

Deep Non-Rigid Structure from Motion

Chen Kong

Carnegie Mellon University

chenk@cs.cmu.edu

Simon Lucey

Carnegie Mellon University

slucey@cs.cmu.edu

Abstract

Current non-rigid structure from motion (NRSfM) algorithms are mainly limited with respect to: (i) the number of images, and (ii) the type of shape variability they can handle. This has hampered the practical utility of NRSfM for many applications within vision. In this paper we propose a novel deep neural network to recover camera poses and 3D points solely from an ensemble of 2D image coordinates. The proposed neural network is mathematically interpretable as a multi-layer block sparse dictionary learning problem, and can handle problems of unprecedented scale and shape complexity. Extensive experiments demonstrate the impressive performance of our approach where we exhibit superior precision and robustness against all available state-of-the-art works by an order of magnitude. We further propose a quality measure (based on the network weights) which circumvents the need for 3D ground-truth to ascertain the confidence we have in the reconstruction.

1. Introduction

Building an AI capable of inferring the 3D structure and pose of an object from a single image is a problem of immense importance. Training such a system using supervised learning requires a large number of labeled images – how to obtain these labels is currently an open problem for the vision community. Rendering [27] is problematic as the synthetic images seldom match the appearance and geometry of the objects we encounter in the real-world. Hand annotation is preferable, but current strategies rely on associating the natural images with an external 3D dataset (e.g. ShapeNet [8], ModelNet [32]), which we refer to as 3D supervision. If the 3D shape dataset does not capture the variation we see in the imagery, it is inherently ill-posed.

Non-Rigid Structure from Motion (NRSfM) offers computer vision a way out of this quandary – by recovering the pose and 3D structure of an object category *solely* from hand annotated 2D landmarks with no need for 3D supervision. Classically [6], the problem of NRSfM has been applied to objects that move non-rigidly over time such as

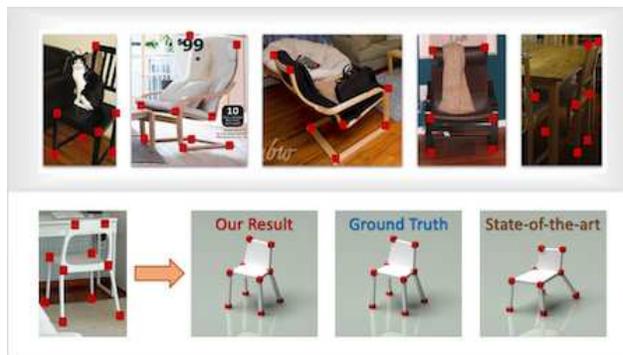


Figure 1: In this paper, we want to reconstruct 3D shapes solely from a sequence of annotated images—shown on top—with no need of 3D ground truth. Our proposed hierarchical sparse coding model and corresponding deep solution outperforms state-of-the-art by an order of magnitude.

the human body and face. But NRSfM is not restricted to non-rigid objects; it can equally be applied to rigid objects whose object categories deform non-rigidly [19, 2, 30]. Consider, for example, the five objects in Figure 1 (top), instances from the visual object category “chair”. Each object in isolation represents a rigid chair, but the set of all 3D shapes describing “chair” is non-rigid. In other words, each object instance can be modeled as a deformation from its category’s general shape.

NRSfM is well-noted in literature as an ill-posed problem due to the non-rigidity. This has been mainly addressed by imposing additional shape priors, e.g. low rank [6, 10], union-of-subspaces [36, 2], and block-sparsity [18, 19]. However, low rank is only applicable to simple non-rigid objects with limited deformations and union-of-subspaces relies heavily on frame clustering which is still an open problem. Block-sparsity, where each shape can be represented by at most K bases out of L , is considered as one of the most promising assumptions in terms of modeling broad shape variations. This is because sparsity can be thought of as a union of $\binom{L}{K}$ subspaces constraint and an over-complete dictionary could be utilized. However, it was noted by Kong *et al.* [18], that searching for the best sub-

space out of $\binom{L}{K}$ is extremely costly and not robust. Based on this observation we propose a novel shape prior that employs hierarchical sparse coding. The introduction of additional layers, in comparison to classical single-layer sparse coding, provides a mechanism for controlling the number of active subspaces. The ability of hierarchical sparse coding to provide a model that is both expressible and robust for general NRSfM forms the central thesis of our paper.

Contributions:

- We propose a novel shape prior based on hierarchical sparse coding and demonstrate that the 2D projections under orthogonal cameras can be represented by the hierarchical dictionaries in a block sparse way.
- We designed a deep neural network to approximately solve the proposed hierarchical block sparse model and show how the network architecture is derived from a classical sparse coding algorithm.
- Finally, extensive experiments are conducted using various datasets. Both quantitative and qualitative results demonstrate our superior performance, outperforming all state-of-the-arts in the order of magnitude.

2. Related Work

Low-rank NRSfM: In rigid structure from motion, the rank of 3D structure is fixed as three [29] since 3D shapes remain the same between frames. Based on this insight, Bregler *et al.* [6] advocated that non-rigid 3D structure could be represented by a linear subspace of low rank. Dai *et al.* [10] developed this prior by proving that low-rank assumption itself is sufficient to address the ill-posedness of NRSfM with no need of additional priors. The low-rank assumption has also been applied temporally [3, 15] – 3D point trajectories can be represented by pre-defined (*e.g.* DCT) or learned bases. Although exhibiting impressive performance, the low-rank assumption has a major drawback. The rank is strictly limited by the number of points and frames (whichever is smaller [10]). This makes low-rank NRSfM infeasible if we want to solve large-scale problems with complex shape variations when the number of points is substantially smaller than the number of frames.

Union-of-subspaces NRSfM: Inspired by an intuition that complex non-rigid deformations could be clustered into a sequence of simple motions, Zhu *et al.* [36] proposed to model non-rigid 3D structure by a union of local subspaces. This was later extended to spatial-temporal domain [1] and applied to rigid object category reconstruction [2]. The major difficulty of union-of-subspaces is how to effectively cluster shape deformations purely from 2D observations and how to estimate affinity matrix when the number of frame is large *e.g.* more than tens of thousand frames.

Sparse NRSfM: Sparse prior [18, 35, 19] is more generic

than union-of-subspaces since it is equivalent to the union of all possible local subspaces. One obvious advantages of this is the large number of subspaces enables the effective modeling of a broader set of 3D structures. However, the sheer number of subspaces that can be entertained by the sparsity prior is its fundamental drawback. Since there are so many possible subspaces to choose from, the approach is sensitive to noise, dramatically limiting its applicability to “real-world” NRSfM problems. In this paper we want to leverage the elegance and expressibility of the sparsity prior without suffering from its inherent sensitivity to noise.

3. Background

Sparse dictionary learning can be considered as an unsupervised learning task and divided into two sub-problems: (i) dictionary learning, and (ii) sparse code recovery. Let us consider sparse code recovery problem, where we estimate a sparse representation \mathbf{z} for a measurement vector \mathbf{x} given the dictionary \mathbf{D} *i.e.*

$$\min_{\mathbf{z}} \|\mathbf{x} - \mathbf{D}\mathbf{z}\|_2^2 \quad \text{s.t.} \quad \|\mathbf{z}\|_0 < \lambda, \tag{1}$$

where λ related to the trust region controls the sparsity of recovered code. One classical algorithm to recover the sparse representation is Iterative Shrinkage and Thresholding Algorithm (ISTA) [11, 26, 5]. ISTA iteratively executes the following two steps with $\mathbf{z}^{[0]} = \mathbf{0}$:

$$\mathbf{v} = \mathbf{z}^{[i]} - \alpha \mathbf{D}^T (\mathbf{D}\mathbf{z}^{[i]} - \mathbf{x}), \tag{2}$$

$$\mathbf{z}^{[i+1]} = \underset{\mathbf{u}}{\operatorname{argmin}} \frac{1}{2} \|\mathbf{u} - \mathbf{v}\|_2^2 + \tau \|\mathbf{u}\|_1, \tag{3}$$

which first uses the gradient of $\|\mathbf{x} - \mathbf{D}\mathbf{z}\|_2^2$ to update $\mathbf{z}^{[i]}$ in step size α and then finds the closest sparse solution using an ℓ_1 convex relaxation. It is well known in literature that the second step has a closed-form solution using the soft thresholding operator. Therefore, ISTA can be summarized as the following recursive equation:

$$\mathbf{z}^{[i+1]} = h_{\tau}(\mathbf{z}^{[i]} - \alpha \mathbf{D}^T (\mathbf{D}\mathbf{z}^{[i]} - \mathbf{x})), \tag{4}$$

where h_{τ} is a soft thresholding operator and τ is related to λ for controlling sparsity.

Recently, Papyan [25] proposed to use ISTA and sparse coding to reinterpret feed-forward neural networks. They argue that feed-forward passing a single-layer neural network $\mathbf{z} = \operatorname{ReLU}(\mathbf{D}^T \mathbf{x} - b)$ can be considered as one iteration of ISTA when $\mathbf{z} \geq 0, \alpha = 1$ and $\tau = b$. Based on this insight, the authors extend this interpretation to feed-forward neural network with n layers

$$\begin{aligned} \mathbf{z}_1 &= \operatorname{ReLU}(\mathbf{D}_1^T \mathbf{x} - b_1) \\ \mathbf{z}_2 &= \operatorname{ReLU}(\mathbf{D}_2^T \mathbf{z}_1 - b_2) \\ &\vdots \\ \mathbf{z}_n &= \operatorname{ReLU}(\mathbf{D}_n^T \mathbf{z}_{n-1} - b_n) \end{aligned} \tag{5}$$

as executing a sequence of single-iteration ISTA, serving as an approximate solution to the multi-layer sparse coding problem: find $\{\mathbf{z}_i\}_{i=1}^n$, such that

$$\begin{aligned} \mathbf{x} &= \mathbf{D}_1 \mathbf{z}_1, & \|\mathbf{z}_1\|_0 &< \lambda_1, \mathbf{z}_1 \geq 0, \\ \mathbf{z}_1 &= \mathbf{D}_2 \mathbf{z}_2, & \|\mathbf{z}_2\|_0 &< \lambda_2, \mathbf{z}_2 \geq 0, \\ &\vdots, & \vdots \\ \mathbf{z}_{n-1} &= \mathbf{D}_n \mathbf{z}_n, & \|\mathbf{z}_n\|_0 &< \lambda_n, \mathbf{z}_n \geq 0, \end{aligned} \quad (6)$$

where the bias terms $\{b_i\}_{i=1}^n$ (in a similar manner to τ) are related to $\{\lambda_i\}_{i=1}^n$, adjusting the sparsity of recovered code. Furthermore, they reinterpret back-propagating through the deep neural network as learning the dictionaries $\{\mathbf{D}_i\}_{i=1}^n$. This connection offers a novel breakthrough for understanding DNNs. In this paper, we extend this to the block sparse scenario and apply it to solving our NRSfM problem.

4. Deep Non-Rigid Structure from Motion

Under orthogonal projection, NRSfM deals with the problem of factorizing a 2D projection matrix $\mathbf{W} \in \mathbb{R}^{p \times 2}$ as the product of a 3D shape matrix $\mathbf{S} \in \mathbb{R}^{p \times 3}$ and camera matrix $\mathbf{M} \in \mathbb{R}^{3 \times 2}$. Formally,

$$\mathbf{W} = \mathbf{S}\mathbf{M}, \quad (7)$$

$$\mathbf{W} = \begin{bmatrix} u_1 & v_1 \\ u_2 & v_2 \\ \vdots & \vdots \\ u_p & v_p \end{bmatrix}, \mathbf{S} = \begin{bmatrix} x_1 & y_1 & z_1 \\ x_2 & y_2 & z_2 \\ \vdots & \vdots & \vdots \\ x_p & y_p & z_p \end{bmatrix}, \mathbf{M}^T \mathbf{M} = \mathbf{I}_2, \quad (8)$$

where $(u_i, v_i), (x_i, y_i, z_i)$ are the image and world coordinates of the i -th point. The goal of NRSfM is to recover simultaneously the shape \mathbf{S} and the camera \mathbf{M} for each projection \mathbf{W} in a given set \mathbb{W} of 2D landmarks. In a general NRSfM including SfC, this set \mathbb{W} could contain deformations of a non-rigid object or various instances from an object category.

4.1. Modeling via multi-layer sparse coding

To alleviate the ill-posedness of NRSfM and also guarantee sufficient freedom on shape variation, we propose a novel prior assumption on 3D shapes via multi-layer sparse coding: The vectorization of \mathbf{S} satisfies

$$\begin{aligned} \mathbf{s} &= \mathbf{D}_1 \boldsymbol{\psi}_1, & \|\boldsymbol{\psi}_1\|_0 &< \lambda_1, \boldsymbol{\psi}_1 \geq 0, \\ \boldsymbol{\psi}_1 &= \mathbf{D}_2 \boldsymbol{\psi}_2, & \|\boldsymbol{\psi}_2\|_0 &< \lambda_2, \boldsymbol{\psi}_2 \geq 0, \\ &\vdots, & \vdots \\ \boldsymbol{\psi}_{n-1} &= \mathbf{D}_n \boldsymbol{\psi}_n, & \|\boldsymbol{\psi}_n\|_0 &< \lambda_n, \boldsymbol{\psi}_n \geq 0, \end{aligned} \quad (9)$$

where $\mathbf{D}_1 \in \mathbb{R}^{3p \times k_1}, \mathbf{D}_2 \in \mathbb{R}^{k_1 \times k_2}, \dots, \mathbf{D}_n \in \mathbb{R}^{k_{n-1} \times k_n}$ are hierarchical dictionaries. In this prior, each non-rigid

shape is represented by a sequence of hierarchical dictionaries and corresponding sparse codes. Each sparse code is determined by its lower-level neighbor and affects the next-level. Clearly this hierarchy adds more parameters, and thus more freedom into the system. We now show that it paradoxically results in a more constrained global dictionary and sparse code recovery.

More constrained code recovery: In a classical single dictionary system, the constraint on the representation is element-wise sparsity. Further, the quality of its recovery entirely depends on the quality of the dictionary. In our multi-layer sparse coding model, the optimal code not only minimizes the difference between measurements \mathbf{s} and $\mathbf{D}_1 \boldsymbol{\psi}_1$ along with sparsity regularization $\|\boldsymbol{\psi}_1\|_0$, but also satisfies constraints from its subsequent representations. This additional joint inference imposes more constraints on code recovery, helps to control the uniqueness and therefore alleviates its heavy dependency on the dictionary quality.

More constrained dictionary: When all equality constraints are satisfied, the multi-layer sparse coding model degenerates to a single dictionary system. From Equation 9, by denoting $\mathbf{D}^{(l)} = \prod_{i=1}^l \mathbf{D}_i$, it is implied that $\mathbf{s} = \mathbf{D}_1 \mathbf{D}_2 \dots \mathbf{D}_n \boldsymbol{\psi}_n = \mathbf{D}^{(n)} \boldsymbol{\psi}_n$. However, this differs from other single dictionary models [36, 37, 18, 19, 34] in terms that a unique structure is imposed on $\mathbf{D}^{(n)}$ [28]. The dictionary $\mathbf{D}^{(n)}$ is composed by simpler atoms hierarchically. For example, each column of $\mathbf{D}^{(2)} = \mathbf{D}_1 \mathbf{D}_2$ is a linear combination of atoms in \mathbf{D}_1 , each column of $\mathbf{D}^{(3)} = \mathbf{D}^{(2)} \mathbf{D}_3$ is a linear combination of atoms in $\mathbf{D}^{(2)}$ and so on. Such a structure results in a more constrained global dictionary and potentially leads to higher quality with lower mutual coherence [14].

4.2. Multi-layer block sparse coding

Given the proposed multi-layer sparse coding model, we now build a conduit from the proposed shape code $\{\boldsymbol{\psi}_i\}_{i=1}^k$ to the 2D projected points. From Equation 9, we reshape vector \mathbf{s} to a matrix $\mathbf{S} \in \mathbb{R}^{p \times 3}$ such that $\mathbf{S} = \mathbf{D}_1^\# (\boldsymbol{\psi}_1 \otimes \mathbf{I}_3)$, where \otimes is Kronecker product and $\mathbf{D}_1^\# \in \mathbb{R}^{p \times 3k_1}$ is a reshape of \mathbf{D}_1 [10]. From linear algebra, it is well known that $\mathbf{A}\mathbf{B} \otimes \mathbf{I} = (\mathbf{A} \otimes \mathbf{I})(\mathbf{B} \otimes \mathbf{I})$ given three matrices \mathbf{A}, \mathbf{B} , and identity matrix \mathbf{I} . Based on this lemma, we can derive that

$$\begin{aligned} \mathbf{S} &= \mathbf{D}_1^\# (\boldsymbol{\psi}_1 \otimes \mathbf{I}_3), & \|\boldsymbol{\psi}_1\|_0 &< \lambda_1, \boldsymbol{\psi}_1 \geq 0, \\ \boldsymbol{\psi}_1 \otimes \mathbf{I}_3 &= (\mathbf{D}_2 \otimes \mathbf{I}_3)(\boldsymbol{\psi}_2 \otimes \mathbf{I}_3), & \|\boldsymbol{\psi}_2\|_0 &< \lambda_2, \boldsymbol{\psi}_2 \geq 0, \\ &\vdots, & \vdots \\ \boldsymbol{\psi}_{n-1} \otimes \mathbf{I}_3 &= (\mathbf{D}_n \otimes \mathbf{I}_3)(\boldsymbol{\psi}_n \otimes \mathbf{I}_3), & \|\boldsymbol{\psi}_n\|_0 &< \lambda_n, \boldsymbol{\psi}_n \geq 0. \end{aligned} \quad (10)$$

Further, from Equation 7, by right multiplying the camera matrix $\mathbf{M} \in \mathbb{R}^{3 \times 2}$ to the both sides of Equation 10 and

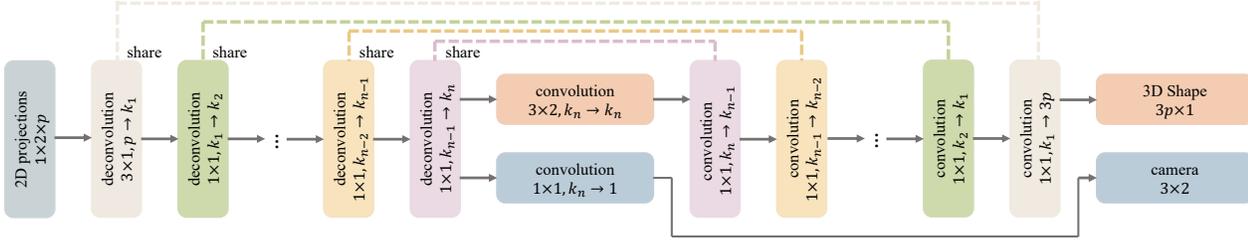


Figure 2: Deep NRSfM architecture. The network can be divided into two parts: encoder and decoder that are symmetric and share convolution kernels (*i.e.* dictionaries). The symbol $a \times b, c \rightarrow d$ refers to the operator using kernel size $a \times b$ with c input channels and d output channels.

denote $\Psi_i = \psi_i \otimes \mathbf{M}$, we obtain that

$$\begin{aligned} \mathbf{W} &= \mathbf{D}_1^\sharp \Psi_1, & \|\Psi_1\|_0^{(3 \times 2)} &< \lambda_1, \\ \Psi_1 &= (\mathbf{D}_2 \otimes \mathbf{I}_3) \Psi_2, & \|\Psi_2\|_0^{(3 \times 2)} &< \lambda_2, \\ &\vdots, & \vdots & \\ \Psi_{n-1} &= (\mathbf{D}_n \otimes \mathbf{I}_3) \Psi_n, & \|\Psi_n\|_0^{(3 \times 2)} &< \lambda_n, \end{aligned} \quad (11)$$

where $\|\cdot\|_0^{(3 \times 2)}$ divides the argument matrix into blocks with size 3×2 and counts the number of active blocks. Since ψ_i has active elements less than λ_i , Ψ_i has active blocks less than λ_i , that is Ψ_i is block sparse. This derivation demonstrates that if the shape vector \mathbf{s} satisfies the multi-layer sparse coding prior described by Equation 9, then its 2D projection \mathbf{W} must be in the format of multi-layer *block* sparse coding described by Equation 11. We hereby interpret NRSfM as a hierarchical *block* sparse dictionary learning problem *i.e.* factorizing \mathbf{W} as products of hierarchical dictionaries $\{\mathbf{D}_i\}_{i=1}^n$ and block sparse coefficients $\{\Psi_i\}_{i=1}^n$.

4.3. Block ISTA and DNNs solution

Before solving the multi-layer block sparse coding problem in Equation 11, we first consider the single-layer problem:

$$\min_{\mathbf{Z}} \|\mathbf{X} - \mathbf{D}\mathbf{Z}\|_F^2 \quad \text{s.t.} \quad \|\mathbf{Z}\|_0^{(3 \times 2)} < \lambda. \quad (12)$$

Inspired by ISTA, we propose to solve this problem by iteratively executing the following two steps:

$$\begin{aligned} \mathbf{V} &= \mathbf{Z}^{[i]} - \alpha \mathbf{D}^T (\mathbf{D}\mathbf{Z}^{[i]} - \mathbf{X}), & (13) \\ \mathbf{Z}^{[i+1]} &= \underset{\mathbf{U}}{\operatorname{argmin}} \frac{1}{2} \|\mathbf{U} - \mathbf{V}\|_F^2 + \tau \|\mathbf{U}\|_{F1}^{(3 \times 2)}, & (14) \end{aligned}$$

where $\|\cdot\|_{F1}^{(3 \times 2)}$ is defined as the summation of Frobenius norm of each 3×2 block, serving as a convex relaxation of block sparsity constraint. It is derived in [13] that the second step has a closed-form solution computing each block separately by $\mathbf{Z}_j^{[i+1]} = (h_\tau(\|\mathbf{V}_j\|_F) / \|\mathbf{V}_j\|_F) \mathbf{V}_j$, where the

subscript j represents the j -th block and h_τ is a soft thresholding operator. However, soft thresholding the Frobenius norms for every block brings unnecessary computational complexity. We show in the supplementary material that an efficient approximation is $\mathbf{Z}_j^{[i+1]} = h_{b_j}(\mathbf{V}_j)$, where b_j is the threshold for the j -th block, controlling its sparsity. Based on this approximation, a single-iteration block ISTA with step size $\alpha = 1$ can be represented by :

$$\mathbf{Z} = h_{\mathbf{b}}(\mathbf{D}^T \mathbf{X}) = \operatorname{ReLU}(\mathbf{D}^T \mathbf{X} - \mathbf{b} \otimes \mathbf{1}_{3 \times 2}), \quad (15)$$

where $h_{\mathbf{b}}$ is a soft thresholding operator using the j -th element b_j as threshold of the j -th block and the second equality holds if \mathbf{Z} is non-negative.

Encoder: Recall from Section 3 that the feed-forward pass through a deep neural network can be considered as a sequence of single ISTA iterations and thus provides an approximate recovery of multi-layer sparse codes. We follow the same scheme: we first assume the multi-layer block sparse coding to be non-negative and then sequentially use single-iteration block ISTA to solve it *i.e.*

$$\begin{aligned} \Psi_1 &= \operatorname{ReLU}((\mathbf{D}_1^\sharp)^T \mathbf{W} - \mathbf{b}_1 \otimes \mathbf{1}_{3 \times 2}), \\ \Psi_2 &= \operatorname{ReLU}((\mathbf{D}_2 \otimes \mathbf{I}_3)^T \Psi_1 - \mathbf{b}_2 \otimes \mathbf{1}_{3 \times 2}), \\ &\vdots \\ \Psi_n &= \operatorname{ReLU}((\mathbf{D}_n \otimes \mathbf{I}_3)^T \Psi_{n-1} - \mathbf{b}_n \otimes \mathbf{1}_{3 \times 2}), \end{aligned} \quad (16)$$

where thresholds $\mathbf{b}_1, \dots, \mathbf{b}_n$ are learned, controlling the block sparsity. This learning is crucial because in previous NRSfM algorithms utilizing low-rank [10], subspaces [36] or compressible [18] priors, the weight given to this prior (*e.g.* rank or sparsity) is hand-selected through a cumbersome cross validation process. In our approach, this weighting is learned simultaneously with all other parameters removing the need for any irksome cross validation process. This formula composes the encoder of our proposed DNN.

Decoder: Let us for now assume that we can extract camera \mathbf{M} and regular sparse hidden code ψ_n from Ψ_n by some

functions *i.e.* $\mathbf{M} = \mathcal{F}(\Psi_n)$ and $\psi_n = \mathcal{G}(\Psi_n)$, which will be discussed in the next section. Then we can compute the 3D shape vector \mathbf{s} by:

$$\begin{aligned} \psi_{n-1} &= \text{ReLU}(\mathbf{D}_n \psi_n - \mathbf{b}'_n), \\ &\vdots \\ \psi_1 &= \text{ReLU}(\mathbf{D}_2 \psi_2 - \mathbf{b}'_2), \\ \mathbf{s} &= \mathbf{D}_1^\# \psi_1, \end{aligned} \quad (17)$$

Note we preserve the ReLU and bias term during decoding to further enforce sparsity and improve robustness. These portion forms the decoder of our DNN.

Variation of implementation: The Kronecker product of identity matrix \mathbf{I}_3 dramatically increases the time and space complexity of our approach. To eliminate it and make parameter sharing easier in modern deep learning environments (*e.g.* TensorFlow, PyTorch), we reshape the filters and features and show that the matrix multiplication in each step of the encoder and decoder can be equivalently computed via multi-channel 1×1 convolution ($*$) and transposed convolution ($*^T$) *i.e.*

$$(\mathbf{D}_1^\#)^T \mathbf{W} = \mathbf{d}_1^\# *^T \mathbf{w}, \quad (18)$$

where $\mathbf{d}_1^\# \in \mathbb{R}^{3 \times 1 \times k_1 \times p}$, $\mathbf{w} \in \mathbb{R}^{1 \times 2 \times p \cdot 1}$.

$$(\mathbf{D}_{i+1} \otimes \mathbf{I}_3)^T \Psi_i = \mathbf{d}_{i+1} *^T \Psi_i, \quad (19)$$

where $\mathbf{d}_{i+1} \in \mathbb{R}^{1 \times 1 \times k_{i+1} \times k_i}$, $\Psi_i \in \mathbb{R}^{3 \times 2 \times k_i}$.

$$\mathbf{D}_i \psi_i = \mathbf{d}_i * \psi_i, \quad (20)$$

where $\mathbf{d}_i \in \mathbb{R}^{1 \times 1 \times k_i \times k_{i-1}}$, $\psi_i \in \mathbb{R}^{1 \times 1 \times k_i}$.

Code and camera recovery: Estimating ψ_n and \mathbf{M} from Ψ_n is discussed in [18] and solved by a closed-form formula. Due to its differentiability, we could insert the solution directly within our pipeline. An alternative solution is using an approximation *i.e.* a fully connected layer connecting Ψ_n and ψ_n and a linear combination among each blocks of Ψ_n to estimate \mathbf{M} , where the fully connected layer parameters and combination coefficients are learned from data. In our experiments, we use the approximate solution and represent them via convolutions, as shown in Figure 2, for conciseness and maintaining proper dimensions. Since the approximation has no way to force the orthonormal constraint on the camera, we seek help from the loss function.

Loss function: The loss function must measure the reprojection error between input 2D points \mathbf{W} and reprojected 2D points \mathbf{SM} while simultaneously encouraging orthonormality of the estimated camera \mathbf{M} . One solution is to use

¹The filter dimension is height \times width \times # of input channel \times # of output channel. The feature dimension is height \times width \times # of channel.

Furnitures	Bed	Chair	Sofa	Table	Mean	Relative
KSTA [16]	0.069	0.158	0.066	0.217	0.128	12.19
BMM [10]	0.059	0.330	0.245	0.211	0.211	20.12
CNR [20]	0.227	0.163	0.835	0.186	0.352	33.55
NLO [12]	0.245	0.339	0.158	0.275	0.243	23.18
RIKS [17]	0.202	0.135	0.048	0.218	0.117	11.13
SPS [18]	0.971	0.946	0.955	0.280	0.788	74.96
SFC [19]	0.247	0.195	0.233	0.193	0.217	20.67
OURS	0.004	0.019	0.005	0.012	0.010	1.00

Table 1: Quantitative comparison against state-of-the-art algorithms using IKEA dataset in normalized 3D error.

spectral norm regularization of \mathbf{M} because spectral norm minimization is the tightest convex relaxation of the orthonormal constraint [34]. An alternative solution is to hard code the singular values of \mathbf{M} to be exact ones with the help of Singular Value Decomposition (SVD). Even though SVD is generally non-differentiable, the numeric computation of SVD is differentiable and most deep learning packages implement its gradients (*e.g.* PyTorch, TensorFlow). In our implementation and experiments, we use SVD to ensure the success of the orthonormal constraint and a simple Frobenius norm to measure reprojection error,

$$\text{Loss} = \|\mathbf{W} - \tilde{\mathbf{S}}\tilde{\mathbf{M}}\|_F, \quad \tilde{\mathbf{M}} = \mathbf{U}\mathbf{V}^T, \quad (21)$$

where $\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T = \mathbf{M}$ is the SVD of the camera matrix.

5. Experiments

We conduct extensive experiments to evaluate the performance of our deep solution for solving NRSfM and SfC problems. For quantitative evaluation, we follow the metric *i.e.* normalized mean 3D error, reported in [4, 10, 16, 2]. A detailed description of our architectures is in the supplementary material. Our implementation and processed data will be publicly accessible for future comparison.

5.1. SfC on IKEA furniture

We first apply our method to a furniture dataset, IKEA dataset [23, 31]. The IKEA dataset contains four object categories: bed, chair, sofa, and table. For each object category, we employ all annotated 2D point clouds and augment them with 2K ones projected from the 3D ground-truth using randomly generated orthonormal cameras². The error evaluated on real images are reported and summarized into Table 1. One can observe that our method outperforms baselines in the order of magnitude, clearly showing the superiority of our model. For qualitative evaluation, we randomly select a frame from each object category and show them in Figure 6 against ground truth and baselines. It shows that our reconstructed landmarks effectively depict

²Augmentation is utilized due to limited valid frames, because the ground-truth cameras are partially missing.

Methods	KSTA [16]		BMM [10]		CNS [20]		MUS [2]		NLO [12]		RIKS [17]		SPS [18]		SfC [19]		OURS	
Aeroplane	0.145	0.175	0.843	1.459	0.263	0.416	0.261	-	-	0.876	-	0.132	-	0.930	-	0.504	-	0.024
Bicycle	0.442	0.245	0.308	1.376	-	0.356	0.178	-	-	0.269	-	0.136	-	1.322	-	0.372	-	0.003
Bus	0.214	0.199	0.300	1.023	-	0.250	0.113	-	-	0.140	-	0.160	-	0.604	-	0.251	-	0.004
Car	0.159	0.152	0.266	1.278	0.099	0.258	0.078	-	-	0.104	-	0.097	-	0.872	-	0.282	-	0.009
Chair	0.399	0.186	0.357	1.297	-	0.170	0.210	-	-	0.146	-	0.192	-	1.046	-	0.226	-	0.007
Diningtable	0.372	0.267	0.422	1.00	-	0.170	0.264	-	-	0.109	-	0.207	-	1.050	-	0.221	-	0.060
Motorbike	0.270	0.255	0.336	0.857	-	0.457	0.222	-	-	0.432	-	0.118	-	0.986	-	0.361	-	0.002
Sofa	0.298	0.307	0.279	1.126	0.214	0.250	0.167	-	-	0.149	-	0.228	-	1.328	-	0.302	-	0.004
Average	0.287	0.223	0.388	1.178	0.192	0.291	0.186	-	-	0.278	-	0.159	-	1.017	-	0.315	-	0.014
Relative	-	15.33	-	80.76	-	19.95	-	-	-	19.09	-	10.92	-	69.74	-	21.61	-	1.00
Aeroplane	0.183	0.207	0.566	1.465	0.294	0.460	0.271	-	-	0.758	-	0.146	-	0.888	-	0.521	-	0.032
Bicycle	0.457	0.232	0.307	1.404	-	0.359	0.188	-	-	0.275	-	0.139	-	0.851	-	0.379	-	0.007
Bus	0.218	0.197	0.255	0.764	-	0.264	0.122	-	-	0.141	-	0.159	-	1.110	-	0.264	-	0.021
Car	0.164	0.139	0.161	1.744	0.122	0.265	0.093	-	-	0.105	-	0.102	-	0.804	-	0.281	-	0.010
Chair	0.396	0.203	0.258	1.197	-	0.171	0.220	-	-	0.145	-	0.193	-	1.016	-	0.223	-	0.017
Diningtable	0.383	0.249	0.358	1.105	-	0.172	0.267	-	-	0.114	-	0.227	-	1.213	-	0.222	-	0.034
Motorbike	0.290	0.227	0.299	1.117	-	0.459	0.233	-	-	0.254	-	0.125	-	0.915	-	0.351	-	0.011
Sofa	0.294	0.436	0.240	1.143	0.228	0.255	0.174	-	-	0.152	-	0.239	-	1.164	-	0.306	-	0.008
Average	0.298	0.236	0.305	1.232	0.215	0.300	0.196	-	-	0.243	-	0.166	-	0.995	-	0.318	-	0.017
Relative	-	16.90	-	88.78	-	21.49	-	-	-	17.39	-	11.90	-	71.11	-	22.78	-	1.27

Table 2: Quantitative evaluation on PASCAL3D+ dataset. We conduct experiments on both original and noisy 2D annotations, listed at the upper and lower half of table respectively. The symbol ‘-’ indicates either algorithm implementation or data is missing. The shaded columns are errors using our processed data and others are copied from Table 2 in [2]. Relative errors are computed with respect to our method, the most accurate solution, without noise perturbation. Our data and implementation will be publicly accessible for future comparison.

the 3D geometry of objects and our method is able to cover subtle geometric details.

5.2. SfC on PASCAL3D+

We then apply our method to PASCAL3D+ dataset [33] which contains twelve object categories and each category is labeled by approximately eight 3D CADs. To compare against more baselines, we follow the experiment setting reported in [2] and use the same normalized 3D error metric. We report our errors in Table 2 emphasized by shading and concatenate the numbers copied from the Table 2 in [2] for comparison. Note that the errors are not exactly reproduced even though using the same dataset and algorithm implementation, because the data preparation details are missing. However, one can clearly see that our proposed method achieves extremely accurate reconstructions with more than ten times of smaller 3D error. This large margin makes the slight difference caused by data preparation even less noticeable. It clearly demonstrates the high precision of our proposed deep neural network and also the superior robustness in noisy situations.

5.3. Large-scale NRSfM on CMU MoCap

We finally apply our method to solving the problem of NRSfM using the CMU motion capture dataset³. We randomly select 10 subjects out of 144 and for each subject

³<http://mocap.cs.cmu.edu/>

we concatenate 80% of motions to form large image collections and remain the left 20% as unseen motions for testing generalization. Note that in this experiment, each subject contains more than ten thousands of frames. We compare our method against state-of-the-art methods, summarized in Table 3. Due to huge volume of frames, KSTA [16], BMM [10], MUS [2], RIKS [17] all fail and thus are omitted in the table. We also report the normalized 3D error on unseen motions, labeled as UNSEEN. One can see that our method obtains impressive reconstruction performance and outperforms others again in every sequences. Moreover, our network also show a well generalization to unseen data which improve the effectiveness in real world applications. For qualitative evaluation, we randomly select a frame from

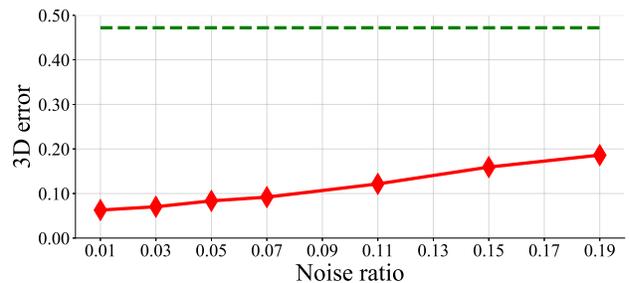


Figure 3: NRSfM with noise perturbation. The red solid line is ours while the green dashed line is CNS [20], the best performance of state-of-the-arts with *no* noise perturbation.

Subjects	01	05	18	23	64	70	102	106	123	127	Average	Relative
CNS [20]	0.613	0.657	0.541	0.603	0.543	0.472	0.581	0.636	0.479	0.644	0.577	5.66
NLO [12]	1.218	1.160	0.917	0.998	1.218	0.836	1.144	1.016	1.009	1.050	1.057	10.37
SPS [18]	1.282	1.122	0.953	0.880	1.119	1.009	1.078	0.957	0.828	1.021	1.025	10.06
OURS	0.175	0.220	0.081	0.053	0.082	0.039	0.115	0.113	0.040	0.095	0.101	1.00
UNSEEN	0.362	0.331	0.437	0.387	0.174	0.090	0.413	0.194	0.091	0.388	0.287	2.81

Table 3: Quantitative comparison on solving large-scale NRSfM problem using CMU MoCap dataset. Each subject contains more than ten thousand of frames. Due to huge volume of frames, KSTA [16], BMM [10], MUS [2], RIKS [17] all fail and thus are omitted in the table. UNSEEN refers to the errors of the motions that are not accessible during training. This is used to demonstrate the well generalization of our proposed network, which is especially important in real world applications.

each subject and render the reconstructed human skeleton in Figure 5. This visually verifies the impressive performance of our deep solution.

Robustness analysis: To analyze the robustness of our method, we re-train the neural network for Subject 70 using projected points with Gaussian noise perturbation. The results are summarized in Figure 3. The noise ratio is defined as $\|\text{noise}\|_F / \|\mathbf{W}\|_F$. One can see that the error increases slowly with adding higher magnitude of noise and when adding up to 20% noise to image coordinates, our method in red still achieves better reconstruction compared to the best baseline with no noise perturbation (in green). This experiment clearly demonstrates the robustness of our model and its high accuracy against state-of-the-art works.

Missing data: Landmarks are not always visible from the camera owing to the occlusion by other objects or itself. In the present paper, we focus on a complete measurement situation not accounting for invisible landmarks. However, thanks to recent progress in matrix completion, our method can be easily extended to missing data. Moreover, in our experiments, we observe that deep neural network shows a well tolerance of missing data. Simply setting missing 2D coordinates as zeros provides satisfactory results. Such technique is widely used in deep-learning-based depth map

reconstruction from sparse observations [9, 24, 21, 22, 7]. These two solutions make our central pipeline of DNN more easily to adapt to handling missing data.

5.4. Coherence as guide

As explained in Section 4.1, every sparse code ψ_i is constrained by its subsequent representation and thus the quality of code recovery depends less on the quality of the corresponding dictionary. However, this is not applicable to the final code ψ_n , making it least constrained with the most dependency on the final dictionary \mathbf{D}_n . From this perspective, the quality of the final dictionary measured by mutual coherence [14] could serve as a lower bound of the entire system. To verify this, we compute the error and coherence in a fixed interval during training in NRSfM experiments. We consistently observe strong correlations between 3D reconstruction error and the mutual coherence of the final dictionary. We plot this relationship in Figure 4. We thus propose to use the coherence of the final dictionary as a measure of model quality for guiding training to efficiently avoid over-fitting especially when 3D evaluation is not available. This improves the utility of our deep NRSfM in future applications without 3D ground-truth.

6. Conclusion

In this paper, we proposed multi-layer sparse coding as a novel prior assumption for representing 3D non-rigid shapes and designed an innovative encoder-decoder neural network to solve the problem of NRSfM using *no* 3D supervision. The proposed network was derived by generalizing the classical sparse coding algorithm ISTA to a block sparse scenario. The proposed network architecture is mathematically interpretable as solving a NRSfM multi-layer sparse dictionary learning problem. Extensive experiments demonstrated our superior performance against the state-of-the-art methods and our generalization to unseen data. Finally, we proposed to use the coherence of the final dictionary as a model quality measure, offering a practical way to avoid over-fitting and select the best checkpoint during training without relying on 3D ground-truth.

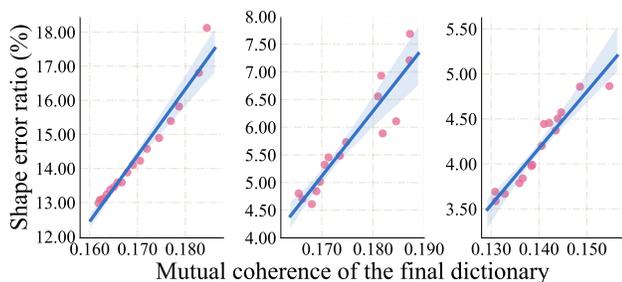


Figure 4: A scatter plot of the shape error ratio in percentage against the final dictionary coherence. A line is fitted based on the data. The left comes from subject 05, the middle from subject 18, the right from subject 64.

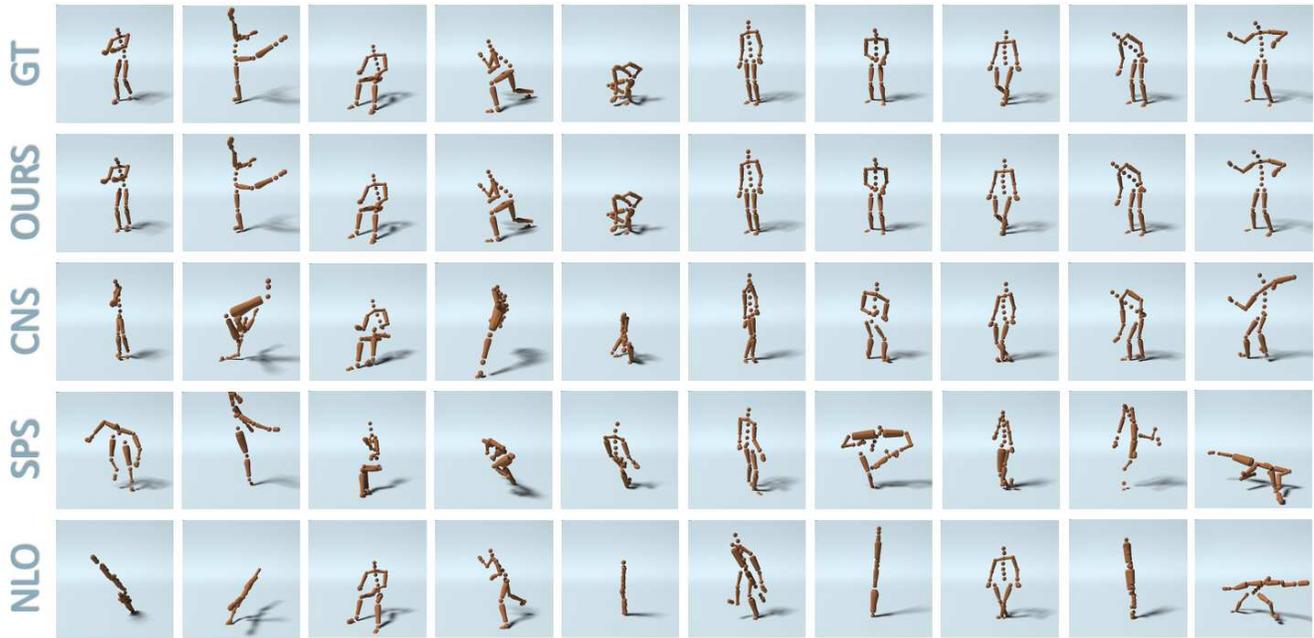


Figure 5: Qualitative evaluation on CMU Mocap dataset. Top to bottom: ground-truth, ours, CNS [20], SPS [18], NLO [12]. Each column corresponds to reconstructions of a certain frame, randomly selected from each subject. Spheres are reconstructed landmarks while bars are for visualization. 3D shapes are already aligned to the ground truth by orthonormal matrix.

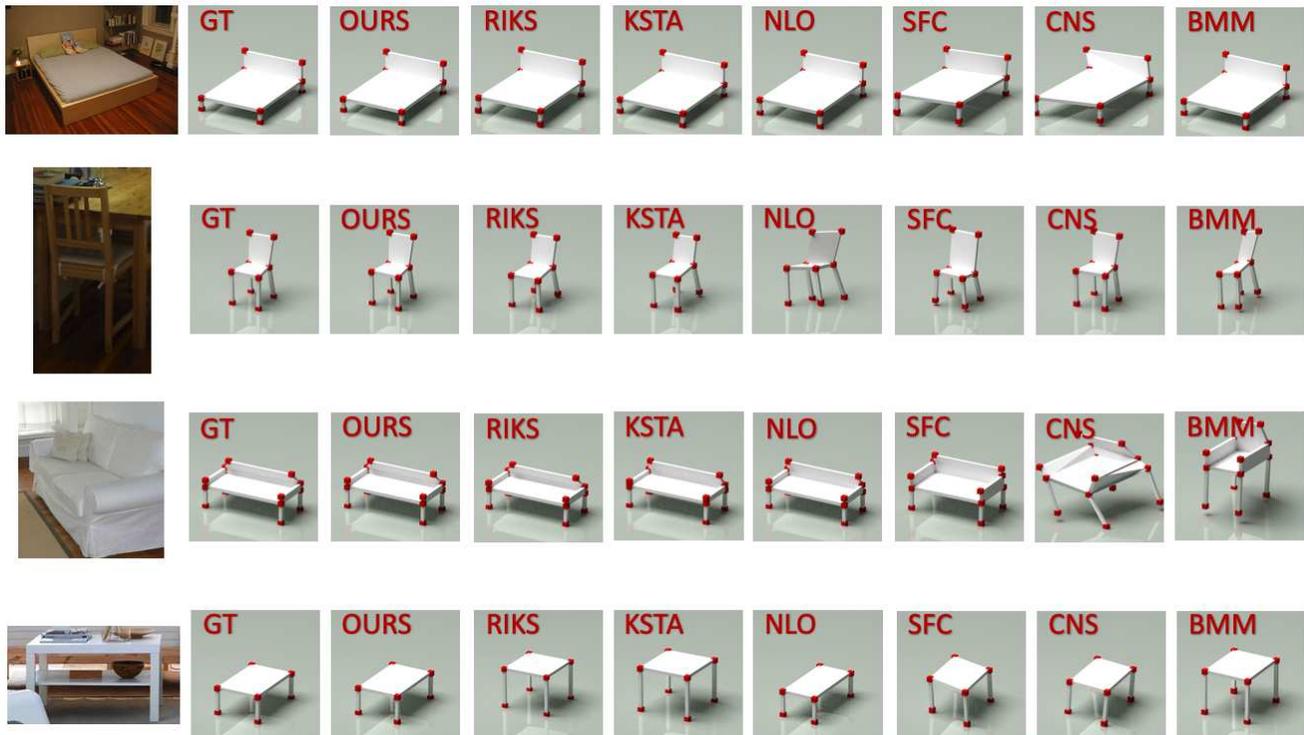


Figure 6: Qualitative evaluation on IKEA dataset. Landmarks projected by annotated cameras are omitted from images. In each rendering, red cubes are reconstructed points while the planes and bars are manually added for descent visualization. Left to right: annotated image, ground-truth, ours, RIKS [17], KSTA [16], NLO [12], SFC [19], CNS [20], BMM [10].

References

- [1] Antonio Agudo and Francesc Moreno-Noguer. Dust: Dual union of spatio-temporal subspaces for monocular multiple object 3d reconstruction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6262–6270, 2017. 2
- [2] Antonio Agudo, Melcior Pijoan, and Francesc Moreno-Noguer. Image collection pop-up: 3d reconstruction and clustering of rigid and non-rigid categories. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2607–2615, 2018. 1, 2, 5, 6, 7
- [3] Ijaz Akhter, Yaser Sheikh, and Sohaib Khan. In defense of orthonormality constraints for nonrigid structure from motion. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1534–1541. IEEE, 2009. 2
- [4] Ijaz Akhter, Yaser Sheikh, Sohaib Khan, and Takeo Kanade. Nonrigid structure from motion in trajectory space. In *Advances in neural information processing systems*, pages 41–48, 2009. 5
- [5] Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm with application to wavelet-based image deblurring. In *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, pages 693–696. IEEE, 2009. 2
- [6] Christoph Bregler, Aaron Hertzmann, and Henning 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, 2
- [7] Cesar Cadena, Anthony R Dick, and Ian D Reid. Multi-modal auto-encoders as joint estimators for robotics scene understanding. In *Robotics: Science and Systems*, 2016. 7
- [8] Angel X. Chang, Thomas A. Funkhouser, Leonidas J. Guibas, Pat Hanrahan, Qi-Xing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiang Xiao, Li Yi, and Fisher Yu. Shapenet: An information-rich 3d model repository. *CoRR*, abs/1512.03012, 2015. 1
- [9] Zhao Chen, Vijay Badrinarayanan, Gilad Drozdov, and Andrew Rabinovich. Estimating depth from rgb and sparse sensing. *European Conference on Computer Vision (ECCV)*, 2018. 7
- [10] Yuchao Dai, Hongdong Li, and Mingyi 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, 3, 4, 5, 6, 7, 8
- [11] Ingrid Daubechies, Michel Defrise, and Christine De Mol. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, 57(11):1413–1457, 2004. 2
- [12] Alessio Del Bue, Fabrizio Smeraldi, and Lourdes Agapito. Non-rigid structure from motion using ranklet-based tracking and non-linear optimization. *Image and Vision Computing*, 25(3):297–310, 2007. 5, 6, 7, 8
- [13] Wei Deng, Wotao Yin, and Yin Zhang. Group sparse optimization by alternating direction method. In *SPIE Optical Engineering+ Applications*, pages 88580R–88580R. International Society for Optics and Photonics, 2013. 4
- [14] David L Donoho, Michael Elad, and Vladimir N Temlyakov. Stable recovery of sparse overcomplete representations in the presence of noise. *IEEE Transactions on information theory*, 52(1):6–18, 2006. 3, 7
- [15] Katerina Fragkiadaki, Marta Salas, Pablo Arbelaez, and Jitendra Malik. Grouping-based low-rank trajectory completion and 3d reconstruction. In *Advances in Neural Information Processing Systems*, pages 55–63, 2014. 2
- [16] Paulo FU Gotardo and Aleix M Martinez. Kernel non-rigid structure from motion. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 802–809. IEEE, 2011. 5, 6, 7, 8
- [17] Onur C Hamsici, Paulo FU Gotardo, and Aleix M Martinez. Learning spatially-smooth mappings in non-rigid structure from motion. In *European Conference on Computer Vision*, pages 260–273. Springer, 2012. 5, 6, 7, 8
- [18] Chen Kong and Simon Lucey. Prior-less compressible structure from motion. *Computer Vision and Pattern Recognition (CVPR)*, 2016. 1, 2, 3, 4, 5, 6, 7, 8
- [19] Chen Kong, Rui Zhu, Hamed Kiani, and Simon Lucey. Structure from category: a generic and prior-less approach. *International Conference on 3D Vision (3DV)*, 2016. 1, 2, 3, 5, 6, 8
- [20] Minsik Lee, Jungchan Cho, and Songhwa Oh. Consensus of non-rigid reconstructions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4670–4678, 2016. 5, 6, 7, 8
- [21] Yaoxin Li, Keyuan Qian, Tao Huang, and Jingkun Zhou. Depth estimation from monocular image and coarse depth points based on conditional gan. In *MATEC Web of Conferences*, volume 175, page 03055. EDP Sciences, 2018. 7
- [22] Yiyi Liao, Lichao Huang, Yue Wang, Sarath Kodagoda, Yinnan Yu, and Yong Liu. Parse geometry from a line: Monocular depth estimation with partial laser observation. In *Robotics and Automation (ICRA), 2017 IEEE International Conference on*, pages 5059–5066. IEEE, 2017. 7
- [23] Joseph J. Lim, Hamed Pirsiavash, and Antonio Torralba. Parsing IKEA Objects: Fine Pose Estimation. *ICCV*, 2013. 5
- [24] Fangchang Mal and Sertac Karaman. Sparse-to-dense: Depth prediction from sparse depth samples and a single image. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1–8. IEEE, 2018. 7
- [25] Vardan Papyan, Yaniv Romano, and Michael Elad. Convolutional neural networks analyzed via convolutional sparse coding. *The Journal of Machine Learning Research*, 18(1):2887–2938, 2017. 2
- [26] Christopher J Rozell, Don H Johnson, Richard G Baraniuk, and Bruno A Olshausen. Sparse coding via thresholding and local competition in neural circuits. *Neural computation*, 20(10):2526–2563, 2008. 2
- [27] Hao Su, Charles R Qi, Yangyan Li, and Leonidas 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. 1
- [28] Jeremias Sulam, Vardan Papyan, Yaniv Romano, and Michael Elad. Multi-layer convolutional sparse modeling: Pursuit and dictionary learning. *arXiv preprint arXiv:1708.08705*, 2017. 3
- [29] Carlo Tomasi and Takeo Kanade. Shape and motion from image streams under orthography: a factorization method. *International Journal of Computer Vision*, 9(2):137–154, 1992. 2
- [30] Sara Vicente, João Carreira, Lourdes Agapito, and Jorge Batista. Reconstructing pascal voc. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 41–48. IEEE, 2014. 1
- [31] Jiajun Wu, Tianfan Xue, Joseph J Lim, Yuandong Tian, Joshua B Tenenbaum, Antonio Torralba, and William T Freeman. Single image 3d interpreter network. *European Conference on Computer Vision (ECCV)*, 2016. 5
- [32] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3d shapenets: A deep representation for volumetric shapes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1912–1920, 2015. 1
- [33] Yu Xiang, Roozbeh Mottaghi, and Silvio 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
- [34] Xiaowei Zhou, Spyridon Leonardos, Xiaoyan Hu, and Kostas 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. 3, 5
- [35] Xiaowei Zhou, Menglong Zhu, Spyridon Leonardos, Kosta Derpanis, and Kostas Daniilidis. Sparseness meets deepness: 3d human pose estimation from monocular video. *arXiv preprint arXiv:1511.09439*, 2015. 2
- [36] Yingying Zhu, Dong Huang, Fernando De La Torre, and Simon 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, 3, 4
- [37] Yingying Zhu and Simon Lucey. Convolutional sparse coding for trajectory reconstruction. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 37(3):529–540, 2015. 3