

# High Fidelity 3D Reconstructions with Limited Physical Views

Mosam Dabhi<sup>1</sup> Chaoyang Wang<sup>1</sup> Kunal Saluja<sup>2</sup> László A. Jeni<sup>1</sup> Ian Fasel<sup>2</sup> Simon Lucey<sup>1,3</sup>

<sup>1</sup> Carnegie Mellon University <sup>2</sup> Apple Inc. <sup>3</sup> The University of Adelaide

<https://sites.google.com/view/high-fidelity-3d-neural-prior>

## Abstract

*Multi-view triangulation is the gold standard for 3D reconstruction from 2D correspondences given known calibration and sufficient views. 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 present a novel approach that leverages recent advances in 2D-3D lifting using neural shape priors while also enforcing multi-view equivariance. We show how our method can achieve comparable fidelity to expensive calibrated multi-view rigs using a limited (2-3) number of uncalibrated camera 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 non-rigid objects [20].

Deep learning has provided an alternate low-cost strategy by posing the 3D reconstruction problem as a supervised 2D-3D lifting problem – allowing for effective reconstructions with as little as a single view. Recently, there have been several breakthrough works – notably Deep NRSfM [51] and C3DPO [37] – allowing this problem to be treated in an unsupervised manner that requires ONLY 2D correspondences (i.e. no 3D supervision) greatly expanding the utility and generality of the approach. These unsupervised methods make up for the lack of physical views by instead leveraging large offline datasets containing 2D correspondences of the object category of interest. Unlike classical triangulation, these correspondences do not need to be rigid or even stem from the same object instance. Although

achieving remarkable results, these deep learning methods to date have not been able to compete with the fidelity and accuracy of multi-view rigs that employ triangulation.

Although the methods using deep learning for single view 2D-3D lifting are of prominent research interest – we argue that multi-view consistency is still crucial for generating 3D reconstructions of high fidelity needed for many real-world applications. To this end, we propose a new multi-view NRSfM architecture that incorporates a neural shape prior while enforcing equivariant view consistency. We demonstrate that this framework is competitive with some of the most complicated multi-view capture rigs – while only requiring a modest number (2-3) of physical camera views. Our effort is the first we are aware of, that utilizes these new advances in neural shape priors for multi-view 3D reconstruction. Figure 1 presents a graphical depiction of our approach. Extensive evaluations are presented across numerous benchmarks and object categories including the human body, human hands, and monkey body. To clarify further, we are interested in a problem setup with multiple views that capture different instances of a deformable object – we deal with non-sequential (atemporal) data.

**Motivation:** In many problems, complex multi-camera rigs may be financially, technologically, or simply practically infeasible. Our work in this paper is motivated by the realization that the simplification of multi-view camera rigs – in terms of (i) the number of physical views, and (ii) the need for calibration – could open the door to a wide variety of applications including entertainment, neuroscience, psychology, ethology, as well as several fields of medicine [10, 23, 12, 7, 15].

**Background:** One of the most notable multi-view rigs for high-fidelity 3D reconstruction is the PanOptic studio [20], which contained 480 VGA cameras, 31 HD Cameras, and 10 RGB+D sensors, distributed over the surface of a geodesic sphere with a 5.49m diameter. This setup also required specialized hardware for storage and gen-lock camera exposures and was aimed initially at human pose reconstruction. 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 [4]. Other notable efforts include [22] for dogs, [16] for human body, and [45, 11] for the human face.

**Limitations:** Classical multi-view triangulation can infer 3D structure solely from the rigid 2D correspondences stemming from the physical cameras at a single time instant. If given sufficient calibrated physical camera views, it remains the gold standard for 3D reconstruction. However, if calibration is unknown or the number of views is sparse, we argue that the proposed approach is of significant benefit. Our approach, though is limited in comparison to triangulation as it requires multi-view 2D correspondences taken at the same (rigid) and different (non-rigid) points in time during the learning/optimization process.

A strength of the proposed approach, however, is that the non-rigid 2D correspondences are treated atemporally (*i.e.* the temporal ordering of the non-rigid correspondences is completely ignored). This means that once the network weights of our approach have been learned – high-fidelity 3D estimates of the structure and cameras can be obtained in real-time from the very first frame. Changes in sampling rate or dynamics between training and run-time have no bearing on performance – for the same reason. Further, just like classical multi-view triangulation, our approach requires no 3D supervision and hence relies only on 2D correspondences. The proposed approach also assumes known 2D projected measurements so it does not directly leverage pixel intensities. Therefore, our approach can be integrated with any available 2D landmark image detector such as HR-Net [47], Stacked Hourglass Networks [36], Integral Pose Regression [48], and others. Finally, the camera is **not** assumed to be static making the proposed approach agnostic to camera movements.

## 2. Related Work

**Multi-view approaches:** Multi-view triangulation [13] has been the method of choice in the context of large-scale complex rigs with multiple cameras [20, 4, 45, 11] 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 also come at considerable cost and effort. Recent work by Isakov et al. [18] and others [43, 21, 49, 41] have explored how supervised learning can be used to enhance multi-view reconstruction. Similarly, work by Rhodin et al. [44] and Kacobas et al. [25] attempted to use supervised and self-supervised learning, respectively, to infer 3D geometry from a single physical camera view. An obvious drawback to these approaches is that one is required to have intimate 3D supervision of the object before deployment –

a limitation that modern multi-view rigs are not faced with. None of these approaches are as general as the one we are proposing. For example, nearly all these prior works deal solely with the reconstruction of the human pose as they are heavily reliant upon peripheral 3D supervision.

**2D to 3D Lifting:** NRSfM [6] aims to reconstruct the 3D structure of a deforming object from 2D correspondences observed from multiple views. While the object deformation has classically been assumed to occur in time [3, 30, 39, 54, 33, 42], the vision community has increasingly drawn attention to *atemporal* applications – commonly known as unsupervised 2D-3D lifting. These temporal approaches rely on the sequential motion of objects, our approach on the other hand is much more unconstrained – accepting uncalibrated atemporal 2D instances. Advances in unsupervised learning based approaches to 2D-3D lifting [27, 37] have seen significant improvements in their robustness and fidelity 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. Preliminary

**Notations** This paper uses the following notations throughout the manuscript.

Variable type	Examples
Scalar	$s, N, K, L$
Vector	$\mathbf{s}, \boldsymbol{\psi}, \boldsymbol{\lambda}$
Matrix	$\mathbf{W}, \mathbf{S}, \mathbf{R}, \mathbf{t}, \mathbf{D}$
Function	$f_e, f_d, g$
$l^{th}$ layer	${}^l\boldsymbol{\psi}, {}^l\mathbf{D}, {}^l\boldsymbol{\lambda}$
$n^{th}$ instance	$\mathbf{W}^{(n)}, \mathbf{S}^{(n)}, \boldsymbol{\psi}^{(n)}, \boldsymbol{\lambda}^{(n)}$
$k^{th}$ view	$\mathbf{W}_k, \mathbf{R}_k, \boldsymbol{\psi}_k$

Any different signs utilized to explain a mathematical phenomenon other than the ones described above would be explicitly defined wherever deemed necessary.

**Problem setup.** We are interested in a camera rig setup with  $K$  synchronized views capturing  $N$  instances (samples) of non-rigid objects from the same category. Specifically, we are given a non-sequential (atemporal) dataset containing  $N$  multi-view 2D observations  $\{\mathbf{W}_1^{(1)}, \dots, \mathbf{W}_1^{(N)}; \dots; \mathbf{W}_K^{(1)}, \dots, \mathbf{W}_K^{(N)}\}$ , where each  $\mathbf{W} \in \mathbb{R}^{P \times 2}$  represents 2D location for  $P$  keypoints. We want to reconstruct the 3D shape  $\mathbf{S}^{(1)}, \dots, \mathbf{S}^{(N)}$ , where each  $\mathbf{S} \in \mathbb{R}^{P \times 3}$  for each of the  $N$  instances of the object.

**Weak perspective projection.** We assume weak perspective projections, *i.e.* for a 3D structure  $\mathbf{S}$  defined at a canonical frame, its 2D projection is approximated as

$$\mathbf{W} \approx s\mathbf{R}\mathbf{S}_{xy} + \mathbf{t}_{xy} \quad (1)$$

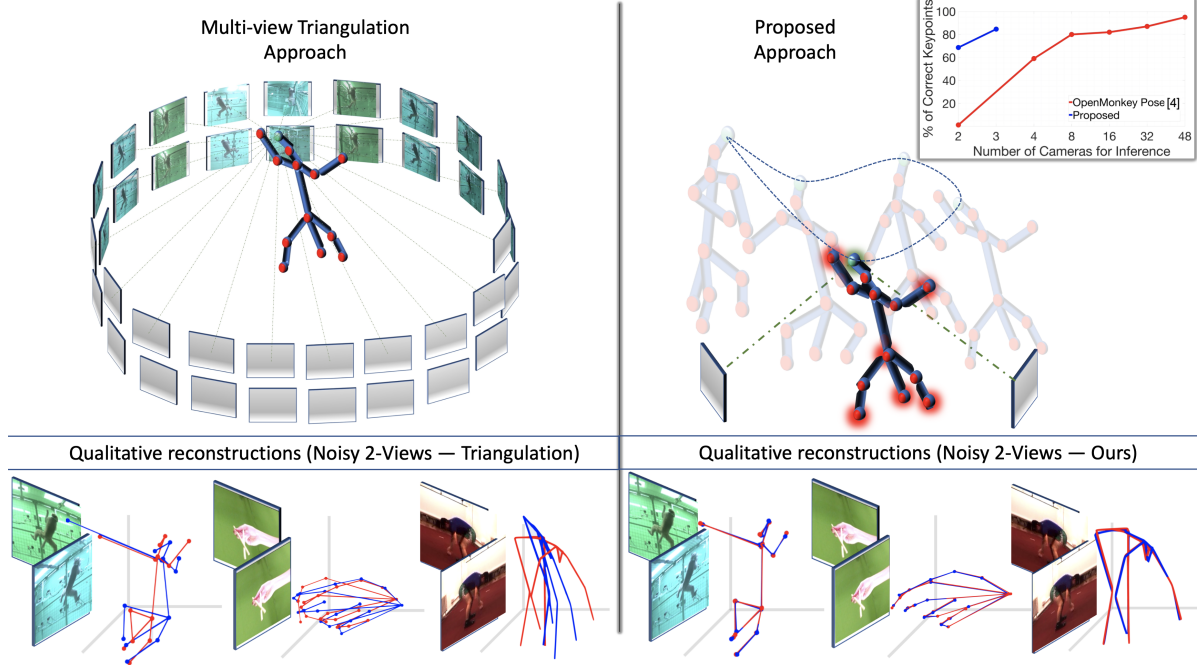


Figure 1: A traditional multi-view setup relies on the concept of triangulation with the assumption that the point being reconstructed is static in time – requiring a large number of physical views (i.e. cameras) to ensure a high fidelity reconstruction. Our approach utilizes multi-view 2D correspondences taken at the same (rigid) and different (non-rigid) points in time via a neural shape prior. Empirically (see the plot in top-right), we demonstrate that our 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 as little as two physical views, and red lines show the corresponding 3D ground-truth.

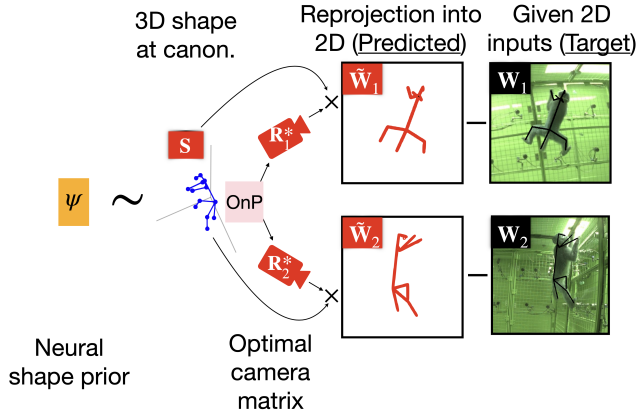


Figure 2: Two views statistical shape prior. The 3D structure  $\mathbf{S}$  is drawn from a statistical shape distribution using neural shape priors and consequently projected to 2 views using the cameras  $\mathbf{R}_k^* \forall k \in [1, 2]$  – calculated through OnP formulation [46]. The proposed approach minimizes the 2D projection error between the predicted 2D projections  $\tilde{\mathbf{W}}_k$  and target (input) 2D projections  $\mathbf{W}_k$ .

where  $\mathbf{R}_{xy} \in \mathbb{R}^{3 \times 2}$ ,  $\mathbf{t}_{xy} \in \mathbb{R}^2$  are the  $x$ - $y$  component of a rigid transformation, and  $s > 0$  is the scaling factor inversely proportional to the object depth if the true camera

model is pin-hole. If all 2D points are visible and centered,  $\mathbf{t}_{xy}$  can be omitted by assuming the origin of the canonical frame is at the center of the object. Due to the bilinear form of (1),  $s$  is ambiguous and becomes up-to-scale recoverable only when  $\mathbf{S}$  is assumed to follow certain prior statistics. We handle scale by approximating with an orthogonal projection and solving an Orthogonal-N-Point (OnP) problem [46] to find the camera pose along with the scale, as discussed in Sec. 4.2.

**Statistical shape model.** We assume a linear model for the 3D shapes  $\mathbf{S}$  to be reconstructed, i.e. at canonical coordinates, the vectorization of  $\mathbf{S}$  in Eq. (1), denoted  $\mathbf{s} = \text{vec}(\mathbf{S}) \in \mathbb{R}^{3P}$  can be written as

$$\mathbf{s} = \mathbf{D}\psi \quad (2)$$

where  $\mathbf{D} \in \mathbb{R}^{3P \times B}$  is the shape dictionary with  $B$  basis and  $\psi \in \mathbb{R}^B$  is the code vector - taking insight from classical sparse dictionary learning methods. The factorization of  $\mathbf{S}$  in Eq. (2) is ill-posed by nature; in order to resolve the ambiguities in this factorization, additional priors are necessary to guarantee the uniqueness of the solution. Notable priors include the assumption of  $\mathbf{S}$  being (i) low rank [8, 6, 2, 9, 31], (ii) lying in a union-of-subspaces [32, 53, 1] (iii) or compressible [26, 52, 28].

The low-rank assumption becomes infeasible when the data exhibits complex shape variations, the Union-of-subspaces NRSfM methods have difficulty clustering shape deformations and estimating affinity matrices effectively just from 2D observations. Finally, the sparsity prior allows more powerful modeling of shape variations with a large number of subspaces but suffers from sensitivity to noise.

**Neural Shape Prior** Our neural shape prior is an approximation to a hierarchical sparsity prior introduced by Kong et al. [27], where each non-rigid shape is represented by a sequence of hierarchical dictionaries and corresponding sparse codes. Other neural shape priors – such as C3PDO [37] – could be entertained as well but we chose to employ Kong et al.’s method due to its simplicity with respect to enforcing multi-view equivariant constraints. The approach in [27] maintains the robustness of sparse code recovery by utilizing overcomplete dictionaries to model highly deformable objects consisting of large-scale shape variation. Moreover, if the subsequent dictionaries in this multi-layered representation are learned properly, they can serve as a filter such that only functional subspaces remain and the redundant are removed. Due to the introduction of multiple levels of dictionaries and codes in the following section, we will abuse the notation of  $\mathbf{D}, \psi$  by adding left superscript 1, *i.e.*  ${}^1\mathbf{D}, {}^1\psi$  indicating that they form the first level of hierarchy. Assuming the canonical 3D shapes are compressible via multi-layered sparse coding with  $l \in L$  layers, the shape code  ${}^1\psi$  is constrained as

$$\begin{aligned} \mathbf{s} &= {}^1\mathbf{D}^1\psi \\ {}^1\psi &= {}^2\mathbf{D}^2\psi \\ &\vdots \\ {}^{L-1}\psi &= {}^L\mathbf{D}^L\psi \\ \text{s.t. } \|\psi\|_1 &\leq {}^l\lambda, {}^l\psi \geq \mathbf{0}, \forall l \in \{1, \dots, L\} \end{aligned} \quad (3)$$

where  ${}^l\mathbf{D} \in \mathbb{R}^{l-1B \times lB}$  are the hierarchical dictionaries,  $l$  is the index of hierarchy level, and  ${}^l\lambda$  is the scalar specifying the amount of sparsity in each level. Thus, the learnable parameters are  $\Theta = \{\dots, {}^l\mathbf{D}, {}^l\lambda, \dots\}$ . The single set of parameters  $\Theta$  are fit *jointly* along with the sparse codes, rotation matrices, and structures  $\mathbf{S}$  for each instance in the dataset. Jointly constraining each instance via a common set of weights (the “neural prior”) makes this work more akin to classic factorization methods, in which both the shared factors and the weightings for each instance are jointly inferred, rather than to network training approaches which aim to find weights that generalize well when later used to perform inference on unseen data.

**Factorization-based NRSfM.** Equivalently, the linear model in Eq. (2) could be rewritten as

$$\mathbf{S} = \mathbf{D}^\#(\psi \otimes \mathbf{I}_3)$$

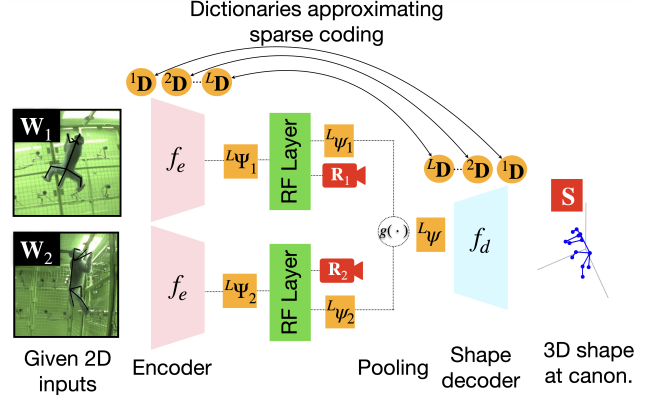


Figure 3: Architecture showing a 2-view 3D reconstruction approach (easily extensible to  $K > 2$  views). The 2D projections from both views  $\mathbf{W}_k \forall k \in [1, 2]$  acts as an input to encoder  $f_e$  that extracts the block sparse code  $\Psi_k$  from the corresponding views. A Rotation Factorization (RF) layer at the bottleneck stage shown in green, factorizes the block sparse code into the respective camera matrix  $\mathbf{R}_k$  and the unrotated vector sparse code  ${}^L\psi_k$ . The codes are then fused via *pooling* function  $g$  into a single code  ${}^L\psi$  that acts as an input to the shape decoder  $f_d$ . The shape decoder predicts the 3D structure  $\mathbf{S}$  in the canonical frame while enforcing equivariant view consistency.

where  $\mathbf{D}^\# \in \mathbb{R}^{P \times 3B}$  is a reshape of  $\mathbf{D}$  and  $\otimes$  denotes a Kronecker product. Applying the camera matrix  $\mathbf{R}_{xy}$  gives the 2D pose. Thus

$$\mathbf{S}\mathbf{R}_{xy} = \mathbf{D}^\#(\psi \otimes \mathbf{R}_{xy})$$

Substituting the input 2D pose  $\mathbf{W}$  from Eq. (1), we have

$$\begin{aligned} \mathbf{W} &= \mathbf{D}^\# \Psi_{xy} \\ \text{s.t. } \Psi_{xy} &= \psi \otimes \mathbf{R}_{xy} \text{ and } \psi \in \mathcal{C} \end{aligned} \quad (4)$$

where  $\Psi_{xy} \in \mathbb{R}^{3B \times 2}$  is the sparse block code denoting the first two columns of  $\Psi \in \mathbb{R}^{3B \times 3}$ ; and  $\mathcal{C}$  denotes the neural shape prior constraints applied on the code  $\psi$ , *e.g.* hierarchical sparsity [27] in our case. Conceptually,  $\Psi$  is a matrix with rotations and sparse code built into it. Under the unsupervised settings,  $\mathbf{D}, \psi, \mathbf{R}, \mathbf{S}$  are all unknowns and are solved under the simplified assumptions that the input 2D poses are obtained through a weak perspective camera projection. We also analytically compute  $\mathbf{R}^*$  as a solution to a Orthographic-n-point (OnP) problem that acts as a supervisory signal to  $\mathbf{R}$ . Corresponding proof for  $\mathbf{R}^*$  is discussed in the supplementary section.

## 4. Approach

### 4.1. Bilevel optimization

Given only the input 2D poses in Eq. (4), two problems remain to address



- How to formulate an optimization strategy to recover  $\mathbf{D}, \psi, \mathbf{R}, \mathbf{S}$ ?
- How to efficiently pool in  $K$  different camera views and enforce equivariance over the predicted  $K$  camera matrices and a single 3D structure in canonical frame?

We choose to impose neural shape priors through hierarchical sparsity constraints [27] literature for approaching a solution to the above problems, with learnable parameters  $\Theta$  (see Sec. 4.2). From Eq. (4), the learning strategy of multi-view NRSfM problem for  $N$  instances with  $K$  views is then interpreted as solving the following bilevel optimization problem. Eq. (3) leads to relaxation of the following lower-level problem

$$\min_{\mathbf{D}, \Theta} \sum_{k=1}^K \sum_{n=1}^N \left( \min_{\psi_k^{(n)}, \mathbf{R}_k^{(n)}} \|\mathbf{W}_k^{(n)} - {}^1\mathbf{D}({}^1\boldsymbol{\Psi}_k^{(n)})\|_F + \sum_{l=1}^L {}^l\lambda \|{}^l\boldsymbol{\Psi}_k^{(n)}\|_F^{(3 \times 2)} + \sum_{l=2}^L \|({}^{l-1}\boldsymbol{\Psi}_k^{(n)}) - {}^l\mathbf{D}({}^l\boldsymbol{\Psi}_k^{(n)})\|_F \right) \quad (5)$$

where the first expression in (5) minimizes the 2D projection error, the second expression enforces sparsity, and the third expression fits each dictionary in the hierarchy to the dictionary representation in the preceding layer. Minimizing the block Frobenius norm of  $\boldsymbol{\Psi}$  is equivalent to minimizing the  $L_1$  norm of the vector sparse code  $\psi$  because  $\|\psi\|_1 = \frac{1}{\sqrt{2}} \|\boldsymbol{\Psi}\|_F^{(3 \times 2)}$ .

## 4.2. Network approximate solution

The optimization problem in Eq. (5) is an instance of dictionary learning problem with sparse codes  $\psi$ . The classical approach to this problem is by solving the Iterative Shrinkage and Thresholding Algorithm (ISTA) [5]. However, Pappayan et al. [38] show that a single layer feedforward network with Rectified Linear Unit (ReLU) activations approximate one step of ISTA, with the bias terms  ${}^l\lambda$  adjusting the sparsity of recovered code for the  $l^{\text{th}}$  layer. Furthermore, the dictionaries  $[{}^1\mathbf{D}, \dots, {}^L\mathbf{D}]$  can be learned by back-propagating through the feedforward network. We devise a network architecture that serves as an approximate solver to the above optimization problem and provide derivations in the following subsections.

**Approximating sparse codes.** We review the sparse dictionary learning problem and consider the single-layer case stated above. To reconstruct an input signal  $\mathbf{X}$ , the optimization problem becomes

$$\min_{\boldsymbol{\Psi}} \|\mathbf{X} - \mathbf{D}\boldsymbol{\Psi}\|_F + \lambda \|\boldsymbol{\Psi}\|_F$$

As stated above, Pappayan et al. [38] propose that one iteration of ISTA gives back the block-sparse codes  $\boldsymbol{\Psi}$  as

$$\boldsymbol{\Psi} = \text{ReLU}(\mathbf{D}^\top \mathbf{X}; \lambda)$$

We interpret ReLU as solving for the block-sparse code and incorporate ReLU as the nonlinearity in our encoder part of the network.

**Encoder architecture.** We propose to devise an encoder network  $f_e$  that takes the 2D poses as input and outputs the block sparse codes  $\boldsymbol{\Psi}$  that has within itself the rotation matrix  $\mathbf{R}$  as well as a rotationally invariant sparse code  $\psi$ , *i.e.*  $f_e(\mathbf{W}_k^{(n)}) \mapsto ({}^L\boldsymbol{\Psi}_k^{(n)})$ . Unrolling one iteration of ISTA for each layer,  $f_e$  takes  $\mathbf{W}_k^{(n)}$  as 2D pose inputs and produces block sparse codes for the last layer  $[{}^1\boldsymbol{\Psi}_k^{(n)}, \dots, {}^L\boldsymbol{\Psi}_k^{(n)}]$  as output, shown in Fig. 3

$$\begin{aligned} {}^1\boldsymbol{\Psi}_k^{(n)} &= \text{ReLU}\left(\left[({}^1\mathbf{D}^\#)^\top \cdot \mathbf{W}_k^{(n)}\right]_{3 \times 2}; {}^1\lambda^{(n)}\right) \\ {}^2\boldsymbol{\Psi}_k^{(n)} &= \text{ReLU}\left(\left({}^2\mathbf{D} \otimes \mathbf{I}_3\right)^\top \cdot {}^1\boldsymbol{\Psi}_k^{(n)}; {}^2\lambda^{(n)}\right) \\ &\vdots \\ {}^L\boldsymbol{\Psi}_k^{(n)} &= \text{ReLU}\left(\left({}^L\mathbf{D} \otimes \mathbf{I}_3\right)^\top \cdot {}^{L-1}\boldsymbol{\Psi}_k^{(n)}; {}^L\lambda^{(n)}\right) \end{aligned} \quad (6)$$

where  ${}^l\lambda^{(n)}$  is the learnable threshold for each layer.  $({}^l\mathbf{D} \otimes \mathbf{I}_3)^\top \cdot {}^{l-1}\boldsymbol{\Psi}_k^{(n)}$  is implemented by a convolution transpose.

**Rotation Factorization layer.** At the bottleneck, our encoder network generates a block sparse code for  $K$ -views  ${}^L\boldsymbol{\Psi}_k^{(n)}$ . As evident in Eq. (4), since the block sparse code has rotations  $\mathbf{R}_k^{(n)}$  as well as an unrotated sparse code  $\psi_k^{(n)}$ , we add a fully-connected layer that factorizes out these quantities, named Rotation Factorization (RF) layer, shown as a green block in Fig. 3. Consequently,  ${}^L\boldsymbol{\Psi}_k^{(n)}$  is then factorized into an unrotated sparse code  ${}^L\psi_k^{(n)}$  and the rotation matrix  $\mathbf{R}_k^{(n)}$  (constraining to  $SO(3)$  using SVD) using this fully-connected RF layer. At this stage, we pool the features from all the rotationally invariant or unrotated sparse codes  ${}^L\psi_k^{(n)}$  using a sum pooling operation  $\mathbf{g}$  that enforces the equivariance consistency within all the views by combining features from multiple views.

$$\mathbf{g}({}^L\psi_1^{(n)}, \dots, {}^L\psi_K^{(n)}) \mapsto ({}^L\psi^{(n)}) \quad (7)$$

as shown in architecture overview Fig. 3, where  $\mathbf{g}$  denotes a **sum** operation. Since the pooled sparse code  ${}^L\psi$  is rotationally invariant, we generate a single canonical 3D structure  $\mathbf{S}$  through a decoder network  $\mathbf{f}_d$ , that remains equivariant to  $K$  camera rotations  $\mathbf{R}_1, \dots, \mathbf{R}_K$ . Thus, the decoder network  $\mathbf{f}_d$  helps supervise the fully-connected RF layer.

**Insight behind multi-view consistency.** For each individual view, we get a block sparse code representation  $\boldsymbol{\Psi}_k^{(n)}$  that has the rotation  $\mathbf{R}_k^{(n)}$  combined with an unrotated sparse code  $\psi_k^{(n)}$ . RF layer disentangles these quantities and generates codes that are consistent with an unrotated or canonicalized view. This architecture thus enforces

equivariance consistency by consequently passing the unrotated sparse code  $\psi$  through a shape decoder to produce a canonicalized 3D structure. When we jointly encode multiple views into a single canonical shape, the equivariance is implicitly enforced after projecting them through the given multiple cameras. These multi-view projections help supervise the multi-view NRSfM network.

**Decoder architecture.** Finally, a decoder  $f_d$  is devised that takes input a pooled bottleneck sparse code (see Eq. 7) and generates a canonical 3D structure  $\mathbf{S}$ . Thus,  $f_d(L\psi^{(n)}) \mapsto (\mathbf{S}^{(n)})$

$$\begin{aligned} L^{-1}\psi^{(n)} &= \text{ReLU}(L\mathbf{D} \cdot L\psi^{(n)}; L\lambda^{(n)}) \\ &\vdots \\ {}^1\psi^{(n)} &= \text{ReLU}({}^2\mathbf{D} \cdot {}^2\psi^{(n)}; {}^2\lambda^{(n)}) \\ \mathbf{S}^{(n)} &= {}^1\mathbf{D} \cdot {}^1\psi^{(n)} \end{aligned} \quad (8)$$

We analytically compute a closed-form solution to  $\mathbf{R}^*$  as a solution to an Orthographic-n-Point (OnP) problem that implicitly acts as supervisory signal for the  $\mathbf{R}_k^{(n)}$ . Detailed proof is shown in the supplementary section.

**Calculating  $\mathbf{R}$  using solution from OnP** We are using a closed-form algebraic solution to  $\mathbf{R}^*$  that gives us an optimal solution for  $\mathbf{R}$  produced by the network at the bottleneck stage. We opt to use an algebraic solution that can be implemented as a differentiable operator and could be easily accomplished via modern autograd packages. The detailed proof for the OnP solution is given in the appendix.

**Loss function** To reemphasize the loss function in our neural architecture, the loss function driving the proposed approach is a reprojection error

$$\mathcal{L} = \frac{1}{KN} \sum_{k=1}^K \sum_{n=1}^N \|\mathbf{W}_k^{(n)} - \mathbf{S}^{(n)}\mathbf{R}_k^{(n)}\|_F \quad (9)$$

## 5. Experiments

Evaluation over objects such as human body, hands, and monkey body is divided into two major categories: (i) Multi-view 3D reconstruction of an input 2D dataset, and (ii) Generation of 3D labels for unseen 2D data. The former compares against classical algorithms to generate high-fidelity 3D reconstruction from multi-view 2D input datasets. The latter discusses the generalization capability of our approach and shows that it does not overfit. For this, we follow one of the standard protocols for a human pose dataset and show results on the validation split. Equally competent 3D reconstructions are obtained for squishy deformable categories such as balloon deflation or paper tearing [19] – making the proposed approach agnostic to deformable object categories. Results from NRSfM Challenge dataset [19] using 2 camera views are shown in the supplementary section.

**Network architecture and implementation details** The same neural prior architecture is used for all the  $K$  views across different datasets. We use  $K$ –encoders and a single shape decoder to generate one 3D structure in a canonicalized frame. The dictionary size (*i.e.* neural units) within each layer of encoder is decreased exponentially:  $\{1024, 512, 256, 128, 64, 32, 16, 8\}$ . Ideally, if a validation set with 3D groundtruth is provided, we could select optimal architecture based on cross-validation. However, due to the unsupervised setting, we rather set the hyperparameters heuristically. For the encoder and decoder architecture discussed in Eq. (6), (8), we use a convolutional network as in Kong et al. [27] and share the convolution kernels (*i.e.* dictionaries) between the encoder and decoder.

**Training details** We keep the same weightings for the re-projection error shown in loss function Eq. (9). We use the Adam optimizer [24] in our implementation.

**Evaluation metrics** Unless otherwise noted, 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 normalizing them to match against the given ground-truth (GT) followed by rigidly aligning these predictions to GT using Procrustes alignment. Lower the better. **PCK**: percentage of correct keypoints after Procrustes alignment. 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*).

**Monkey body dataset** OpenMonkeyStudio [4] is a huge Rhesus Macaque monkey pose dataset in a setup similar to PanOptic Studio where 62 cameras capture the markerless pose of Rhesus Macaque monkeys. We use the provided 2D annotations over the Batch (7, 9, 9a, 9b, 10, and 11). This dataset also provides the groundtruth 3D labels for the given batches to evaluate the 3D reconstruction performance.

**Human body dataset** Human 3.6 Million (H3.6M) [16] is a large-scale human pose dataset with images featuring actors performing daily activities from 4 camera views - annotated by motion capture systems. The 2D keypoint annotations of H3.6M preserve the perspective effect, and thus is a realistic dataset for evaluating the practical usage of generating 3D labels for unseen data as well as test the generalization capability of our approach. The results obtained on this dataset supports our hypothesis that the weak-perspective is a reasonable preliminary assumption; we plan to account for perspective effects as part of future work. We use this dataset for both quantitative and qualitative evaluation. For generating 3D reconstruction of an input dataset (see Sec. 5.1), we pick 5 subjects (1, 5, 6, 7, 8) and compare against the classical multi-view triangulation baselines. For generating 3D labels over unseen 2D data to showcase the

generalization capability (see Sec. 5.2), we follow the standard protocol on H3.6M and use the subjects (1, 5, 6, 7, 8) during the training stage and the subjects (9, 11) for evaluation stage. Evaluation is performed on every 64th frame of the test set. We include average errors for each method.

**Human hands dataset** Finally, we use an open-source hands dataset - FreiHand [55] - a large-scale open-source dataset with varied movements of hands with 3D pose annotated by motion capture systems. It consists of 32560 samples with their corresponding camera intrinsics. We generate random camera extrinsics and randomly create multiple camera views to generate multi-view 2D inputs for evaluating the proposed approach.

### 5.1. 3D reconstruction of an input 2D dataset

Like multi-view triangulation or bundle adjustment, our approach jointly infers the unknown 3D shape and camera rotations from 2D keypoints. By simultaneously fitting the shared network parameters used to recover shape and pose, our approach constrains the possible reconstructions much more strongly than multi-view triangulation or bundle-adjustment approaches. Although we showcase the generalization capability of the setup by applying the fitted network to generate 3D labels for unseen 2D data (see Sec. 5.2), the major contribution of our approach is the optimization process for multi-view 3D reconstruction of an input 2D dataset. The goal is to evaluate the robustness of the proposed multi-view neural shape prior across different shape variations and hence as part of the evaluation, we report how well our method is able to reconstruct different datasets compared to the baseline methods.

**Baseline** We use an implementation of iterative multi-view triangulation with robust outlier rejection [14, 13], referred to as **TRNG** – a method of choice for multi-view 3D human pose learning by Kocabas et al. [25]. They also provide an open-source implementation for this baseline method. A more recent method doing classical optimization on triangulation is proposed by Lee and Civera [34], however, their method is not necessarily optimal in terms of accuracy, but more in terms of computation time. **TRNG** first finds the points which minimize the distance from all the rays and removes the rays which are the furthest away from that point. It then re-evaluates the triangulation and this iteration is repeated 2-3 times. Empirically, we find that increasing the iteration leads us to predict near-perfect 3D reconstruction if we have exact camera calibration parameters and exact, clean 2D projections. 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 accurate 3D reconstructions used to train 3D regression methods. We evaluate our approach on the above three datasets with substantial non-rigid deformities. For all the given experiments the cameras are chosen at random and the

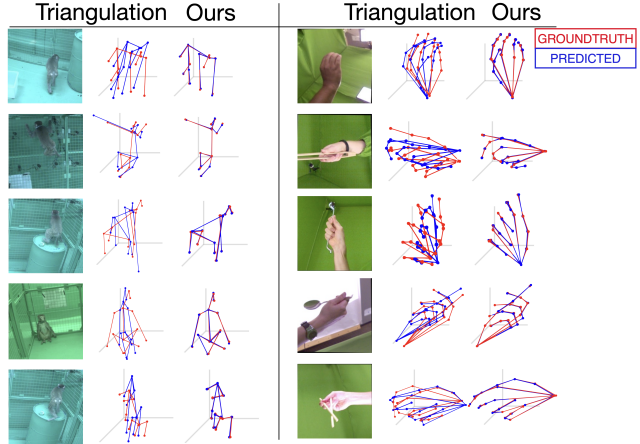


Figure 4: Qualitative 3D reconstruction comparison between the multi-view triangulation technique and our technique for Monkey body [4] and human hands [55] when operated over noisy 2D keypoints.

same set of cameras are used in the comparative baselines for a fair comparison.

**Evaluation analysis** For the Monkey body dataset, multi-view 3D reconstruction with 2– or 3– view using our approach significantly outperforms the given results in [4] and achieves comparable fidelity with limited physical views. We consider all the keypoints as correct if their reconstruction is within 10cm of the groundtruth in the **PCK** protocol. Table 3 and top-right plot in Fig. 1 shows that we outperform the given results of 2-Views by a significant margin (1.2% vs. 68.63%). The fidelity of 3D reconstructions using our method continues to rise as we add in more views - evident by the uptick in performance from 3–views. Qualitative performance of Monkey and Human hands dataset is shown in Fig. 4 and quantitative performance of Monkey body is given in Tab. 1 when operated over noisy 2D keypoints. For the Human body dataset, we inject noise in the camera extrinsics, intrinsics, and 2D keypoints separately and compare the performance in Fig. 5 and Tab. 2. The baseline method fails when noise with a small standard deviation is added, degrading the fidelity of the 3D reconstruction. Since our approach is only dependent on the quality of 2D keypoints, it shows slightly degraded performance when the noise is injected over the input 2D keypoints. Qualitative 3D reconstruction of our approach in Fig. 4, 5 shows the visual improvement over the classical multi-view triangulation approaches when operated over noisy 2D keypoints.

### 5.2. Generation to unseen 2D data

Several multi-view approaches exist for 3D human pose estimation that leverage either full or weak 3D supervision [18, 43, 21, 49, 41, 44, 25]. None of these references, however, directly tackle the unsupervised multi-view 3D reconstruction problem and hence are not as general as our so-

Method	Batch#7	Batch#9	Batch#9a	Batch#9b	Batch#10	Batch#11
<b>TRNG</b>	21.21	24.32	30.67	24.50	26.10	22.77
<b>MV NRSfM</b>	<b>8.36</b>	<b>8.25</b>	<b>9.12</b>	<b>11.52</b>	<b>8.203</b>	<b>8.17</b>

Table 1: **PA-MPJPE** error values for Monkey body dataset shows substantial improvement over the baseline multi-view triangulation approach while using only two views over noisy 2D keypoints. **PA-MPJPE** values are in **cm**.

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

Table 2: Robustness to camera calibration and 2D annotations noise for Human 3.6M dataset. Values are in **mm**.

Method	PCK
2 Views [4]	1.2%
4 Views [4]	59%
8 Views [4]	80%
16 Views [4]	82%
32 Views [4]	87%
48 Views [4]	95%
2-views (ours)	<b>68.63%</b>
3-views (ours)	<b>84.63%</b>

Table 3: Percentage of Correct Keypoint (PCK) % for OpenMonkeyStudio dataset. Following [4], the threshold for considering a keypoint to be correct is set at 10 cm.

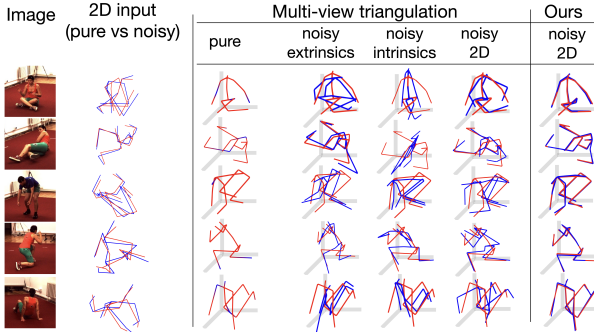


Figure 5: Qualitative results on Human 3.6M dataset with  $\sigma = [0.5, 0.5, 25]$  as intrinsics, extrinsics, and 2D keypoints Gaussian noise, respectively.

lution. However, to showcase the generalization capability of our approach, we include these approaches in our evaluation, shown in Tab. 4 (the values are in **mm**). Furthermore, we also compare against recent monocular unsupervised 3D reconstruction methods. We leverage the processed datasets by Dovotny et al. [37] as the detected 2D keypoints for a fair evaluation and use the evaluation split of H3.6M dataset for this comparison. We find that our approach outperforms all other unsupervised approaches, and is on-par with many supervised methods.

Method	Detected 2D	GT 2D
Iskakov et al. [18]	20.8	-
Remelli et al. [43]	30.2	-
Kadkhodamohammadi et al. [21]	49.1	-
Tome et al. [49]	52.8	-
Pavlakos et al. [41]	56.9	-
Multi-view Martinez [35]	57.0	-
Rhodin et al. [44]	51.6	-
Kocabas et al [25]	45.04	-
Kocabas et al. (SS w/o R) [25]	70.67	-
PRN [40]	124.5	86.4
RepNet [50]	65.1	38.2
Iqbal et al. [17]	69.1	-
Pose-GAN [29]	173.2	130.9
Deep NRSfM [27]	-	104.2
C3DP0 [37]	153.0	95.6
<b>MV NRSfM (Ours)</b>	<b>45.2</b>	<b>30.2</b>

Table 4: Generalization experiments. Red tint rows have 3D supervision. Yellow tint are unsupervised 3D reconstruction methods. Our method is on par with most 3D supervised methods, and outperforms all unsupervised methods.

## 6. Discussion and Conclusion

We propose a multi-view 2D-3D lifting architecture that incorporates neural shape prior using the recent advances of modern deep learning methods. Our contribution combines the ideas from multi-view geometry and recent monocular deep 3D lifting approaches – essentially leveraging the best features of both worlds. We also show the generalization capability of the proposed approach by generating accurate 3D reconstructions on unseen data. Although we require limited rigid views at any instant of time, our approach still requires multiple non-rigid atemporal views to enforce the proposed neural shape prior during training/optimization. Literature in domain of neural shape priors is extensive [37, 51] and new innovations are proposed constantly, and we believe we could leverage these innovations in our framework as part of future direction.



## 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] Ijaz Akhter, Yaser Sheikh, Sohaib Khan, and Takeo Kanade. Nonrigid structure from motion in trajectory space. In *Advances in neural information processing systems*, pages 41–48, 2009. 2
- [4] 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. 2, 6, 7, 8
- [5] 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. 5
- [6] 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
- [7] 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
- [8] 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
- [9] 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
- [10] Richard A Gibbs, Jeffrey Rogers, et al. Evolutionary and biomedical insights from the rhesus macaque genome. *science*, 316(5822):222–234, 2007. 1
- [11] Ralph Gross, Iain Matthews, Jeffrey Cohn, Takeo Kanade, and Simon Baker. Multi-pie. *Image and Vision Computing*, 28(5):807–813, 2010. 2
- [12] 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
- [13] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003. 2, 7
- [14] Richard I Hartley and Peter Sturm. Triangulation. *Computer vision and image understanding*, 68(2):146–157, 1997. 7
- [15] Brian Hrotenok, Tucker Balch, David Byrd, Rebecca Roberts, Chanh 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
- [16] 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. 2, 6
- [17] Umar Iqbal, Pavlo Molchanov, and Jan Kautz. Weakly-supervised 3d human pose learning via multi-view images in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5243–5252, 2020. 8
- [18] Karim Isakov, Egor Burkov, Victor Lempitsky, and Yuriy Malkov. Learnable triangulation of human pose. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7718–7727, 2019. 2, 7, 8
- [19] Sebastian Hoppe Nesgaard Jensen, Mads Emil Brix Doest, Henrik Aanæs, and Alessio Del Bue. A benchmark and evaluation of non-rigid structure from motion. *International Journal of Computer Vision*, 129(4):882–899, 2021. 6
- [20] 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
- [21] Abdolrahim Kadhodamohammadi and Nicolas Padoy. A generalizable approach for multi-view 3d human pose regression. *Machine Vision and Applications*, 32(1):1–14, 2021. 2, 7, 8
- [22] Sinead Kearney, Wenbin Li, Martin Parsons, Kwang In Kim, and Darren Cosker. Rgb-dog: Predicting canine pose from rgb-d sensors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8336–8345, 2020. 2
- [23] 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
- [24] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6
- [25] Muhammed Kocabas, Salih Karagoz, and Emre Akbas. Self-supervised learning of 3d human pose using multi-view geometry. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1077–1086, 2019. 2, 7, 8
- [26] 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

- [27] 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. 2, 4, 5, 6, 8
- [28] 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
- [29] Yasunori Kudo, Keisuke Ogaki, Yusuke Matsui, and Yuri Odagiri. Unsupervised adversarial learning of 3d human pose from 2d joint locations. *arXiv preprint arXiv:1803.08244*, 2018. 8
- [30] Suryansh Kumar. Non-rigid structure from motion: Prior-free factorization method revisited. In *Winter Conference on Applications of Computer Vision (WACV 2020)*, 2020. 2
- [31] 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
- [32] 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
- [33] Suryansh Kumar, Yuchao Dai, and Hongdong Li. Super-pixel soup: Monocular dense 3d reconstruction of a complex dynamic scene. *IEEE transactions on pattern analysis and machine intelligence*, 2019. 2
- [34] Seong Hun Lee and Javier Civera. Closed-form optimal two-view triangulation based on angular errors. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2681–2689, 2019. 7
- [35] 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. 8
- [36] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hour-glass networks for human pose estimation. In *European conference on computer vision*, pages 483–499. Springer, 2016. 2
- [37] 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, 2, 4, 8
- [38] 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. 5
- [39] Hyun Soo Park, Takaaki Shiratori, Iain Matthews, and Yaser Sheikh. 3d reconstruction of a moving point from a series of 2d projections. In *European conference on computer vision*, pages 158–171. Springer, 2010. 2
- [40] Sunghoon Park, Minsik Lee, and Nojun Kwak. Procrustean regression networks: Learning 3d structure of non-rigid objects from 2d annotations. In *European Conference on Computer Vision*, pages 1–18. Springer, 2020. 8
- [41] 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, 7, 8
- [42] Marc Pollefeys, Reinhard Koch, and Luc Van Gool. Self-calibration and metric reconstruction inspite of varying and unknown intrinsic camera parameters. *International Journal of Computer Vision*, 32(1):7–25, 1999. 2
- [43] Edoardo Remelli, Shangchen Han, Sina Honari, Pascal Fua, and Robert Wang. Lightweight multi-view 3d pose estimation through camera-disentangled representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6040–6049, 2020. 2, 7, 8
- [44] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8437–8446, 2018. 2, 7, 8
- [45] 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. 2
- [46] Carsten Steger. Algorithms for the orthographic-n-point problem. *Journal of Mathematical Imaging and Vision*, 60(2):246–266, 2018. 3
- [47] 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
- [48] 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
- [49] Denis Tome, Matteo Toso, Lourdes Agapito, and Chris Russell. Rethinking pose in 3d: Multi-stage refinement and recovery for markerless motion capture. In *2018 international conference on 3D vision (3DV)*, pages 474–483. IEEE, 2018. 2, 7, 8
- [50] Bastian Wandt and Bodo Rosenhahn. Repnet: Weakly supervised training of an adversarial reprojection network for 3d human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7782–7791, 2019. 8
- [51] Chaoyang Wang, Chen-Hsuan Lin, and Simon Lucey. Deep nrsfm++: Towards 3d reconstruction in the wild. *arXiv preprint arXiv:2001.10090*, 2020. 1, 8
- [52] 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
- [53] 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
- [54] Yingying Zhu and Simon Lucey. Convolutional sparse coding for trajectory reconstruction. *IEEE transactions on*

*pattern analysis and machine intelligence*, 37(3):529–540, 2013. [2](#)

- [55] 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. [7](#)