Two views are all you need

Mosam Dabhi
Carnegie Mellon University
mdabhi@andrew.cmu.edu

Chaoyang Wang
Carnegie Mellon University
chaoyanw@andrew.cmu.edu

Kunal Saluja
Apple Inc.
kunal_saluja@apple.com

ShiangYong Looi
Apple Inc.
shiangyong_looi@apple.com

Laszlo Jeni
Carnegie Mellon University
laszlojeni@cmu.edu

Ian Fasel
Apple Inc.
ifasel@apple.com

Simon Lucey
Carnegie Mellon University
slucey@andrew.cmu.edu

Abstract

Triangulating a point in 3D space should only require two corresponding camera projections. However in practice, expensive multi-view setups – involving tens sometimes hundreds of cameras – are required in order to obtain the high fidelity 3D reconstructions necessary for many modern applications. In this paper we argue that similar fidelity can be obtained from only two views by breaking the tenet of rigidity which is central to much of modern multi-view geometry. Our approach instead leverages recent advances in non-rigid monocular 2D-3D lifting using deep learning. We show how our method can achieve comparable fidelity to expensive multi-view rigs using only two views.

1. Introduction

Triangulation refers to determining the location of a point in 3D space from projected 2D correspondences across multiple views. In theory, only two calibrated camera views should be necessary to accurately reconstruct the 3D position of a point. However, in practice the effectiveness of triangulation is heavily dependent upon the accuracy of the measured 2D correspondences, baseline, and occlusions. As a result expensive and cumbersome multi-view rigs, sometimes involving hundreds of cameras and specialized hardware, are currently the method of choice to obtain high fidelity 3D reconstructions of nonrigid objects [16].

In this paper, we challenge the need to have such expensive, cumbersome multi-view rigs to obtain high-fidelity 3D reconstructions. We argue that comparable fidelity to these high complexity rigs can be obtained using only two uncalibrated views. Such a simplification would enable data collection in unstructured, "in-the-wild" environments, opening the door to a wide variety of applications ranging from entertainment, neuroscience, psychology, ethology, and several fields of medicine [9, 18, 11, 6, 14], where complex multi-camera rigs may be financially, technologically, or simply practically infeasible.

One of the most notable multi-view rigs for human pose reconstruction is the PanOptic studio [16], which contained 480 VGA cameras, 31 HD Cameras, and 10 RGB+D sensors, distributed over the surface of geodesic sphere with a 5.49m diameter. This setup also required specialized hardware for storage and to gen-lock camera exposures. Despite its cost and complexity, the fidelity of the 3D reconstructions from PanOptic studio has motivated similar efforts across industry and academia. Of particular note is a recent effort that employed 62 hardware synchronized cameras to capture the pose of Rhesus Macaque monkeys [3]. Other notable efforts include [17] for dogs, [15] for human body, and [26, 10] for the human face.

While standard structure from motion methods can combine multiple views of a *rigid* object over time [29], reconstruction of nonrigid objects that *change their shape over time* has required complex multi-view setups as in [16, 3, 17, 26, 10] to provide synchronized observations on *every* frame. The core insight in this paper is to leverage recent innovations in non-rigid monocular 2D-3D lifting using deep learning [20, 31, 25, 30] to enable views from different moments in time to be used in reconstruction. Notably, despite allowing points to change their position over time, our proposed approach is atemporal and does not make any assumptions about the temporal ordering or smoothness of the 2D projected measurements. Our effort is the first we are aware of that utilizes these new advances for multi-view re-

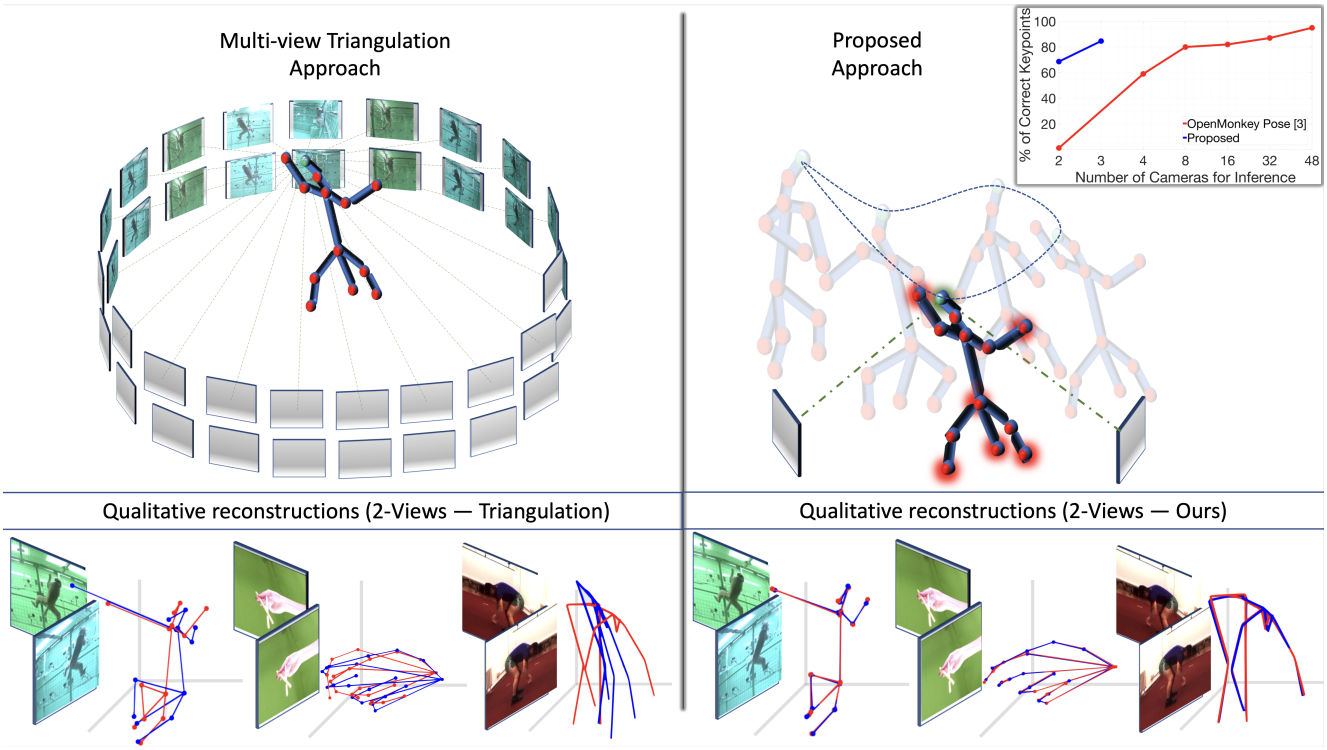

Figure 1: A traditional multi-view setup relies on the concept of triangulation with the assumption that the point being reconstructed is static in time – requiring a large number of physical views (i.e. cameras) to ensure a high fidelity reconstruction. Our approach breaks this triangulation assumption by allowing the reconstructed points to move over time (i.e. non-rigidity). Empirically (see plot in top-right), we demonstrate that our proposed approach can achieve comparable fidelity to expensive multi-view rigs using only two physical views. Blue lines depict the triangulation and proposed approaches (left vs. right, respectively) with just two-physical views and red lines show the corresponding ground-truth 3D pose of the non-rigid object.

construction. Figure 1 presents a graphical depiction of our approach.

Contributions: Our approach draws inspiration from the monocular 2D-3D lifting method Deep NRSfM[20, 31]. We make the following contributions:

- We demonstrate how the hierarchical sparse coding shape prior of Deep NRSfM can be re-interpreted for the problem of multi-view reconstruction using equi-variant constraints.
- A novel network architecture is proposed which utilizes an elegant multi-view pooling step to circumvent the need for explicitly enforcing the equi-variant constraints.
- Extensive evaluations are presented across numerous benchmarks and object categories including the human body, human hands, and monkey body. Across all evaluations our approach significantly outperforms rigid multi-view methods achieving comparable fidelity with only two physical views.

We should note that our proposed approach assumes known 2D projected measurements so does not directly leverage pixel intensities. This being said the approach can be integrated with any available 2D landmark image detector such as HR-Net [27], Stacked Hourglass Networks [24], Integral Pose Regression [28], and others.

2. Related Work

Classical multi-view approaches: Multi-view triangulation [12] has been the method of choice in the context of large scale complex rigs with multiple cameras [16, 3, 26, 10] for obtaining 3D reconstruction from 2D measurements. The number of views, 2D measurement noise, baseline, and occlusions bound the fidelity of these 3D reconstructions. These time synchronized multiple physical views, however, come at the considerable cost and effort described in the introduction.

Monocular 2D-3D lifting: The problem of lifting 3D pose of the object from monocular projected 2D measurements has received substantial attention recently. The task – a subset of Non-Rigid Structure from Motion (NRSfM [5]) where the temporal ordering of frames is ignored – is to simultaneously recover the non-rigid 3D structure and cam-

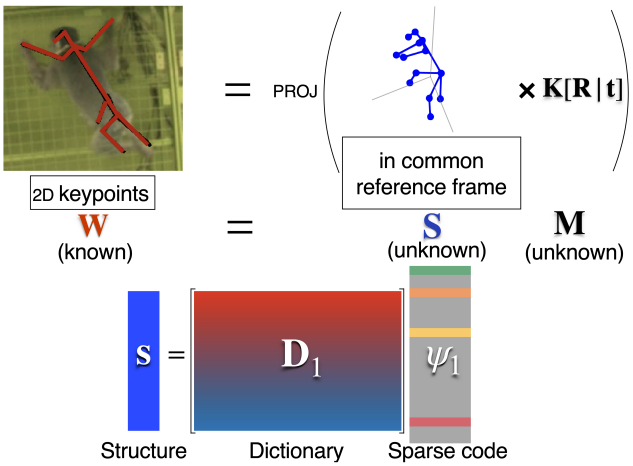

Figure 2: A pictorial representation of the factorization assumed under an orthogonal camera projection, showing 2D annotations $\mathbf{W}$ are orthogonal projections of the transformed 3D structure $\mathbf{S}$. The lower part of the figure depicts the priors enforced upon the 3D structure as a linear combination of shape dictionary $\mathbf{D}$ and sparse code vector $\boldsymbol{\psi}$.

era pose from an ensemble of monocular 2D measurements captured at different points in time. Advances in deep learning based approaches to 2D-3D lifting [20, 25] has seen significant improvements in the robustness and fidelity of these non-rigid 3D reconstructions across a broad set of object categories and scenarios. These recent advances to date have only been applied to problems where there is only a single view (i.e. monocular) of the object at a particular point in time. Our approach is the first – to our knowledge – to leverage these advancements for 3D reconstruction when there are multi-view measurements taken at the same instance in time.

3. Methodology

3.1. Problem Setup

K-view 2D-3D lifting involves the problem of recovering N non-rigid 3D structures from $K \times N$ 2D measurements. For each frame n, the $N\times$ 2D measurements stem from the same rigid shape. For each view k, the $K\times$ 2D measurements stem from an ensemble of non-rigid shapes. The frame indexes are atemporal, meaning they are not in any coherent temporal order, separating the problem subtlety from NRSfM. Thus we have

$$\mathbf{W}^{(\mathrm{k,n})} = \mathbf{S}^{(\mathrm{n})}\mathbf{P}^{(\mathrm{k})} \qquad \textbf{s.t.} \quad \mathbf{P}^{(\mathrm{k})\top}\mathbf{P}^{(\mathrm{k})} = \mathbf{I}_2 \qquad (1)$$

For a shape of P points and focusing on a specific single frame instance and hence dropping the superscript n, $\mathbf{I}$ in the above expression is the identity matrix, and the p-th row of $\mathbf{W}^{(\mathrm{k})}$ and $\mathbf{S}$ corresponds respectively to the k-th image coordinates $(u_p^{(k)}, v_p^{(k)})$ and the world coordinates (x_p, y_p, z_p) of the p-th point for the specific n-th frame. This factorization problem is ill-posed by nature; in order to resolve the ambiguities in solution, additional priors are necessary to guarantee the uniqueness of the solution. These priors include the assumption of shape matrices being (i) low rank [7, 5, 2, 8, 22], (ii) compressible [19, 32, 21], or (iii) lying in a union-of-subspaces [23, 33, 1].

Classical monocular 2D−3D lifting methods, generally referred to as NRSfM, encounter limitations in large-scale datasets (*e.g.* with a large number of frames). The low-rank assumption becomes infeasible when the data exhibits complex shape variations, while the union-of-subspaces NRSfM methods have difficulty clustering shape deformations and estimating affinity matrices effectively. The sparsity prior allows more powerful modeling of shape variations with large number of subspaces, but suffers from sensitivity to noise.

3.2. Bilinear Factorization

The proposed $K-$view 2D−3D lifting method assumes a linear model for the 3D shapes to be reconstructed, *i.e.* at canonical coordinates, the vectorization of $\mathbf{S}$ in Eq. (1), denoted $\mathbf{s} = \mathrm{vec}(\mathbf{S}) \in \mathbb{R}^{3P}$ can be written as $\mathbf{s} = \mathbf{D}\boldsymbol{\psi}$ where $\mathbf{D} \in \mathbb{R}^{3P\times B}$ is the shape dictionary with B basis and $\boldsymbol{\psi} \in \mathbb{R}^B$ is the code vector. Equivalently, this linear model could be rewritten as $\mathbf{S} = \mathbf{D}^{\#}(\boldsymbol{\psi} \otimes \mathbf{I}_3)$, where $\mathbf{D}^{\#} \in \mathbb{R}^{P\times 3B}$ is a reshape of $\mathbf{D}$ and $\otimes$ denotes a Kronecker product.

Applying the camera extrinsics (rotation $\mathbf{R}^{(\mathrm{k})} \in \mathbb{SO}(3)$ and translation $\mathbf{t}^{(\mathrm{k})} \in \mathbb{R}^3$) and camera intrinsics for k-th camera view gives the 2D projection in the k-th camera frame. Thus

$$\mathbf{S}\mathbf{M}^{(\mathrm{k})} = \mathbf{D}^{\#}(\boldsymbol{\psi} \otimes \mathbf{M}^{(\mathrm{k})}) \qquad (2)$$

$$\mathbf{W}^{(\mathrm{k})} = \mathbf{D}^{\#}\boldsymbol{\Psi}_{\mathrm{xy}}^{(\mathrm{k})} \quad \textbf{s.t.}\ \boldsymbol{\Psi}^{(\mathrm{k})} = \boldsymbol{\psi} \otimes \mathbf{M}^{(\mathrm{k})} \text{ and } \boldsymbol{\psi} \in \mathcal{C} \qquad (3)$$

where $\boldsymbol{\Psi}_{\mathrm{xy}}^{(\mathrm{k})} \in \mathbb{R}^{3\times 2}$ denoted the first two columns of $\boldsymbol{\Psi}^{(\mathrm{k})} \in \mathbb{R}^{3B\times 3}$ is the block code (as it is a Kronecker product) and $\mathcal{C}$ denotes the prior constants applied on the code $\boldsymbol{\psi}$, *e.g.* low rank [7] or hierarchical sparsity [20]. Eq. (2) was written by ignoring the translation component $\mathbf{t}$ and was removed assuming the input 2D points were pre-centered under the implicit object-centric assumption that the origin of the canonical coordinates is placed at the object center leading to bilinear factorizaton shown in Eq. (3). Under the unsupervised settings, $\mathbf{D}, \boldsymbol{\psi}, \mathbf{M}$ are all unknowns and are usually solved under the simplified assumptions, *i.e.* complete $K-$view 2D points under the orthogonal camera projection.

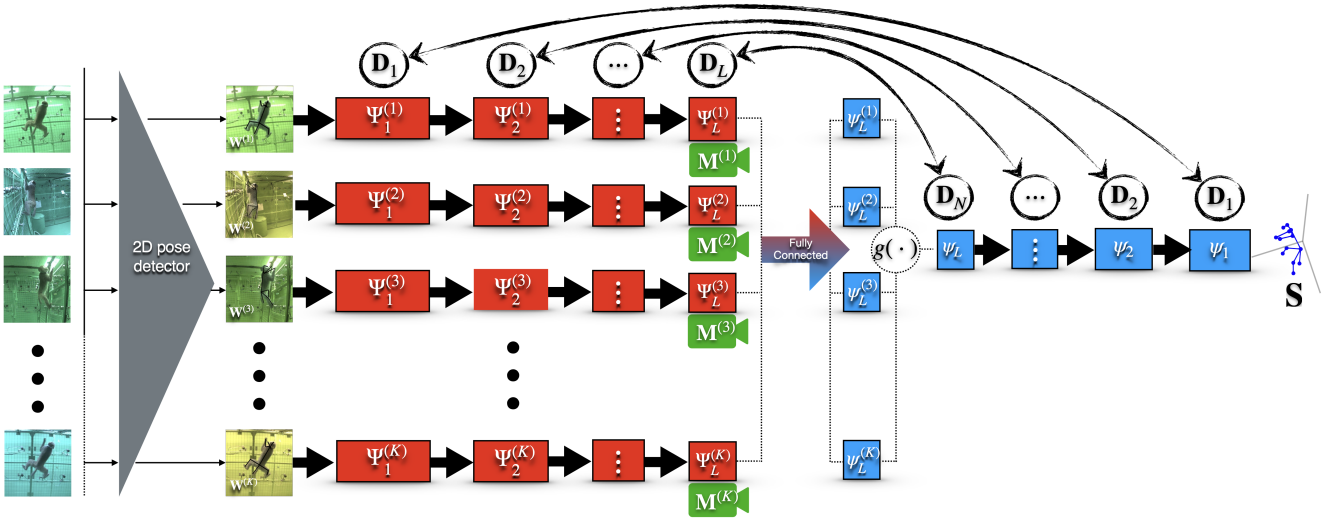

Figure 3: $K-$view 2D-3D lifting approach using hierarchical sparse coding shape prior for finding the solution to multi-view reconstruction using the equivariance constraints enforced is within this architecture. We take as input, 2D keypoints in our pipeline for all the given views $\mathbf{W}^{(\mathrm{k})}\ \forall k = [1, \cdots, K]$. We follow the encoder-decoder network with $K-$ parallel encoders and a single decoder. The $K-$view encoder-decoder network derived from the hierarchical block-sparse coding jointly predicts Kcamera matrices and a single 3D structure $\mathbf{S}$. Since the generated 3D structure in its canonicalized frame is equivariant to the camera rotations, equivariance is enforced within this architecture. The training objective is to minimize the difference between the 2D projection generated over Kcamera views and the input 2D keypoint annotations.

3.3. 2D-3D Lifting

To solve this problem above for $K-$views 2D$-$3D lifting given KN tuples of $(\mathbf{W}^{(\mathrm{k,n})}, \mathbf{M}^{(\mathrm{k,n})})$ as the dataset (with n indexing the frame and k indexing the camera view), three problems remain to address

- How to define heuristics $\mathcal{C}$, that is required to have an accurate solution
- How to formulate an optimization strategy
- How to efficiently pool in K different camera views and enforce equivariance over the K generated camera matrices and a canonicalized 3D structure.

We choose to follow Deep NRSfM [20] that instead of leveraging simple handcrafted priors (*e.g.* low rank or sparsity), instead imposes hierarchical sparsity constraint $\mathcal{C}_\Theta$ with learnable parameters Θ (see Sec. 3.4). Learning strategy of $K-$view 2D$-$3D lifting is then interpreted as solving the following bilevel optimization problem

$$\min_{\mathbf{D},\Theta} \sum_{k=1}^{K}\sum_{n=1}^{N} \min_{\boldsymbol{\psi}^{(\mathrm{n})},\mathbf{M}^{(\mathrm{k,n})}} \|\mathbf{W}^{(\mathrm{k,n})} - \mathbf{D}^{\#}\boldsymbol{\Psi}_{xy}^{(\mathrm{k,n})}\|_F \qquad (4)$$

where the lower level problem is to solve single-frame 2D$-$3D lifting with given $\mathbf{D}, \Theta$, and the upper level problem is to find optimal $\mathbf{D}, \Theta$ for the whole input dataset. Descent method is then employed for solving this bilevel problem.

We first approximate the solver of the lower level problem as a feed-forward network, *i.e.* $f(\mathbf{W}^{(\mathrm{k,n})}; \mathbf{D}, \Theta) \mapsto ((\mathbf{M}^*)^{(\mathrm{k,n})}_{\mathbf{D},\Theta}), (\boldsymbol{\psi}^*)^{(\mathrm{n})}_{\mathbf{D},\Theta})$. Architecture of the network (shown in Fig. 3) is induced from unrolling one iteration of Iterative Shrinkage and Thresholding Algorithm (ISTA) [4] for all the given K views. Then with $(\boldsymbol{\Psi}^*)^{(\mathrm{k,n})}_{\mathbf{D},\Theta} = (\boldsymbol{\psi}^*)^{(\mathrm{n})}_{\mathbf{D},\Theta} \otimes (\mathbf{M}^*)^{(\mathrm{k,n})}_{\mathbf{D},\Theta}$, the original bilevel problem is reduced to a single level unconstrained problem, *i.e.*

$$\min_{\mathbf{D},\Theta} \sum_{k=1}^{K}\sum_{n=1}^{N} \|\mathbf{W}^{(\mathrm{k,n})} - \mathbf{D}^{\#}(\boldsymbol{\Psi}^*)^{(\mathrm{k})}_{\mathbf{D},\Theta}\|_F \qquad (5)$$

allowing the use of solvers such as gradient descent.

Finally, with $\mathbf{D}, \Theta$ learned, $f(\mathbf{W}^{(\mathrm{k})}; \mathbf{D}, \Theta)$ is the $K-$view 2D$-$3D lifting network applicable to unseen data. Due to the introduction of multiple levels of dictionaries and codes in the following section, we will abuse the notation of $\mathbf{D}, \boldsymbol{\psi}, \boldsymbol{\Psi}$ by adding subscript 1, *i.e.* $\mathbf{D}_1, \boldsymbol{\psi}_1, \boldsymbol{\Psi}_1^{(\mathrm{k})}$ indicating that they form the first level of hierarchy for k-th camera view.

3.4. Multi-layered Sparse Shape Prior

Assuming the canonical 3D shapes are compressible via multi-layered sparse coding, the shape code $\boldsymbol{\psi}_1$ is constrained by $\mathcal{C}_\Theta$ as

$$
\begin{aligned}
\boldsymbol{\psi}_1 &= \mathbf{D}_2\boldsymbol{\psi}_2 \\
&\vdots \\
\boldsymbol{\psi}_{L-1} &= \mathbf{D}_L\boldsymbol{\psi}_L \\
\textbf{s.t. } &\|\boldsymbol{\psi}_l\|_1 \leq \boldsymbol{\lambda}_l\,,\, \boldsymbol{\psi}_l \geq \mathbf{0}\,, \forall l \in \{1, \cdots, L\}
\end{aligned} \tag{6}
$$

where $\mathbf{D}_l \in \mathbb{R}^{B_{l-1}\times B_l}$ are the hierarchical dictionaries, l is the index of hierarchy level, and $\boldsymbol{\lambda}_l$ is the scalar specifying the amount of sparsity in each level. Thus, the learnable parameters for $\mathcal{C}_\Theta$ is $\Theta = \{\cdots, \mathbf{D}_l, \boldsymbol{\lambda}_l, \cdots\}$

Constraints on multi-layer sparsity not only preserves sufficient freedom on shape variation, but it also results in more constrained code recovery. Multi-layer sparse coding induces a hierarchical block sparsity on the block codes $\boldsymbol{\Psi}_l^{(\mathrm{k})}$ $\big($equal to $\boldsymbol{\psi}_l \otimes \mathbf{M}_{\mathrm{xy}}^{(\mathrm{k})}\big)$, leading to relaxation of the lower-level problems in Eq. (4)

$$
\begin{aligned}
\sum_{k=1}^{K} \min_{\boldsymbol{\Psi}_1^{(\mathrm{k})},\ldots,\boldsymbol{\Psi}_L^{(\mathrm{k})}} \|\mathbf{W}^{(\mathrm{k})} - \mathbf{D}_1\boldsymbol{\Psi}_1^{(\mathrm{k})}\|_F^2 + \sum_{l=1}^{L} \boldsymbol{\lambda}_l \|\boldsymbol{\Psi}_l^{(\mathrm{k})}\|_F^{(3\times 2)} \\
+ \sum_{l=2}^{L} \|\boldsymbol{\Psi}_{l-1}^{(\mathrm{k})} - (\mathbf{D}_l \otimes \mathbf{I}_3)\boldsymbol{\Psi}_l^{(\mathrm{k})}\|_F^2
\end{aligned} \tag{7}
$$

where $\|\cdot\|_F^{(3\times 2)}$ denotes thee Frobeinus norm of each 3×2 block.

Taking insight from Deep NRSfM [20], we further relax the block sparsity constraint in Eq. (7) to L_1 sparsity with a nonnegative constraint to allow for the use of ReLU activations in the network architecture.

3.5. Encoder Network and Equivariant Structure

For $K-$ views, by unrolling one iteration of block ISTA for each layer of each views, our encoder takes $\mathbf{W}^{(\mathrm{k})}$ as $K-$view inputs and produces block codes for each of the $K-$views for the last layer $\boldsymbol{\Psi}_L^{(\mathrm{k})}$ as output

$$
\begin{aligned}
\boldsymbol{\Psi}_1^{(\mathrm{k})} &= \mathrm{ReLU}\Big(\big[(\mathbf{D}_1^{\#})^T\mathbf{W}^{(\mathrm{k})}\big]_{3B_1\times 2}; \boldsymbol{\lambda}_1^{(\mathrm{k})}\Big) \\
\boldsymbol{\Psi}_2^{(\mathrm{k})} &= \mathrm{ReLU}\Big((\mathbf{D}_2 \otimes \mathbf{I}_3)^\top\boldsymbol{\Psi}_1^{(\mathrm{k})}; \boldsymbol{\lambda}_2^{(\mathrm{k})}\Big) \\
&\vdots \\
\boldsymbol{\Psi}_L^{(\mathrm{k})} &= \mathrm{ReLU}\Big((\mathbf{D}_L \otimes \mathbf{I}_3)^\top\boldsymbol{\Psi}_{L-1}^{(\mathrm{k})}; \boldsymbol{\lambda}_L^{(\mathrm{k})}\Big)
\end{aligned} \tag{8}
$$

where $\boldsymbol{\lambda}_l^{(\mathrm{k})}$ is the learnable threshold for each $B_l^{(\mathrm{k})}$ block and $[\cdot]_{3B_1\times 2}$ is a $3B_1\times 2$ reshape. $(\mathbf{D}_l \otimes \mathbf{I}_3)^\top\boldsymbol{\Psi}_{l-1}^{(\mathrm{k})}$ are implemented by convolution transpose. At the last, innermost (bottleneck) layer $\boldsymbol{\Psi}_L^{(1)}, \cdots, \boldsymbol{\Psi}_L^{(K)}$ is then factorized into $\boldsymbol{\psi}_L^{(1)}, \cdots, \boldsymbol{\psi}_L^{(K)}$ and $\mathbf{M}^{(1)}, \cdots, \mathbf{M}^{(K)}$ (constraining to $\mathbb{SO}(3)$ using SVD [20]).

Enforcing equivariance: Eq. (8) forms the encoder part of our framework and we generate bottleneck for $K-$views. We enforce equivariance implicitly by amalgamating the available features at the bottleneck where we leverage pooling function $g(\boldsymbol{\psi}_L^{(1)}, \cdots, \boldsymbol{\psi}_L^{(K)}) \mapsto (\boldsymbol{\psi}_L)$ as *max pooling* as shown in architecture overview Fig. 3. We generate the final shared pool of bottleneck features $\boldsymbol{\psi}_L$ for generating a single structure $\mathbf{S}$ that should remain equivariant to different camera rotations $\mathbf{M}^{(1)}, \cdots, \mathbf{M}^{(K)}$ generated subsequently at the bottleneck stage. Thus, we generate all camera matrices corresponding to $K-$views and leverage them in the reprojection error shown in Eq. (3) that enforces the equivariance constraint implicitly, *i.e.* the 3D structure in canonical view is equivariant under different $\mathbf{M}^{(1)}, \cdots, \mathbf{M}^{(K)}$ transformations.

3.6. Recovering 3D shape – Decoder Network

The 3D shape $\mathbf{S}$ is recovered from *max-pooled* feature bottleneck sparse code $\boldsymbol{\psi}_L$ via the decoder as

$$
\begin{aligned}
\boldsymbol{\psi}_{L-1} &= \mathrm{ReLU}(\mathbf{D}_L\boldsymbol{\psi}_L + \mathbf{b}_L) \\
\boldsymbol{\psi}_1 &= \mathrm{ReLU}(\mathbf{D}_2\boldsymbol{\psi}_2 + \mathbf{b}_2) \\
&\vdots \\
\mathbf{S} &= \mathbf{D}_1\boldsymbol{\psi}_1
\end{aligned} \tag{9}
$$

4. Experiments

In this section we are mainly trying to discuss the following results and claims: **R1**: Two-physical views are enough and can achieve comparable fidelity to expensive multi-view rigs; **R2**: Proposed approach significantly outperforms rigid multi-view methods such as Iterative triangulation with robust outlier rejection *or* classical stereo triangulation mentioned in [12] when operated over real-world data; **R3**: Proposed approach is generalized enough to conduct extensive evaluations across various object categories including human body, human hands, and monkey body; **R4**: Proposed approach substantially outperforms the existing monocular deep 2D−3D lifting methods such as [20].

We have partitioned this section with respect to different datasets to showcase the claims mentioned above. Within each subsections, we discuss the dataset details, evaluation metrics, and corresponding result claims that each of them represents.

Architectural details: The most significant hyperparameter of the proposed approach is the dictionary size at the last layer, which relies upon the shape variety that displays inside the dataset. We find that setting it to 8 gives reasonable reconstruction over all the assessed datasets.

Evaluation metrics: We utilize the following metrics to assess the prediction accuracy of 3D reconstruction. **PA-MPJPE**: prior to computing the mean per-joint position error, we standardize the scale of the predictions by normaliz-

ing them to match against the given ground-truth (GT) followed by rigidly aligning these predictions to GT. Lower the better **PCK**: percentage of correct keypoints. The predicted joint is viewed as correct if the separation between the predicted and the GT joint is within a specific range (usually in *cm* or *mm*).

Baseline methods for comparison: We leverage the widely used rigid-triangulation approach mentioned in [12] that is available within OpenCV's open source library and call the predictions generated using this approach as **CV-TRNG**. We require perfect camera calibration parameters for the camera matrices of $K-$views and very accurate K 2D projections to use this baseline accurately. We further use a baseline implementation of iterative triangulation with robust outlier rejection [13] referred as **ITR-TRNG**. ITR-TRNG first finds the points which minimizes the distance from all the rays and removes the rays which is the furthest away from that point. It then re-evaluates the triangulation and this iteration is repeated for 2-3 times. Empirically, we find that increasing the iteration leads us to predict the correct 3D position in space. We find that this approach gives near-perfect 3D reconstruction if we have exact camera calibration parameters and exact, clean 2D projections since it is robust to outliers. We consider this to be a very strong baseline comparison since this approach is being widely used in industry as well as academia to generate very accurate 3D reconstructions that are further used to train 3D regression methods. We evaluate our approach on three types of datasets with substantial non-rigid deformities. For all the given experiments we should note that the 2-View cameras are chosen at random and for maintaining uniformity and valid comparisons, same set of cameras are being used in the comparative baseline approaches for a fair comparison.

4.1. Monkey body [3]

OpenMonkeyStudio [3] is a huge Rhesus Macaque monkey pose dataset in a setup similar to PanOptic Studio where 62 cameras capture the markerless pose of the Rhesus Macaque monkey subjects. We leverage the provided 2D annotations over the Batch7, Batch9, Batch9a, Batch9b, Batch10, and Batch11. This dataset has also provided the 3D GT for these batches for us to evaluate the 3D reconstruction performance. This subsection is responsible for claiming **R1, R3**.

We evaluate the proposed $2-$view and $3-$view $2D-3D$ lifting method and we see that compared to the given results in [3], the proposed approach significantly outperforms them and achieves comparable fidelity with only two physical views. For the **PCK** evaluation we use the same metrics as used in [3] where we consider all the keypoints as correct if their reconstruction is within $10cm$ of the GT. As shown in Table 2 and top-right plot in Fig. 1 we outperform the given results of 2-Views by a significant margin (1.2% vs. 68.63%). As we expected, the accuracy of the proposed method continues to rise as we add in more views clearly evident by the uptick in performance from $3-$views. Similar performance metrics are evident when viewing the **PA-MPJPE** quantitative performance in Table 1. For the rigid multi-view baseline we used **CV-TRNG** method. The results over a convoluted object category such as a monkey shows the benefit of using the proposed multi-view $2D-3D$ lifting method over the rigid multi-view methods. Qualitative 3D reconstruction performance of the proposed approach in Fig. 4a shows the substantial improvement visually over the rigid multi-view triangulation approach.

4.2. Human body [15]

This subsection is responsible for claiming **R1, R2, R3**. We leverage Human 3.6M dataset [15] – a large-scale human pose dataset annotated by motion capture systems and 4 cameras view the subject with perspective projection.

Evaluation protocol (Human 3.6M): In this protocol, we pick 5 subjects (1, 5, 6, 7, 8), 2 cameras (1, 3) and carry out unsupervised 3D pose lifting using the proposed approach as well as using the iterative triangulation with robust outlier rejection **ITR-TRNG** method. As we have the accurate 3D pose labels we render the 2D keypoints using the provided camera parameters. Since the camera parameters as well as 3D pose labels are very accurate we have very accurate 2D keypoints. As we should expect, the triangulation approach **ITR-TRNG** provides perfect 3D reconstruction if the camera parameters and 2D keypoints are clean. However, to mimic the real world data we inject noise in camera extrinsics, camera intrinsics, as well as 2D keypoints separately and compare the performance between the multi-view rigid baseline **ITR-TRNG** and the proposed approach. We observe that the strong baseline completely fails as soon as noise with even a very small standard deviation is added degrading the fidelity of the 3D reconstruction by a huge factor. On the other hand, since the proposed approach is not dependent on camera calibration parameters and instead only dependent on the quality of 2D keypoint annotations shows slightly degraded performance only when the noise is injected over the input 2D keypoint annotations. Qualitatively (Fig. 5) as well as quantitatively(Table 4, the proposed approach is better than the multi-view rigid triangulation approach used generally within the industry as well as academia community.

4.3. Human hands [34]

Finally, we use an open-source hands dataset - FreiHand [34] to support **R1, R3, R4**. FreiHand is a large-scale open-source dataset with varied movements of hands with 3D pose annotated by motion capture systems. It consists of 32560 3D pose samples with their corresponding

Method	Batch#7	Batch#9	Batch#9a	Batch#9b	Batch#10	Batch#11
2-Views **CV-TRNG**	21.21	24.32	30.67	24.50	26.10	22.77
2-views (ours)	**8.36**	**8.25**	**9.12**	**11.52**	**8.203**	**8.17**

Table 1: **PA-MPJPE** error values for the given Monkey body dataset shows substantial improvement of the proposed method over the baseline rigid multi-view triangulation approach while using only two views. **PA-MPJPE** values are in **cm**.

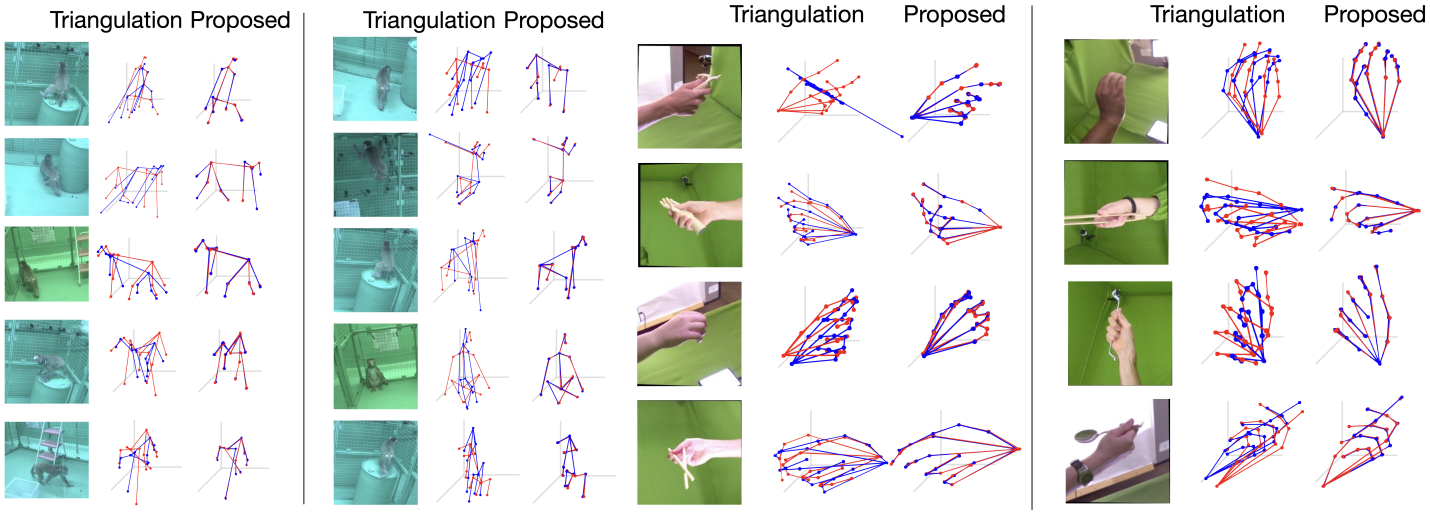

Figure 4: Figure (a) and Fig. (b) shows qualitative 3D reconstruction comparison between the rigid multi-view triangulation technique and the proposed technique for Monkey body and Human hands, respectively. The blue lines represent the calculated/predicted 3D structure and the red lines represent the given GT.

Method	All batch
2 Views [3]	1.2%
4 Views [3]	59%
8 Views [3]	80%
16 Views [3]	82%
32 Views [3]	87%
48 Views [3]	95%
2-views (ours)	**68.63**%
3-views (ours)	**84.63**%

Table 2: Percentage of Correct Keypoint (PCK) % for OpenMonkeyStudio dataset. Following [3], the threshold for considering a keypoint to be correct is set at 10cm.

camera intrinsics. We generate random camera extrinsics to form a camera projection matrix and randomly create camera views to use this dataset for the proposed approach. Table 3 and Fig. 4b shows the quantitative as well as qualitative improvement of the proposed approach over the rigid multi-view triangulation approaches. We use this dataset to compare against the leading monocular 2D−3D lifting method [20]. As shown in Fig. 6 there is a substantial improvement and an obvious need to use multi-view or even 2-View 2D−3D lifting approach over the the single-view 2D-3D lifting approach. Further quantification of this claim is being supported by Table 3 that shows the improvement of 2-View and 3-View 2D-3D lifting approaches over the single-view 2D−3D lifting approach.

Method	Training Set
1-view (Oracle)	9.58
2-views (random)	**4.1**
3-views (random)	**3.8**

Table 3: Quantitative comparison. **FreiHand** dataset. The proposed approach with multi-views is better than triangulation. Reconstruction error values for the training set are the standard PA - Mean Per Joint Position Error (MPJPE) in **mm**.

5. Conclusion and Future Directions

We propose a multi-view atemporal approach for 2D-3D lifting using the recent advances of modern deep learning stream of methods. We break the rigidity assumption that most of the complex data capturing rigs having hun-

	S1, S5, S6, S7, S8								
	Extrinsics Noise			Intrinsics Noise			2D keypoints Noise		
	$\sigma = 0.1$	$\sigma = 0.5$	$\sigma = 0.9$	$\sigma = 0.1$	$\sigma = 0.5$	$\sigma = 0.9$	$\sigma = 15$	$\sigma = 25$	$\sigma = 35$
ITR-TRNG	65.49	131.66	145.94	69.57	188.63	234.47	70.08	114.06	154.41
2-Views (ours)	**30.53**			**30.53**			**54.22**	**65.74**	**77.82**

Table 4: Robustness to camera calibration and 2D annotations noise for Human 3.6M dataset.

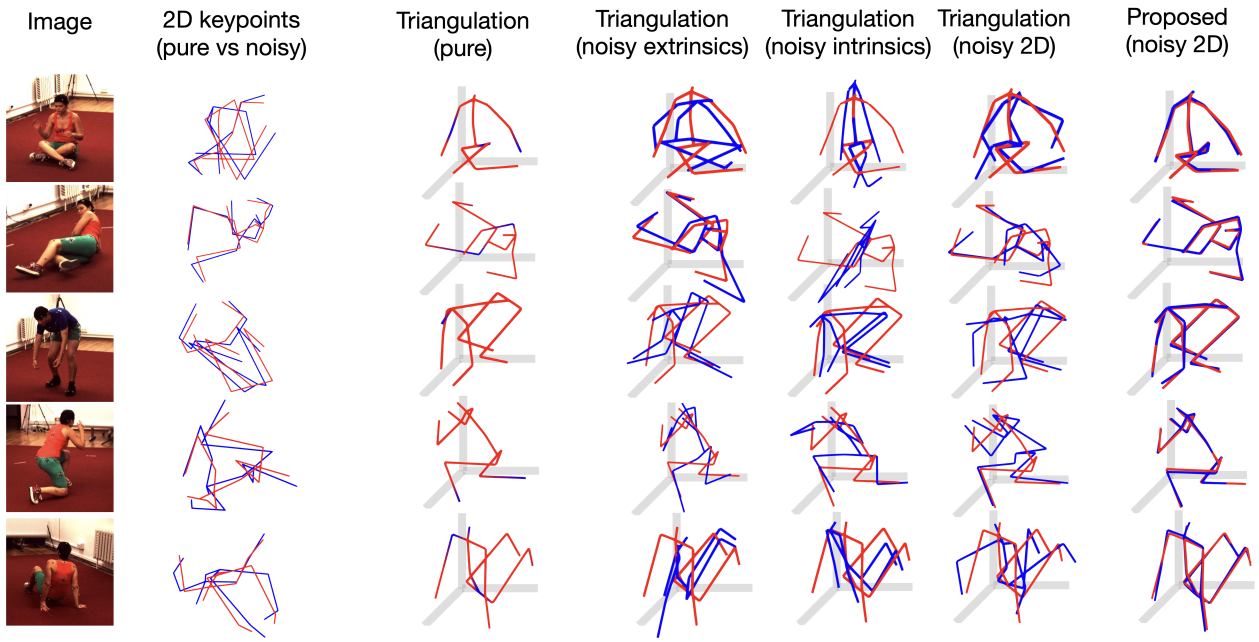

Figure 5: Qualitative results on Human 3.6M dataset with $\sigma = [0.5, 0.5, 25]$ as intrinsics, extrinsics, and 2D keypoints Gaussian noise, respectively.

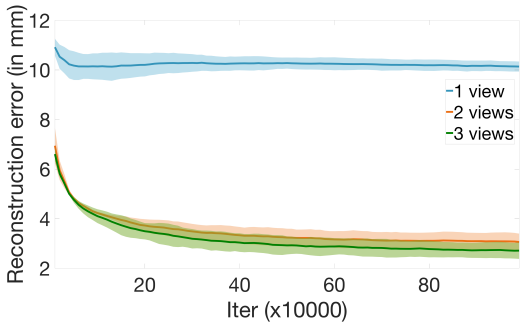

Figure 6: **R4**: Dark center line represents the mean over different trials while the surrounding shadow represents the standard deviation of the corresponding trials.

dreds of cameras assumes – allowing us to dramatically reduce the number of cameras required at any specific instant of time. We observe that two-physical views are able to achieve comparable fidelity to these complex, expensive setups allowing us to generate accurate 3D reconstruction of points for downstram application tasks. The **limitations** of this work follows the fact that we require two rigid views at any instant of time; however our approach requires multiple non-rigid atemporal views for our approach to enforce the proposed shape prior.

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. 3

[2] Ijaz Akhter, Yaser Sheikh, and Sohaib Khan. In defense of orthonormality constraints for nonrigid structure from motion. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1534–1541. IEEE, 2009. 3

[3] Praneet C Bala, Benjamin R Eisenreich, Seng Bum Michael Yoo, Benjamin Y Hayden, Hyun Soo Park, and Jan Zimmermann. Openmonkeystudio: automated markerless pose estimation in freely moving macaques. *bioRxiv*, 2020. 1, 2, 6, 7

[4] Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm with application to wavelet-based image deblurring. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 693–696. IEEE, 2009. 4

[5] Christoph Bregler, Aaron Hertzmann, and Henning Biermann. Recovering non-rigid 3d shape from image streams. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2000 (Cat. No. PR00662)*, volume 2, pages 690–696. IEEE, 2000. 2, 3

[6] Hristos S Courellis, Samuel U Nummela, Michael Metke, Geoffrey W Diehl, Robert Bussell, Gert Cauwenberghs, and Cory T Miller. Spatial encoding in primate hippocampus during free navigation. *PLoS biology*, 17(12):e3000546, 2019. 1

[7] 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. 3

[8] Katerina Fragkiadaki, Marta Salas, Pablo Arbelaez, and Jitendra Malik. Grouping-based low-rank trajectory completion and 3d reconstruction. In *Advances in Neural Information Processing Systems*, pages 55–63, 2014. 3

[9] Richard A Gibbs, Jeffrey Rogers, et al. Evolutionary and biomedical insights from the rhesus macaque genome. *science*, 316(5822):222–234, 2007. 1

[10] Ralph Gross, Iain Matthews, Jeffrey Cohn, Takeo Kanade, and Simon Baker. Multi-pie. *Image and Vision Computing*, 28(5):807–813, 2010. 1, 2

[11] Darcy L Hannibal, Eliza Bliss-Moreau, Jessica Vandeleest, Brenda McCowan, and John Capitanio. Laboratory rhesus macaque social housing and social changes: implications for research. *American Journal of Primatology*, 79(1):e22528, 2017. 1

[12] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003. 2, 5, 6

[13] Richard I Hartley and Peter Sturm. Triangulation. *Computer vision and image understanding*, 68(2):146–157, 1997. 6

[14] Brian Hrolenok, Tucker Balch, David Byrd, Rebecca Roberts, Chanho Kim, James M Rehg, Scott Gilliland, and Kim Wallen. Use of position tracking to infer social structure in rhesus macaques. In *Proceedings of the Fifth International Conference on Animal-Computer Interaction*, pages 1–5, 2018. 1

[15] 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, 2013. 1, 6

[16] 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. 1, 2

[17] Sinead Kearney, Wenbin Li, Martin Parsons, Kwang In Kim, and Darren Cosker. Rgbd-dog: Predicting canine pose from rgbd sensors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8336–8345, 2020. 1

[18] Matt J Kessler, John D Berard, and Richard G Rawlins. Effect of tetanus toxoid inoculation on mortality in the cayo santiago macaque population. *American journal of primatology*, 15(2):93–101, 1988. 1

[19] 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. 3

[20] Chen Kong and Simon Lucey. Deep non-rigid structure from motion. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1558–1567, 2019. 1, 2, 3, 4, 5, 7

[21] Chen Kong, Rui Zhu, Hamed Kiani, and Simon Lucey. Structure from category: a generic and prior-less approach. In *2016 Fourth International Conference on 3D Vision (3DV)*, pages 296–304. IEEE, 2016. 3

[22] Suryansh Kumar. Non-rigid structure from motion: Prior-free factorization method revisited. In *The IEEE Winter Conference on Applications of Computer Vision*, pages 51–60, 2020. 3

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

[24] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *European conference on computer vision*, pages 483–499. Springer, 2016. 2

[25] David Novotny, Nikhila Ravi, Benjamin Graham, Natalia Neverova, and Andrea Vedaldi. C3dpo: Canonical 3d pose networks for non-rigid structure from motion. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7688–7697, 2019. 1, 3

[26] Terence Sim, Simon Baker, and Maan Bsat. The cmu pose, illumination, and expression (pie) database. In *Proceedings of Fifth IEEE International Conference on Automatic Face Gesture Recognition*, pages 53–58. IEEE, 2002. 1, 2

[27] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5693–5703, 2019. 2

[28] 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

[29] Carlo Tomasi and Takeo Kanade. Shape and motion from image streams under orthography: a factorization method. *International Journal of Computer Vision*, 9(2):137–154, 1992. 1

[30] Chaoyang Wang, Chen Kong, and Simon Lucey. Distill knowledge from nrsfm for weakly supervised 3d pose learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 743–752, 2019. 1

[31] Chaoyang Wang, Chen-Hsuan Lin, and Simon Lucey. Deep nrsfm++: Towards 3d reconstruction in the wild. *arXiv preprint arXiv:2001.10090*, 2020. 1, 2

[32] Xiaowei Zhou, Menglong Zhu, Spyridon Leonardos, Konstantinos G Derpanis, and Kostas Daniilidis. Sparseness meets deepness: 3d human pose estimation from monocular video. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4966–4975, 2016. 3

[33] Yingying Zhu, Dong Huang, Fernando De La Torre, and Simon Lucey. Complex non-rigid motion 3d reconstruction by union of subspaces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1542–1549, 2014. 3

[34] Christian Zimmermann, Duygu Ceylan, Jimei Yang, Bryan Russell, Max Argus, and Thomas Brox. Freihand: A dataset for markerless capture of hand pose and shape from single rgb images. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 813–822, 2019. 6