Using Locally Corresponding CAD Models for Dense 3D Reconstructions from a Single Image

Chen Kong Chen-Hsuan Lin Simon Lucey
Carnegie Mellon University
{chenk, chenhsul, slucey}@andrew.cmu.edu

Abstract

We investigate the problem of estimating the dense 3D shape of an object, given a set of 2D landmarks and silhouette in a single image. An obvious prior to employ in such a problem is a dictionary of dense CAD models. Employing a sufficiently large enough dictionary of CAD models, however, is in general computationally infeasible. A common strategy in dictionary learning to encourage generalization is to allow for linear combinations of dictionary elements. This too, however, is problematic as most CAD models cannot be readily placed in global dense correspondence. In this paper, we propose a two-step strategy. First, we employ orthogonal matching pursuit to rapidly choose the "closest" single CAD model in our dictionary to the projected image. Second, we employ a novel graph embedding based on local dense correspondence to allow for sparse linear combinations of CAD models. We validate our framework experimentally in both synthetic and real world scenario and demonstrate the superiority of our approach to both 3D mesh reconstruction and volumetric representation.

1. Introduction

Reconstructing the 3D geometry of objects from 2D images is a fundamental task in computer vision. With the remarkable success in Structure from Motion (SfM), which is now capable of reconstructing entire cities using large-scale photo collections [1] and real-time visual SLAM on embedded and mobile devices [14], the computer vision community is starting to explore the possibility of constructing a 3D model of an object from a single image [17, 20, 11, 21, 18]. Since estimating 3D geometry from a single view is an inherently ill-posed problem, all of these approaches need to employ some sort of 3D prior. An increasingly popular type of 3D prior are CAD models due to: (i) their ubiquity online (e.g. 3D warehouse), and (ii) their dense level of detail.

A 3D CAD model is made up of verticies and edges. Unfortunately, most CAD models, even those coming from the same object category, (e.g. chair, table, etc.), do not have verticies in correspondence or even typically share the same edge topology. This can be somewhat alleviated through the manual annotation of common 3D landmarks, but even this can be cumbersome due to the inherent lack of global correspondences for certain object categories. For example, as shown in Figure 1, the table C and G share the same coarse landmarks (i.e. shape skeleton) but differ substantially on their finer details such that no meaningful dense correspondence can be established. Figure 1 shows eight tables from A to H and the solid lines depict which pairs of CAD models that can be brought into dense correspon-

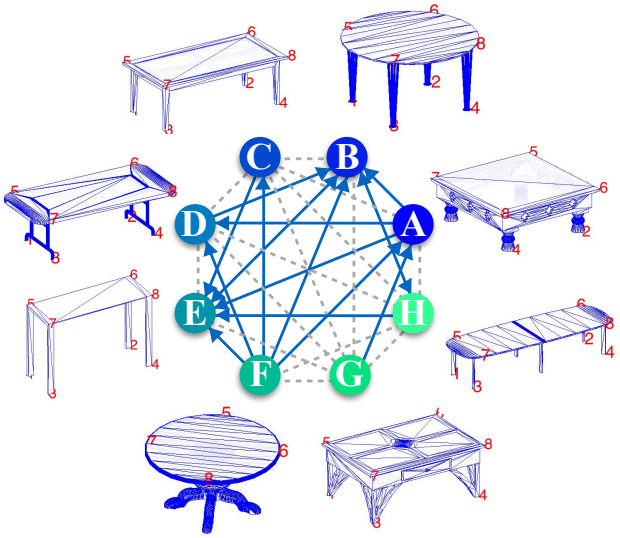

Figure 1. A toy example of our proposed local dense correspondence graph. The node represents an CAD model shown besides it and the edge (solid) denotes dense correspondence. The graph is not fully connected (shown in dashed lines) since global dense correspondence do not exist. However, in some subgraphs, all nodes are connected by a center, *e.g.* F connects to all nodes in subgraph {A, B, C, D, E, F}. This indicates that a dictionary with CAD models as elements can be established in such subgraphs using local dense correspondence, which is at the heart of our paper.

dence with one another. One can see that the graph is not fully connected, verifying the nonexistence of global dense correspondence for this object category. However, we observe that some subgraphs (*e.g*. subgraph {A, B, C, D, E, F}) are fully connected by a center (F in this case). This indicates that a dense dictionary can be established in such subgraphs based on these local dense correspondence. This insight is the core of our paper.

In this paper, we make the following contributions:

- We propose a novel graph embedding based on the local dense correspondence between 3D models, and we demonstrate that in each subgraph a dense shape dictionary can be established and a sparse linear combination can be used to create a deformable dense model.
- We propose a two-step coarse-to-fine strategy which can first rapidly select a subgraph by landmark registration and second refine camera position and create a dense model by fitting both landmarks and silhouette.
- Finally, we show empirically the utility of our approach for estimating fine geometry for various object categories from a single image. Qualitative and quantitative results are reported on both synthetic and natural images compared with volumetric representation.

2. Related Work

Reconstructing the 3D geometry of a 2D projected object from a known category is receiving increasing attention in the field of computer vision. Due to intra-category variation (*e.g*. sedan, coupe, SUV in car category) we consider the shape of each instance to be inherently deformable and non-rigid. A common strategy for representing this deformable prior is through the employment of a 3D dictionary of shape instances. Kong *et al*. [9, 10] recently proposed a method for learning such a 3D dictionary solely from an ensemble of 2D landmark projections stemming from a known object category. Although this work was capable of handling highly deformable objects, due to the non-convex characteristics of the group-sparse dictionary learning problem, it was overly sensitive to initialization and landmark noise.

An alternative strategy for learning a robust 3D dictionary is to leverage the increasing availability of 3D CAD models. Zhou *et al*. [20] learned a dictionary from a 3D shape dataset and proposed a convex relaxation technique to estimate the pose and shape parameters simultaneously given a single image of 2D projected landmarks. Non-dictionary strategies have also been entertained, most notably Wu *et al*. [18] trained a deep network to infer 3D points from 2D landmarks. Despite their promising results, all these works, are only capable to reconstruct the sparse 3D skeleton for the objects of interest. This limitation can be mostly attributed to the difficulty of establishing global dense correspondence between CAD models of the same object category (as argued in the introduction).

For the problem of dense 3D shape reconstruction, notable examples include [7] and [13] both of which extended the above methods by employing optical flow in order to establish denser correspondences. Both of these methods restricted themselves to less-deformable objects, *e.g*. faces and surfaces, thus limiting generalization to the task of 3D object reconstruction.

More recently, Vicente *et al*. [16] proposed a framework utilizing landmarks and silhouette to establish dense convex hulls for the Pascal VOC imageset. The approach first initializes camera positions using rigid structure from motion, and then applies a novel visual hull reconstruction method to the set of images which are considered to share the same shape but different viewpoints by assuming such image surrogates always exist in large image sets. A fundamental issue to this approach, however, is that the inferred 3D reconstruction is rigid. To handle this drawback, Kar *et al*. [8] employed a novel dense surface model originating from Active Shape Models (ASMs) [6] to estimate a deformable dense hull for each single image. Although impressive, their approach is limiting as the process smooths over important fine details in the reconstructed 3D geometry. Further both the works require a large number of images stemming from the same object category.

Another family of works related to our paper is to employ a volumetric representation through the employment of deep neural networks [5, 12]. Despite achieving remarkable results, the volumetric representation itself is problematic due to its low spatial resolution caused by the computational complexity of training a network with such large 3D signals. A major advantage of our proposed approach is that the 3D mesh representation of CAD models (using edges and vertices) is more able to preserve important 3D detail - as compared to such volumetric methods. A full comparison of these two representations can be found in our experimental portion.

3. Overview

Given a single image of a certain object, our method seeks to estimate camera position and reconstruct a 3D dense model by utilizing the 2D landmark and silhouette information. We assume that the landmark positions have been labeled/detected and the image has been segmented beforehand. To achieve the goal, we first leverage the 3D CAD models in the object category by building up a graph, which we refer to as the Local Dense Correspondence (LDC) graph, to describe the dense correspondence between each pair of CAD models. We then propose a two-step approach: (1) estimate a coarse camera position and select the CAD model from the LDC graph to best register the 2D landmarks; (2) refine the camera position and de-

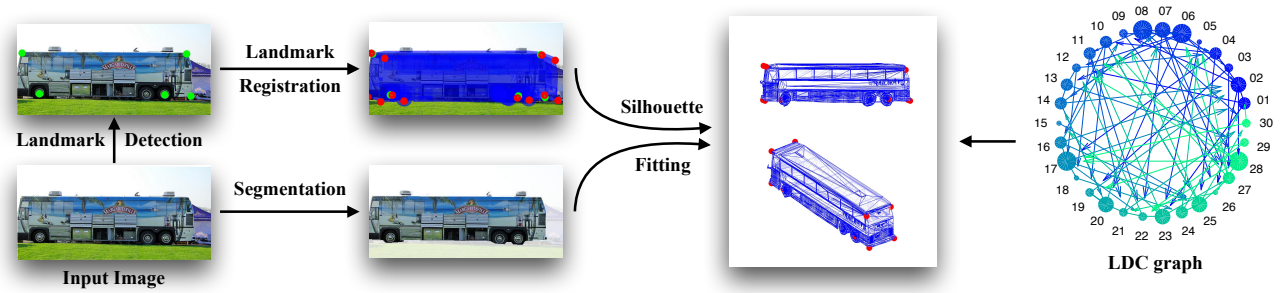

Figure 2. Overview of our method. Given a single image with annotated/detected landmarks and silhouette, we build up a local dense correspondence graph (right), followed by a coarse estimate of camera position and CAD model by landmark registration, and final refinement creating a *deformable* dense model to fit both landmarks and silhouette. Best viewed in color.

form the best CAD model by linear combination with its neighbors in the LDC graph to fit both the landmarks and the silhouette. Figure 2 shows the proposed LDC graph and our two-step coarse-to-fine method.

4. Local Dense Correspondence Graph

Local Dense Correspondence (LDC) graph is a directed graph with CAD models as nodes and the dense correspondence as edges. As each model here is manually and independently designed and does not necessarily share the same number of vertices or the same structure of meshes, dense correspondence based on vertex matching is not feasible. Instead, to build up the dense correspondence from model $\mathcal{S}_1$ to $\mathcal{S}_2$, we find a matching point on the surface of $\mathcal{S}_2$ for each vertices of $\mathcal{S}_1$. Therefore, such correspondence from $\mathcal{S}_1$ to $\mathcal{S}_2$ is not identical to that of $\mathcal{S}_2$ to $\mathcal{S}_1$, implying that the LDC graph is directed.

4.1. Creating graph

To create the LDC graph, we exploit the non-rigid ICP algorithm [2] to find a matching point for each vertex. We propose a distance metric to establish match quality. More specially, to build up dense correspondence from $\mathcal{S}_1(\mathbf{V}_1, \mathbf{E}_1)$ to $\mathcal{S}_2(\mathbf{V}_2, \mathbf{E}_2)$ where $\mathbf{V}, \mathbf{E}$ indicates the vertices and triangulation respectively, we warp the source, $\mathcal{S}_1$ in this case, to the target, $\mathcal{S}_2$ by non-rigid ICP, such that the warped $\mathcal{S}_1$ can represent the same shape as $\mathcal{S}_2$. For convenience, we denote the positions of warped vertices as $\mathbf{V}_1^2$, and the warped surface as $\mathcal{S}_1^2(\mathbf{V}_1^2, \mathbf{E}_1)$, since the triangulation should not change during the warp. We define that if the warped surface $\mathcal{S}_1^2$ represents the target $\mathcal{S}_2$ successfully, the warped vertices $\mathbf{V}_1^2$ are the dense correspondence from $\mathcal{S}_1$ to $\mathcal{S}_2$. To estimate the success of the warping, in other words, the similarity between the warped surface and the target, we propose the following metric[1]:

$$E_{12} = \frac{1}{|\mathbf{V}_1^2|} \sum_{\mathbf{v}_i \in \mathbf{V}_1^2} e(\mathbf{v}_i, \mathcal{S}_2; \theta) + \frac{1}{|\mathbf{V}_2|} \sum_{\mathbf{v}_i \in \mathbf{V}_2} e(\mathbf{v}_i, \mathcal{S}_1^2; \theta), \tag{1}$$

where function

$$e(\mathbf{v}, \mathcal{S}; \theta) = \begin{cases} 1 & \text{if dist}(\mathbf{v}, \mathcal{S}) > \theta \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

The value θ was chosen through a cross-validation such that consistent LDC graphs are formed across object categories.

Once we have measured warping quality, we establish dense correspondence according to a predefined threshold. Warps which score below the threshold are ignored. Note that, due to failure of nonrigid ICP in some cases, we found that indirect warping typically does not improve the performance. More specifically, given $\mathcal{S}_i, i = 1, 2, 3$, the indirect warping sequence $\mathcal{S}_1 \rightarrow \mathcal{S}_2 \rightarrow \mathcal{S}_3$ typically does not outperform the direct deformation $\mathcal{S}_1 \rightarrow \mathcal{S}_3$. Therefore, when creating our LDC graph, we only consider direct warping.

Figure 2 (right) shows an example of the proposed LDC graph, which has 30 nodes and 87 edges. The numbers besides the nodes are the indices of the corresponding CAD models. The size of nodes showed in the figure is proportional to the number of edges starting from that node. Note that due to the difficulty of non-rigid matching, not every edges are bidirectional. This can be caused by many factors: unbalanced numbers of vertices/meshes between two nodes, unbalanced sizes between two models and *etc*.

4.2. LDC subgraphs

Due to the nonexistence of global dense correspondence, the LDC graph is never fully connected. Therefore, we explore the local properties and sparsity structure of the LDC graph in this section. We divide the LDC graph into multiple subgraphs such that each subgraph has a node as center and contains all nodes that have dense correspondence from the center. Specifically, denote Ω as an index set pointing to the nodes in a certain subgraph with $\mathcal{S}_c$ as the center. The

[1] As measuring the similarity between two surfaces is not a main contribution of our paper, we utilize this simple metric. More accurate metrics including 3D descriptors could be used to boost performance.

definition of subgraph implies that dense correspondence $\mathbf{V}_c^i$ always exist for any $i \in \Omega$. Therefore, a deformable model $\mathcal{S}(\mathbf{V}, \mathbf{E})$ can be created by linear combination:

$$\mathbf{V} = \omega_c \mathbf{V}_c + \sum_{i \in \Omega} \omega_i \mathbf{V}_c^i, \quad \mathbf{E} = \mathbf{E}_c, \tag{3}$$

where $\mathbf{V}$'s, $\mathbf{E}$'s are matrices containing the vertex position and triangulation respectively and ω's are combination weights. As a result, each LDC subgraph actually defines one deformable dense model controlled by the combination weights ω's. This insight is at the heart of our paper. Benefiting from this insight, the dense 3D reconstruction task could be addressed by first searching all subgraphs to find the best one, and then estimating the weights ω's.

Before visiting all possible subgraphs, we are curious how many subgraphs exist there, and how big these subgraphs are. From its definition, one can learn that the number of subgraphs equals to the number of nodes in the LDC graph[2], while the size of it varies from the smallest 1 (the center itself) to the number of nodes (the whole graph.)

5. Landmark Registration

Given the LDC graph and a single image, we now want to decide which subgraph or which deformable dense model is the best option for a dense 3D reconstruction task. Even though exhaustively searching all possible subgraphs is a strategy to achieve the best performance, it is, however, computationally infeasible, especially when using large-scale 3D model dataset, like ShapeNet [4]. Therefore, instead of visiting all possible subgraphs, we propose a landmark registration algorithm to rapidly select the "closest" single node to the given image, and use the subgraph extended from this node as the optimal LDC subgraph.

Give the single image $\mathbf{I}$, we assume a certain landmark detection algorithm has been exploited, such that the 2D positions of landmarks on the image plane are known as $\mathbf{w}_p$, for $p = 1, ..., P$. Since some landmarks may not be visible, due to occlusion or self-occlusion, we denote an index set, $\mathcal{P}$, to indicate landmark visibility. By using the weak-perspective projection, we denote $\mathbf{R} \in \mathbb{R}^{2\times 3}$ as the first two rows of rotation matrix, $\mathbf{t}$ as the translation, s as the scale of camera. We define the i-th column in matrix $\mathbf{Y}_p \in \mathbb{R}^{3\times N}$, where N is the number of models, as the 3D position of p-th landmark in i-th model. Instead of trying all possible candidates exhaustively, we propose to use sparsity constraint for simultaneously selecting the best CAD model and estimating the camera paramters:

$$\begin{aligned} &\underset{\mathbf{R},s,\mathbf{t},\mathbf{c}}{\operatorname{argmin}} \frac{1}{2} \sum_{p\in\mathcal{P}} \left\| s\mathbf{R}\mathbf{Y}_p\mathbf{c} + \mathbf{t} - \mathbf{w}_p \right\|_2^2 \\ &\text{s.t. } \mathbf{R}\mathbf{R}^T = \mathbf{I}_2, \quad \|\mathbf{c}\|_0 = 1, \end{aligned} \tag{4}$$

[2]As the subgraph is the largest subset of nodes connected from its center, one node has and only has one subgraph expanded from it.

where $\|\cdot\|_0$ is the ℓ_0 norm and $\mathbf{c}$ contains either zero or one, indicating which model is active. This objective can be minimized efficiently by Alternating Direction Method of Multipliers (ADMMs) [3].

From ADMMs, an auxiliary variable $\mathbf{Z}$ is introduced and the Equation 4 can be identically expressed as:

$$\begin{aligned} &\underset{\mathbf{M},\mathbf{Z},\mathbf{t},\mathbf{c}}{\operatorname{argmin}} \frac{1}{2} \sum_{p\in\mathcal{P}} \left\| \mathbf{Z}\mathbf{Y}_p\mathbf{c} + \mathbf{t} - \mathbf{w}_p \right\|_2^2 \\ &\text{s.t. } \mathbf{M}\mathbf{M}^T = s^2\mathbf{I}_2, \quad \|\mathbf{c}\|_0 = 1, \quad \mathbf{Z} = \mathbf{M}, \end{aligned} \tag{5}$$

where $\mathbf{M} = s\mathbf{R}$ for convenience. The augmented Lagrangian of Equation 5 is formulated as:

$$\begin{aligned} \mathcal{L} = &\frac{1}{2} \sum_{p\in\mathcal{P}} \left\| \mathbf{Z}\mathbf{Y}_p\mathbf{c} + \mathbf{t} - \mathbf{w}_p \right\|_2^2 + \\ &\langle \mathbf{\Lambda}, \mathbf{M} - \mathbf{Z} \rangle + \frac{\rho}{2} \|\mathbf{M} - \mathbf{Z}\|_F^2, \end{aligned} \tag{6}$$

where $\mathbf{\Lambda}$ is the lagrangian multiplier, ρ is a penalty factor to control the convergence behavior, and $< \cdot, \cdot >$ is Frobenius product of two matrices. ADMMs decomposes an objective into several sub-problems and iteratively solves them till convergence occurs [3]. We update $\mathbf{Z}$ by:

$$\begin{aligned} \mathbf{Z}^+ &= \underset{\mathbf{Z}}{\operatorname{argmin}} \mathcal{L} \\ &= \Big(\sum_{p\in\mathcal{P}} (\mathbf{w}_p - \mathbf{t})\mathbf{c}^T\mathbf{Y}_p^T + \mathbf{\Lambda} + \rho\mathbf{M} \Big) \Big(\sum_{p\in\mathcal{P}} \mathbf{Y}_p\mathbf{c}\mathbf{c}^T\mathbf{Y}_p^T + \rho\mathbf{I} \Big)^\dagger, \end{aligned} \tag{7}$$

and update $\mathbf{M}$ by:

$$\mathbf{M}^+ = \underset{\mathbf{M}}{\operatorname{argmin}} \mathcal{L} = \mathbf{U} \begin{bmatrix} (\sigma_1 + \sigma_2)/2 & \\ & (\sigma_1 + \sigma_2) \end{bmatrix} \mathbf{V}^T, \tag{8}$$

where

$$\mathbf{Z} - \frac{1}{\rho}\mathbf{\Lambda} = \mathbf{U} \begin{bmatrix} \sigma_1 & \\ & \sigma_2 \end{bmatrix} \mathbf{V}^T, \tag{9}$$

and update $\mathbf{c}$ by:

$$\begin{aligned} \mathbf{c} = \underset{\mathbf{c}}{\operatorname{argmin}} \mathcal{L} = \underset{\mathbf{c}}{\operatorname{argmin}} \frac{1}{2} \sum_{p\in\mathcal{P}} \left\| \mathbf{Z}\mathbf{Y}_p\mathbf{c} + \mathbf{t} - \mathbf{w}_p \right\|_2^2, \\ \text{s.t. } \|\mathbf{c}\|_0 = 1, \end{aligned} \tag{10}$$

which can be solved by Orthogonal Matching Pursuit (OMP) [15] efficiently, and update $\mathbf{t}$ by:

$$\mathbf{t} = \underset{\mathbf{t}}{\operatorname{argmin}} \mathcal{L} = \frac{\sum_{p\in\mathcal{P}} \mathbf{w}_p - \mathbf{Z}\mathbf{Y}_p\mathbf{c}}{|\mathcal{P}|}, \tag{11}$$

where $|\mathcal{P}|$ indicate the number of visible points, and update Lagrangian multipliers and penalty factor by:

$$\mathbf{\Lambda} = \mathbf{\Lambda} + (\mathbf{M} - \mathbf{Z}), \quad \rho = \min(\rho * \tau, \rho_{max}), \tag{12}$$

where τ is the updating rate, and ρ_{max} is the upper bound of ρ. The whole algorithm is shown in Algorithm 1.

Algorithm 1: Landmark registration by ADMMs

Initialize variables: $\mathbf{M} = \begin{bmatrix}1,0,0\\0,1,0\end{bmatrix}, t = \begin{bmatrix}0\\0\end{bmatrix}, \mathbf{Z} = \mathbf{M}, \mathbf{\Lambda} = \mathbf{0}$;

while *not converge* **do**

  Update $\mathbf{Z}$ by Equation 7;

  Update $\mathbf{M}$ by Equation 8;

  Update $\mathbf{c}$ by Equation 10;

  Update $\mathbf{t}$ by Equation 11;

  Update lagrangian multiplier $\mathbf{\Lambda}$ and penalty ρ;

end

6. Silhouette Fitting

After landmark registration, we now have a rough estimate of camera position and a selected node which is considered to be "closest" to the given image. By treating this node as center, we extend an LDC subgraph to undertake our silhouette fitting step. We assume that a certain segmentation method has been executed so that the given image $\mathbf{I}$ has been segmented into foreground and background which means the silhouette is known. The main idea of this step is to simultaneously refine camera position and estimate combination weights such that as many vertices of the created model as possible are projected inside the silhouette.

In particular, by denoting the center as $\mathcal{S}_c$, Ω as the index set pointing to the nodes in the LDC subgraph, the deformable model $\mathcal{S}(\mathbf{V}, \mathbf{E})$ can be represented by Equation 3 with landmark positions $\mathbf{X} = \omega_c \mathbf{X}_c + \sum_{i\in\Omega} \omega_i \mathbf{X}_c^i$, where $\mathbf{X}_c, \mathbf{X}_c^i$ are the 3D position of landmarks on model $\mathcal{S}_c$ and $\mathcal{S}_c^i$ respectively. The silhouette fitting problem can then be written as minimizing the energy function, with respect to the camera estimate $\mathbf{R}, \mathbf{t}$[3] and combination weights ω's:

$$
\begin{aligned}
&E(\mathbf{R}, \mathbf{t}, \boldsymbol{\omega}) \\
&= \frac{1}{2} \sum_{p\in\mathcal{P}} \Big\| s\mathbf{R}\big(\omega_c[\mathbf{X}_c]_p + \sum_{i\in\Omega} \omega_i[\mathbf{X}_c^i]_p\big) + \mathbf{t} - \mathbf{w}_p \Big\|_2^2 + \\
&\mu \sum_{p=1}^{N} \mathbf{C}\Big(s\mathbf{R}\big(\omega_c[\mathbf{V}_c]_p + \sum_{i\in\Omega} \omega_i[\mathbf{V}_c^i]_p\big) + \mathbf{t}\Big) + \frac{\gamma}{2} \sum_{i\in\Omega} \omega_i^2,
\end{aligned} \tag{13}
$$

where $[\cdot]_p$ is the p-th column of the matrix, N is the number of vertices, and μ, γ are penalty weights. The first term is the reprojection error as in Equation 4, the second term penalizes the vertices whose projection is outside of silhouette, where $\mathbf{C}$ is the Chamfer distance map from the segmentation of $\mathbf{I}$, and the third term is an ℓ_2 regularization.

By using exponential map to depict the change of rotation, we can identically express the energy function as

$$
\begin{aligned}
&E(\boldsymbol{\xi}, \mathbf{t}, \boldsymbol{\omega}) \\
&= \frac{1}{2} \sum_{p\in\mathcal{P}} \Big\| s\mathbf{R}e^{[\boldsymbol{\xi}]_\times}\big(\omega_c[\mathbf{X}_c]_p + \sum_{i\in\Omega} \omega_i[\mathbf{X}_c^i]_p\big) + \mathbf{t} - \mathbf{w}_p \Big\|_2^2 + \\
&\mu \sum_{p=1}^{N} \mathbf{C}\Big(s\mathbf{R}e^{[\boldsymbol{\xi}]_\times}\big(\omega_c[\mathbf{V}_c]_p + \sum_{i\in\Omega} \omega_i[\mathbf{V}_c^i]_p\big) + \mathbf{t}\Big) + \frac{\gamma}{2} \sum_{i\in\Omega} \omega_i^2,
\end{aligned} \tag{14}
$$

where $[\cdot]_\times$ is the skew-symmetric matrix. To minimize the proposed energy, we use gradient descent.

The gradient of the energy with respect to ω_i's is

$$
\begin{aligned}
&\sum_{p\in\mathcal{P}} \Big(s\mathbf{R}\big(\omega_c[\mathbf{X}_c]_p + \sum_{i\in\Omega} \omega_i[\mathbf{X}_c^i]_p\big) + \mathbf{t} - \mathbf{w}_p\Big)^T s\mathbf{R}[\mathbf{X}_c^i]_p + \\
&\mu \sum_{p=1}^{N} \nabla\mathbf{C}^T s\mathbf{R}[\mathbf{V}_c^i]_p + \gamma\omega_i,
\end{aligned} \tag{15}
$$

where $\nabla\mathbf{C}$ is the derivative of Chamfer distance.

The gradient of the energy with respect to $\boldsymbol{\xi}$ is

$$
\begin{aligned}
&\sum_{p\in\mathcal{P}} \Big(s\mathbf{R}\big(\omega_c[\mathbf{X}_c^*]_p + \sum_{i\in\Omega} \omega_i[\mathbf{X}_c^i]_p\big) + \mathbf{t} - \mathbf{w}_p\Big)^T \\
&\Big(s\mathbf{R}\frac{\partial[\boldsymbol{\xi}]_\times}{\xi_j}\big(\omega_c[\mathbf{X}_c]_p + \sum_{i\in\Omega} \omega_i[\mathbf{X}_c^i]_p\big)\Big) + \\
&\mu \sum_{p=1}^{N} \nabla\mathbf{C}^T\Big(s\mathbf{R}\frac{\partial[\boldsymbol{\xi}]_\times}{\xi_j}\big(\omega_c[\mathbf{V}_c]_p + \sum_{i\in\Omega} \omega_i[\mathbf{V}_c^i]_p\big)\Big).
\end{aligned} \tag{16}
$$

The gradient of the energy with respect to translation $\mathbf{t}$ is

$$
\sum_{p\in\mathcal{P}} \Big(s\mathbf{R}\big(\omega_c[\mathbf{X}_c]_p + \sum_{i\in\Omega} \omega_i[\mathbf{X}_c^i]_p\big) + \mathbf{t} - \mathbf{w}_p\Big) + \mu \sum_{p=1}^{N} \nabla\mathbf{C}. \tag{17}
$$

We use backtracking to decide step sizes in each iteration[4].

7. Experiments

We evaluate our method by three metrics: (i) 2D landmark reprojection error, (ii) pose error, and (iii) structure error. More specifically, the 2D landmark reprojection error measures the accuracy of reprojected landmarks, which is computed as mean Euclidean distance between the projected landmarks of estimated dense model and the labeled landmarks on the image plane: $\text{err}_{2d} = \frac{1}{|\mathcal{P}|} \sum_{p\in\mathcal{P}} \Big\| s\mathbf{R}\big(\omega_c[\mathbf{X}_c]_p + \sum_{i\in\Omega} \omega_i[\mathbf{X}_c^i]_p\big) + \mathbf{t} - \mathbf{w}_p \Big\|_2$, following the same notations in Section 6. The pose error measures the accuracy of estimated pose (rotation): $\text{err}_{rot} = \|\mathbf{R}^* - \mathbf{R}\|_F$, where $\mathbf{R}^*, \mathbf{R}$ are the estimated and ground truth rotation matrices respectively. The structure error measures the quality of reconstructed dense model against the ground truth, following the same metrics shown in Equation 1.

[3] The scale in camera position is absorbed by ω's

[4] To be general, we initialized $\omega_c = 1$ and $\omega_i = 0$ for $i \in \Omega$

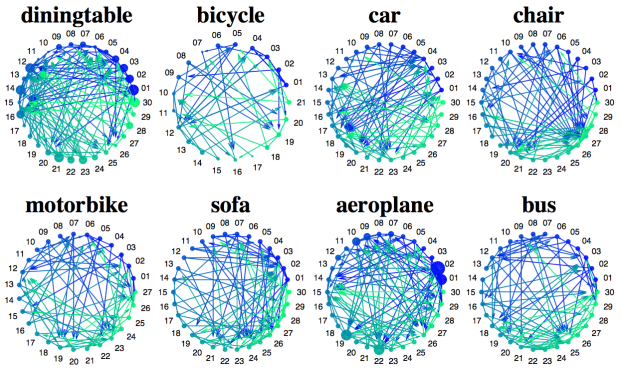

Figure 3. Visualization of learned LDC graphs from ShapeNet dataset for eight object categories.

As described in previous sections, our method consists of two steps where the silhouetting fitting involves three key components: (i) ℓ_2 regularization, (ii) refining camera position, and (iii) estimating the weights of linear combination. To show the performance boost introduced by the silhouetting fitting against the landmark registration with/without each components, extensive experiments are conducted using both synthetic and real images[5]. To our best knowledge, the most related work [8, 16] reconstruct 3D dense models purely from 2D images without any access to 3D CAD models. Therefore, a direct comparison of our method against theirs is not fair. However, from visual evaluation[6] (Figure 6), one can clearly observe the deformation of CAD models and their detailed geometry, which outperforms the state-of-the-art dense reconstruction algorithms.

7.1. Learned LDC Graphs

To show the generalization of the proposed method in various object categories, we learn LDC graphs for eight categories: diningtable, bicycle, car, chair, motorbike, sofa, aeroplane, and bus. To learn the graph, we randomly sample approximately 30 CAD models from the ShapeNet dataset [4] in each object category and manually annotate landmarks on each CAD models [7].

The learned LDC graphs are shown in Figure 3. One can observe that the density of connection varies significantly among different object categories, *e.g.* bicycle is the sparsest and diningtable is the densest. This connection density actually reflects the intracategory variations. Moreover, by visualizing the size of nodes in the graph proportional to the number of edges starting from that node, one can see that in some categories, like aeroplane, some nodes have an obviously larger size against others. This implies that these categories are more likely to share the same basic structures which is consistent to the common sense that aeroplanes have similar structure (wings in the middle of body, *etc.*) due to the same functionality.

[5]The full noise performance analysis is in the supplementary material.

[6]The videos showing 360 degree views and how dense models deform to fit silhouette are released on the GitHub page.

[7]The annotated CAD models and annotation tools will be released to public on our GitHub page.

7.2. Synthetic Experiments

We first evaluate the performance of our method using synthetic images projected by weak-perspective cameras. To generate these synthetic images, we visit all CAD models used as ground truth in PASCAL3D+ dataset [19]. By randomly generating weak-perspective camera positions, we project these CAD models into the image plane and estimate the corresponding segmentation and landmark positions. The results of this experiment is shown in Figure 4, demonstrating the performance increased by silhouette fitting and dense model combination. This evaluation shows that the silhouette fitting step with all components not only creates a deformable dense model closer to actual object geometry by LDC graph but also balances well between camera refinement and model combination.

7.3. Pascal3D+

To evaluate the performance of our framework over perspective projection and missing landmarks, we apply our proposed method to reconstruct 3D dense models of the PASCAL3D+ [19] natural images. For evaluation, we utilize the ground truth camera position, CAD models, and their annotated landmarks associated with the dataset to compute the pose, structure, and reprojection errors. The results are summarized in Figure 5 and Table 1. For all these

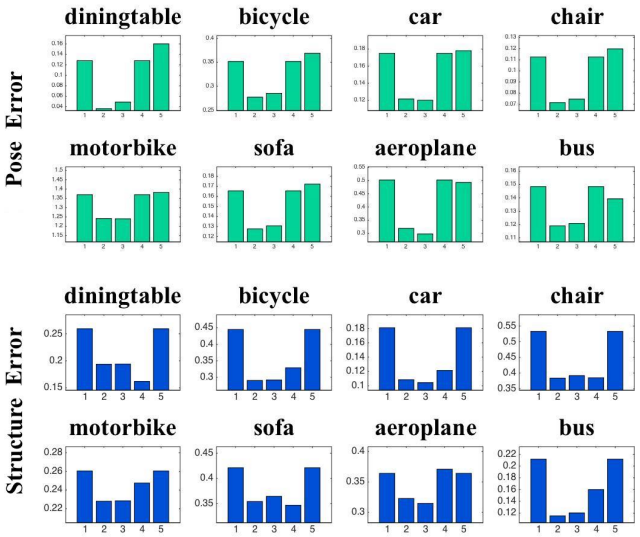

Figure 4. Evaluating our method using synthetic images in terms of pose error (top), and structure error (bottom). The x-axis shows the results of (1) landmark registration, (2) silhouette fitting with all components, (3) silhouette fitting without ℓ_2 regularization, (4) silhouette fitting without camera refinement, and (5) silhouette fitting without dense model combination.

	Component	din-ingtable	bicycle	car	chair	motor-bike	sofa	aero-plane	bus
Pose Error	LR	0.2227	**0.3216**	**0.2484**	0.1964	0.8674	0.3430	0.4527	**0.1699**
	Full SF	**0.1948**	0.3944	0.2777	**0.1858**	**0.8037**	**0.2682**	0.3507	0.2148
	SF-l2	0.1966	0.4036	0.2842	0.1901	0.8138	0.2918	**0.3401**	0.2116
	SF-Cam	0.2227	0.3216	0.2484	0.1964	0.8674	0.3430	0.4527	0.1699
	SF-Ome	0.2434	0.4081	0.3009	0.1917	0.9070	0.3495	0.5355	0.2198
Struct Error	LR	1.2936	0.3424	0.2541	0.3316	0.1954	0.4838	0.3709	**0.0998**
	Full SF	**0.6441**	**0.3314**	**0.2004**	**0.3046**	**0.1830**	**0.3872**	0.3098	0.1197
	SF-l2	0.7912	0.3409	0.2012	0.3130	0.1934	0.4831	0.3058	0.1164
	SF-Cam	0.9265	0.3479	0.2217	0.3137	0.1862	0.4486	**0.2997**	0.1267
	SF-Ome	1.2936	0.3424	0.2541	0.3236	0.1954	0.4839	0.3709	0.0998
Reproj Error	LR	30	**32**	45	21	**33**	29	**39**	**25**
	Full SF	23	41	40	19	33	22	44	38
	SF-l2	**22**	42	**39**	**17**	33	**20**	43	38
	SF-Cam	25	39	43	19	35	24	45	38
	SF-Ome	29	40	52	22	36	29	46	37

Table 1. Pose, structure and reprojection error obtained by landmark registration (LR), silhouette fitting with all components (Full SF), silhouette fitting without ℓ_2 regularization (SF-l2), silhouette fitting without refining camera position (SF-Cam), and silhouette fitting without model deformation (SF-Ome) for eight object categories.

eight categories except "bus", our method with full components achieves the best performance in terms of dense 3D models, camera positions and balancing between them.

Some qualitative results are shown in Figure 6. The models estimated by landmark registration (the second and forth columns) shows that landmark registration itself is not sufficient to select a correct model or estimate precise pose due to the limited information offered by sparse points. Further, the comparison between models reconstructed by landmark registration and silhouette fitting in both 2D and 3D shows that our proposed method not only refines the object pose but also deforms the dense model to be consistent with 2D images, *e.g.* changing the length-width ratio of table, diminishing arms of a chair, and even warping a van into a sedan. We also compare our results again volumetric representation which is directly voxelized from ground truth 3D models. It is clear to see that the volumetric representation suffers from low resolution and is too coarse to represent any finer geometry. As shown in Figure 5 and Figure 6, our method fails in "bus" category. This is caused by either the strong perspective effect or the high occlusion of buses in PASCAL3D+ image set. Note that our method would fail if the large amount of object silhouette is broken or invisible.

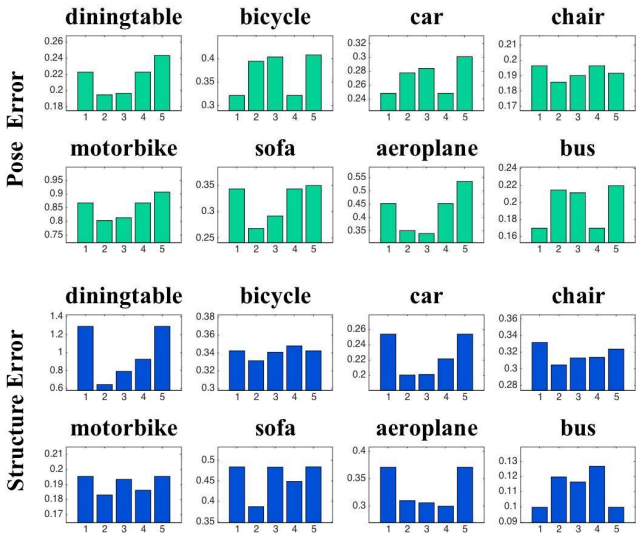

Figure 5. Evaluating our method using PASCAL3D+ natural images in terms of pose error (top), and structure error (bottom). The x-axis shows the results of (1) landmark registration, (2) silhouette fitting with all components, (3) silhouette fitting without ℓ_2 regularization, (4) silhouette fitting without camera refinement, and (5) silhouette fitting without dense model combination.

8. Conclusion

In this paper, we demonstrated that a deformable, dense 3D model can be inferred only from local dense correspondence. Our method eschews the need for global correspondence prior. In this regard, we proposed a two-step strategy using landmarks and silhouette to reconstruct a deformable dense model from a single image. Impressive results were shown on both synthetic and real-world natural images.

Acknowledgement

This material is based upon work supported by the National Science Foundation under Grant No.1526033.

Figure 6. Visual evaluation of estimated 3D models by our proposed methods for eight object categories including diningtable, bicycle, car, chair, motorbike, sofa, aeroplane, and bus. We denote green nodes as labelled landmarks, red nodes as projected landmarks, and red number as landmarks index. The columns here shows respectively (1) the input images with landmarks and silhouette, (2) projection of dense model estimated by landmark registration, (3) projection of dense model estimated by silhouette fitting with all components, (4) the dense model estimated by landmark registration, (5) the dense model estimated by silhouette fitting with all components, (6) ground truth, and (7) volumetric representation of ground truth. The failure case is shown in red in the last row. Best viewed in color.

References

[1] S. Agarwal, Y. Furukawa, N. Snavely, I. Simon, B. Curless, S. M. Seitz, and R. Szeliski. Building rome in a day. *Communications of the ACM*, 54(10):105–112, 2011. 1

[2] B. Amberg, S. Romdhani, and T. Vetter. Optimal step nonrigid icp algorithms for surface registration. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2007. 3

[3] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning*, 3(1):1–122, 2011. 4

[4] A. X. Chang, T. A. Funkhouser, L. J. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su, J. Xiao, L. Yi, and F. Yu. Shapenet: An information-rich 3d model repository. *CoRR*, abs/1512.03012, 2015. 4, 6

[5] C. B. Choy, D. Xu, J. Gwak, K. Chen, and S. Savarese. 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. *European Conference on Computer Vision (ECCV)*, 2016. 2

[6] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham. Active shape models-their training and application. *Computer vision and image understanding*, 61(1):38–59, 1995. 2

[7] R. Garg, A. Roussos, and L. Agapito. Dense variational reconstruction of non-rigid surfaces from monocular video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1272–1279, 2013. 2

[8] A. Kar, S. Tulsiani, J. Carreira, and J. Malik. Category-specific object reconstruction from a single image. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1966–1974. IEEE, 2015. 2, 6

[9] C. Kong and S. Lucey. Prior-less compressible structure from motion. *Computer Vision and Pattern Recognition (CVPR)*, 2016. 2

[10] C. Kong, R. Zhu, H. Kiani, and S. Lucey. Structure from category: a generic and prior-less approach. *International Conference on 3DVision (3DV)*, 2016. 2

[11] V. Ramakrishna, T. Kanade, and Y. Sheikh. Reconstructing 3d human pose from 2d image landmarks. In *European Conference on Computer Vision*, pages 573–586. Springer, 2012. 1

[12] D. J. Rezende, S. Eslami, S. Mohamed, P. Battaglia, M. Jaderberg, and N. Heess. Unsupervised learning of 3d structure from images. *arXiv preprint arXiv:1607.00662*, 2016. 2

[13] C. Russell, J. Fayad, and L. Agapito. Dense non-rigid structure from motion. In *2012 Second International Conference on 3D Imaging, Modeling, Processing, Visualization & Transmission*, pages 509–516. IEEE, 2012. 2

[14] P. Tanskanen, K. Kolev, L. Meier, F. Camposeco, O. Saurer, and M. Pollefeys. Live metric 3d reconstruction on mobile phones. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 65–72, 2013. 1

[15] J. A. Tropp and A. C. Gilbert. Signal recovery from random measurements via orthogonal matching pursuit. *Information Theory, IEEE Transactions on*, 53(12):4655–4666, 2007. 4

[16] S. Vicente, J. Carreira, L. Agapito, and J. Batista. Reconstructing pascal voc. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 41–48. IEEE, 2014. 2, 6

[17] C. Wang, Y. Wang, Z. Lin, A. L. Yuille, and W. Gao. Robust estimation of 3d human poses from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2361–2368, 2014. 1

[18] J. Wu, T. Xue, J. J. Lim, Y. Tian, J. B. Tenenbaum, A. Torralba, and W. T. Freeman. Single image 3d interpreter network. *European Conference on Computer Vision (ECCV)*, 2016. 1, 2

[19] Y. Xiang, R. Mottaghi, and S. Savarese. Beyond pascal: A benchmark for 3d object detection in the wild. In *IEEE Winter Conference on Applications of Computer Vision*, pages 75–82. IEEE, 2014. 6

[20] X. Zhou, S. Leonardos, X. Hu, and K. Daniilidis. 3d shape estimation from 2d landmarks: A convex relaxation approach. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4447–4455, 2015. 1, 2

[21] X. Zhou, M. Zhu, S. Leonardos, K. Derpanis, and K. Daniilidis. Sparseness meets deepness: 3d human pose estimation from monocular video. *arXiv preprint arXiv:1511.09439*, 2015. 1