Distill Knowledge from NRSfM for Weakly Supervised 3D Pose Learning

Chaoyang Wang Chen Kong Simon Lucey
Carnegie Mellon University
{chaoyanw, chenk, slucey}@cs.cmu.edu

Abstract

We propose to learn a 3D pose estimator by distilling knowledge from Non-Rigid Structure from Motion (NRSfM). Our method uses solely 2D landmark annotations. No 3D data, multi-view/temporal footage, or object specific prior is required. This alleviates the data bottleneck, which is one of the major concern for supervised methods. The challenge for using NRSfM as teacher is that they often make poor depth reconstruction when the 2D projections have strong ambiguity. Directly using those wrong depth as hard target would negatively impact the student. Instead, we propose a novel loss that ties depth prediction to the cost function used in NRSfM. This gives the student pose estimator freedom to reduce depth error by associating with image features. Validated on H3.6M dataset, our learned 3D pose estimation network achieves more accurate reconstruction compared to NRSfM methods. It also outperforms other weakly supervised methods, in spite of using significantly less supervision.

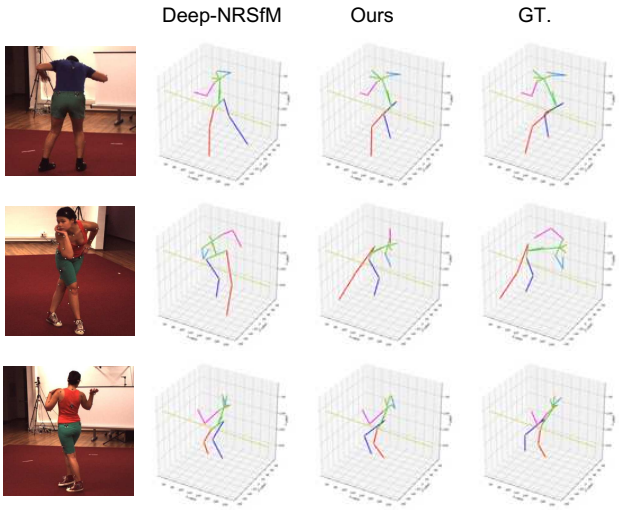

Figure 1. NRSfM methods often achieve poor reconstructions when the 2D projections have strong ambiguity. Our proposed knowledge distilling method lets the student pose estimation network (3rd column) correct some of the mistakes made by its NRSfM teacher (2nd column).

1. Introduction

Learning to estimate 3D pose from images is bottlenecked by the availability of abundant 3D annotated data. Weakly supervised methods that reduce the amount of required annotation is of high practical value. Prior works approach this problem by supplementing their training set with: (i) extra 2D annotated data [47]; (ii) aligning 3D models to 2D annotations [35, 43, 37]; (iii) exploiting geometric cues from multi-view footage [33, 32, 38]; or (iv) utilizing adversarial framework to impose a prior on the 3D structure [12]. These methods, however, are either restricted to laboratory settings or still requires a 3D training set – which limits the type of target objects they can work with. This paper addresses a more general setting – we utilize image datasets with solely 2D landmark annotations (i.e. no 3D supervision). This allows our method to be applied to a wider scope of objects, not limited by the availability of 3D models, kinematic priors, or sequential/multi-view footage.

Our work is made possible by some recent advances in Non-Rigid Structure from Motion (NRSfM). NRSfM methods reconstruct 3D shapes and camera positions from multiple 2D projections of articulated 3D points. These points do not have to belong to the same object, but can be from multiple instances of the same object category, which naturally applies to our problem. Prior NRSfM methods are restricted by the number of frames and the type of shape variability they can handle, which limits their usage to many real world problems. Kong and Lucey [21] recently proposed a neural network architecture (Deep-NRSfM) interpreted as solving a multi-layer block sparse dictionary learning problem, and can handle problems of unprecedented scale and shape complexity. Our *modified* version of Deep-NRSfM achieves state-of-the-arts accuracy on H3.6M [18] dataset, outperforming other NRSfM methods by a significant margin.

Despite this progress, NRSfM still has difficulty in predicting correct depth for shapes with strong ambiguity in terms of 2D projection, e.g. identifying if a leg is stretching

towards/away from the camera, even though these are distinguishable with texture features. Therefore, directly using the depth output from NRSfM as labels to train a pose estimation network is affected by those errors. Instead of this hard assignment of training labels, we propose a softer approach – we want to penalize less when there's high ambiguity in 2D projection, so as to leave room for the pose estimation network to correct errors made by NRSfM through associating image features (see Fig. 1).

To design our learning objective, we review the dictionary learning problem used to solve NRSfM. Assuming the camera matrix fixed, a depth hypothesis defines a subspace of codes – any codes in this subspace is to have the same depth reconstruction as the hypothesis, but have different cost (2D reprojection error + regularizer). A natural way to characterize the quality of a depth hypothesis is by the minimum cost of codes in its subspace. However, directly using this as a learning objective leads to solving a constrained optimization problem numerically per SGD iteration, which is computationally intractable. Instead, we derive a convex upper bound by evaluating the cost at the projection of the NRSfM solution on the subspace. Experiments show that pose network trained by this loss noticeably reduces error on the training set compared to our already strong NRSfM baseline, and consequently leads to lower validation error as a weakly supervised learning task.

Another benefit of the proposed knowledge distilling loss is that, it poses no restriction on the architecture of the student pose estimation network, as long as it outputs the depth value for the landmarks. This is not the case for some of the prior works [43, 13], where the pose estimation network has to output the coefficients associated to some external shape dictionary.

In conclusion, contributions of this paper are:

- We propose a weakly supervised pose estimation method using solely 2D landmark annotations. We do not use any 3D labels, multi-view footage, or target specific shape prior. In spite of using weaker supervision, we achieve the best results compared to other weakly supervised methods.
- We establish a strong NRSfM baseline modified from Deep-NRSfM [21], which outperforms current published state-of-the-art NRSfM methods on H3.6M dataset.
- We propose a new knowledge distilling algorithm applicable to NRSfM methods based on dictionary learning. We demonstrate that our learned network gets significantly lower error on the training set compared to its NRSfM teacher.

2. Related Works

Non-rigid structure from motion NRSfM is a classical ill-posed problem since the 3D shapes can vary between images, resulting in more variables than equations. To alleviate the ill-posedness, various constraints are exploited including 1) temporal smoothness [2, 15, 24, 23], 2) fixed articulation [31] and more commonly used 3) shape priors. The first statistical shape prior—non-rigid objects can be modeled by a local subspace in low rank—is first proposed by Bregler *et al*. [5] and later developed by Dai *et al*. [9]. Following this direction, increasing works are reported to model more complex objects while still maintaining a well-conditioned system. Among them, representatives are union-of-subspaces [48, 1], and block-sparsity [20, 22]. Of particular interest to this paper is the most recent work [21] that introduces deep neural network to accurately solving large scale NRSfM problem. Even though great success, majority NRSfM algorithms rely heavily on 2D annotation-based priors. However, as pointed in the introduction, much broader information are embedded under image itself, under pixel values. In this paper, we impose a novel image prior such that NRSfM is no longer trapped at 2D coordinates of landmarks but also learn from origin images.

Weakly supervised 3D pose learning Most 3D pose estimation methods [36, 30, 29, 47, 45, 44, 28, 8, 26, 6] are fully supervised. One bottleneck for the supervised methods is that data coming from multi-view motion capture systems [19, 18] includes limited number of human subject, and has simple backgrounds. This would affect the generalization ability of a trained model. Weakly supervised methods aim to alleviate this problem by limiting the requirement for labeled data. They can be loosely categorized as: using synthetic datasets [7, 40] to increase the training set size. These methods face the problem of generalizing to new motions and environments that are different from the simulated data; On the other hand, given the existing large-scale image datasets with 2D annotation, Zhou *et al*. [47] train their model with 2D labeled images together with motion capture data. To further reduce dependency on paired 3D annotation, 3D interpreter network [43], multi-modal model [37] and generative adversarial networks [13, 41] are trained on external 3D data; multi-view footage is also used to enforce geometric constraints [38, 33]; However, these methods still require a large enough 3D training set to properly initialize and constraint their learning process.

Recently, Rhodin *et al*. [32] propose a method based on geometric-aware representation learning, which requires only a small amount of annotation. Its performance however is limited, which restricts its practical usage. A concurrent work of Drover *et al*. [12] propose to use adversarial framework to impose a prior on the 3D structure, learned

solely from 2D projections. Yet they still utilize the ground-truth 3D poses to generate a large number of synthetic 2D poses for training, which augments the original 1.5M 2D poses in Human3.6M by almost 10 times.

3. Non-rigid Structure from Motion

Under weak perspective camera assumption, 2D projection $\mathbf{W} \in \mathbb{R}^{P\times 2}$ is the product of 3D shape $\mathbf{S} \in \mathbb{R}^{P\times 3}$ and camera matrix $\mathbf{M} \in \mathbb{R}^{3\times 2}$:

$$\mathbf{W} = \mathbf{S}\mathbf{M}, \quad \mathbf{W} = \begin{bmatrix} \vdots & \vdots \\ u_p & v_p \\ \vdots & \vdots \end{bmatrix}, \quad \mathbf{S} = \begin{bmatrix} \vdots & \vdots & \vdots \\ x_p & y_p & z_p \\ \vdots & \vdots & \vdots \end{bmatrix}, \tag{1}$$

where (u_p, v_p) and (x_p, y_p, z_p) are the image and world coordinate of p-th point, and $\mathbf{M}$ is required to be orthonormal. The goal of NRSfM is to recover 3D shape $\mathbf{S}$ and camera matrix $\mathbf{M}$ given the observed 2D projections $\mathbf{W}$. This is an inherent ill-posed problem. Finding a unique solution requires sufficient regularization and prior knowledge.

One type of NRSfM methods approach the problem through dictionary learning. Denote $\mathbf{s} \in \mathbb{R}^{3P}$ is the vectorization of $\mathbf{S}$, it satisfies: $\mathbf{s} = \mathbf{D}\boldsymbol{\varphi}$, where $\mathbf{D} \in \mathbb{R}^{3P\times K}$ is a dictionary with K bases; and $\boldsymbol{\varphi} \in \mathbb{R}^K$ is a code vector. Given multiple observation of 2D projections $\mathbf{W}^{(i)}$ from an articulated object deforming over time, or different objects of the same category, these methods can be loosely interpreted as minimizing the following objective:

$$\min_{\mathbf{D},\{\boldsymbol{\varphi}^{(i)}\},\{\mathbf{M}^{(i)}\}} \sum_i \|[\mathbf{D}\boldsymbol{\varphi}^{(i)}]_{P\times 3}\mathbf{M}^{(i)} - \mathbf{W}^{(i)}\| + h(\boldsymbol{\varphi}^{(i)}) \tag{2}$$

where operator $[\,]_{P\times 3}$ is defined as reshaping the vecorized 3D shape into matrix form with dimension $P \times 3$; $h(\boldsymbol{\varphi})$ is a regularizer introduced to improve uniqueness of solution, e.g. low rank [9], sparsity [20], *etc*.

Our knowledge distilling method (see Section 4) is designed for this general type of NRSfM method, and in principal, it is agnostic to the type of regularizor they use, as long as the dictionary is overcomplete.

Deep NRSfM Kong and Lucey[21] propose a prior assumption that 3D shapes are compressible via multi-layer sparse coding:

$$\begin{aligned} \mathbf{s} &= \mathbf{D}_1\boldsymbol{\varphi}_1, \quad \|\boldsymbol{\varphi}_1\|_1 \le \lambda_1, \quad \boldsymbol{\varphi}_1 \ge 0, \\ \boldsymbol{\varphi}_1 &= \mathbf{D}_2\boldsymbol{\varphi}_2, \quad \|\boldsymbol{\varphi}_2\|_1 \le \lambda_2, \quad \boldsymbol{\varphi}_2 \ge 0, \\ &\vdots, \ \vdots \\ \boldsymbol{\varphi}_{n-1} &= \mathbf{D}_n\boldsymbol{\varphi}_n, \quad \|\boldsymbol{\varphi}_n\|_1 \le \lambda_n, \quad \boldsymbol{\varphi}_n \ge 0, \end{aligned} \tag{3}$$

where $\mathbf{D}_i$ are hierarchical dictionaries, and code vectors $\boldsymbol{\varphi}_i \in \mathbb{R}^{K_i}$ are constrained to be sparse and non-negative. Compared to single level sparse coding, codes in multi-layer sparse coding not only minimizes the reconstruction error at their individual levels, but is also regularized by the codes from other levels. This helps to impose more constraints on code recovery while maintaining similar shape expressibility versus single level sparse coding with the same dictionary size.

To recover sparse codes, one of the classical method to use is Iterative Shrinkage and Thresholding Algorithm (ISTA) [10, 4, 34]. Papyan *et al*. [27] find that feed-forward neural netorks can be interpreted as approximating one iteration of inferencing sparse codes by ISTA, and the dictionaries $\mathbf{D}_1, \mathbf{D}_2, \ldots, \mathbf{D}_n$ serves as the neural network weights. Based on this insight, Chen *et al*. derived a novel neural network architecture which approximates the solution of sparse codes $\boldsymbol{\varphi}_1$ and camera matrix $\mathbf{M}$. In this paper, we made significant modification to their original architecture, which we find important to get good result in experiment. Limited by space, we put description about our version of camera matrix estimation network $q_{\mathbf{M}}(\mathbf{W}) : \mathbb{R}^{P\times 2} \mapsto \mathbb{R}^{3\times 2}$, and sparse code estimation network $q_{\boldsymbol{\varphi}}(\mathbf{W}, \mathbf{M}) : \mathbb{R}^{P\times 2} \times \mathbb{R}^{3\times 2} \mapsto \mathbb{R}^{K_1}$ in the supplementary material.

With the feed-forward code/camera estimation networks parameterized by the dictionaries, we can now learn the dictionaries through minimizing reprojection error of all samples in the dataset. Denote $\tilde{\boldsymbol{\varphi}}_1^{(i)}$, $\tilde{\mathbf{M}}^{(i)}$ to be the output of networks $q_{\boldsymbol{\varphi}}$, $q_{\mathbf{M}}$ given ith 2D projection $\mathbf{W}^{(i)}$, the loss function is:

$$\min_{\mathbf{D}_1,\mathbf{D}_2,\ldots,\mathbf{D}_n} \sum_i \|[\mathbf{D}_1\tilde{\boldsymbol{\varphi}}_1^{(i)}]_{P\times 3}\tilde{\mathbf{M}}^{(i)} - \mathbf{W}^{(i)}\|_2 + \lambda\|\tilde{\boldsymbol{\varphi}}_1\|_1. \tag{4}$$

In this loss function, in addition to reprojection error, we add sparisity penalty using a small weighting, which we find helpful to improve results.

4. Distilling Knowledge from NRSfM

Problem setup: Given an image dataset paired with annotated 2D locations of landmarks on target objects: $\{(\mathbf{I}^{(i)}, \mathbf{W}^{(i)})\}$, we want to train a 3D pose estimation network able to predict 3D landmark positions from image input. The main difficulty of this task is how to learn to predict depth of landmarks without any depth supervision. Our cue is from dictionary learning-based NRSfM method (Deep-NRSfM in our experiment), which gives us a 3D shape dictionary $\mathbf{D}$, and recovered camera matrices $\mathbf{M}^{(i)}$ and codes $\boldsymbol{\varphi}_{\text{nrsfm}}^{(i)}$.

With the dictionary, camera matrices and codes from NRSfM, depth in the image coordinate can be computed by simply rotating the 3D shape reconstruction $\mathbf{D}\boldsymbol{\varphi}_{\text{nrsfm}}^{(i)}$. Given this, a simple baseline for this task would be: we use the depth reconstruction as labels to train the 3D pose estima-

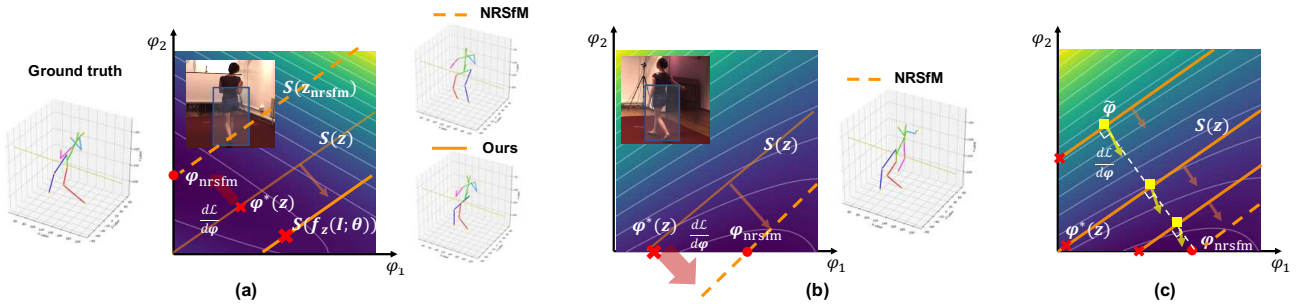

Figure 2. Illustration of the proposed knowledge distilling algorithm. **(a)** For illustration purpose, we assume the code φ is 2-dimensional. We plot the cost function (Eq. 9) as a 2D heatmap. The NRSfM solution φ_{nrsfm} is approximately the minima of this heat map (represented as red dot). Given a depth hypothesis $\mathbf{z}$, all the codes satisfies $\mathbf{z}$ forms a subspace $\mathcal{S}(\mathbf{z})$, which is shown as the orange line. The quality of a depth hypothesis is evaluated by the best point on its subspace, denoted as $\varphi^*(\mathbf{z})$ (red cross). Given different depth hypothesis is equivalent to parallel translate the line. Suppose $\mathbf{z}$ is free to have any value, then minimizing our loss function (Eq. 10) would push the line to cross φ_{nrsfm}(see the dashed orange line). This gives the same wrong depth reconstruction as the NRSfM method. (**b**) Suppose we get another image of similar pose but with less 2D projection ambiguity. In this case, NRSfM gives correct shape recovery. Since texture features are similar for both images, the pose estimation network is implicitly constrained to make similar depth predictions. Then minimizing our loss for both images would lead to a better solution for image 1 (shown as solid orange line), because gradients are larger from the 2nd image due to the fact that it has less ambiguity. **(c)** We approximate the loss by evaluating at the projection of φ_{nrsfm} on the subspace (yellow square). This approximation is a convex upper bound for the original loss. It would still reflect the degree of projection ambiguity, and push the subspace (lines) torwards φ_{nrsfm}.

tion network. However, as shown in Fig. 1, we find that NRSfM tends to make wrong estimation due to strong ambiguity in 2D projections. Using those as hard target for regression would bottleneck the accuracy of learned pose estimation network. We propose a better approach - we want to establish a direct relation between depth prediction and the cost function (Eq. 2) we used in NRSfM, which is the better metric to evaluate the quality of predicted 3D shapes. In this way, we can avoid confusing our student network with wrong labels, and allow them to implicitly associate image features to disambiguate difficult poses for NRSfM. This intuition is inline with other geometric self-supervised learning, e.g. self-supervised depth estimation [46, 14, 42], in which photometric loss is used to train a depth estimation network.

Outline: The core problem is how to design a loss function which properly evaluates the quality of a depth hypothesis produced by the pose estimator. To derive our loss function, We first show that a depth hypothesis associates with a subspace of codes (see Section 4.1). We then advocate that the loss should be the minimum cost value of codes in the subspace (see Section 4.2). Finally, we derive a convex upper bound for the loss, which is computationally trackable for SGD training (see Section 4.3). A 2D illustration is given in Fig. 2 to help decipher the text.

4.1. Depth hypothesis defines a subspace of codes

From NRSfM, we get the dictionary $\mathbf{D}$, and per example camera matrix $\mathbf{M}^{(i)}$. We find that the camera matrices from our modified Deep-NRSfM are accurate, thus we treat them as oracle and fixed in our learning algorithm. With this, we can simplify our notation by absorbing camera matrix into dictionary through rotation. Rotation matrix $\mathbf{R}^{(i)} \in \mathbb{R}^{3\times3}$ is formed from camera matrix by:

$$\mathbf{R}^{(i)} = [\mathbf{m}_1^{(i)}, \mathbf{m}_2^{(i)}, \mathbf{m}_1^{(i)} \times \mathbf{m}_2^{(i)}], \tag{5}$$

where $\mathbf{m}_1^{(i)}, \mathbf{m}_2^{(i)}$ are columns of camera matrix $\mathbf{M}^{(i)}$. Then the dictionary is rotated by multiplying every 3D coordinates inside $\mathbf{D}$ with $\mathbf{R}^{(i)}$:

$$\mathbf{B}^{(i)} = \left[[\mathbf{d}_x^1, \mathbf{d}_y^1, \mathbf{d}_z^1]\mathbf{R}^{(i)} \quad \dots \quad [\mathbf{d}_x^P, \mathbf{d}_y^P, \mathbf{d}_z^P]\mathbf{R}^{(i)}\right]^T \tag{6}$$

We further split $\mathbf{B}^{(i)}$ into two matrices – one matrix takes all the x, y coordinate elements of $\mathbf{B}^{(i)}$, while the other takes all the rest z coordinate elements.

$$\begin{aligned} \mathbf{B}_{xy}^{(i)} &= \left[\mathbf{b}_x^{1(i)} \quad \mathbf{b}_y^{1(i)} \quad \dots \quad \mathbf{b}_x^{P(i)} \quad \mathbf{b}_y^{P(i)} \quad \right]^T, \\ \mathbf{B}_z^{(i)} &= \left[\mathbf{b}_z^{1(i)} \quad \dots \quad \mathbf{b}_z^{P(i)} \quad \right]^T, \end{aligned} \tag{7}$$

With this, $\mathbf{B}_{xy}^{(i)}\varphi^{(i)}$ computes 2D projection of shape reconstructed by code $\varphi^{(i)}$; and $\mathbf{B}_z^{(i)}\varphi^{(i)}$ is reconstructed depth in the image coordinate.

For a depth hypothesis $\mathbf{z}' = f_z(\mathbf{I}^{(i)};\boldsymbol{\theta})$ produced by the pose estimation network, codes giving depth reconstruction equal to $\mathbf{z}'$ forms a subspace:

$$\mathcal{S}^{(i)}(\mathbf{z}') = \{\varphi : \mathbf{B}_z^{(i)}\varphi = \mathbf{z}'\}. \tag{8}$$

The subspace is not empty assuming that dictionary is overcomplete. In Fig. 2, the subspaces are visualized as orange lines in 2D.

4.2. Loss = minimum cost on subspace

The quality of a depth hypothesis $\mathbf{z}'$ could be represented by the best code inside its subspace. As in NRSfM, the quality of a code is measured by the cost function = reprojection error + some regularizer, i.e.:

$$\mathcal{C}^{(i)}(\boldsymbol{\varphi}) = \|\mathbf{B}_{xy}^{(i)}\boldsymbol{\varphi} - \mathbf{w}^{(i)}\| + h(\boldsymbol{\varphi}), \tag{9}$$

where $\mathbf{w}^{(i)}$ is the vectorization of $\mathbf{W}^{(i)}$. To keep formulation general, we don't specify the type of norm and regularizer here. Thereby we have the following definition of quality function for $\mathbf{z}'$, which we use as the loss function for knowledge distilling:

$$\mathcal{L}^{(i)}(\mathbf{z}') = \min_{\boldsymbol{\varphi}\in\mathcal{S}^{(i)}(\mathbf{z}')} \mathcal{C}^{(i)}(\boldsymbol{\varphi}). \tag{10}$$

This computes the minimum cost value of codes inside the subspace defined by the depth hypothesis $\mathbf{z}'$.

To evaluate this loss function, we need to first solve for the minima $\boldsymbol{\varphi}^*$ of the constrained convex optimization problem in Eq. 10 (red cross in Fig. 2). Suppose we can express $\boldsymbol{\varphi}^*$ as a differentiable function of $\mathbf{z}'$, i.e. $\boldsymbol{\varphi}^* = q^{(i)}(\mathbf{z}')$, Eq. 10 becomes:

$$\mathcal{L}^{(i)}(\mathbf{z}') = \|\mathbf{B}_{xy}^{(i)} q^{(i)}(\mathbf{z}') - \mathbf{w}^{(i)}\| + h(q^{(i)}(\mathbf{z}')). \tag{11}$$

This loss is explicitly a function of $\mathbf{z}'$, and thus allows the gradients to be propagated to the pose estimation network.

As a side note, suppose the pose network has unlimited capacity, in other words, able to overfit any depth values, then the end result of minimizing this loss function would be a network predicting the same depth as the NRSfM algorithm (illustrated in Fig. 2(a)). We argue that this would not be the case in practice, since convolution networks constrained by their structure, is equivalent to have a deep image prior [39] imposed on their output. This image prior provides extra constraint to disambiguate confusing 2D projections, thus is the key source for our improvement over the NRSfM teacher.

4.3. Convex upper bound of Eq. 11

Using Eq. 11 requires to form the (sub)differentiable function $q^{(i)}(\mathbf{z}')$ which produces the solution to the constrained optimization problem in Eq. 10. However, solving this constrained optimization problem requires iterative numerical method due to the existence of regularizer. As a result, it's computationally intractable to solve it exactly per SGD iteration during training. Therefore we derive an approximate solution as follow:

Suppose $\boldsymbol{\varphi}_{\text{nrsfm}}^{(i)}$ is the solution we get from NRSfM, and it approximates the minima of the optimization problem in Eq. 10 without the subspace constraint, then an approximate solution for the constrained problem could be the projection of $\boldsymbol{\varphi}_{\text{nrsfm}}^{(i)}$ onto the subspace $\mathcal{S}^{(i)}(\mathbf{z}')$:

$$\tilde{\boldsymbol{\varphi}}^{(i)}(\mathbf{z}') = \arg\min_{\boldsymbol{\varphi}\in\mathcal{S}^{(i)}(\mathbf{z}')} \frac{1}{2}\|\boldsymbol{\varphi} - \boldsymbol{\varphi}_{\text{nrsfm}}^{(i)}\|_2^2 \tag{12}$$

The closed form solution to Eq. 12 is:

$$\tilde{\boldsymbol{\varphi}}^{(i)}(\mathbf{z}') = \boldsymbol{\varphi}_{\text{nrsfm}}^{(i)} + (\mathbf{B}_z^{(i)})^{\dagger}(\mathbf{z}' - \mathbf{B}_z^{(i)}\boldsymbol{\varphi}_{\text{nrsfm}}^{(i)}), \tag{13}$$

where $(\mathbf{B}_z^{(i)})^{\dagger} = \mathbf{B}_z^{(i)^T}(\mathbf{B}_z^{(i)}\mathbf{B}_z^{(i)^T})^{-1}$ is the right inverse of $\mathbf{B}_z^{(i)}$. Eq. 13 is implemented as a differentiable operator thanks to modern deep learning library.

Substitute the exact solution $q^{(i)}(\mathbf{z}')$ in Eq. 11 by the approximate solution $\tilde{\boldsymbol{\varphi}}^{(i)}(\mathbf{z}')$ gives a convex upper bound of Eq. 11:

$$\tilde{\mathcal{L}}^{(i)}(\mathbf{z}') = \|\mathbf{B}_{xy}^{(i)}\tilde{\boldsymbol{\varphi}}^{(i)}(\mathbf{z}') - \mathbf{w}^{(i)}\| + h(\tilde{\boldsymbol{\varphi}}^{(i)}(\mathbf{z}')) \tag{14}$$

In our experiment, we find that using this convex upper bound as training loss, is sufficient to give lower error on the training set compared to our already strong NRSfM baseline.

4.4. Learning the 3D pose estimator

We use the state-of-the-art integral regression network [36] as our student pose estimator. The network directly predicts 3D coordinates of landmarks in the image coordinate. During training, the (x, y) coordinate is directly supervised by 2D landmark annotations; while z coordinate is supervised by our knowledge distilling loss (Eq. 14). The proposed learning objective is:

$$\min_{\boldsymbol{\theta}} \sum_i \|f_{xy}(\mathbf{I}^{(i)};\boldsymbol{\theta}) - \mathbf{w}^{(i)}\|_1 + \tilde{\mathcal{L}}^{(i)}(f_z(\mathbf{I}^{(i)};\boldsymbol{\theta})), \tag{15}$$

where f_{xy},f_z denote the output of the network at (x, y) and z coordinates; and $\boldsymbol{\theta}$ refers to the network weights. For the knowledge distilling loss $\tilde{\mathcal{L}}$, we use L_2 norm for the reprojection error, and L_1 norm for the regularizer in our experiment. The regularizer is weighted by an empirically found coefficient, which is 0.3 in our experiment.

5. Experiment

5.1. Implementation details

Data preprocessing: We assume no knowledge of 3D label in both training and testing. We crop the image according to the 2D human bounding box, and then resize and pad such that it is 256x256 resolution. The 2D points are then represented by the patch coordinate. In evaluation, we follow the same procedure as in [36], which aligns the scale of the prediction by average bone length before computing the metrics.

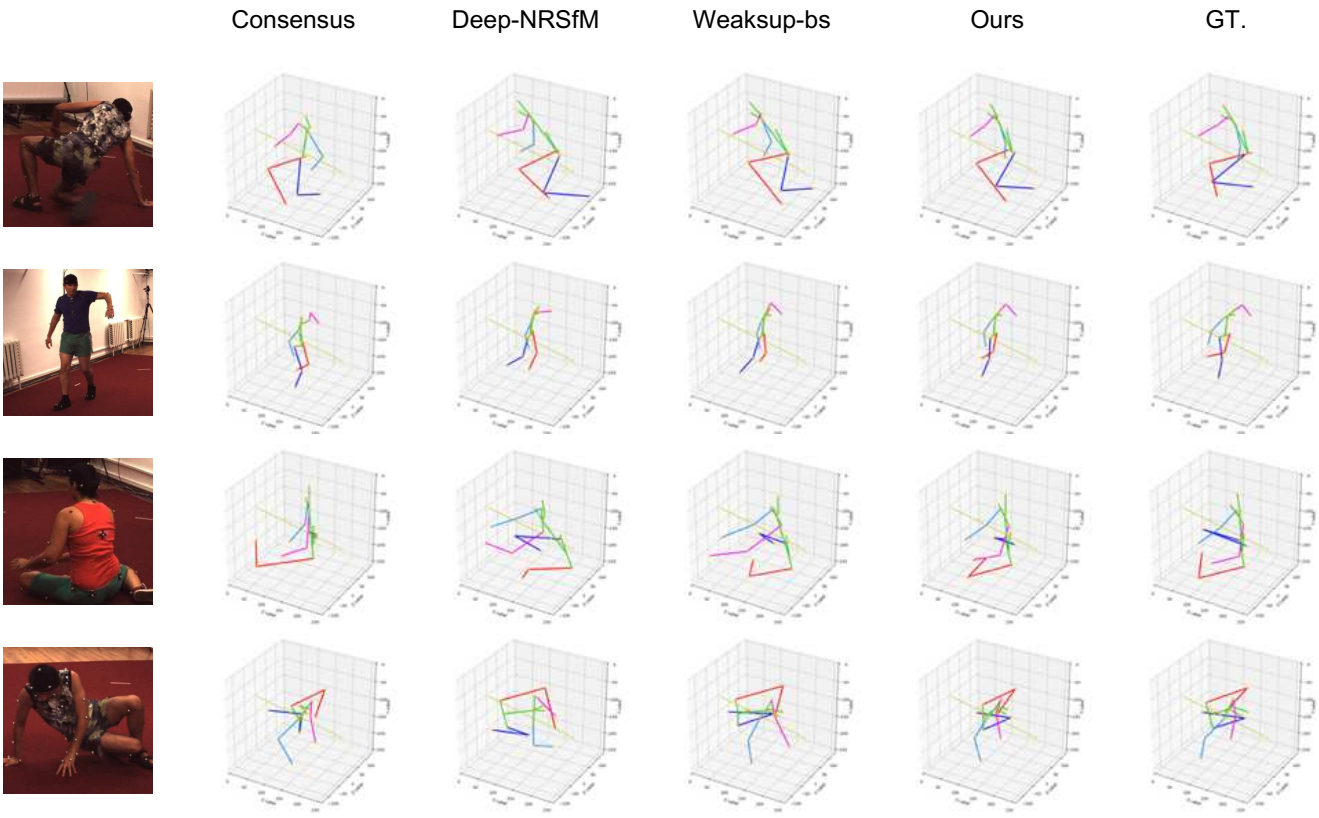

Figure 3. Visual comparison of NRSfM methods versus methods which include image as extra constraint (i.e. our weakly supervised baseline and our knowledge distilling method) on the training set. Our method shows significant improvement over its teacher, i.e. deep-NRSfM. Skeletons are rendered from side view for better visualization of the difference in depth reconstruction. We use red and magenta to color left leg and arm, while blue and dodgerblue are used to color right leg and arm.

	P-MPJPE	MPJPE	depth error
Ranklet [11]	281.1	-	-
Sparse [20]	217.4	-	-
SPM(2k) [9]	209.5	-	-
SFC [22]	167.1	218.0	135.6
KSTA(5k) [16]	123.6	-	-
RIKS(5k) [17]	103.9	-	-
Consensus [25]	79.6	120.1	111.5
Deep-NRSfM* [21]	73.2	101.6	76.5
Weaksup-bs	61.2	86.2	75.3
Ours	56.4	80.9	71.2

Table 1. Compare with NRSfM methods on the training set of H3.6M ECCV18 challenge dataset. KSTA, RIKS are evaluated on a subset of 5k images, and SPM is evaluated on 2k images. $*$ Our implementation of Deep-NRSfM has significant difference compared to the original paper.

	2D	3D	MV	P-MPJPE	MPJPE
Sun *et al.* [36]	-	-	-	-	86.4
Rhodin *et al.* [32]		✓	✓	98.2	131.7
Tung *et al.* [38]	✓	✓	✓	98.4	-
3Dinterp. [43]	✓	✓		98.4	-
AIGN [13]	✓	✓		97.2	-
Tome *et al.* [37]	✓	✓		-	88.4
Drover *et al.* [12]	✓	✓		64.6	-
Weaksup-bs	✓			67.3	95.0
Ours	✓			62.8	86.4
+ MPII	✓			**57.5**	**83.0**

Table 2. Compare with weakly supervised methods on H3.6M validation set. Supervision source used by each method is marked: '2D' refers to 2D landmark annotation; '3D' represents any training source with 3D annotation, including synthetic 3D dataset, external human 3D model, etc.; 'MV' is the abbreviation for multi-view.

3D pose estimation network: We select the integral regression network [36] due to its state-of-the-art performance in human pose estimation. Throughout our experiment, we use ResNet50 as the backbone for the regression network, and the input image resolution is set as 256×256. Using deeper backbone network (e.g. ResNet152) and higher image resolution would improve result, as already shown in [36]. We choose this cheaper setting for a fairer comparison with other weakly supervised methods which use ResNet50.

	Direct.	Disc.	Eat	Greet	Phone	Photo	Pose	Purch.	Sit	SitD	Smoke	Wait	Walk	WalkD	WalkP
3Dinterp. [43]	78.6	90.8	92.5	89.4	108.9	112.4	77.1	106.7	127.4	139.0	103.4	91.4	79.1	-	-
AIGN [13]	77.6	91.4	89.9	88.0	107.3	110.1	75.9	107.5	124.2	137.8	102.2	90.3	78.6	-	-
Drover *et al.* [12]	60.2	60.7	59.2	65.1	65.5	63.8	59.4	59.4	69.1	88.0	64.8	60.8	64.9	63.9	65.2
Weaksup-bs	58.8	62.4	56.7	59.8	68.6	60.8	59.7	81.0	93.4	68.5	75.8	65.9	61.5	67.6	65.0
Ours	54.7	57.7	54.8	55.8	61.6	56.3	52.7	73.7	95.5	62.3	68.5	60.8	55.5	64.0	58.0
+MPII	**50.3**	**48.9**	**52.7**	**53.9**	**59.9**	**50.7**	**48.3**	70.9	82.6	**58.0**	65.3	**54.7**	**50.8**	**57.7**	**55.6**

Table 3. Per action PA-MPJPE reported on H3.6M validation set. Our approach performs favorably compared to other weakly supervised methods.

During training, we follow most of the settings in [36], i.e. the base learning rate is 1e-3, and it drops to 1e-5 when the loss on the validation set saturates. Limited by our computational resources, we use a smaller batch size of 32.

Deep-NRSfM: We use dictionaries with 6 levels. The size for the dictionaries from lower level to higher is: 256, 128, 64, 32, 16, 8. When learning the dictionaries, the sparsity weight (λ in Eq. 2) is selected through cross validation and set as 0.01. For more details of our modified version of Deep-NRSfM, we refer the reader to our supplementary material.

5.2. Experiment setup

Dataset: We validate our method on Human3.6M dataset (H3.6M) [18], which is the major dataset used in current 3D human pose estimation research. Despite our experiment is focused on human pose estimation, we'd like to emphasize that the proposed method is a general algorithm. Unlike other weakly supervised methods which are deeply coupled with external 3D human model, our method doesn't require any target specific prior knowledge, thus should be applicable to other type of objects without restriction.

H3.6M includes sequences of 11 actors performing 15 type of actions captured from 4 camera locations. Footage of 7 out of 11 actors are released for training/validation. We follow the experiment convention conducted by prior papers: 5 subjects (S1, S5, S6, S7, S8) are used as training set, and 2 subjects (S9, S11) for testing. Although H3.6M dataset comes with 3D annotation, we use only 2D annotation during training, and 3D labels are kept for validation.

Strategies to sample frames from the training footage can have a direct impact on validation accuracy. For reproducibility, we use the subset (35k+ images) selected by H3.6M ECCV18 Challenge for training. We augment the training set through random image warping and perturbation as in [36].

Evaluation metric: We follow the two common evaluation protocols used in literature, and report both of them.

- MPJPE: mean per joint positioning error measures the mean euclidean distance between the reconstructed and ground truth joints after shifting them to have the same root joint coordinate.
- PA-MPJPE: Align the reconstructed joints to the ground truth through rigid transformation before evaluating MPJPE. This metric is more often used in NRSfM to measure the correctness of the reconstructed shape.

In addition, we also report 'depth error' which measures the mean difference along z-axis. This is the most important metric to validate our method, because the core problem of weakly supervised learning is how to recover depth without annotation.

Weakly supervised learning baseline: As previously mentioned, a simple weakly supervised learning baseline is using the depth output from our Deep-NRSfM method as training labels. We use this baseline (refer as "Weaksup-bs") to validate the contribution of our novel knowledge distilling loss. To train the pose estimation network, we employ L1 regression loss which has been proven effective in [36].

Weighting value for the L_1 regularizer: We study the effect of different weighting values for the L_1 regularizer in the propoesed knowledge distilling loss (Eq. 14). As shown in Table 4, under a reasonable range (0.1-0.5) of the weights, our method consistently outperforms the baseline.

L_1 weight	0.01	0.1	0.3	0.5	Weaksup-bs
depth error (mm)	79.0	74.6	**73.1**	76.7	78.0
PA-MPJPE (mm)	73.0	73.6	**70.5**	71.0	75.8

Table 4. Comparing different weighting values for the L_1 regularizer in Eq. 14. Numbers reported on the validation set of H3.6M ECCV18 challenge.

Using extra data from MPII: Prior works [47] has shown that including external 2D data such as MPII [3] as training source can improve generalization ability of the learned 3D pose estimator. Thus, we also report result of our method trained with H3.6M+MPII. Due to our current method does not handle missing joints, we apply our proposed knowledge distilling loss only to those MPII images with complete 2D skeleton annotation; for images with occluded/out-of-view joints, we only use 2D regression loss as in [36].

5.3. Compare with NRSfM methods

We compare with 7 state-of-the-art NRSfM methods on our training set (35k+ images from H3.6M ECCV18 Chal-

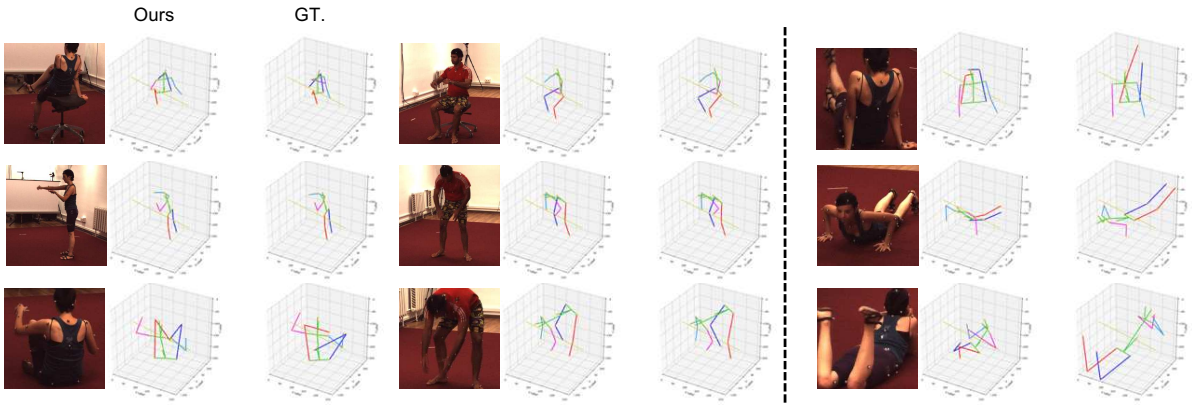

Figure 4. Qualitative results of ours on H3.6M validation set. The right part shows some of our failure cases. Our method may fail under severe occlusion and rare body poses.

lenge). We find this dataset is challenging to the compared methods due to: 1) large variation in camera positions; 2) difficult poses such as sitting and prone occupy a significant portion of the dataset; 3) variation in scale is large, due to the fact that without the knowledge of 3D, we cannot normalize 2D projections by distance or calculating bone length. The best we can do is to normalize 2D points by the size of 2D bounding box. This leads to certain pose e.g. sitting appears larger compared to others after normalization; 4) some of the methods fails to cope with a large number of samples (e.g. $>$5k). For those methods, we report result on the largest subset they can handle. We also try to compare with the recently proposed MUS [1], but their implementation fails to handle H3.6M dataset with large number of frames.

Despite of these difficulties, our implementation of Deep-NRSfM outperforms all of them. As shown in Table. 1, it reduces depth error by more than 33% compared to the second best. This means that switching to other NRSfM method is bound to inferior result of training a 3D pose estimator.

More interestingly, although our weakly supervised learning baseline (Weaksup-bs) is trained to reconstruct the same depth value produced by deep NRSfM, it actually gets slightly lower depth error compared to its regression target. This indicates that the deep image prior is taking effect, but still restricted by the noisy labels from Deep-NRSfM.

Finally, the pose estimation network learned by our knowledge distilling loss reduces the depth error from Deep-NRSfM's 76.5mm to 71.2mm. As shown in Fig. 3 and 1, this 5.3mm average difference includes a huge improvement in cases such as identifying if a leg is stretching towards or away from the camera.

5.4. Compare with weakly supervised methods

We compare with other weakly supervised 3D pose learning methods on the H3.6M validation set. In Table. 2, we first list the performance of Integral regression network by Sun *et al*. [36] as a supervised learning baseline. We copied its MPJPE (corresponding to ResNet50 with 256×256 input size and I_1 loss) from their paper. Since in our experiment, we're using exactly the same pose estimation network architecture, this serves as the upper bound of accuracy, which a weakly supervised learning method can achieve.

Next, we list results from 7 weakly supervised methods, and the type of their training source is marked. '2D' refers to 2D landmark annotation; '3D' represents any external 3D training source, including 3D human models, unpaired 3D skeleton dataset, synthetic dataset with 3D annotations, etc.; MV is the abbreviation for multi-view footage. We find that our method outperforms all the compared methods, while using the least amount of supervision. We also experiment with including MPII as extra training source, which leads to more error reduction. Fig. 4 shows some qualitative results of our method on the validation set. For per action error break down, we list PA-MPJPE of 13 different actions in Table 3.

6. Conlusion

In this paper, we presented a weakly supervised 3D pose learning algorithm requires zero 3D annotation. We proposed a novel loss to distill knowledge from a general type of NRSfM method based on dictionary learning. We also established a strong NRSfM baseline on a challenging dataset, beating all the state-of-the-arts. Despite its current sucess, the limitations of our method are: 1) we require weak perspective projection, thus objects with strong perspective change is not ideal for the proposed method; 2) we do not model missing labels yet, thus another iteration is needed to extend the method to datasets with lots of occluded/out-of-view objects. We leave these for future work.

References

[1] 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. 2, 8

[2] Ijaz Akhter, Yaser Sheikh, Sohaib Khan, and Takeo Kanade. Trajectory space: A dual representation for nonrigid structure from motion. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 33(7):1442–1456, 2011. 2

[3] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014. 7

[4] Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm with application to wavelet-based image deblurring. 2009. 3

[5] 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. 2

[6] Ching-Hang Chen and Deva Ramanan. 3d human pose estimation = 2d pose estimation + matching. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 2

[7] Wenzheng Chen, Huan Wang, Yangyan Li, Hao Su, Zhenhua Wang, Changhe Tu, Dani Lischinski, Daniel Cohen-Or, and Baoquan Chen. Synthesizing training images for boosting human 3d pose estimation. In *2016 Fourth International Conference on 3D Vision (3DV)*, pages 479–488. IEEE, 2016. 2

[8] Rishabh Dabral, Anurag Mundhada, Uday Kusupati, Safeer Afaque, Abhishek Sharma, and Arjun Jain. Learning 3d human pose from structure and motion. In *The European Conference on Computer Vision (ECCV)*, September 2018. 2

[9] 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. 2, 3, 6

[10] 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. 3

[11] 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. 6

[12] Dylan Drover, Rohith MV, Ching-Hang Chen, Amit Agrawal, Ambrish Tyagi, and Cong Phuoc Huynh. Can 3d pose be learned from 2d projections alone? In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 0–0, 2018. 1, 2, 6, 7

[13] Hsiao-Yu Fish Tung, Adam W. Harley, William Seto, and Katerina Fragkiadaki. Adversarial inverse graphics networks: Learning 2d-to-3d lifting and image-to-image translation from unpaired supervision. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 2, 6, 7

[14] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 270–279, 2017. 4

[15] Paulo FU Gotardo and Aleix M Martinez. Computing smooth time trajectories for camera and deformable shape in structure from motion with occlusion. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 33(10):2051–2065, 2011. 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. 6

[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. 6

[18] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339, jul 2014. 1, 2, 7

[19] Hanbyul Joo, Hao Liu, Lei Tan, Lin Gui, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. Panoptic studio: A massively multiview system for social motion capture. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3334–3342, 2015. 2

[20] Chen Kong and Simon Lucey. Prior-less compressible structure from motion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4123–4131, 2016. 2, 3, 6

[21] Chen Kong and Simon Lucey. Deep interpretable non-rigid structure from motion. *arXiv preprint arXiv:1902.10840*, 2019. 1, 2, 3, 6

[22] Chen Kong, Rui Zhu, Hamed Kiani, and Simon Lucey. Structure from category: a generic and prior-less approach. *International Conference on 3DVision (3DV)*, 2016. 2, 6

[23] Suryansh Kumar, Anoop Cherian, Yuchao Dai, and Hongdong Li. Scalable dense non-rigid structure-from-motion: A grassmannian perspective. *arXiv preprint arXiv:1803.00233*, 2018. 2

[24] Suryansh Kumar, Yuchao Dai, and Hongdong Li. Multi-body non-rigid structure-from-motion. In *3D Vision (3DV), 2016 Fourth International Conference on*, pages 148–156. IEEE, 2016. 2

[25] Minsik Lee, Jungchan Cho, and Songhwai Oh. Consensus of non-rigid reconstructions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4670–4678, 2016. 6

[26] Julieta Martinez, Rayat Hossain, Javier Romero, and James J Little. A simple yet effective baseline for 3d human pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2640–2649, 2017. 2

[27] 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. 3

[28] Georgios Pavlakos, Xiaowei Zhou, and Kostas Daniilidis. Ordinal depth supervision for 3d human pose estimation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 2

[29] Georgios Pavlakos, Xiaowei Zhou, Konstantinos G Derpanis, and Kostas Daniilidis. Coarse-to-fine volumetric prediction for single-image 3d human pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7025–7034, 2017. 2

[30] Georgios Pavlakos, Xiaowei Zhou, Konstantinos G Derpanis, and Kostas Daniilidis. Harvesting multiple views for marker-less 3d human pose annotations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6988–6997, 2017. 2

[31] Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Reconstructing 3d human pose from 2d image landmarks. In *European Conference on Computer Vision*, pages 573–586. Springer, 2012. 2

[32] Helge Rhodin, Mathieu Salzmann, and Pascal Fua. Unsupervised geometry-aware representation for 3d human pose estimation. In *The European Conference on Computer Vision (ECCV)*, September 2018. 1, 2, 6

[33] Helge Rhodin, Jrg Sprri, Isinsu Katircioglu, Victor Constantin, Frdric Meyer, Erich Mller, Mathieu Salzmann, and Pascal Fua. Learning monocular 3d human pose estimation from multi-view images. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 1, 2

[34] 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. 3

[35] Xibin Song, Peng Wang, Dingfu Zhou, Rui Zhu, Chenye Guan, Yuchao Dai, Hao Su, Hongdong Li, and Ruigang Yang. Apollocar3d: A large 3d car instance understanding benchmark for autonomous driving, 2018. 1

[36] Xiao Sun, Bin Xiao, Fangyin Wei, Shuang Liang, and Yichen Wei. Integral human pose regression. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 529–545, 2018. 2, 5, 6, 7, 8

[37] Denis Tome, Chris Russell, and Lourdes Agapito. Lifting from the deep: Convolutional 3d pose estimation from a single image. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 1, 2, 6

[38] Hsiao-Yu Tung, Hsiao-Wei Tung, Ersin Yumer, and Katerina Fragkiadaki. Self-supervised learning of motion capture. In *Advances in Neural Information Processing Systems*, pages 5236–5246, 2017. 1, 2, 6

[39] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Deep image prior. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9446–9454, 2018. 5

[40] Gul Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J Black, Ivan Laptev, and Cordelia Schmid. Learning from synthetic humans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 109–117, 2017. 2

[41] Bastian Wandt and Bodo Rosenhahn. Repnet: Weakly supervised training of an adversarial reprojection network for 3d human pose estimation. *CoRR*, abs/1902.09868, 2019. 2

[42] Chaoyang Wang, José Miguel Buenaposada, Rui Zhu, and Simon Lucey. Learning depth from monocular videos using direct methods. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2022–2030, 2018. 4

[43] Jiajun Wu, Tianfan Xue, Joseph J Lim, Yuandong Tian, Joshua B Tenenbaum, Antonio Torralba, and William T Freeman. Single image 3d interpreter network. In *European Conference on Computer Vision*, pages 365–382. Springer, 2016. 1, 2, 6, 7

[44] Wei Yang, Wanli Ouyang, Xiaolong Wang, Jimmy Ren, Hongsheng Li, and Xiaogang Wang. 3d human pose estimation in the wild by adversarial learning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 2

[45] Andrei Zanfir, Elisabeta Marinoiu, and Cristian Sminchisescu. Monocular 3d pose and shape estimation of multiple people in natural scenes - the importance of multiple scene constraints. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 2

[46] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G Lowe. Unsupervised learning of depth and ego-motion from video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1851–1858, 2017. 4

[47] Xingyi Zhou, Qixing Huang, Xiao Sun, Xiangyang Xue, and Yichen Wei. Towards 3d human pose estimation in the wild: a weakly-supervised approach. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 398–407, 2017. 1, 2, 7

[48] 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. 2