2016 Fourth International Conference on 3D Vision

Structure from Category: A Generic and Prior-less Approach

Chen Kong Rui Zhu Hamed Kiani Simon Lucey
Carnegie Mellon University
{chenk, rz1, hamedk, slucey}@andrew.cmu.edu

Abstract

Inferring the motion and shape of non-rigid objects from images has been widely explored by Non-Rigid Structure from Motion (NRSfM) algorithms. Despite their promising results, they often utilize additional constraints about the camera motion (e.g. *temporal order) and the deformation of the object of interest, which are not always provided in real-world scenarios. This makes the application of NRSfM limited to very few deformable objects* (e.g. *human face and body). In this paper, we propose the concept of Structure from Category (SfC) to reconstruct 3D structure of generic objects solely from images with no shape and motion constraint* (i.e. *prior-less). Similar to the NRSfM approaches, SfC involves two steps: (i) correspondence, and (ii) inversion. Correspondence determines the location of key points across images of the same object category. Once established, the inverse problem of recovering the 3D structure from the 2D points is solved over an augmented sparse shape-space model. We validate our approach experimentally by reconstructing 3D structures of both synthetic and natural images, and demonstrate the superiority of our approach to the state-of-the-art low-rank NRSfM approaches.*

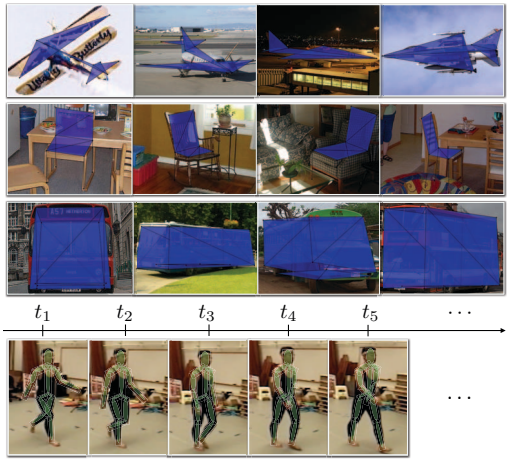

Figure 1. (top) Structure from Category. The first row shows four instances drawn from the visual object category "aeroplane". Each instance in isolation represents a rigid aeroplane, however, the space of all 3D shapes describing aeroplane category is non-rigid (same for chairs and buses). The goal of S*f*C is to reconstruct the 3D structure of generic objects with no a priori assumption and constraint on the object shape and camera motion. 3D shapes of these samples inferred by S*f*C is shown in blue. (bottom) NRS*f*M, however, imposes constraints (such as temporal order in this figure) to deal with its ill-conditioned objective. This limits the application of NRS*f*M for real-world unconstrained situations.

1. Introduction

Reconstructing the shape and motion of objects from images is a central goal of computer vision, which is generally known as *Structure from Motion* (S*f*M) in the vision literature [1, 16]. S*f*M has been broadly explored for rigid objects whose 3D shape is fixed between images. However, this is not the case for many objects in the real world with time-varying shapes, including human, animal and deformable objects. This has encouraged the vision community to introduce non-rigid S*f*M algorithms [5, 7, 25, 10, 11], assuming that an object's shape may vary over the time.

It is, however, well-noted in the literature that non-rigid S*f*M is an inherently ill-posed problem, if arbitrary deformations are allowed. In such case, the solution is not unique and might be very sensitive to initialization and the noise of key points correspondence. This has been mainly addressed by imposing additional constraints on the object shape and camera motion, which, however, comes at the cost of poor scalability for larger problems when the number of shape bases increases [17].

This paper introduces the method of Structure from Category (S*f*C) to infer 3D structures of images from the same object category. S*f*C is built upon the insight that the shape space describing an object category (*e.g.* aeroplane) is inherently non-rigid, even though individual instances of the category may be rigid, Fig. 1 (top). In other words, the shape of each instance can be modelled as a deformation from its category's general shape. Based on this observation, we frame S*f*C through an augmented sparse shape-

978-1-5090-5407-7/16 $31.00 © 2016 IEEE
DOI 10.1109/3DV.2016.38

space model that estimates the 3D shape of an object as a sparse linear combination of a set of rotated shape bases.

The proposed S*f*C is a *generic* and *prior-less* 3D reconstruction algorithm. Unlike current NRS*f*M methods which are mainly limited to very few deformable objects (*e.g*. human body and face), S*f*C can be generally applied on any object category, due to the non-rigid assumption of objects shape space. Moreover, all parameters including shape bases, sparse coefficients and (scaled) camera motion are jointly learned though an iterative manner, with *no* constraint on camera motion, 3D shape structure, temporal order and deformation patterns (prior-less). Being generic and prior-less with no learning procedure in advance offers robust large scale 3D reconstruction for unseen object images and categories.

Contribution. In this paper we make the following contributions:

- We introduce the concept of Structure from Category (S*f*C) to infer 3D structure of images from the same object category with no additional constraint and assumption about the shape space and camera motion.
- We formulate S*f*C over an augmented sparse shape-space model, and we demonstrate that the proposed S*f*C objective can be optimized in an iterative manner using the Alternating Direction Method of Multipliers (ADMM) algorithm [4].
- We conduct extensive experiments to evaluate our proposed framework on both synthetic images and challenging PASCAL3D+ natural images. The results demonstrate the superior performance of S*f*C for the task of 3D reconstruction, compared to well-know NRS*f*M methods.

2. Related Work

The field of computer vision has made significant progress for inferring 3D shape and camera motion of rigid scenes/objects over the last three decades, with rigid SfM algorithms now capable of reconstructing entire cities using large-scale photo collections [2], and real-time visual SLAM on embedded and mobile devices [15]. Current rigid S*f*M, however, assumes that the 3D structure of the object/scene of interest does not change over the time, an assumption that limits the application of this class of approaches for deformable objects. This led to the development of non-rigid S*f*M algorithms were carefully tailored for elastic objects with time-varying 3D shapes [7, 25, 11].

Despite promising results of these family of approaches, NRS*f*M algorithms are inherently ill-conditioned, since the structure can vary between images, resulting in more variables than equations. The main focus of existing NRS*f*M works has been addressing this drawback by introducing additional priors and constraints to make the NRS*f*M problem less ambiguous. Notable examples of additional priors include: basis [21], temporal [3, 17, 25], articulation [13, 18], and camera motion [10] constraints. These priors, although useful for making the NRSfM problem tractable, considerably limit its applicability to scenarios where these constraints do not hold. For example, many of the aforementioned priors/constraints do not hold for commonly used object recognition datasets such as ImageNet [8] or PASCAL VOC [9], which contain images taken from disjoint points in space and time.

The recent work of Dai *et al*. [7] were devoted to answer this question: what is the minimal set of constraints/priors required to find a unique solution to the problem? In this work, they proposed an approach to NRS*f*M assuming that the non-rigid 3D structure could be represented by a linear subspace of *known* rank K, with no more prior knowledge and additional constraint. However, the rank K is bounded by the number of points and frames, which, in most cases, may drastically degrade the performance of this work confronting object categories with large intra-class variations. To address this drawback, Kong and Lucey [11] proposed a block-sparse coding approach solely assuming that the non-rigid 3D structure could be represented by an over-complete dictionary sparsely with no more prior knowledge. Although this work is capable of handling highly deformable objects, due to the non-convex characteristics of dictionary learning procedure, it is sensitive to initialization and the noise of key points detection.

There are few methods developed for 3D reconstruction of object categories purely from large-scale 2D image datasets [19, 24]. Vicente *et al*. [19] proposed a novel strategy for obtaining dense per-object 3D reconstructions using only ground-truth segmentations and a small set of annotated key points. The approach first initializes camera positions using rigid S*f*M, and then applies a novel visual hull reconstruction method using both the hand-labeled key points and figure-ground segmentations. This work was able to successfully infer 3D structures from large-scale 2D image dataset with minimal amounts of hand labelled ground-truth, however, through a number of simplifying assumptions that inhibit the future progression of model-based methods applied to large-scale image sets. The authors initialize the camera estimation using rigid S*f*M even though, as discussed earlier, 3D shapes adhering to the same object category will, in general, form a non-rigid 3D set. They also assume that a subset of corresponding key points across all images within the same object category are manually annotated by "a few clicks per image" [19] in order to apply rigid S*f*M. This is impractical across a dataset containing millions of images in thousands of object categories.

Very recently, Zhou *et al*. [24] proposed an augmented

sparse shape-space model to estimate the 3D shape of an object from a single image. They assume that a set of key points within the query image is annotated, and a huge set of training shapes (*e.g.* thousands of annotated 3D CAD models) describing the 3D structure of the target category is given to learn a shape dictionary (*i.e.* set of shape bases). The shape dictionary together with annotated key points of the image will be used over the proposed augmented sparse shape-space model to estimate the object's 3D shape. Since the sparse model is non-convex, they utilized the convex relaxation of orthogonality constraints to convert the sparse objective into a convex spectral-norm regularized linear inverse problem with globally optimal solution [24]. This method performs well if adequate amount of training 3D shapes is available to learn a well-generalized shape dictionary. However, this situation rarely happens especially for large scale 3D reconstruction of object categories with huge intra-class and deformation variations.

3. S*f*C: Formulation

Inspired by the augmented sparse shape-space model [24], the 3D shape of instance f, $\mathbf{S}_f \in \mathbb{R}^{3\times P}$, can be well-approximated as a linear combination of a set of L rotated 3D *shape bases* $\{\mathbf{B}_l\}_{l=1}^{L}$:

$$\mathbf{S}_f = \sum_{l=1}^{L} c_{fl}\mathbf{R}_{fl}\mathbf{B}_l, \tag{1}$$

where $\mathbf{B}_l \in \mathbb{R}^{3\times P}$, represented by the location of P key points in the 3D space, describe the object's shape space. $\mathbf{R}_{fl} \in \mathbb{R}^{3\times 3}$ and c_{fl} respectively refer to the rotation matrix and the coefficient of the l-th shape base and the f-th instance.

Given a set of F instances of the same object category, Eq(1) can be written as :

$$\begin{bmatrix}\mathbf{S}_1\\ \vdots \\ \mathbf{S}_F\end{bmatrix} = \begin{bmatrix} c_{11}\mathbf{R}_{11} & \cdots & c_{1L}\mathbf{R}_{1L} \\ \vdots & \vdots & \vdots \\ c_{F1}\mathbf{R}_{F1} & \cdots & c_{FL}\mathbf{R}_{FL}\end{bmatrix}\begin{bmatrix}\mathbf{B}_1\\ \vdots \\ \mathbf{B}_L\end{bmatrix}. \tag{2}$$

The projection of $\{\mathbf{S}_f\}_{f=1}^{F}$ into the image plane, $\{\mathbf{W}_f\}_{f=1}^{F}$, is computed by:

$$\begin{aligned}\begin{bmatrix}\mathbf{W}_1\\ \vdots \\ \mathbf{W}_F\end{bmatrix} &= \begin{bmatrix}\mathbf{K}\mathbf{S}_1\\ \vdots \\ \mathbf{K}\mathbf{S}_F\end{bmatrix} + \begin{bmatrix}\mathbf{T}_1\\ \vdots \\ \mathbf{T}_F\end{bmatrix} \\ &= \begin{bmatrix} c_{11}\mathbf{K}\mathbf{R}_{11} & \cdots & c_{1L}\mathbf{K}\mathbf{R}_{1L} \\ \vdots & \vdots & \vdots \\ c_{F1}\mathbf{K}\mathbf{R}_{F1} & \cdots & c_{FL}\mathbf{K}\mathbf{R}_{FL}\end{bmatrix}\begin{bmatrix}\mathbf{B}_1\\ \vdots \\ \mathbf{B}_L\end{bmatrix} + \begin{bmatrix}\mathbf{T}_1\\ \vdots \\ \mathbf{T}_F\end{bmatrix}\end{aligned} \tag{3}$$

where we denote translation by $\mathbf{T}_f$, and projection matrix by $\mathbf{K}$. $\mathbf{W}_f \in \mathbb{R}^{2\times P}$ contains the 2D locations of P key points projected into the image plane. We consider weak-perspective cameras, which is a reasonable assumption for objects whose variation in depth is small compared to their distance from the camera, *i.e.* $\mathbf{K} = \begin{bmatrix}1 & 0 & 0\\ 0 & 1 & 0\end{bmatrix}$.

Denoting $\mathbf{M}_{fl} = c_{fl}\mathbf{K}\mathbf{R}_{fl}$, Eq(3) can be written as:

$$\begin{bmatrix}\mathbf{W}_1\\ \vdots \\ \mathbf{W}_F\end{bmatrix} = \begin{bmatrix} \mathbf{M}_{11} & \cdots & \mathbf{M}_{1L} \\ \vdots & \vdots & \vdots \\ \mathbf{M}_{F1} & \cdots & \mathbf{M}_{FL}\end{bmatrix}\begin{bmatrix}\mathbf{B}_1\\ \vdots \\ \mathbf{B}_L\end{bmatrix} + \begin{bmatrix}\mathbf{T}_1\\ \vdots \\ \mathbf{T}_F\end{bmatrix} \tag{4}$$

and more concisely in the matrix form as,

$$\mathbf{W} = \mathbf{M}\mathbf{B} + \mathbf{T} \tag{5}$$

The goal of S*f*C is to jointly compute $\mathbf{M}$ (projected rotation matrix), $\mathbf{B}$ (shape bases), and $\mathbf{T}$ (translation), using $\mathbf{W}$ (location of corresponding key points in a set of 2D images). This is performed by minimizing the *projection error* subject to the scaled orthogonality constraint on each $\mathbf{M}_{fl}$ and the sparsity constraint on the number of shape bases activated for each instance, which is framed as:

$$\begin{aligned}\min_{\mathbf{M},\mathbf{B},\mathbf{T}} \quad & \frac{1}{2}\left\|\mathbf{\Gamma}\odot\left(\mathbf{M}\mathbf{B}+\mathbf{T}\right)-\mathbf{W}\right\|_F^2 + \lambda\|\mathbf{C}\|_1 \\ \text{s.t.} \quad & \mathbf{M}_{fl}\mathbf{M}_{fl}^T = c_{fl}^2\mathbf{I}_2,\ f=1,...,F,\ l=1,...,L, \\ & \|\mathbf{B}_l\|_F = 1,\ f=1,...,F,\end{aligned} \tag{6}$$

where $\mathbf{C} = [c_{fl}]$ and $\|\mathbf{C}\|_1$ computes the summation of ℓ_1-norm of each row in $\mathbf{C}$. $\|.\|_F$ denotes the Frobenius norm of a matrix, and $\mathbf{\Gamma}$ is a binary matrix that encodes the visibility (1) and occlusion (0) of each key point. The objective in Eq(6) is non-convex due to the multiplication of $\mathbf{M}$ and $\mathbf{B}$ and the orthogonality constraint on each $\mathbf{M}_{fl}$. To make the problem more convex, we utilize the relaxation strategy proposed by Zhou *et al.* [24] that eliminates the orthogonality constraint by replacing it with a spectral norm regularization. In such case, Eq(6) is relaxed as:

$$\begin{aligned}\min_{\mathbf{M},\mathbf{B},\mathbf{T}} \quad & \frac{1}{2}\left\|\mathbf{\Gamma}\odot\left(\mathbf{M}\mathbf{B}+\mathbf{T}\right)-\mathbf{W}\right\|_F^2 + \lambda\sum_{l,f}\|\mathbf{M}_{fl}\|_2 \\ \text{s.t.} \quad & \|\mathbf{B}_l\|_F = 1,\ l=1,...,L,\end{aligned} \tag{7}$$

where $\|.\|_2$ here is the spectral norm of a matrix. The Alternating Direction Method of Multipliers (ADMM) [4] will be utilized to solve the objective in Eq(7).

4. S*f*C: Optimization

, Our proposed approach for solving Eq(7) involves the introduction of two auxiliary variables $\mathbf{Z}$ and $\mathbf{A}$. In this case, Eq(7) can be identically expressed as:

$$\begin{aligned}\min_{\mathbf{M},\mathbf{B},\mathbf{T},\mathbf{Z},\mathbf{A}} \quad & \frac{1}{2}\left\|\mathbf{\Gamma}\odot\left(\mathbf{ZB}+\mathbf{T}\right)-\mathbf{W}\right\|_F^2+\lambda\sum_{f,l}\|\mathbf{M}_{fl}\|_2 \\ \text{s.t.} \quad & \mathbf{M}=\mathbf{Z},\ \mathbf{A}=\mathbf{B}, \\ & \|\mathbf{A}_l\|_F=1,\ l=1,...,L.\end{aligned} \tag{8}$$

The augmented Lagrangian of Eq(8) is formulated as:

$$\begin{aligned}\mathcal{L}(\mathbf{M},\mathbf{Z},\mathbf{B},\mathbf{A},\mathbf{T},\mathbf{\Lambda},\mathbf{\Pi}) &= \frac{1}{2}\left\|\mathbf{\Gamma}\odot\left(\mathbf{ZB}+\mathbf{T}\right)-\mathbf{W}\right\|_F^2 \\ &+\lambda\sum_{f,l}\|\mathbf{M}_{fl}\|_2+\frac{\mu}{2}\left\|\mathbf{M}-\mathbf{Z}\right\|_F^2+\frac{\rho}{2}\left\|\mathbf{A}-\mathbf{B}\right\|_F^2 \\ &+\left\langle\mathbf{\Lambda},\mathbf{M}-\mathbf{Z}\right\rangle_F+\left\langle\mathbf{\Pi},\mathbf{A}-\mathbf{B}\right\rangle_F \\ \text{s.t.} \quad & \|\mathbf{A}_l\|_F=1,\ l=1,...,L,\end{aligned} \tag{9}$$

where $\mathbf{\Pi}$, $\mathbf{\Lambda}$ are Lagrangian multipliers, and μ, ρ are penalty factors to control the convergence behavior, and $<\cdot,\cdot>_F$ is Frobenius product of two matrices.

Particularly, we utilize the Alternating Direction Method of Multipliers (ADMM) to optimize Eq(9). ADMM decomposes an objective into several sub-problems, and iteratively solves them till convergence occurs [4]. We detail each of the sub-problem as follows.

Sub-problem M:

$$\begin{aligned}\mathbf{M}^* &= \operatorname{argmin}\mathcal{L}(\mathbf{M};\mathbf{Z},\mathbf{B},\mathbf{A},\mathbf{T},\mathbf{\Lambda},\mathbf{\Pi}) \\ &= \operatorname{argmin}\lambda\sum_{f,l}\|\mathbf{M}_{fl}\|_2+\frac{\mu}{2}\left\|\mathbf{M}-\mathbf{Z}\right\|_F^2+\left\langle\mathbf{\Lambda},\mathbf{M}-\mathbf{Z}\right\rangle\end{aligned} \tag{10}$$

Following [24], each $\mathbf{M}_{fl}$ can be computed by using soft-thresholding:

$$\mathbf{M}^*_{fl}=\mathcal{D}_{\lambda/\mu}\left(\mathbf{Z}_{fl}-\frac{1}{\mu}\mathbf{\Lambda}_{fl}\right) \tag{11}$$

Sub-problem Z:

$$\begin{aligned}\mathbf{Z}^* &= \operatorname{argmin} \quad \mathcal{L}(\mathbf{Z};\mathbf{M},\mathbf{B},\mathbf{A},\mathbf{T},\mathbf{\Lambda},\mathbf{\Pi}) \\ &= \operatorname{argmin} \quad \frac{1}{2}\left\|\mathbf{\Gamma}\odot\left(\mathbf{ZB}+\mathbf{T}\right)-\mathbf{W}\right\|_F^2 \\ &\qquad +\frac{\mu}{2}\left\|\mathbf{M}-\mathbf{Z}\right\|_F^2+\left\langle\mathbf{\Lambda},\mathbf{M}-\mathbf{Z}\right\rangle_F\end{aligned} \tag{12}$$

$\mathbf{Z}^*$ is updated iteratively by gradient descent several times, where the gradient is $\left(\mathbf{\Gamma}\odot\mathbf{\Gamma}\odot\left(\mathbf{ZB}+\mathbf{T}\right)-\mathbf{W}\right)\mathbf{B}^T-\mathbf{\Lambda}+\mu(\mathbf{Z}-\mathbf{M})$. If $\mathbf{\Gamma}$ is all ones (all key points are visible), we can compute $\mathbf{Z}^*$ easily by pseudo-inverse:

$$\mathbf{Z}^*=\left(\mathbf{BB}^T+\mu\mathbf{I}\right)^\dagger\left((\mathbf{W}-\mathbf{T})\mathbf{B}^T+\mathbf{\Lambda}+\mu\mathbf{M}\right) \tag{13}$$

Sub-problem B:

$$\begin{aligned}\mathbf{B}^* &= \operatorname{argmin} \quad \mathcal{L}(\mathbf{B};\mathbf{M},\mathbf{Z},\mathbf{A},\mathbf{T},\mathbf{\Lambda},\mathbf{\Pi}) \\ &= \operatorname{argmin} \quad \frac{1}{2}\left\|\mathbf{\Gamma}\odot\left(\mathbf{ZB}+\mathbf{T}\right)-\mathbf{W}\right\|_F^2 \\ &\qquad +\left\langle\mathbf{\Pi},\mathbf{A}-\mathbf{B}\right\rangle_F+\frac{\rho}{2}\left\|\mathbf{A}-\mathbf{B}\right\|_F^2\end{aligned} \tag{14}$$

Each column of $\mathbf{B}$, corresponded to each key point p, can be independently optimized as:

$$\begin{aligned}\mathbf{B}_p^* &= \operatorname{argmin} \quad \frac{1}{2}\left\|\operatorname{diag}\left(\mathbf{\Gamma}_p\right)\mathbf{ZB}_p+\mathbf{\Gamma}_p\odot\mathbf{T}_p-\mathbf{W}_p\right\|_2^2 \\ &\qquad +\left\langle\mathbf{\Pi}_p,\mathbf{A}_p-\mathbf{B}_p\right\rangle_F+\frac{\rho}{2}\left\|\mathbf{A}_p-\mathbf{B}_p\right\|_2^2\end{aligned} \tag{15}$$

We utilized a gradient descent solver to optimize Eq(15) when ρ is small (Eq(15) is poorly conditioned). Once ρ becomes big enough, we solve $\mathbf{B}_p$ directly using a least square solver. If all entries of $\mathbf{\Gamma}$ are one, *i.e.* all key points are visible, $\mathbf{B}^*$ can efficiently computed by:

$$\mathbf{B}^*=\left(\mathbf{Z}^T\mathbf{Z}+\rho\mathbf{I}\right)^\dagger\left(\mathbf{Z}^T\left(\mathbf{W}-\mathbf{T}\right)+\mathbf{\Pi}+\rho\mathbf{A}\right) \tag{16}$$

Sub-problem A:

$$\begin{aligned}\mathbf{A}^* &= \operatorname{argmin}\mathcal{L}(\mathbf{A};\mathbf{M},\mathbf{Z},\mathbf{B},\mathbf{T},\mathbf{\Lambda},\mathbf{\Pi}) \\ &= \operatorname{argmin}\left\langle\mathbf{\Pi},\mathbf{A}-\mathbf{B}\right\rangle_F+\frac{\rho}{2}\left\|\mathbf{A}-\mathbf{B}\right\|_F^2 \\ &\text{s.t.} \quad \left\|\mathbf{A}_l\right\|_F=1,\ l=1,...,L.\end{aligned} \tag{17}$$

^{F}The optimal solution for Eq(17) can be obtained as [6],

$$\mathbf{A}_l^*=\frac{\mathbf{B}_l-1/\rho\mathbf{\Pi}_l}{\left\|\mathbf{B}_l-1/\rho\mathbf{\Pi}_l\right\|_F} \tag{18}$$

Sub-problem T:

$$\begin{aligned}\mathbf{T}^* &= \operatorname{argmin}\mathcal{L}(\mathbf{T};\mathbf{M},\mathbf{Z},\mathbf{B},\mathbf{A},\mathbf{\Lambda},\mathbf{\Pi}) \\ &= \operatorname{argmin}\frac{1}{2}\left\|\mathbf{\Gamma}\odot\left(\mathbf{ZB}+\mathbf{T}\right)-\mathbf{W}\right\|_F^2.\end{aligned} \tag{19}$$

Since all columns of $\mathbf{T}\in\mathbb{R}^{2F\times P}$, $\boldsymbol{\tau}$'s, are identical, we compute a $\boldsymbol{\tau}\in\mathbb{R}^{2F\times 1}$ by minimizing the above objective:

$$\boldsymbol{\tau}^*= \operatorname{argmin}\frac{1}{2}\sum_{p=1}^{P}\left\|\mathbf{\Gamma}_p\odot\left(\mathbf{ZB}_p+\boldsymbol{\tau}\right)-\mathbf{W}_p\right\|_2^2, \tag{20}$$

and optimal $\boldsymbol{\tau}$ is computed by:

$$\boldsymbol{\tau}^*=\left(\sum_{p=1}^{P}\mathbf{W}_p-\sum_{p=1}^{P}\mathbf{\Gamma}_p\odot\mathbf{\Gamma}_p\odot\mathbf{ZB}_p\right)\oslash\left(\sum_{p=1}^{P}\mathbf{\Gamma}_p\odot\mathbf{\Gamma}_p\right) \tag{21}$$

where $\oslash$ denotes the element-wise division.

Lagrange Multiplier Update: The lagrange multipliers $\mathbf{\Pi}, \mathbf{\Lambda}$ at each iteration are updated as,

$$\begin{aligned} \mathbf{\Lambda}^{[i+1]} &= \mathbf{\Lambda}^{[i]} + \mu\big(\mathbf{M}^{[i+1]} - \mathbf{Z}^{[i+1]}\big) \\ \mathbf{\Pi}^{[i+1]} &= \mathbf{\Pi}^{[i]} + \rho\big(\mathbf{A}^{[i+1]} - \mathbf{B}^{[i+1]}\big) \end{aligned} \tag{22}$$

Penalty Update: Superlinear convergence of ADMM may be achieved by $\mu, \rho \to \infty$. In practice, we limit the value of μ, ρ to avoid poor condition and numerical errors. Specifically, we adopt the following update strategy:

$$\begin{aligned} \mu^{[i+1]} &= \min(\mu_{max}, \beta_1 \mu^{[i]}) \\ \rho^{[i+1]} &= \min(\rho_{max}, \beta_2 \rho^{[i]}) \end{aligned} \tag{23}$$

We found experimentally $\mu^{[0]} = 10^{-2}$, $\rho^{[0]} = 10^{-1}$, $\beta_1(\beta_2) = 1.1$, and $\mu_{max}(\rho_{max}) = 10^5$ to perform well.

5. Experiments

5.1. Evaluation setup

We compare the proposed method against the most notable NRS*f*M algorithms: Tomasi-Kanade factorization [16], and the state-of-the-art Dai *et al.*'s prior-less NRS*f*M method [7], in terms of reprojection and reconstruction errors. The reprojection error measures the accuracy of reprojected key points: $\frac{1}{F}\sum_{i=1}^{F} \|\mathbf{W}_i - \hat{\mathbf{W}}_i\|_F$. The reconstruction error, on the other hand, evaluates the quality of estimated 3D shapes: $\frac{1}{F}\sum_{i=1}^{F} \min_\kappa \|\mathbf{S}_i - \kappa\hat{\mathbf{S}}_i\|_F$. κ (scalar) handles the scale ambiguity in camera projection.

Extensive experiments are conducted to evaluate the performance of our framework using both synthetic and natural images. For the synthetic images, we downloaded 70 CAD models of aeroplane category from Sketchup 3D warehouse [1], and manually annotated their 3D key points. The synthetic images are simply generated by projecting random poses of these 3D models under weak-perspective camera into the image plane. The PASCAL3D+ dataset [20] is used for the natural image experiment, which consists of 12 object categories, and each category comes with a set of annotated 3D CAD models and corresponding natural images. We utilize most of images from all categories except those displaying highly occluded objects. More details of the PASCAL3D+ dataset can be found in [20].

The main differences between synthetic and PASCAL3D+ images come from the camera projection and object occlusion. We utilize random *weak*-perspective projection to generate the synthetic images of the aeroplane dataset, which follows the weak-projection assumption in this paper, whilst, the camera projection in the PASCAL3D+ is perspective. Moreover, all key points in synthetic images are visible, while, some key points in the PASCAL3D+ may be occluded by object itself or other objects.

[1] https://3dwarehouse.sketchup.com/

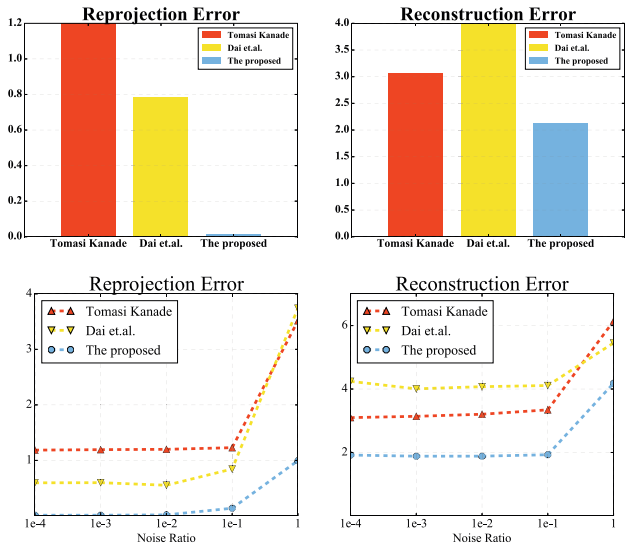

Figure 2. Comparing our method with Tomasi-Kanade [16] and Dai *et al.* [7] methods using the synthetic images. (top) The reconstruction and reprojection errors. (bottom) Noise performance.

5.2. 3D reconstruction from synthetic images

The first experiment evaluates the performance of the proposed method on synthetic images, comparing with the Tomasi-Kanade factorization [16] and Dai *et al.*'s prior-less NRS*f*M approaches [7]. The synthetic images are randomly generated from all 3D CADs of the aeroplane dataset under weak perspective projection, and these approaches are applied to reconstruct the 3D shape of each image. The predicted shapes, then, are projected into the 2D plane to compute the key points reprojection error. The result of this experiment is shown in Fig. 2 (top), demonstrating the superior performance of our method to the other approaches. This evaluation shows that the 3D shapes reconstructed by the proposed S*f*C not only represent the actual geometry of the objects in 3D space, but also preserve the objects' spatial configuration when projected in the image plane. The result also verifies the sensitivity of the low-rank factorization NRS*f*M algorithm, *e.g.* Dai *et al.*'s method in the real world uncontrolled circumstances, when the shape of an object can not be modeled by very few shape bases [19].

5.3. Noise performance

To analyse the robustness of our method against inaccurate key point detection, which is inevitable in real-world circumstances, we repeat the first experiment (using synthetic aeroplane images) with different levels of Gaussian noise added to the ground truth 2D locations. The average reconstruction and reprojection errors of ten random runs for each noise ratio is reported in Fig. 2 (bottom), showing that, compared to the other methods, the S*f*C method is more robust against inaccurate key point detections.

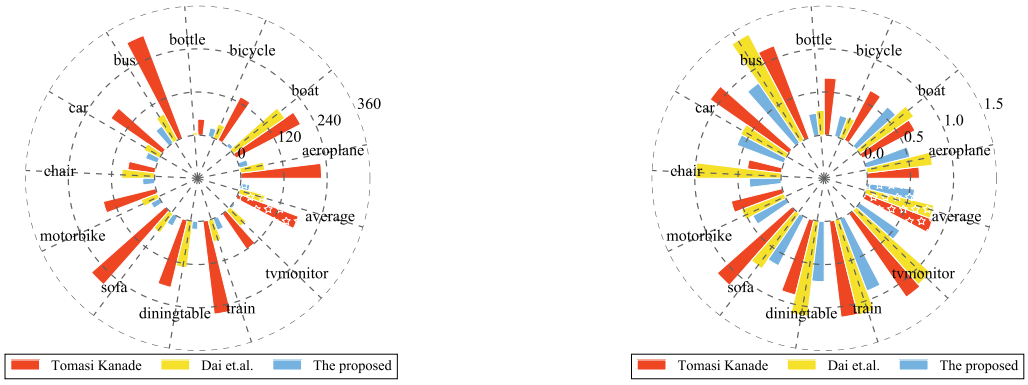

Figure 3. The reprojection (left) and reconstruction (right) performance of the proposed method, Tomasi-Kanade factorization [16] and Dai *et al.*'s method [7] on natural images (the PASCAL3D+ dataset) with ground truth key points.

5.4. 3D reconstruction of PASCAL3D+ dataset

To evaluate the performance of our framework over perspective projection and missing key points, we apply the proposed S*f*C approach to reconstruct 3D shapes of the PASCAL3D+ natural images. There is no additional shape and camera motion assumption given in this experiment, and images of all 12 object categories are taken under uncontrolled real-world circumstances. All images and their corresponding ground truth 3D CAD models are represented by a set of 2D and 3D annotated key points, respectively, which together with the predicted 3D structures and their reprojected 2D key points will be used to compute the reconstruction and reprojection errors. Since the Tomasi-Kanade factorization and Dai *et al.*'s method are not capable of handling occluded objects, we utilize the non-convex matrix completion via iterated soft thresholding [12] to predict the missing points for these approaches. This experiment is conducted over two different settings. In the first setting, we use the ground truth key points of each image provided by the PASCAL3D+. In the other setting, however, we adapt the SDM [22] approach for key point detection, and the predicted points are used for 3D reconstruction.

Using ground truth key points: The reprojection and reconstruction errors for each object category are summarized in Table 1 and showed by Fig. 3, where our approach outperforms the competitors and achieves the lowest reconstruction and reprojection error for each object category.

Using predicted key points: We adapt the Supervised Descent Method (SDM) [22], originally proposed for the task of facial landmarks alignment, to detect key points of generic objects within natural images. The main assumption of the SDM is that training samples fall into a Domain of Homogeneous Descent (DHD)[2], due to their limited pose space and appearance variation [23]. This assumption, however, is rarely valid in an object category with large intra-class appearance and pose variations that lies in multiple DHDs. To deal with this situation, we propose to employ a subset of training images with homogeneous gradient directions to train an SDM in an "on-the-fly" manner. Particularly, given a test image, we use fc_7 feature from the ConvNet [14] to retrieve its M most similar samples from training images and use them to train an SDM. The training set is generated by adding Gaussian noise to the ground truth locations. After training the SDM regressors, we run them independently from M different initializations (the ground truth landmark locations of the M retrieved samples). This returns M sets of predicted key points, which will be further pruned by the mean-shift algorithm. More details of SDM training/testing can be found in [22].

The results are shown in Fig. 4 and Table 1. For both two settings, using ground truth and predicted key points, our method achieves the best reconstruction and reprojection performance. The results also state that the performance of using ground truth key points is much better than the detected key points. Some qualitative results are shown in Fig. 5, illustrating the 3D reconstruction of two instances of each object category using ground truth key points and detected key points respectively. During the experiments, we observed that most of the failure cases are caused by severe perspective effect (*e.g.* train), missing key points (*e.g.* sofa), and inaccurate key point detection (*e.g.* chair).

6. Conclusion

In this paper, we introduce the concept of Structure from Category to reconstruct 3D shapes of generic object categories from images. We argued that 3D shapes of an object category, in general, form a non-rigid space. Thus, we formulate the method of S*f*C as a NRS*f*M algorithm. Unlike most existing NRS*f*M methods, our approach requires no additional constraint on the shape or camera motion. Instead, all shape and camera motion parameters (including shape bases) are jointly estimated through an augmented

[2] A DHD refers to optimization spaces of a function that share similar directions of gradients.

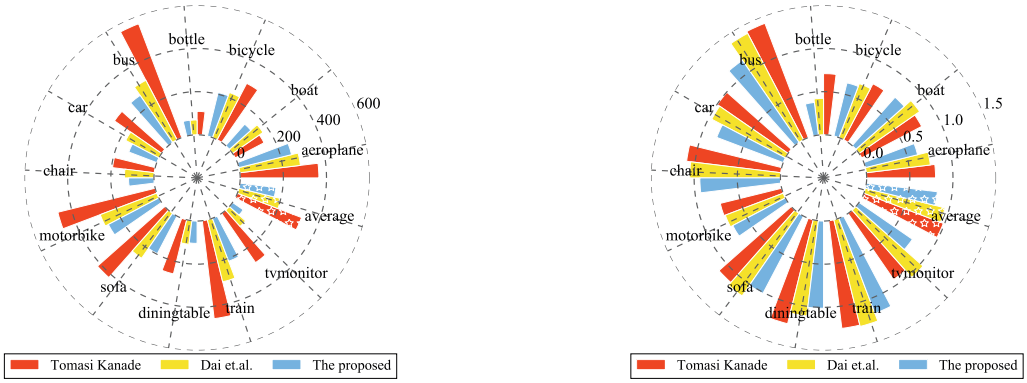

Figure 4. The reprojection (left) and reconstruction (right) performance of the proposed method, Tomasi-Kanade factorization [16] and Dai *et al.*'s method [7] on natural images (the PASCAL3D+ dataset) with detected key points.

category	key points	Reprojection Error			Reconstruction Error		
		Tomasi Kanade	Dai *et al.*	Our method	Tomasi Kanade	Dai *et al.*	Our Method
aeroplane	GT	224.3925	67.7078	**24.5695**	0.6035	0.7631	**0.5257**
	detected	364.7172	282.5179	**251.0064**	0.7986	0.7465	**0.6223**
boat	GT	202.9794	174.2009	**11.1862**	0.6892	0.7609	**0.6061**
	detected	150.7320	171.5790	**133.1670**	0.7844	0.8531	**0.7497**
bicycle	GT	135.9651	41.8621	**24.7112**	0.6490	0.2568	**0.2495**
	detected	295.2249	223.5721	**207.6959**	0.7327	0.6695	**0.6351**
bottle	GT	44.4231	6.4836	**2.8315**	0.6609	0.2865	**0.2590**
	detected	108.4824	**68.6833**	69.7238	0.7087	0.4220	**0.3812**
bus	GT	304.8072	82.1719	**56.0355**	1.1427	1.3839	**0.8396**
	detected	564.3329	311.0550	**264.9117**	1.4164	1.3924	**1.1562**
car	GT	173.6506	49.5333	**35.4720**	1.1062	0.5943	**0.5808**
	detected	265.4429	173.8730	**138.6603**	0.9959	0.9636	**0.8242**
chair	GT	75.9437	91.5107	**33.0905**	0.3958	0.9887	**0.3671**
	detected	194.7178	136.9023	**117.6726**	1.0985	1.0511	**0.9338**
motorbike	GT	150.6358	48.3516	**27.1717**	0.6096	0.5252	**0.4344**
	detected	464.8820	280.3500	**264.5549**	0.7333	0.7185	**0.6887**
sofa	GT	274.9890	64.2714	**30.0575**	1.1561	0.7727	**0.6438**
	detected	416.9723	253.0140	**196.6783**	1.1198	1.1617	**1.0126**
diningtable	GT	192.5072	130.5157	**21.9391**	0.8924	1.1084	**0.6982**
	detected	258.3700	110.2296	**103.4765**	1.2404	1.1124	**1.0107**
train	GT	260.5996	61.7900	**34.2347**	1.1215	1.1316	**0.8957**
	detected	457.0754	296.3881	**213.2750**	1.2568	1.2728	**1.1799**
tvmonitor	GT	119.8794	59.2110	**6.6706**	1.1740	1.1454	**0.5653**
	detected	277.1977	100.6167	**60.0780**	0.9307	1.0412	**0.7516**
average	GT	180.0644	73.1342	**25.6642**	0.8501	0.8098	**0.5554**
	detected	318.1790	200.7318	**168.4084**	0.9847	0.9504	**0.8288**

Table 1. Reprojection and Reconstruction errors obtained by Tomasi Kanade factorization [16], Dai *et al.*'s method [7], and our method using ground truth key points (GT) and detected key points (detected).

sparse shape-space model. Since all the key points are automatically detected by an adapted SDM method, our framework can be applied for large scale 3D reconstruction with no limitation on the number of object categories, number of images per category, object shape and camera motion. We demonstrated the proposed S*f*C method outperformed the state-of-the-art NRS*f*M methods, using both synthetic and natural images.

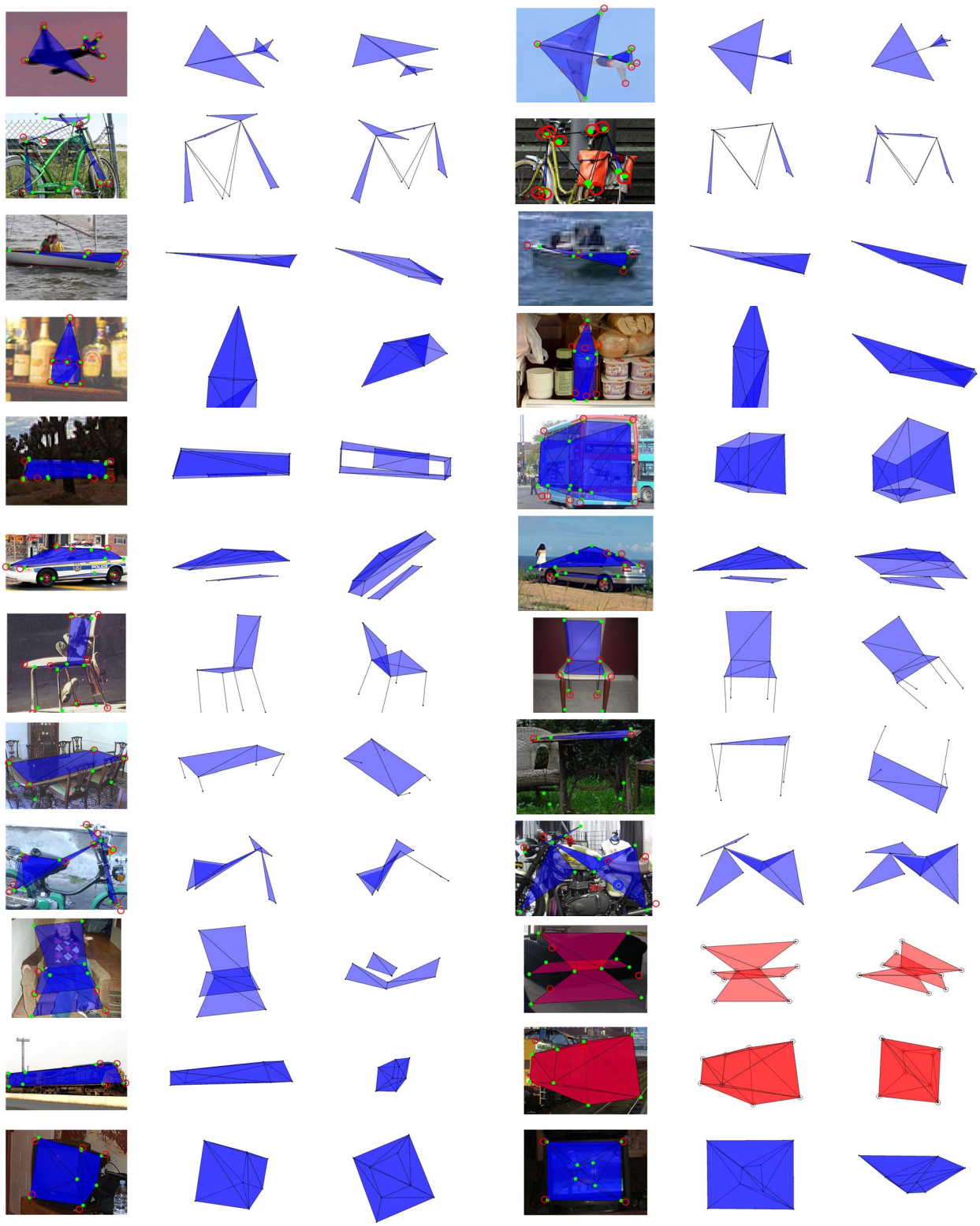

Figure 5. Visual evaluation of estimated structures for every category including aeroplane, bicycle, boat, bottle, bus, car, chair, diningtable, motorbike, sofa, train, and tvmonitor. The first 3 columns use ground truth key points, while the last 3 columns use detected key points. In each triplet columns, the left columns show the images, projection of estimated 3D shapes, projection of estimated landmarks (green), and the ground truth landmarks (red, some are missing due to occlusion); The middle ones show the estimated 3D shapes in the same viewpoint as camera; The right ones show a new viewpoint of the estimated 3D shapes. Two failure cases are shown in red. 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] S. Agarwal, N. Snavely, I. Simon, S. M. Seitz, and R. Szeliski. Building rome in a day. In *2009 IEEE 12th international conference on computer vision*, pages 72–79. IEEE, 2009. 2

[3] I. Akhter, Y. Sheikh, S. Khan, and T. Kanade. Nonrigid structure from motion in trajectory space. In *Advances in neural information processing systems*, pages 41–48, 2009. 2

[4] 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. 2, 3, 4

[5] C. Bregler, A. Hertzmann, and H. Biermann. Recovering non-rigid 3d shape from image streams. In *Computer Vision and Pattern Recognition, 2000. Proceedings. IEEE Conference on*, volume 2, pages 690–696. IEEE, 2000. 1

[6] H. Bristow, A. Eriksson, and S. Lucey. Fast convolutional sparse coding. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 391–398. IEEE, 2013. 4

[7] Y. Dai, H. Li, and M. He. A simple prior-free method for non-rigid structure-from-motion factorization. *International Journal of Computer Vision*, 107(2):101–122, 2014. 1, 2, 5, 6, 7

[8] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. IEEE, 2009. 2

[9] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010. 2

[10] P. F. Gotardo and A. M. Martinez. Kernel non-rigid structure from motion. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 802–809. IEEE, 2011. 1, 2

[11] C. Kong and S. Lucey. Prior-less compressible structure from motion. *Computer Vision and Pattern Recognition (CVPR)*, 2016. 1, 2

[12] A. Majumdar and R. K. Ward. Some empirical advances in matrix completion. *Signal Processing*, 91(5):1334–1338, 2011. 6

[13] H. S. Park and Y. Sheikh. 3d reconstruction of a smooth articulated trajectory from a monocular image sequence. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 201–208. IEEE, 2011. 2

[14] H. Su, C. R. Qi, Y. Li, and L. J. Guibas. Render for cnn: Viewpoint estimation in images using cnns trained with rendered 3d model views. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2686–2694, 2015. 6

[15] 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. 2

[16] C. Tomasi and T. Kanade. Shape and motion from image streams under orthography: a factorization method. *International Journal of Computer Vision*, 9(2):137–154, 1992. 1, 5, 6, 7

[17] L. Torresani, A. Hertzmann, and C. Bregler. Nonrigid structure-from-motion: Estimating shape and motion with hierarchical priors. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 30(5):878–892, 2008. 1, 2

[18] J. Valmadre, Y. Zhu, S. Sridharan, and S. Lucey. Efficient articulated trajectory reconstruction using dynamic programming and filters. In *Computer Vision–ECCV 2012*, pages 72–85. Springer, 2012. 2

[19] 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, 5

[20] 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. 5

[21] J. Xiao, J. Chai, and T. Kanade. A closed-form solution to non-rigid shape and motion recovery. *International Journal of Computer Vision*, 67(2):233–246, 2006. 2

[22] X. Xiong and F. De la Torre. Supervised descent method and its applications to face alignment. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 532–539, 2013. 6

[23] X. Xiong and F. De la Torre. Global supervised descent method. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2664–2673, 2015. 6

[24] 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. 2, 3, 4

[25] Y. Zhu, D. Huang, F. D. L. Torre, and S. Lucey. Complex non-rigid motion 3d reconstruction by union of subspaces. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 1542–1549. IEEE, 2014. 1, 2