sizes. The keyboard dataset consists of seven keyboards. Centroid localization error (relative to object length) and orientation estimation error are within 5% and 40° for the fruits and within 10% and 30° for the keyboards.

To quantitatively demonstrate the efficacy of our framework, we perform a rigorous evaluation against ground truth. Reconstruction error (relative to ground truth scan) is computed using the metric in [7] (other metrics such as [27] are equally applicable). For each test instance of the car category, we perform reconstructions using 48, 15 and 5 images. The baseline method is MVS [15], with the reconstructed patches meshed using Poisson Surface Reconstruction (PSR) [20]. We also evaluate errors for intermediate results of our pipepine. See Table 1. It is clear that each stage of our framework leads to significant improvement, with an over 40% improvement in final quality over traditional MVS. Also note that our reconstruction error in the challenging situation of 5 images is even lower than the baseline method with 15 images.

The efficacy of using anchor points and their learned weights can be demonstrated by Table 2. Using anchor points can greatly reduce the reconstruction error compared to only using object detection for alignment. Learning anchor point weights further enhances the reconstruction accuracy.

We also use our reconstruction method for scenes with multiple objects in a cluttered environment (Figure 10). The method of [1] is used to detect multiple objects in the 3D scene and our framework is individually applied to each object. Note that our reconstructed objects are aligned in the same

# img	Base %	RGD %	WP %	Full %
48	1.22	1.00	0.88	0.71
15	2.72	2.39	2.29	1.88
5	4.66	2.91	2.86	2.47

Table 1. Reconstruction error in car dataset. Base: [15]+[20]. RGD: Rigidly align mean shape to test object using matched anchor points. WP: Align and warp mean shape using matched anchor points (without refinement). Full: Our complete algorithm. Errors are reported in the metric of [7]. Note a 40% improvement between Base and Full.

Base	IA+RF	RGD+RF	WP (No ω)+RF	WP+RF
1.22%	1.94%	0.85%	0.75%	0.71%

Table 2. Reconstruction error of alternative designs of our pipeline. Base: [15]+[20]. IA: Initial alignment using object detection (Section 5.1). RF: Refinement (Section 5.2). RGD: Rigidly align the mean shape to a test object by using matched anchor points. WP: Align and warp the mean shape by using matched anchor points (Section 5.2). No ω: Using anchor points with equal weights. Errors are computed by using the car dataset with 48 images available for each car.

coordinate system as the SFM point cloud of the scene. This allows us to automatically overlay the 3D objects reconstructed using our method with the point cloud of the background.

In Figure 11, we show several comparisons of our reconstructions against state-of-the-art MVS [15, 20]. Note the lack of texture and specularities in the sample images shown in (a). Diffuse photo-consistency is not a metric well-suited to these situations, so the MVS output in (b) is visibly noisy and contains a large number of artifacts in the form of holes and outliers. Consequently, the resulting PSR mesh in (c) is distorted. In contrast, we successfully learn meaningful semantic priors across shape variations and use them in our reconstruction, to produce the much higher quality reconstructions in (d), that closely resemble the ground truth (e).

7. Discussion and Future Work

We have presented a comprehensive framework for dense object reconstruction that uses data-driven semantic priors to recover shape in situations unfavorable to traditional MVS. Our learned priors, combined with robust anchor point matching and refinement mechanisms, are shown to produce visually high quality and quantitatively accurate results.

The success of this framework also opens up directions for future research. While semantic information for objects such as cars is easily correlated to shape, many categories such as chairs show shape variation at finer granularities. Thus, ongoing research efforts in fine-grained recognition and detection of object parts may also benefit our semantic reconstruction framework. In our future work, we seek to demonstrate our system in an MRF-based MVS framework like [18], since it provides the flexibility to combine our shape prior with silhouette information from object detectors like [12].

Acknowledgements This research was conducted as the first author's summer internship at NEC Labs America. We thank Ishan Mittal at University of Michigan for helping with the data acquisition. Y. Bao and S. Savarese also acknowledge the support of NSF CAREER grant 1054127.

Figure 10. Reconstruction of multi-object scenes. (Left) 1 out of 10 input images. (Middle) MVS [14]. (Right) Our reconstruction.

References

[1] S. Bao and S. Savarese. Semantic structure from motion. In *CVPR*, pages 2025 –2032, 2011.

[2] S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. *PAMI*, 24(4):509–522, 2002.

[3] A. Berg, T. Berg, and J. Malik. Shape matching and object recognition using low distortion correspondences. In *CVPR*, volume 1, pages 26–33, 2005.

[4] V. Blanz and T. Vetter. A morphable model for the synthesis of 3D faces. In *SIGGRAPH*, pages 187–194, 1999.

[5] F. L. Bookstein. Principal warps: Thin-plate splines and the decomposition of deformations. *PAMI*, 11(6):567–585, 1989.

[6] H. Chui and A. Rangarajan. A new point matching algorithm for non-rigid registration. *CVIU*, 89(2-3):114–141, 2003.

[7] P. Cignoni, C. Rocchini, and R. Scopigno. Metro: Measuring error on simplified surfaces. *CGF*, 17:167–174, 1998.

[8] T. Cootes, C. Taylor, D. Cooper, and J. Graham. Active shape models: Their training and application. *CVIU*, 61(1):38 – 59, 1995.

[9] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, pages 886–893, 2005.

[10] I. Dryden and K. Mardia. *Statistical Shape Analysis*. John Wiley and Sons, 1998.

[11] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *PAMI*, 32(9):1627 –1645, 2010.

[12] V. Ferrari, F. Jurie, and C. Schmid. Accurate object detection with deformable shape models learnt from images. In *CVPR*, pages 1–8, 2007.

[13] Y. Furukawa, B. Curless, S. Seitz, and R. Szeliski. Manhattan-world stereo. In *CVPR*, pages 1422 –1429, 2009.

[14] Y. Furukawa, B. Curless, S. Seitz, and R. Szeliski. Towards internet-scale multi-view stereo. In *CVPR*, pages 1434–1441, 2010.

[15] Y. Furukawa and J. Ponce. Accurate, dense and robust multiview stereopsis. *PAMI*, 32(8):1362–1376, 2010.

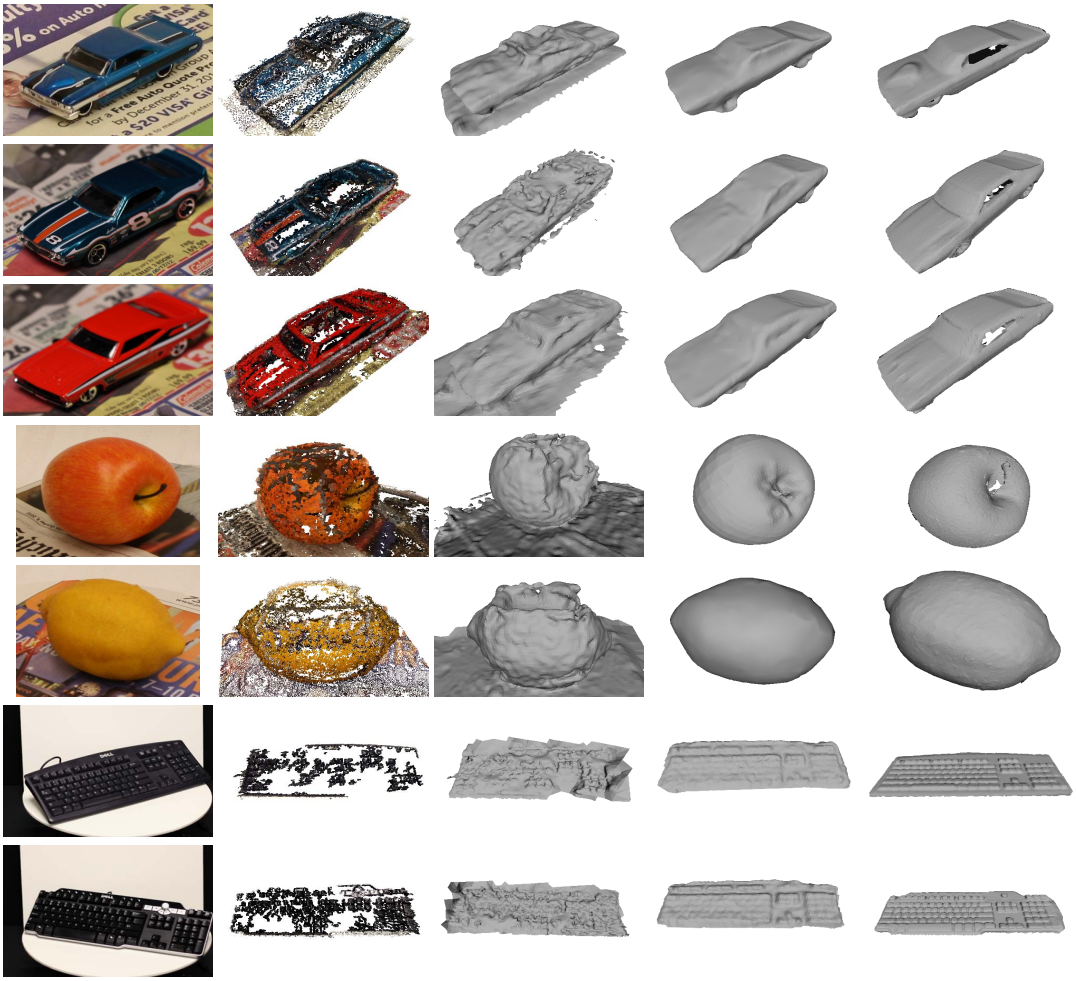
[16] D. Gallup, J.-M. Frahm, and M. Pollefeys. Piecewise planar and non-planar stereo for urban scene reconstruction. In *CVPR*, pages 1418–1425, 2010.

[17] M. Goesele, J. Ackermann, S. Fuhrmann, R. Klowsky, F. Langguth, P. Müandcke, and M. Ritz. Scene reconstruction from community photo collections. *IEEE Computer*, 43:48–53, 2010.

[18] C. Hernández and G. Vogiatzis. Shape from photographs: A multi-view stereo pipeline. In *Computer Vision*, volume 285 of *Studies in Comp. Intell.*, pages 281–311. Springer, 2010.

[19] T. Jiang, F. Jurie, and C. Schmid. Learning shape prior models for object matching. In *CVPR*, pages 848 –855, 2009.

[20] M. Kazhdan, M. Bolitho, and H. Hoppe. Poisson surface reconstruction. In *SGP*, pages 61–70, 2006.



(a) Sample Image (b) MVS Patches [15] (c) MVS + PSR [20] (d) Our Method (e) Ground Truth

Figure 11. Examples of reconstructed objects. Notice the lack of texture and presence of specularities in sample images (a). MVS reconstruction from 48 images using the method of [14] produces clearly visible holes and extremely noisy reconstructed patches (b). Poisson surface reconstruction fails to produce a reasonable mesh under such scenarios (c). Our semantic framework, on the other hand, yields a high quality reconstruction (d), which closely resembles the ground truth (e), both visually and quantitatively. The results are obtained by using 48 images for cars and fruits, and 5 images for keyboards.

[21] V. Kolmogorov and R. Zabih. Multi-camera scene reconstruction via graph cuts. In *ECCV*, pages 82–96, 2002.

[22] B. Leibe, A. Leonardis, and B. Schiele. An implicit shape model for combined object categorization and segmentation. In *Toward Category-Level Object Recognition*, volume 4170 of *LNCS*, pages 508–524. Springer, 2006.

[23] M. J. Leotta and J. L. Mundy. Predicting high resolution image edges with a generic, adaptive, 3D vehicle model. In *CVPR*, pages 1311–1318, 2009.

[24] B. Munsell, P. Dalal, and S. Wang. Evaluating shape correspondence for statistical shape analysis: A benchmark study. *PAMI*, 30(11):2023 –2039, 2008.

[25] M. Pauly, N. J. Mitra, J. Giesen, M. Gross, and L. J. Guibas. Example-based 3D scan completion. In *SGP*, pages 23–32, 2005.

[26] K. Rohr, H. S. Stiehl, R. Sprengel, W. Beil, T. M. Buzug, J. Weese, and M. H. Kuhn. Point-based elastic registration of medical image data using approximating thin-plate splines. In *Int. Conf. on Vis. in Biomed. Comp.*, pages 297–306, 1996.

[27] S. M. Seitz, B. Curless, J. Diebel, D. Scharstein, and R. Szeliski. A comparison and evaluation of multi-view stereo reconstruction algorithms. In *CVPR*, pages 519–528, 2006.

[28] M. Sun, G. Bradski, B.-X. Xu, and S. Savarese. Depth-encoded hough voting for coherent object detection, pose estimation, and shape recovery. In *ECCV*, pages 658–671, 2010.

[29] A. Thomas, V. Ferrari, B. Leibe, T. Tuytelaars, and L. Van Gool. Depth-from-recognition: Inferring meta-data by cognitive feedback. In *ICCV*, pages 1–8, 2007.

[30] G. Vogiatzis, C. Hernandez, P. Torr, and R. Cipolla. Multiview stereo via volumetric graph-cuts and occlusion robust photo-consistency. *PAMI*, 29(12):2241 –2246, 2007.

[31] C. Wu, S. Agarwal, B. Curless, and S. Seitz. Schematic surface reconstruction. In *CVPR*, pages 1498–1505, 2012.

[32] Y. Xiang and S. Savarese. Estimating the aspect layout of object categories. In *CVPR*, pages 3410–3417, 2012.