C3DPO: Canonical 3D Pose Networks for Non-Rigid Structure From Motion

David Novotny* Nikhila Ravi* Benjamin Graham Natalia Neverova Andrea Vedaldi

{dnovotny,nikhilar,benjamingraham,nneverova,vedaldi}@fb.com

Facebook AI Research

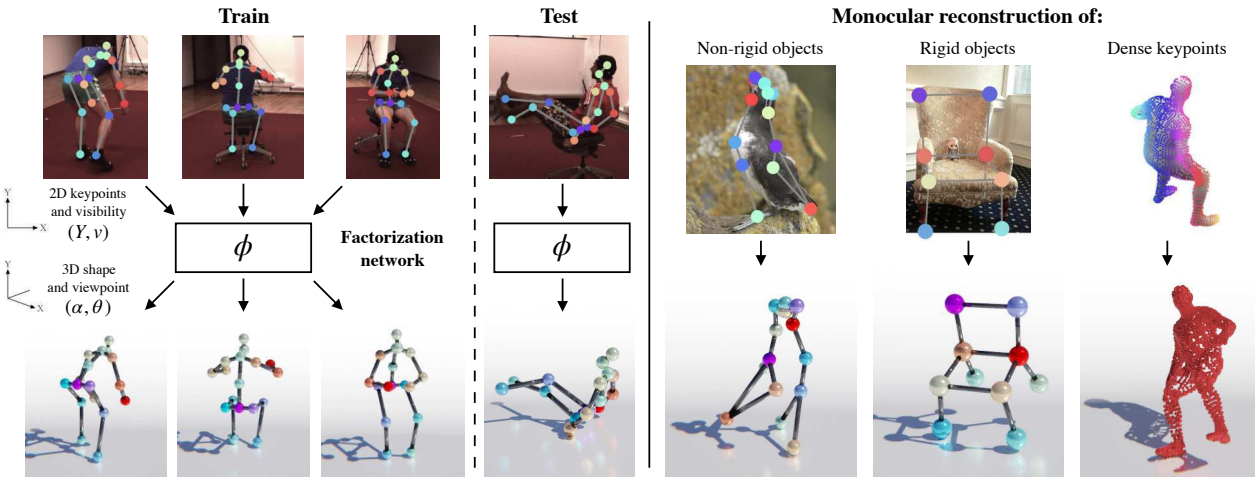

Figure 1: Our method learns a 3D model of a deformable object category from 2D keypoints in unconstrained images. It comprises a deep network that learns to factorize shape and viewpoint and, at test time, performs monocular reconstruction.

Abstract

We propose C3DPO, a method for extracting 3D models of deformable objects from 2D keypoint annotations in unconstrained images. We do so by learning a deep network that reconstructs a 3D object from a single view at a time, accounting for partial occlusions, and explicitly factoring the effects of viewpoint changes and object deformations. In order to achieve this factorization, we introduce a novel regularization technique. We first show that the factorization is successful if, and only if, there exists a certain canonicalization function of the reconstructed shapes. Then, we learn the canonicalization function together with the reconstruction one, which constrains the result to be consistent. We demonstrate state-of-the-art reconstruction results for methods that do not use ground-truth 3D supervision for a number of benchmarks, including Up3D and PASCAL3D+.

1. Introduction

3D reconstruction of static scenes is mature, but the problem remains challenging when objects can deform due to articulation and intra-class variations. In some cases, deformations can be avoided by capturing multiple simultaneous images of the object. However, this requires expensive hardware comprising several imaging sensors and only provides instantaneous 3D reconstructions of the objects without modelling their deformations. Extracting deformation models requires establishing correspondences between the instantaneous 3D reconstructions, which is often done by means of physical markers. Modern systems such as the Panoptic Studio [14] can align 3D reconstructions without markers, but require complex specialized hardware, making them unsuitable for use outside a specialized laboratory.

In this paper, we thus consider the problem of reconstructing and modelling 3D deformable objects given only unconstrained monocular views and keypoint annotations. Traditionally, this problem has been regarded as a generalization of static scene reconstruction, and approached by extending Structure from Motion (SFM) techniques. Due to their legacy, such Non-Rigid SFM (NR-SFM) methods have often focused on the geometric aspects of the problem, but the quality of the reconstructions also depends on the ability to *model statistically* the object shapes and deformations.

*Authors contributed equally.

We argue that modern deep learning techniques may be used in NR-SFM to capture much better statistical models of the data than the simple low-rank constraints employed in traditional approaches. We thus propose a method that reconstructs the object in 3D while learning a deep network that models it. This network is inspired by recent approaches [21, 16, 30, 10, 18] that accurately lift 2D keypoints to 3D given a single view of the object. The difference is that our network does not require 3D information for supervision, but is instead trained jointly with 3D reconstruction from 2D keypoints.

Our model, named C3DPO, has two important innovations. First, it performs 3D reconstruction by *factoring* the effects of viewpoint changes and object deformations. Hence, it reconstructs the 3D object in a canonical frame that registers the overall 3D rigid motion and leaves as residual variability only the motion "internal" to the object.

However, achieving this factorization correctly is non-trivial, as noted extensively in the NR-SFM literature [40]. Our second innovation is a solution to this problem. We observe that, if two 3D reconstructions overlap up to a rigid motion, they must coincide (since the reconstruction network should remove the effect of a rigid motion). Hence, any class of 3D shapes equivalent up to a rigid motion must contain at most one canonical reconstruction. If so, there exits a "canonicalization" function that maps elements in each equivalent class to this canonical reconstruction. We exploit this fact by learning, together with the reconstruction network, a second network that performs this canonicalization, which regularizes the solution.

Empirically, we show that these innovations lead to a very effective and robust approach to non-rigid reconstruction and modelling of deformable 3D objects from unconstrained 2D keypoint data. We compare C3DPO against several traditional NR-SFM baselines as well as other approaches that use deep learning [16, 21]. We test on a number of benchmarks, including Human3.6M, PASCAL3D+, and Synthetic Up3D, showing superior results for methods that make no use of ground-truth 3D information.

2. Related work

There are several lines of work which address the problem of 3D shape and viewpoint recovery of a deforming object from 2D observations. This section covers relevant work in NR-SFM and recent deep-learning based methods.

NR-SFM. There are several solutions to the NR-SFM problem which can recover the viewpoint and 3D shape of a deforming object from 2D keypoints across multiple frames [4, 6, 5, 9], the majority of which are based on Bregler's factorization framework [6]. However the NR-SFM problem is severely under constrained as both the camera and 3D object are moving along with the object deforming. This poses a challenge in correctly factoring the viewpoint and shape [40], and additional problems with missing values in the observations. Priors about the shape and the camera motion are employed to improve conditioning of the problem, including the use of low-rank subspaces in the spatial domain [3, 11, 9, 43], temporal domain, for example, fitting 2D keypoint trajectories to a set of predefined DCT basis functions [4, 5], spatio-temporal domain [1, 12, 22, 23], multiple unions of low-rank subspaces [43, 2], learning an overcomplete dictionary of basis shapes from 3D motion capture data and imposing an L1 penalty on basis coefficients [41, 42] and imposing Gaussian priors on the shape coefficients [33].

Many of these approaches however, as we have empirically verified, are not scalable and can only reliably reconstruct datasets of few thousands of images and hundreds of keypoints. Furthermore, many of them require keypoint correspondences for the *same* instance from *multiple* images from a monocular view or from multi-view cameras. Finally, in contrast to our method, using the listed approaches it is difficult or computationally expensive to reconstruct new test samples after training on a fixed collection of training shapes.

Category specific 3D shapes. Also related are methods that reconstruct shapes of a visual object category, such as cars or birds. [8] is an early work that learns a morphable model of dolphins from 2D keypoints and segmentation masks. Using similar supervision, Vicente *et al*. [37, 7] reconstruct the categories of PASCAL VOC. An important part of the pipeline is an initial SFM algorithm which returns a mean shape and camera matrices of each object category. Similarly, Kar *et al*. [18] utilize an NR-SFM method for reconstructing categories from PASCAL3D+. [27] proposed the first purely image-driven method for single-view reconstruction of rigid object categories. Most recently, Kanazawa *et al*. [16] train a deep network capable of learning both shape and texture of deformable objects. The commonality among the aforementioned methods is their reliance on the initial SFM/NR-SFM step which can often fail. Our method overcomes this problem by learning a monocular shape predictor in a single step without any additional, potentially unreliable, preprocessing steps.

Weakly supervised 3D human pose estimation. Our approach is related to weakly supervised methods that lift 2D human skeleton keypoints to 3D given a single input view. Besides the fully supervised methods [25, 26], several works have explored multi-view supervision [20, 29, 31], ordinal depth supervision [28], unpaired 2D-3D data [30, 36, 41, 15] or videos [17] to alleviate the need for full 2D-3D annotations. While these auxiliary sources of supervision allow for compelling 3D predictions, in this work we use only inexpensive 2D keypoint labels.

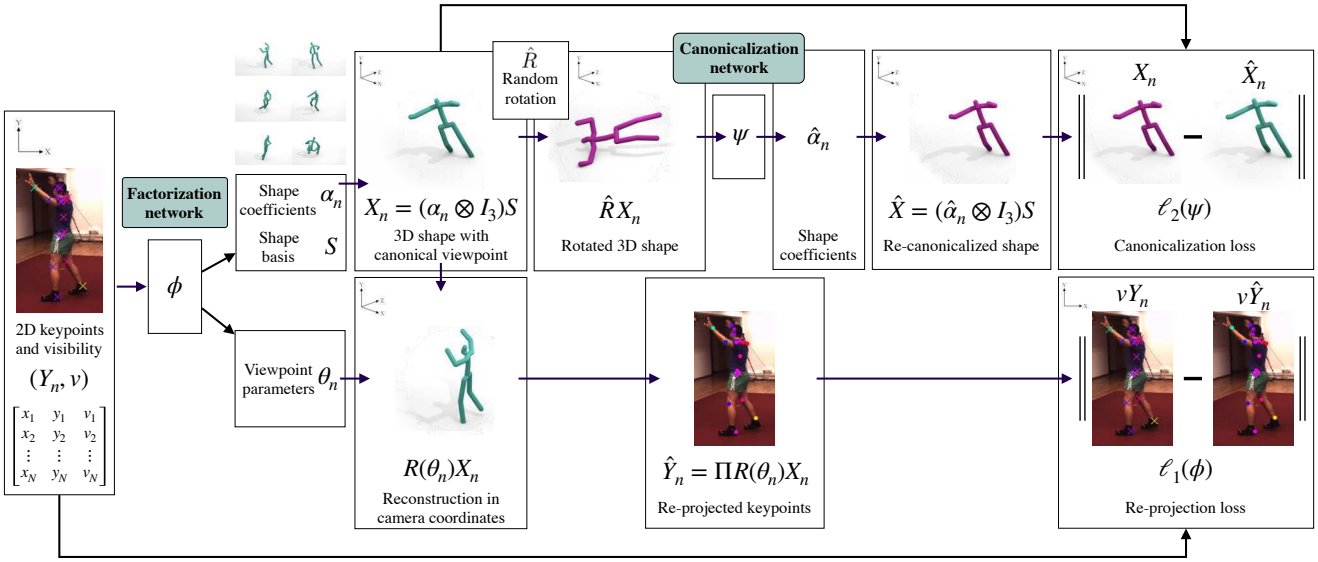

Figure 2: **An overview of C3DPO.** The lower branch learns monocular 3D reconstruction by minimizing the re-projection error ℓ_1. The upper branch learns to factorize viewpoints and internal deformations by the means of the canonicalization loss.

Closer to our supervisory scheme, [21, 10] recently proposed a method that rotates the 3D-lifted keypoints into new views and validates the resulting projections with an adversarial network that learns the distribution of plausible 2D poses. However, both methods require *all* keypoints to be visible in every frame. This restricts their use to 'multi-view' datasets such as Human3.6M. In addition to the 2D keypoints, [10] use the intrinsic camera parameters, and 3D ground truth data to generate new synthetic 2D views, which leads to substantially better quantitative results at the cost of a greater level of supervision.

To conclude, our contribution differs from prior work as it 1) recovers both 3D canonical shape and viewpoint using only 2D keypoints in a single image at test time, 2) uses a novel self-supervised constraint to correctly factorize 3D shape and viewpoint, 3) can handle occlusions and missing values in the observations, 4) works effectively across multiple object categories.

3. Method

We start by summarizing some background facts about SFM and NR-SFM and then we introduce our method.

3.1. Structure from motion

The input to *structure from motion* (SFM) are tuples $y_n = (y_{n1}, \ldots, y_{nK}) \in \mathbb{R}^{2\times K}$ of 2D keypoints, representing N *views* $y_1, \ldots, y_n$ of a rigid object. The views are generated from a single tuple of 3D points $X = (X_1, \ldots, X_K) \in \mathbb{R}^{3\times K}$, called the *structure*, and N *rigid motions* $(R_n, T_n) \in SO(3) \times T(3)$. The views, the structure, and the motions are related by equations $y_{nk} = \Pi(R_n X_k + T_n)$ where $\Pi : \mathbb{R}^3 \to \mathbb{R}^2$ is the *camera projection* function. For simplicity of exposition we consider an orthographic camera. In this case, the projection function is linear and given by matrix $\Pi = \begin{bmatrix} I_2 & 0 \end{bmatrix}$ where $I_2 \in \mathbb{R}^{2\times 2}$ is the 2D identity matrix and the projection equation $y_{nk} = \Pi R_n X_k + \Pi T_n$ is also linear. If all keypoints are visible, they can be centered together with the structure, eliminating the translation from this equation (details in the supplementary material). This yields the simplified system of equations $y_{nk} = M_n X_k$, where $M_n = \Pi R_n$ are the camera view matrices, or *viewpoints*. The equations can be written in matrix form as

$$Y = \begin{bmatrix} y_{11} & \cdots & y_{1K} \\ \vdots & \ddots & \vdots \\ y_{N1} & \cdots & y_{NK} \end{bmatrix}, \; M = \begin{bmatrix} M_1 \\ \vdots \\ M_N \end{bmatrix}, \; \underbrace{Y}_{2N\times K} = \underbrace{M}_{2N\times 3} \underbrace{X}_{3\times K}. \tag{1}$$

Hence, SFM can be formulated as factoring the views Y into viewpoints M and structure X. This factorization is not unique, resulting in a mild reconstruction ambiguity, as discussed in supplementary material.

3.2. Non-rigid structure from motion

The *non-rigid* SFM (NR-SFM) problem is similar to the SFM problem, except that the structure X_n is allowed to deform from one view to the next. Obtaining a non-trivial solution is only possible if such deformations are constrained in some manner. The simplest constraint is a linear model $X_n = X(\alpha_n; S)$, expressing the structure X_n as a small vector of view-specific *pose* parameters $\alpha_n \in \mathbb{R}^D$ and a view-invariant *shape basis* $S \in \mathbb{R}^{3D\times K}$:

$$X(\alpha_n; S) = (\alpha_n \otimes I_3)S \tag{2}$$

where α_n is a row vector and $\otimes$ is the Kronecker product. We can expand the equation for individual points as $X_{nk} = \sum_{d=1}^{D} \alpha_{nd} S_{dk}$ where $S_{dk} \in \mathbb{R}^3$ is a shorthand for the subvector $S_{3d-2:3d,k}$. We can also extend it to all points and poses as $X = (\alpha \otimes I_3)S \in \mathbb{R}^{3N \times K}$ where $\alpha \in \mathbb{R}^{N \times D}$ encodes a pose per row.

Given multiple views of the points, the goal of NR-SFM is to recover the views, the poses, and the shape basis from observations $y_{nk} = \Pi(R_n \sum_{d=1}^{D} \alpha_{nd} S_{dk} + T_n)$. As in SFM, for orthographic projection the translation can be removed from the equation by centering, and NR-SFM can be expressed as a multi-linear matrix factorization problem:

$$\underbrace{Y}_{2N \times K} = \underbrace{\bar{M}}_{2N \times 3N \text{ (sparse)}} \; (\underbrace{\alpha}_{N \times D} \otimes I_3) \; \underbrace{S}_{3D \times K}, \qquad (3)$$

where the N camera view matrices are contained in the block-diagonal matrix $\bar{M} = \mathrm{diag}(M_1, \ldots, M_N)$. Like SFM, this factorization has mild ambiguities, discussed in the supplementary material.

3.3. Monocular motion and structure estimation

Once the shape basis S is learned, model (3) can be used to reconstruct viewpoint and pose given a single view Y of the object, yielding monocular reconstruction. However, this still requires solving a matrix factorization problem.

For C3DPO, we propose to instead *learn* a mapping Φ that performs this factorization in a feed-forward manner, recovering the view matrix M and the pose parameters α from the keypoints Y:

$$\Phi : \mathbb{R}^{2K} \times \{0,1\}^K \to \mathbb{R}^D \times \mathbb{R}^3, \quad (Y, v) \mapsto (\alpha, \theta).$$

Here, v is a (row) vector of boolean flags denoting whether a keypoint is visible in that particular view or not (if the keypoint is not visible, the flag as well as the spatial coordinates of the point are set to zero). The function outputs the D pose parameters α and the three parameters $\theta \in \mathbb{R}^3$ of the camera view matrix $M(\theta) = \Pi R(\theta)$, where the rotation is given by $R(\theta) = \mathrm{expm}[\theta]_\times$, expm is the matrix exponential and $[\cdot]_\times$ is the hat operator.

The benefit of using a learned mapping, besides speed, is the fact that it can embody prior information on the structure of the object which is not apparent in the linear model. The mapping itself is learned by minimizing the *re-projection loss* obtained by averaging the loss over visible keypoints:

$$\ell_1(Y, v; \Phi, S) = \frac{1}{K} \sum_{k=1}^{K} v_k \cdot \| Y_k - M(\theta)(\alpha \otimes I_3) S_{:,k} \|_\epsilon, \qquad (4)$$

where $(\alpha, \theta) = \Phi(Y, v)$ and $\|z\|_\epsilon = (\sqrt{1 + (\|z\|/\epsilon)^2} - 1)\epsilon$ is the pseudo-huber loss with soft threshold ϵ[1]. Given a dataset $(Y, v) \in \mathcal{D}$ of views of an object category, the neural network Φ is trained by minimizing the empirical average of this loss. This setup is illustrated in the bottom half of fig. 2.

[1]We set $\epsilon = 0.01$ in all experiments.

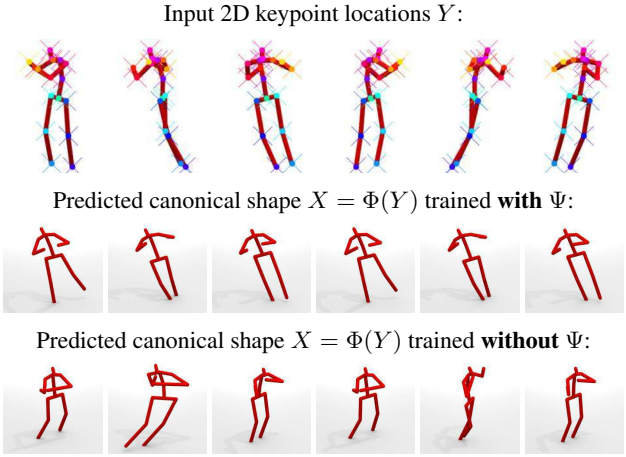

Figure 3: **Effects of the canonicalization network Ψ.** Each column shows a 2D pose Y input to the pose prediction network Φ (top) and the predicted 3D canonical shape $X = \Phi(Y)$ when Φ is trained with (middle) and without (bottom) the canonicalization network Ψ. Observe that training with Ψ provides significantly more stable canonical shape predictions X as the input pose rotates around the camera y-axis.

3.4. Consistent factorization via canonicalization

A challenge with NR-SFM is the ambiguity in decomposing variations in the 3D shape of an object into viewpoint changes (rigid motions) and internal object deformations [40]. In this section, we propose a novel approach to directly encourage the reconstruction network Φ to be *consistent* in the way reconstructions are performed. This means that it must not be possible for the network to produce two different 3D reconstructions that differ only by a rigid motion, because such a difference should have been instead explained as a viewpoint change.

Formally, let $\mathcal{X}_0$ be the set of all reconstructions $X(\alpha; S)$ obtained by the network, where the parameters $(\alpha, \theta) = \Phi(Y, v)$ are obtained by considering all possible views (Y, v) of the object. If the network factorizes viewpoint and pose consistently, then there cannot be two different reconstructions $X, X' \in \mathcal{X}_0$ related by a mere viewpoint change $X' = RX$. This is formalized by the following definition:

Definition 1. *The set $\mathcal{X}_0 \subset \mathbb{R}^{3 \times K}$ has the* transversal property *if, for any pair $X, X' \in \mathcal{X}_0$ of structures related by a rotation $X' = RX$, then $X = X'$.*

Transversality can also be interpreted as follows: rotations partition the space of structures $\mathbb{R}^{3 \times K}$ into equivalence classes. We would like reconstructions to be unique

within each equivalence class. A set that has a unique or *canonical* element for each equivalent class is also called a *transversal*. Definition 1 captures this idea for the set of reconstructions $\mathcal{X}_0$.

For the purpose of learning, we propose to enforce transversality via the following characterizing property (proofs in the supplementary material):

Lemma 1. *The set $\mathcal{X}_0 \subset \mathbb{R}^{3\times K}$ has the transversal property if, and only if, there exists a* canonicalization function *$\Psi : \mathbb{R}^{3\times K} \to \mathbb{R}^{3\times K}$ such that, for all rotations $R \in SO(3)$ and structures $X \in \mathcal{X}_0$, $X = \Psi(RX)$.*

Intuitively, this lemma states that, if $\mathcal{X}_0$ has the transversal property, then any rotation of its elements can be undone unambiguously. Otherwise stated, we can construct a canonicalization function with range in the set of reconstructions $\mathcal{X}_0$ if, and only if, this set contains only canonical elements, i.e. it has the transversal property (definition 1).

For C3DPO, the lemma is used to enforce a consistent decomposition in viewpoint and pose via the following loss:

$$\ell_2(X, R; \Psi) = \frac{1}{K}\sum_{k=1}^{K} \|X_{:,k} - \Psi(RX)_{:,k}\|_\epsilon, \qquad (5)$$

where $R \in SO(3)$ is a randomly-sampled rotation, and Ψ is a regressor *canonicalization* network trained in parallel with the factorization network Φ.

Regularizer ℓ_2 (eq. (5)) is combined with the re-projection loss ℓ_1 (eq. (4)) as follows (fig. 2): given an input sample Y_n, we first pass it through $\Phi(Y_n, v)$ to generate viewpoint and pose parameters θ_n and α_n, which enter the re-projection loss ℓ_2. In addition, a random rotation $\hat{R}$ is applied to the generated structure $X_n = X(\alpha_n; S)$, and $\hat{R}X_n$ is passed to the auxiliary canonicalization neural network Ψ. Ψ then undoes $\hat{R}$ by predicting shape coefficients $\hat{\alpha}_n$ that produce a shape $\hat{X}_n = X(\hat{\alpha}_n; S)$ which should reconstruct the unrotated input shape X_n as precisely as possible. This is enforced by passing $\hat{X}_n$ and X_n to the loss ℓ_2. The two networks Φ and Ψ are trained in parallel by minimizing $\ell_1 + \ell_2$, which encourages learning consistent viewpoint-pose factorization. The effect of the loss is illustrated in fig. 3.

3.5. In-plane rotation invariance

Rotation equivariance is another property of the factorization network that can be used to constrain learning. Let $Y = \Pi RX$ be a view of the 3D structure X. Rotating the camera around the optical axis has the effect of applying a rotation $r_z \in SO(2)$ to the keypoints. Hence, the two reconstructions $\Phi(Y, v) = (\alpha, \theta)$ and $\Phi(r_z Y, v) = (\alpha', \theta')$ must yield the same 3D structure $\alpha = \alpha'$. This is captured via a modified reprojection loss that exchanges α for α':

$$\ell_3(Y, v; \Phi, S) = \frac{1}{K}\sum_{k=1}^{K} v_k \cdot \|r_z Y_k - M(\theta')(\alpha \otimes I_3)S_{:,k}\|_\epsilon \qquad (6)$$

This yields the combined loss $\ell_2 + \ell_3$ (the range of losses are comparable are combined with equal weight).

4. Experiments

In this section, we compare our method against several strong baselines. First, the employed benchmarks are described followed by quantitative and qualitative evaluations.

4.1. Datasets

We consider three diverse benchmarks containing images of objects with 2D keypoints annotations. The datasets differ by keypoint density, object type, deformations, and intra-class variations.

Synthetic Up3D (S-Up3D) We first validate C3DPO in a noiseless setting using a large synthetic 2D/3D dataset of dense human keypoints based on the Unite the People 3D (Up3D) dataset [24]. For each Up3D image, the SMPL body shape and pose parameters are provided and are used to produce a mesh with 6890 vertices. Each of the 8515 meshes is randomly rotated into 30 different views and the orthographic projection of each vertex is recorded along with its visibility (computed using a ray tracer). The goal is then to recover the 3D shapes given the set of 2D keypoint renders. We maintain the same train/test split as in the Up3D dataset.

Similar to [24], performance is evaluated on the 79 representative vertices of the SMPL model. Although C3DPO can reconstruct the original set of 6890 SMPL model keypoints effortlessly, we evaluate on a subset of points due to a limited scalability of some of the baselines [33, 11]. For the same reason, we further randomly sampled the generated test poses to 15k images. Performance is measured by averaging a 3D reconstruction error metric (see below) over all frames in the test set.

PASCAL3D+ [39] Similar to [16, 35], we evaluate our method on the the PASCAL3D+ dataset which consists of PASCAL VOC and ImageNet images for 12 rigid object categories with a set of sparse keypoints annotated on each image (deformations still arise due to intra-class shape variations). There are up to 10 CAD models available for each category, from which one is manually selected and aligned for each image, providing an estimate of the ground truth 3D keypoint locations. To maintain consistency between the 2D and 3D keypoints, we use the 2D orthographic projections of the aligned CAD model keypoints as opposed to the per-image 2D keypoint annotations, and update the visibility indicators based on the CAD model annotations.

Human3.6M [13] is perhaps the largest dataset of human poses annotated with 3D ground truth extracted using

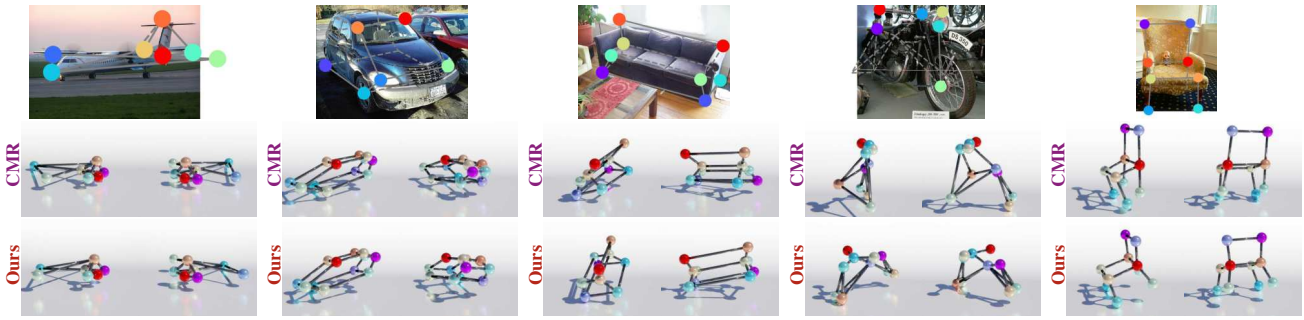

Figure 4: **Qualitative results on PASCAL3D+** comparing our method **C3DPO-HRNet** (red) with **CMR** [16] (violet). Each column contains the input monocular 2D keypoints (top) and lifting of the 2D keypoints into 3D by CMR (middle) and by our method (bottom) viewed from 2 different angles.

Method	MPJPE	Stress
EM-SfM [33]	0.107	0.061
GbNrSfM [11]	0.093	0.062
C3DPO-base	0.160	0.105
C3DPO-equiv	0.154	0.102
C3DPO	**0.068**	**0.040**

Table 1: **Results on the synthetic Up-3D (S-Up3D)** comparing our method (C3DPO), NRSfM baselines [33, 11] and two variants of our method (C3DPO-equiv, C3DPO-base) which ablate effects of individual components of C3DPO.

Method	MPJPE	Stress
GbNrSfM [11]	184.6	111.3
EM-SfM [33]	131.0	116.8
C3DPO-base	53.5	46.8
C3DPO-equiv	50.1	44.5
C3DPO	**38.0**	**32.6**
CMR [16]†	74.4	53.7
C3DPO + HRNet†	**57.5**	**41.4**

Table 2: **Average reconstruction error (MPJPE) and ℓ_1 stress over the 12 classes of Pascal3D** comparing our method C3DPO with two ablations of our approach (C3DPO-equiv, C3DPO-base) and the methods from [11, 16, 33]. Approaches marked with † predict 3D shape without knowledge of the ground-truth 2D keypoints at test time.

MoCap systems. As in [21], two variants of the dataset are used: the first contains ground-truth 2D keypoints during both train and test time and in the second, 2D keypoint locations are obtained by the Stacked Hourglass network of [34]. We closely follow the evaluation protocol of [21] and report absolute errors measured over 17 joints without any procrustes alignment. We maintain the same train and test split as [21], and report an average over errors attained for each frame in a given MoCap sequence of an action type.

CUB-200-2011 [38] consists of 11,788 images of 200 bird species. Each image is annotated with 2D locations of 15 semantic keypoints and corresponding visibility indicators. There are no ground truth 3D keypoints for this dataset so we only perform a qualitative evaluation. We use the 2D annotations from [16].

4.2. Evaluation metrics

As common practice, the absolute mean per joint position error is reported: $\mathbf{MPJPE}(X^*, X) = \sum_{k=1}^{K} \|X_k - X_k^*\|/K$, where $X_k \in \mathbb{R}^3$ is the predicted 3D location of the k-th keypoint and X_k^* is its corresponding ground-truth 3D location (both in the 3D frame of the camera).

In order to evaluate MPJPE properly, two types of projection ambiguities have to be handled. To deal with the **absolute depth** ambiguity, for Human3.6M we follow [21] and normalize each pose by applying a translation that puts the skeleton root to the origin of the coordinate system. For PASCAL3D+ and S-Up3D, the mean depth of predicted and ground truth point clouds is zero centered before evaluation. The second, **depth flip** ambiguity, is resolved as in [33] by evaluating MPJPE twice for the original and depth-flipped point cloud, retaining the better of the two.

We also report the ℓ_1 $\mathbf{Stress}(X, X^*) = \sum_{i<j} | \|X_i - X_j\| - \|\hat{X}_i^* - X_j^*\| |_1/(K(K-1))$. This metric is invariant to camera pose and the absolute depth and z-flip ambiguities.

4.3. Baselines

C3DPO is compared to several strong baselines. **EM-SfM** [33] and **GbNrSfM** [11] are NR-SFM methods with publicly available code. Because, when using [11, 33], it is difficult to make predictions on previously unseen data, we run the two methods directly on the test set and report results after convergence. This gives the two baselines an advantage over our method. On Human3.6M, out of several available methods, we compare with [21] (**Pose-GAN**) which is a current state-of-the-art approach for unsupervised 3D pose estimation that does not require any 3D, multiview or video annotations. Unlike other weakly supervised methods [10, 20], Pose-GAN does not assume knowl-

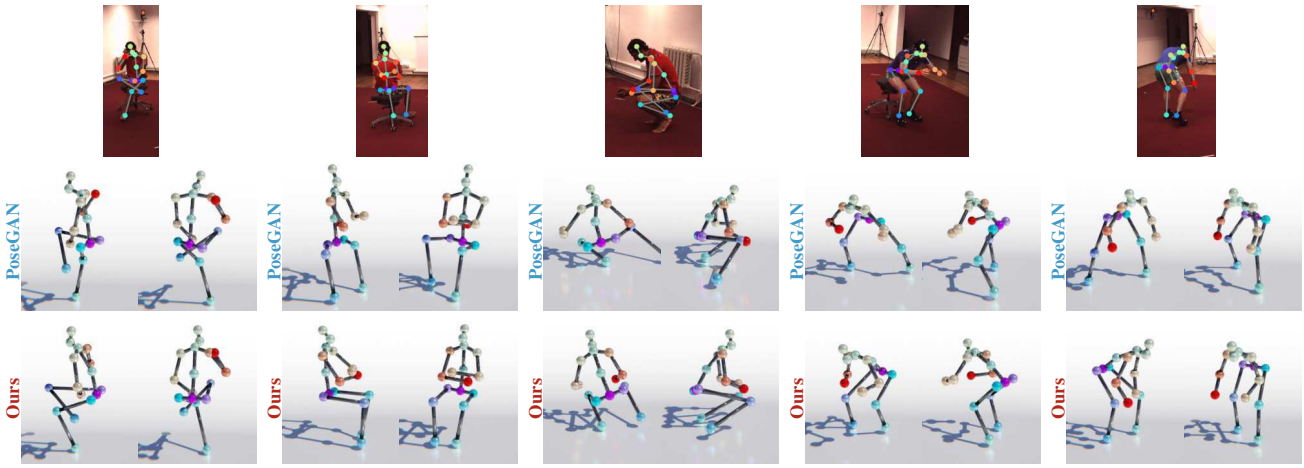

Figure 5: **3D poses on Human3.6M** predicted from monocular keypoints. Each column contains the input 2D keypoints (top) and a comparison between **PoseGAN** [21] (blue, middle), and our method **C3DPO** (bottom, red) from two 3D viewpoints.

Method	Ground truth pose		Predicted pose	
	MPJPE	Stress	MPJPE	Stress
Pose-GAN [21]	130.9	51.8	173.2	-
C3DPO-base	135.2	56.9	201.6	101.4
C3DPO-equiv	128.2	53.0	190.4	93.9
C3DPO	**101.8**	**43.5**	**153.0**	**86.0**

Table 3: **Results on Human3.6M** reporting average per joint position error (MPJPE) and ℓ_1 stress over the set of test actions (follows the evaluation protocol from [21]). We compare performance, when ground truth pose keypoints are available during test-time (2nd and 3rd column) and when the keypoints are predicted using the Stacked Hourglass network [34] (4th and 5th column).

edge of the camera intrinsic parameters, hence it is the most comparable to our approach. To ensure fair comparison, we use their public evaluation code together with the provided keypoint detections of the Stacked Hourglass model. PoseGAN was not tested on other datasets as the method cannot handle inputs with occluded keypoints. On PASCAL3D+, our method is compared with Category-Specific Mesh Reconstruction (**CMR**) from [16]. CMR provides results for 2 categories out of the 12 of PASCAL3D+, but we trained models for all 12 categories using the public code. Note that CMR additionally uses segmentation masks during training, hence has a higher level of supervision than our method.

The effects of individual components of our method are evaluated by ablation and recording the change in performance. This generates three variants of our method: (1) **C3DPO-base** only optimizes the re-projection loss $\ell_1(\Phi)$ from eq. (4), (2) **C3DPO-equiv** replaces $\ell_1(\Phi)$ with the optimization of the z-invariant loss $\ell_3(\Phi)$ (section 3.5), (3) **C3DPO** extends C3DPO-equiv with the secondary canonicalization network Ψ (section 3.4).

4.4. Technical details

Networks Ψ and Φ share the same core architecture and consist of 6 fully connected residual layers each with 1024/256/1024 neurons (please refer to the supplementary material for architecture details). Residual skip connections were found important since they prevented networks from converging to a rigid average shape.

Keypoints Y_n are first zero-centered before being passed to Ψ. We further scale each set of centered 2D locations by the same scalar factor so their extent is roughly $[-1, 1]$ on average in the axis of the highest variance. The network is trained with a batched SGD optimizer with momentum with an initial learning rate of 0.001, decaying 10 fold whenever the training objective plateaued. The batch size was set to 256. The training losses $\ell_3(\Phi)$ and $\ell_2(\Psi)$ were weighted equally.

For Human3.6M, we did not model the translation T of the camera as the centroid of the input 2D keypoints coincides with the centroid of the 3D shape (due to the lack of occluded keypoints). For the other datasets, which contain occlusions, we estimate the camera translation as the difference vector between the mean of the input visible points and the re-projected visible 3D shape keypoints.

In order to adapt our method for the multiclass setting of PASCAL3D+, which has different sets of keypoints for each of the 12 object categories, we adjust the keypoint annotations as follows. For each object category $\mathcal{C} \in \{1, \ldots, 12\}$ with a set of $K_\mathcal{C}$ keypoints $Y_n^\mathcal{C} \in \mathcal{R}^{2 \times K_\mathcal{C}}$ in an image n, we form a multiclass keypoint annotation $Y_n = \left[\mathbf{0}, \ldots, Y_n^\mathcal{C}, \ldots, \mathbf{0}\right]$ by assigning $Y_n^\mathcal{C}$ to the $\mathcal{C}$-th block of Y_n and padding with zeros. The visibility indicators v_n are expanded in a similar fashion. This avoids reconstructing each class separately, allowing our method to train only

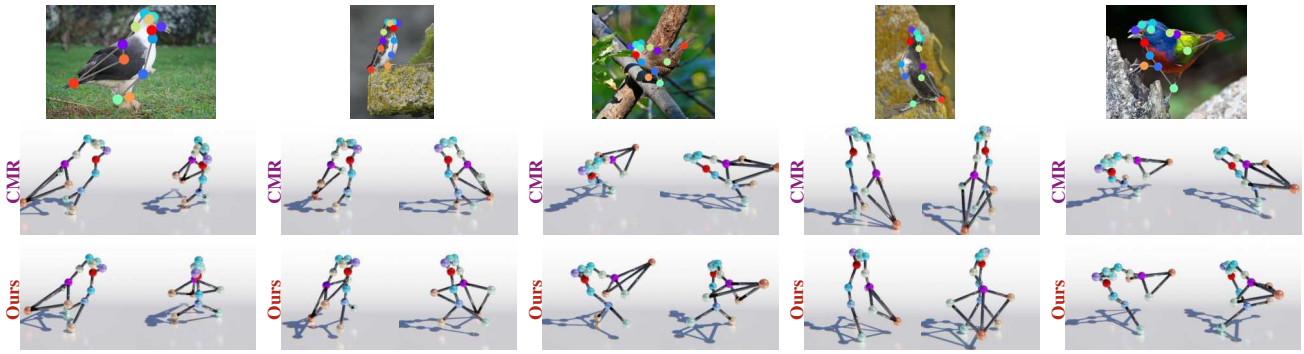

Figure 6: **Qualitative results on CUB-200-2011** comparing our method **C3DPO-HRNet** (red) with **CMR** [16] (violet). Each column contains the input monocular 2D keypoints (top), lifting of the 2D keypoints into 3D by CMR (middle) and by our method (bottom) from 2 different 3D viewpoints (the same view and a view offset by 90° along camera y-axis).

once for all classes. This also tests the ability of the model to capture non-rigid deformations not only within, but also across object categories. While this expanded version of keypoint annotations was also tested for GbNrSfM, for EM-SfM, we could not obtain satisfactory performance and reconstructed each class independently. Similarly for CMR, 12 class-specific models were trained separately.

4.5. Results

Synthetic Up3D. Table 1 reports the results on the S-Up3D dataset. Our method outperforms both EM-SfM and GbNrSfM, which validates our approach as a potential replacement for existing NR-SFM methods based on matrix factorization. The table also shows that C3DPO performs substantially better than C3DPO-base, highlighting the importance of the canonicalization network Ψ.

PASCAL3D+. For PASCAL3D+ we consider two types of methods. Methods of the first type include GbNrSfM and EM-SfM and take as input 2D ground truth keypoint annotations on the PASCAL3D+ test set, reconstructing it directly. The second type is CMR which uses ground truth annotations for training, but does not use keypoint annotations for evaluation on the test data. In order to make our method comparable with CMR, we used as a detector the High Resolution Residual network (HRNet [19]), training it on the 2D keypoint annotations from the PASCAL3D+ training set. The trained HRNet is applied to the test set to extract the 2D keypoints Y and these are lifted to 3D by applying C3DPO (abbreviated as **C3DPO+HRNet**).

The results are reported in table 2. C3DPO performs better than EM-SfM and GbNrSfM when ground truth keypoints are available during testing. Our method also outperforms CMR by 16%. On several classes (motorbike, train), we obtain significantly better results due to the reliance of CMR on an initial off-the-shelf rigid SFM algorithm that fails to obtain satisfactory reconstructions. This result is especially interesting since, unlike CMR, C3DPO is trained for all classes at once without ground truth segmentation masks. Figure 4 contains qualitative evaluation.

Human3.6M. Results on the Human3.6M dataset are summarized in table 3. C3DPO outperforms Pose-GAN for both ground truth and predicted keypoint annotations. Again, C3DPO improves over baseline C3DPO-base by a significant margin. Example reconstructions are in Figure 5.

CUB-200-2011. Similar to PASCAL3D+, in order to make our method comparable with CMR, HRNet is trained on keypoints from the CUB-200-2011 train set and used to predict keypoints on unseen test images which are then input to C3DPO. Figure 6 compares qualitatively our reconstructions to CMR. Our method is capable of modelling more flexible poses than CMR. We hypothesise this is because of the reliance of CMR on an estimate of the camera matrices obtained using rigid SFM which limits the flexibility of the learned deformations. On the other hand, CMR does not use a keypoint detector.

5. Conclusions

We have proposed a new approach to learn a model of a 3D object category from unconstrained monocular views with 2D keypoints annotations. Compared to traditional solutions that cast this as NR-SFM and solve it via matrix factorization, our solution is based on learning a deep network that performs monocular 3D reconstruction and factorizes internal object deformations and viewpoint changes. While this factorization is an ambiguous task, we have shown a novel approach that constrains the solution recovered by the learning algorithm to be as consistent as possible by means of an auxiliary canonicalization network. We have shown that this leads to considerably better performance, enough to outperform strong baselines on benchmarks that contain large non-rigid deformations within a category (Human3.6M, Up3D) and across categories (PASCAL3D+).

References

[1] Antonio Agudo and Francesc Moreno-Noguer. Dust: Dual union of spatio-temporal subspaces for monocular multiple object 3d reconstruction. In *Proc. CVPR*, pages 6262–6270, 2017.

[2] Antonio Agudo and Francesc Moreno-Noguer. Deformable motion 3D reconstruction by union of regularized subspaces. In *Proc. ICIP*, pages 2930–2934. IEEE, 2018.

[3] Antonio Agudo, Melcior Pijoan, and Francesc Moreno-Noguer. Image collection pop-up: 3D reconstruction and clustering of rigid and non-rigid categories. In *Proc. CVPR*, pages 2607–2615, 2018.

[4] Ijaz Akhter, Yaser Sheikh, Sohaib Khan, and Takeo Kanade. Nonrigid structure from motion in trajectory space. In *Proc. NIPS*, 2009.

[5] Ijaz Akhter, Yaser Sheikh, Sohaib Khan, and Takeo Kanade. Trajectory space: A dual representation for nonrigid structure from motion. *PAMI*, 33(7):1442–1456, 2011.

[6] Christoph Bregler, Aaron Hertzmann, and Henning Biermann. Recovering non-rigid 3D shape from image streams. In *Proc. CVPR*, page 2690. IEEE, 2000.

[7] Joao Carreira, Abhishek Kar, Shubham Tulsiani, and Jitendra Malik. Virtual view networks for object reconstruction. In *Proc. CVPR*, 2015.

[8] Thomas J Cashman and Andrew W Fitzgibbon. What shape are dolphins? building 3d morphable models from 2d images. *PAMI*, 35(1):232–244, 2013.

[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.

[10] 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 *Proc. ECCV*, 2018.

[11] Katerina Fragkiadaki, Marta Salas, Pablo Arbelaez, and Jitendra Malik. Grouping-based low-rank trajectory completion and 3D reconstruction. In *Proc. NIPS*, pages 55–63, 2014.

[12] Paulo FU Gotardo and Aleix M Martinez. Non-rigid structure from motion with complementary rank-3 spaces. In *Proc. CVPR*, pages 3065–3072. IEEE, 2011.

[13] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *PAMI*, 36(7):1325–1339, 2014.

[14] Hanbyul Joo and Hao Liu. Panoptic studio: A massively multiview system for social motion capture. 2015.

[15] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Proc. CVPR*, pages 7122–7131, 2018.

[16] Angjoo Kanazawa, Shubham Tulsiani, Alexei A Efros, and Jitendra Malik. Learning category-specific mesh reconstruction from image collections. In *Proc. ECCV*, pages 371–386, 2018.

[17] Angjoo Kanazawa, Jason Y Zhang, Panna Felsen, and Jitendra Malik. Learning 3d human dynamics from video. In *Proc. CVPR*, 2019.

[18] Abhishek Kar, Shubham Tulsiani, Joao Carreira, and Jitendra Malik. Category-specific object reconstruction from a single image. In *Proc. CVPR*, 2015.

[19] Dong Liu Ke Sun, Bin Xiao and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proc. CVPR*, 2019.

[20] Muhammed Kocabas, Salih Karagoz, and Emre Akbas. Self-supervised learning of 3d human pose using multi-view geometry. In *Proc. CVPR*, 2019.

[21] Yasunori Kudo, Keisuke Ogaki, Yusuke Matsui, and Yuri Odagiri. Unsupervised adversarial learning of 3D human pose from 2D joint locations. *Proc. ECCV*, 2018.

[22] Suryansh Kumar, Anoop Cherian, Yuchao Dai, and Hongdong Li. Scalable dense non-rigid structure from motion: A grassmannian perspective. In *Proc. CVPR*. IEEE, 2018.

[23] Suryansh Kumar, Yuchao Dai, and Hongdong Li. Spatial-temporal union of subspaces for multi-body non-rigid structure-from-motion. *Pattern Recognition Journal*, 2017.

[24] Christoph Lassner, Javier Romero, Martin Kiefel, Federica Bogo, Michael J. Black, and Peter V. Gehler. Unite the people: Closing the loop between 3D and 2D human representations. In *Proc. CVPR*, July 2017.

[25] Julieta Martinez, Rayat Hossain, Javier Romero, and James J Little. A simple yet effective baseline for 3d human pose estimation. In *Proc. ICCV*, pages 2640–2649, 2017.

[26] Francesc Moreno-Noguer. 3d human pose estimation from a single image via distance matrix regression. In *Proc. CVPR*, 2017.

[27] David Novotny, Diane Larlus, and Andrea Vedaldi. Learning 3d object categories by looking around them. In *Proc. ICCV*, 2017.

[28] Georgios Pavlakos, Xiaowei Zhou, and Kostas Daniilidis. Ordinal depth supervision for 3d human pose estimation. In *Proc. ICCV*, 2018.

[29] Georgios Pavlakos, Xiaowei Zhou, Konstantinos G Derpanis, and Kostas Daniilidis. Harvesting multiple views for marker-less 3d human pose annotations. In *Proc. CVPR*, 2017.

[30] Georgios Pavlakos, Luyang Zhu, Xiaowei Zhou, and Kostas Daniilidis. Learning to estimate 3D human pose and shape from a single color image. In *Proc. CVPR*, pages 459–468, 2018.

[31] Helge Rhodin, Jörg Spörri, Isinsu Katircioglu, Victor Constantin, Frédéric Meyer, Erich Müller, Mathieu Salzmann, and Pascal Fua. Learning monocular 3d human pose estimation from multi-view images. In *Proc. CVPR*, 2018.

[32] Lorenzo Torresani, Aaron Hertzmann, and Christoph Bregler. Learning non-rigid 3D shape from 2D motion. In *Proc. NIPS*, pages 1555–1562, 2004.

[33] Lorenzo Torresani, Aaron Hertzmann, and Chris Bregler. Nonrigid structure-from-motion: Estimating shape and motion with hierarchical priors. *PAMI*, 30(5):878–892, 2008.

[34] Alexander Toshev and Christian Szegedy. Deeppose: Human pose estimation via deep neural networks. In *Proc. CVPR*, 2014.

[35] Shubham Tulsiani, Abhishek Kar, Joao Carreira, and Jitendra Malik. Learning category-specific deformable 3D models for object reconstruction. *PAMI*, 39(4):719–731, 2017.

[36] 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 *Proc. ICCV*, 2017.

[37] Sara Vicente, Joao Carreira, Lourdes Agapito, and Jorge Batista. Reconstructing PASCAL VOC. In *Proc. CVPR*, 2014.

[38] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-UCSD birds-200-2011 dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

[39] Yu Xiang, Roozbeh Mottaghi, and Silvio Savarese. Beyond PASCAL: A benchmark for 3d object detection in the wild. In *WACV*, 2014.

[40] Jing Xiao, Chai Jin-xiang, and Takeo Kanade. A closed-form solution to non-rigid shape and motion recovery. In *Proc. ECCV*, pages 573–587, 2004.

[41] Xiaowei Zhou, Menglong Zhu, Kosta Derpanis, and Kostas Daniilidis. Sparseness meets deepness: 3d human pose estimation from monocular video. In *Proc. CVPR*, 2016.

[42] Xiaowei Zhou, Menglong Zhu, Spyridon Leonardos, and Kostas Daniilidis. Sparse representation for 3D shape estimation: A convex relaxation approach. *PAMI*, 2016.

[43] Yingying Zhu, Dong Huang, Fernando De La Torre, and Simon Lucey. Complex non-rigid motion 3D reconstruction by union of subspaces. In *Proc. CVPR*, pages 1542–1549, 2014.