PR-RRN: Pairwise-Regularized Residual-Recursive Networks for Non-rigid Structure-from-Motion

Haitian Zeng[1], Yuchao Dai[2], Xin Yu[3], Xiaohan Wang[1,3], Yi Yang[4*]
Baidu Research[1], Northwestern Polytechnical University[2],
University of Technology Sydney[3], Zhejiang University[4]
zenghaitian@baidu.com; daiyuchao@gmail.com; xin.yu@uts.edu.au;
xiaohan.wang-3@student.uts.edu.au; yee.i.yang@gmail.com

Abstract

We propose PR-RRN, a novel neural-network based method for Non-rigid Structure-from-Motion (NRSfM). PR-RRN consists of Residual-Recursive Networks (RRN) and two extra regularization losses. RRN is designed to effectively recover 3D shape and camera from 2D keypoints with novel residual-recursive structure. As NRSfM is a highly under-constrained problem, we propose two new pairwise regularization to further regularize the reconstruction. The Rigidity-based Pairwise Contrastive Loss regularizes the shape representation by encouraging higher similarity between the representations of high-rigidity pairs of frames than low-rigidity pairs. We propose minimum singular-value ratio to measure the pairwise rigidity. The Pairwise Consistency Loss enforces the reconstruction to be consistent when the estimated shapes and cameras are exchanged between pairs. Our approach achieves state-of-the-art performance on CMU MOCAP and PASCAL3D+ dataset.

1. Introduction

The reconstruction of 3D object shapes and camera motions from 2D observations is an important problem in computer vision. When the object is rigid, this problem is defined as rigid Structure-from-Motion (SfM) and it can be solved reliably using existing methods like [40]. Non-Rigid Structure-from-Motion (NRSfM) relaxes the assumption of a rigid object in SfM to a deforming one, leading to a more general and challenging problem.

NRSfM is known to be an under-constrained problem if the shape is allowed to deform arbitrarily in each observation. To make this problem tractable, a standard assumption is that in each frame the 3D shape is a linear combination of a small number of basis shapes [4]. With this assumption, NRSfM is formulated as factorizing the stacked observation matrix into three component matrices: *cameras*, *coefficients* and *basis*. Previous researches exploit various constraints to solve this factorization problem, involving orthogonal constraint on the camera matrix [11, 47], restricting the basis to 3D shapes [46]. Different from those constraints on cameras or basis, another important category of approaches applied constraints to the coefficient matrix, including smooth trajectories over time in original coeffi-

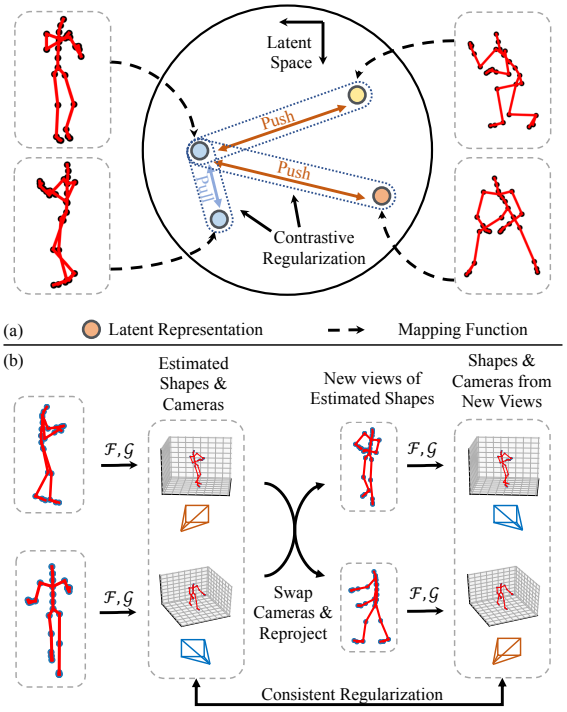

Figure 1. Illustration of pairwise losses. (a) Proposed pairwise contrastive regularization 'pushes' or 'pulls' the representations based on pairwise rigidity (similarity) of 2D shapes. (b) Consistent regularization forces the networks to produce consistent shape and camera estimation given new views of estimated shapes.

*Corresponding author.

cient matrix [15, 3] or in low-dimensional manifold [16], prior distributions [25, 41] and spatial smoothness [17]. In neural-network based models, the latent representation can be thought of as the 'coefficients', and Sidhu *et al.* [37] first apply latent space constraints for sequential dense reconstruction. These constraints reduce the indeterminacy of the NRSfM task and potentially lead to better reconstruction.

However, regularizing the reconstruction is difficult when the data is large-scale and orderless. In such cases, assuming a representation manifold or using temporal smoothness is not possible. To tackle this, we propose to regularize the non-rigid shape reconstruction in a *pairwise* manner. Compared to a strong global assumption of shapes, pairwise information are much easier to obtain, therefore the regularization can be achieved effectively.

In this paper, we introduce Pairwise-Regularized Residual-Recursive Networks (PR-RRN), a novel neural-network based model for NRSfM. PR-RRN consists of a Residual-Recursive Network (RRN) and two novel losses: Pairwise Contrastive Loss and Pairwise Consistency Loss. RRN alone can reconstruct the non-rigid shapes accurately, and it is further improved by pairwise losses. RRN contains a shape estimation network and a rotation estimation network, and the shape estimation network is constructed with a novel Residual-Recursive module, which is capable to enhance the reconstruction compared to a standard convolution layer with the same number of parameters. And the rotation estimation network is designed to estimate the camera matrix from the 2D input. Furthermore, two pairwise losses regularize the reconstruction in two different aspects, as shown in Fig. 1. Inspired by recent advances in unsupervised representation learning [18, 42, 39, 48, 36, 35], the proposed Pairwise Contrastive Loss encourages higher similarity between the latent representations of high-rigidity pairs of inputs than low-rigidity pairs. The pairwise rigidity is obtained by a novel measurement *minimal singular-value ratio*. The Pairwise Consistency Loss enforces the reconstruction to be consistent when the estimated shapes and cameras are exchanged between pairs and reprojected as new inputs. The experimental results show that PR-RRN achieves state-of-the-art reconstruction performance on large-scale human motion and categorical objects datasets.

Our contributions are summarized as following:

- We introduce a novel Residual-Recursive Network for non-rigid shapes reconstruction, which achieves state-of-the-art performance on CMU MOCAP Dataset.
- We propose Pairwise Contrastive Loss and Consistency Loss to further improve RNN. These two losses can regularize the reconstruction without assuming a global shape distribution.
- We design a novel pairwise rigidity measurement *minimal singular-value ratio*. It is easy to compute and can be used to test the rigidity of a pair of 2D observations.

2. Related Works

NRSfM. The problem of non-rigid structure from motion is first introduced in the research of recovering a sequence of 3D face landmarks and camera positions by Bregler *et al.* [4]. This research proposes a widely accepted assumption that the deforming 3D shapes can be compactly represented as a linear combination of a small number of basis shapes. Although this low rank assumption has been made, the deformation of shapes still remains under-constrained, which makes the NRSfM a challenging task for years. Various types of constraints [14, 47, 13, 44] have been explored to restrict the deforming 3D structure. Xiao *et al.* [46] propose a basis constraint to reach a close-form solution of NRSfM factorization. Torresani *et al.* [41] develop a Gaussian prior of the shape coefficients and reconstruct the shapes and cameras with probabilistic principal components analysis. With a sequence as the input, temporal smoothness can be leveraged to improve the reconstruction. Akhter *et al.* [3] introduce a dual representation of the NRSfM problem. Gotardo *et al.* [15] formulate the temporal deformation of shapes as a smooth trajectory over the coefficients of shape basis, and this idea is improved by modeling the shape trajectory and basis shape in low-dimensional manifold [33] and using kernel to measure the distance [16]. There is a milestone that Dai *et al.* [11] propose a block matrix method and achieve the outstanding performance with low rank priors. There are more breakthroughs [23, 24, 22, 31, 1] in the field like inextensible [9, 44], piecewise [12] methods, metric projection [32].

Further researches have been made to extend the NRSfM problem to more challenging situations. Zhu *et al.* [52] show that complex non-rigid human motions adhere to a union of subspace and solve it by a combined optimization of NRSfM and Low Rank Representation [28]. Li *et al.* [27] exploit grouped recurrent shapes and perform rigid SfM. Deep models have been applied to NRSfM [34, 37]. A recent work of Kong *et al.* [20] proposes to solve the NRSfM problem by learning a multi-layer sparse dictionary which is approximated with a deep neural networks. Novotny *et al.* [30] introduce a factorization network and a canonicalization network to learn shape basis with transversal property. Sidhu *et al.* [37] build a deep model for sequential dense non-rigid shape reconstruction and show that the latent space constraints are useful.

Unsupervised Representation Learning. Researches on unsupervised representation learning have achieved remarkable success. He *et al.* [18] propose Momentum Contrast (MoCo) for learning representations from unlabeled images, and the learned features are shown to be useful for

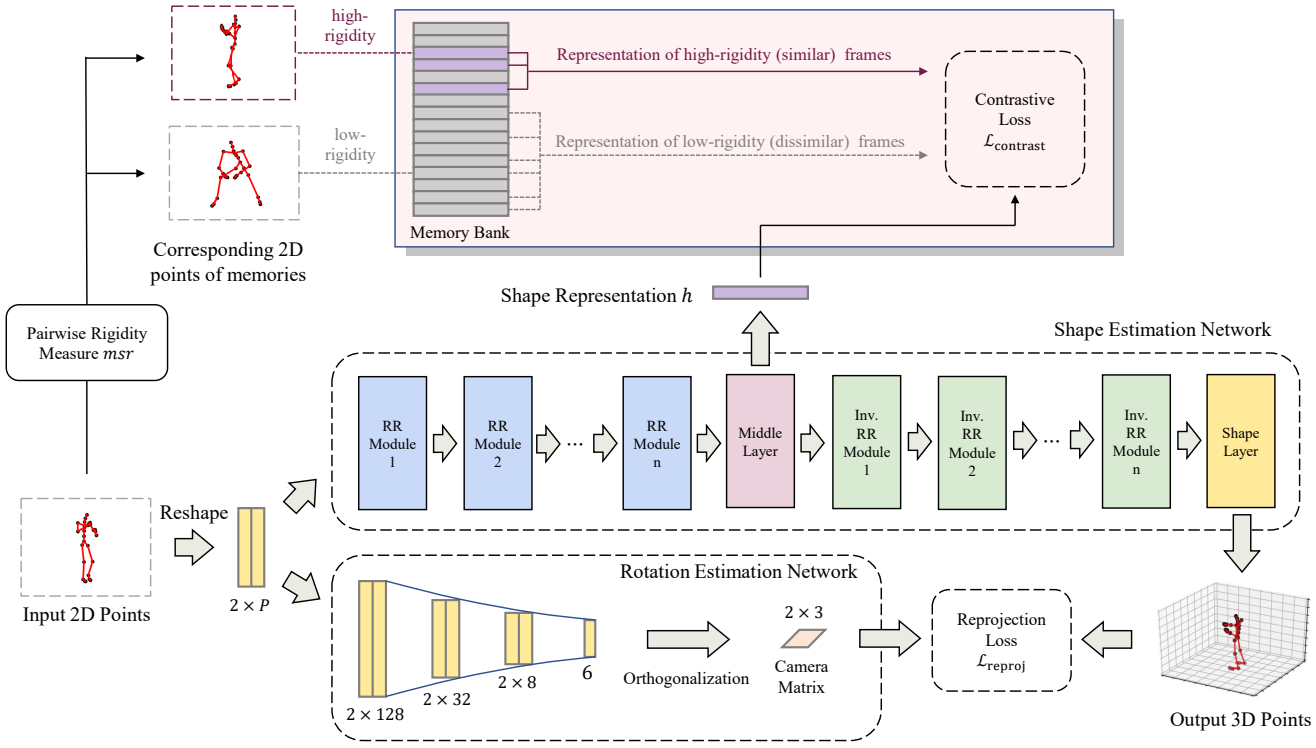

Figure 2. An overview of the proposed Residual-Recursive Networks and Pairwise Contrastive Loss. The RRN consists of two sub-networks: the shape estimation network and the rotation estimation network. The input 2D points are reshaped and fed into the two networks respectively. In the shape estimation network, a shape representation h is produced. In the contrative loss, representations of other frames (memory) in the memory bank are divided into positive and negative examples using rigidity measure msr, and are used to contrast with h. When the current training step is finished, h will be stored in the memory bank, replacing the oldest memories.

downstream tasks. Oord *et al.* [42] present the noise contrastive learning with InfoNCE loss, and show that InfoNCE loss maximizes the lower bound of mutual information between related representations. Tian *et al.* [39] introduce a Contrastive Multiview Coding method for unsupervised learning of multi-view (or multi-modal) data by using the multiple views of a same example as positive pairs. Contrastive learning is also exploited for learning representations of 3D objects by Sanghi [35] *et al.*, where the learned representation is shown to be useful for retrieving different views of a rigid object or similar objects.

3. NRSfM Recap

We first briefly review the classic Non-Rigid Structure-from-Motion problem. The inputs of NRSfM problem are F frames of P keypoints, which are 2D views of a deformable object. Let the i-th frame to be $\mathrm{W}_i \in \mathbb{R}^{2\times P}$, containing P 2D coordinates. Under the condition of orthographic projection, the camera matrix of i-th frame is $\mathrm{M}_i \in \mathbb{R}^{2\times 3}$, and satifies $\mathrm{M}_i\mathrm{M}_i^{\mathrm{T}} = \mathrm{I}_2$. The reconstructed 3D shape of i-th frame denotes $\mathrm{S}_i \in \mathbb{R}^{3\times P}$, and it is related to M_i and W_i by the following projection equation:

$$\mathrm{W}_i = \mathrm{M}_i\mathrm{S}_i. \tag{1}$$

The NRSfM problem is known [47] to be ill-posed if no assumption is made on S_i. Bregler *et al.* [4] make a widely-accepted assumption that S_i of all frames are a linear combination of K basis shapes $\mathrm{B}_k \in \mathbb{R}^{3\times P}$, which is:

$$\mathrm{S}_i = \sum_{k=1}^{K} (c_{i,k} \otimes \mathrm{I}_3)\mathrm{B}_k, \tag{2}$$

where $K \ll F$, $\otimes$ is the Kronecker product, and $c_{i,k}$ stands for the coefficient (weight) of B_k in S_i.

In case of a rigid object, *i.e.* the 3D shape does not deform across frames, NRSfM degrades into a Structure from Motion (SfM) problem, which can be formulated as:

$$\mathbf{W} = \begin{bmatrix} \mathrm{W}_1 \\ \vdots \\ \mathrm{W}_F \end{bmatrix} = \begin{bmatrix} \mathrm{M}_1 \\ \vdots \\ \mathrm{M}_F \end{bmatrix} \mathrm{S}^r, \tag{3}$$

where $\mathbf{W} \in \mathbb{R}^{2F\times P}$ is the stacked matrix of W_i, $\mathrm{S}^r \in \mathbb{R}^{3\times P}$ is the rigid shape. So that in this case $\mathrm{rank}(\mathbf{W}) \leq 3$,

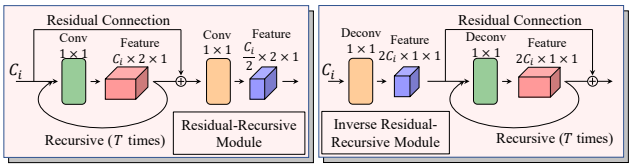

Figure 3. Design of the Residual-Recursive Module. Suppose the channel of input feature is C_i. The feature is repeatedly fed into the recursive layer for T times before fed to the next layer.

and Tomasi & Kanade [40] use a truncated Singular Value Decomposition (SVD) of $\mathbf{W}$ to recover the cameras and the rigid shape. This property is used in Sec. 4.2 to derive a measure of rigidity.

4. Method

In this section, we introduce the Pairwise-Regularized Residual-Recursive Networks. The neural-network model is described in Sec. 4.1 and the two pairwise-regularization losses are explained in Sec. 4.2 and Sec. 4.3. We suppose that the 2D keypoints in each frames are zero-centered so that the transition term is cancelled.

4.1. Residual-Recursive Networks

The Residual-Recursive Networks (RRN) consist of two sub-networks: the shape estimation network and the rotation estimation network. With a reprojection loss described in Sec. 4.1.2, RRN can be trained to reconstruct 3D shapes from 2D keypoints.

4.1.1 Shape estimation network

The role of shape estimation network is to map a single 2D input to a 3D shape. Let $\mathrm{W}_i \in \mathbb{R}^{2\times P}$ be the i-th 2D input, the output of shape estimation network is the 3D shape $\hat{\mathrm{S}}_i \in \mathbb{R}^{3\times P}$, which can be written as:

$$\hat{\mathrm{S}}_i = \mathcal{F}(\mathrm{W}_i). \tag{4}$$

Network Input. The input W_i is reshaped into a $P \times 2 \times 1$ feature tensor which can be fed into convolution layers. Here, the number of channels is P, the width is 2 and the height is 1. We empirically find that it works better than vectorizing W_i into $2P \times 1 \times 1$ like in [30, 8].

Network Structure. The shape estimation network is an autoencoder consisting of n Residual-Recursive (RR) modules. The overall framework is shown in Fig. 2. In each RR module, the feature is first fed into a recursive layer with residual connection, then it is processed with a fully connected layer where the number of output channels is reduced to half. The details of RR module is illustrated in Fig. 3. After a middle layer, the feature is mapped to a 3D shape with Inverse Residual-Recursive modules. The two types of RR modules are illustrates in Fig. 3. A tied-weight strategy [20] is used here. Finally, the 3D shape is produced by a shape layer. Motivated by previous works of NRSfM where the compactness of model is emphasized [41, 16], we choose the residual-recursive structure [50, 49, 10] to enhance the representation power of a standard convolution layer without increasing the number of parameters. We empirically find that this structure is more effective than a standard convolution layer in an autoencoder for learning difficult 2D-3D mapping.

Shape Representation. We designate the output of the middle layer as the shape representation h_i. The regularization on the shape representation is explained in Sec. 4.2.

4.1.2 Rotation estimation network

In order to estimate the camera, we design a rotation estimation network to output an orthographic projection matrix $\hat{\mathrm{M}}_i \in \mathbb{R}^{2\times 3}$ for a given 2D input, that is:

$$\hat{\mathrm{M}}_i = \mathcal{G}(\mathrm{W}_i). \tag{5}$$

We suppose that the rotation is arbitrary in each frame. A recent work [51] shows that one needs 6D representation as well as a good mapping to avoid discontinuity while using the polar decomposition for enforcing orthogonality. The rotation estimation network is designed to output a 6D vector. The network is constructed with multiple linear layers, as shown in Fig. 2. The output of the linear layers is reshaped into $\tilde{\mathrm{M}}_i \in \mathbb{R}^{2\times 3}$.

The output $\tilde{\mathrm{M}}_i$ should be further turned into an orthographic projection matrix, *i.e.* $\hat{\mathrm{M}}_i\hat{\mathrm{M}}_i^{\mathrm{T}} = \mathrm{I}_2$. The orthogonalization can be done in several ways, like Gram-Schmidt procedure [51], projections onto $\mathrm{SO}(3)$ [32, 19]. We follow [8, 20] to use Singular Value Decomposition (SVD) as the orthogonalization method, as it is shown that SVD produces better rotation estimation than Gram-Schmidt in many supervised and unsupervised tasks [26]. The orthogonalization process can be expressed as:

$$\hat{\mathrm{M}}_i = \mathrm{U}\mathrm{V}^{\mathrm{T}} \quad \text{s.t.} \quad \tilde{\mathrm{M}}_i = \mathrm{U}\Sigma\mathrm{V}^{\mathrm{T}}, \tag{6}$$

where $\mathrm{U}\Sigma\mathrm{V}^{\mathrm{T}}$ is the SVD of $\tilde{\mathrm{M}}_i$.

Finally, the 3D shape $\hat{\mathrm{S}}_i$ estimated by the shape estimation network is reprojected to a 2D shape using the rotation $\hat{\mathrm{M}}_i$. The reprojection loss is calculated as:

$$\mathcal{L}_{\text{reproj}} = \left\| \mathrm{W}_i - \hat{\mathrm{M}}_i\hat{\mathrm{S}}_i \right\|_F, \tag{7}$$

where $\|\cdot\|_F$ is the Frobenius norm.

4.2. Rigidity-based Pairwise Contrastive Loss

We introduce a Rigidity-based Contrastive loss to improve the performance of deformable shape reconstruction. First, we define a new rigidity measure. Then, a Rigidity-based Contrastive loss is proposed.

4.2.1 Minimal Singular Value Ratio

Given two 2D frames, W_i and W_j, we consider to measure the rigidity of them, *i.e.*, how similar are the corresponding 3D shapes of the two frames. Note that W_i and W_j are randomly selected frames. Let $\mathrm{A} \in \mathbb{R}^{4\times P}$ to be the stacked matrix of W_i and W_j:

$$\mathrm{A} = \begin{bmatrix} \mathrm{W}_i \\ \mathrm{W}_j \end{bmatrix}. \tag{8}$$

Inspired by Hamsici *et al.* [17], we use the ratio of the minimal (*i.e.* fourth) singular value of A to define a novel rigidity measure msr:

$$\mathrm{msr}(\mathrm{W}_i, \mathrm{W}_j) = \frac{\sigma_4^2}{\sum_{l=1}^{4} \sigma_l^2}, \tag{9}$$

where σ_l is the l-th singular value of A in descending order. As $\mathrm{rank}(\mathrm{A}) \leq 4$, the range of msr is $[0, 0.25]$. Intuitively, msr measures how much A is away from a rank-3 matrix. If $\mathrm{rank}(\mathrm{A}) \leq 3$, then $\mathrm{msr} = 0$ and it means that $\mathrm{W}_i, \mathrm{W}_j$ are two views of a rigid 3D structure. On the contrary, the rigidity of $\mathrm{W}_i, \mathrm{W}_j$ becomes lower when msr grows. More qualitative examples of msr are provided in Fig. 7.

4.2.2 Rigidity-based Contrastive Loss

We now introduce the Rigidity-based Contrastive Loss. This loss aims to regularize the representation of shapes by encouraging high similarity between similar shapes. The similarity of shapes can be found using the rigidity measure msr proposed previously. This regularization can be performed without assuming a global distribution or manifold of representation.

For a given frame W_i, we calculate a *positive set* $\mathcal{P}_i$ and a *negative set* $\mathcal{N}_i$. The positive set contains the indices of frames that are (near) rigid with W_i measured by msr, and *vice versa*. These two sets are defined as:

$$\mathcal{P}_i = \{j | \mathrm{msr}(\mathrm{W_i}, \mathrm{W_j}) < \tau, \forall \mathrm{j}\}, \tag{10}$$

$$\mathcal{N}_i = \{k | \mathrm{msr}(\mathrm{W_i}, \mathrm{W_k}) > \xi, \forall \mathrm{k}\}, \tag{11}$$

where τ, ξ are threshold parameters.

The Rigidity-based Contrastive Loss is:

$$\mathcal{L}_{\mathrm{contrast}} = -\mathbb{E}\left[\log \frac{\sum_j \exp(h_i \cdot h_j)}{\sum_j \exp(h_i \cdot h_j) + \sum_k \exp(h_i \cdot h_k)}\right], \tag{12}$$

where $\cdot$ is the dot product, $j \in \mathcal{P}_i$ and $k \in \mathcal{N}_i$. Intuitively, this loss is minimized when $h_i \cdot h_j$ have high values and $h_i \cdot h_k$ have low values. We normalize all h to unit norm before calculating the loss.

In practice, the networks are trained with mini-batches, therefore the frames outside the current mini-batch are not available. To deal with that, we use a memory bank [18] to store representations from previous mini-batches. The size of the memory bank is N_{mem}. After each training step, the current batch of representations are stored in memory bank, replacing the oldest ones. In other words, the memory bank works as a queue of representation. This allows the representation to be regularized by as much pairs as possible, and it is proved to be beneficial to learning good representation [42, 18].

4.3. Pairwise Consistency Loss

In this subsection we propose a novel pairwise consistency constraint. Given two random 2D frames $\mathrm{W}_i, \mathrm{W}_j$, the shapes and rotations are estimated using the $\mathcal{F}$ and $\mathcal{G}$ of RRN:

$$\hat{\mathrm{S}}_i = \mathcal{F}(\mathrm{W}_i), \quad \hat{\mathrm{M}}_i = \mathcal{G}(\mathrm{W}_i), \tag{13}$$

$$\hat{\mathrm{S}}_j = \mathcal{F}(\mathrm{W}_j), \quad \hat{\mathrm{M}}_j = \mathcal{G}(\mathrm{W}_j). \tag{14}$$

If the positions of estimated camera motion $\hat{\mathrm{M}}_i$, $\hat{\mathrm{M}}_j$ are exchanged and further reprojected with $\hat{\mathrm{S}}_i$, $\hat{\mathrm{S}}_j$, two new observations $\mathrm{W}_i^{'}$, $\mathrm{W}_j^{'}$ can be obtained, that is:

$$\mathrm{W}_i^{'} = \hat{\mathrm{M}}_j \hat{\mathrm{S}}_i, \quad \mathrm{W}_j^{'} = \hat{\mathrm{M}}_i \hat{\mathrm{S}}_j. \tag{15}$$

The proposed Pairwise Consistency enforces $\mathcal{F}$ and $\mathcal{G}$ to estimate $\mathrm{W}_i^{'}$, $\mathrm{W}_j^{'}$ consistently back to $[\hat{\mathrm{M}}_i, \hat{\mathrm{S}}_j]$.

This idea can be easily extended from two frames to a mini-batch of L frames. Given the output $\{\hat{\mathrm{M}}_i, \hat{\mathrm{S}}_i\}_L$ of $\mathcal{G}, \mathcal{F}$, we produce a new batch of 2D observations $\{\mathrm{W}_i^{'}\}_L$ by performing the following reprojections:

$$\mathrm{W}_i^{'} = \hat{\mathrm{M}}_{r_i} \hat{\mathrm{S}}_i, \tag{16}$$

where $r_1 \ldots r_L$ is a random permutation of $1 \ldots L$. In order to enforce the Pairwise Consistency, the Pairwise Consistency Loss $\mathcal{L}_{\text{consist}}$ is applied to the training process of the model, and it is calculated as:

$$\mathcal{L}_{\text{consist}} = \sum_{i=1}^{L} \left\| \hat{\mathrm{S}}_i - \hat{\mathrm{S}}_i^{'} \right\|_F + \left\| \hat{\mathrm{M}}_i - \hat{\mathrm{M}}_i^{'} \right\|_F, \tag{17}$$

where $\hat{\mathrm{S}}_i^{'} = \mathcal{F}(\mathrm{W}_i^{'})$, $\hat{\mathrm{M}}_{r_i}^{'} = \mathcal{G}(\mathrm{W}_i^{'})$, and $\{\hat{\mathrm{M}}_{r_i}^{'}\}$ is rearranged to the original order $\{\hat{\mathrm{M}}_i^{'}\}$ using the inverse permutation of $r_1 \cdots r_L$. We notice that there are better measurements of rotation distance than the second term in (17), but we empirically find that the Frobenius norm is also feasible and simple to implement. In addition, we find that replacing $\hat{\mathrm{M}}_i$ with a random rotation matrix is a good alternative which can slightly improve the performance.

Methods	Subj. 07	Subj. 20	Subj. 23	Subj. 33	Subj. 34	Subj. 38	Subj. 39	Subj. 43	Subj. 93
CSF [15]	1.231	1.164	1.238	1.156	1.165	1.188	1.172	1.267	1.117
URN [7]	1.504	1.770	1.329	1.205	1.305	1.303	1.550	1.434	1.601
CNS [6]	0.310	0.217	0.184	0.177	0.249	0.223	0.312	0.266	0.245
C3DPO [30]	0.226	0.235	0.342	0.357	0.354	0.391	0.189	0.351	0.246
DNRSFM [20]	0.045	0.137	0.053	0.137	0.062	0.053	0.041	0.125	0.214
PR-RRN (Ours)	**0.024**	**0.034**	**0.039**	**0.043**	**0.039**	**0.034**	**0.025**	**0.028**	**0.152**
PR-RRN (Unseen)	0.061	0.167	0.249	0.254	0.265	0.108	0.028	0.080	0.242

Table 1. The reconstruction error $\mathrm{e}_{3\mathrm{D}}$ on CMU MOCAP.

4.4. Alternative Training

The final training objective of PR-RRN is:

$$\mathcal{L} = \mathcal{L}_{\text{reproj}} + \lambda_1 \mathcal{L}_{\text{contrast}} + \lambda_2 \mathcal{L}_{\text{consist}}, \tag{18}$$

where λ_1 and λ_2 are weighting parameters. We empirically find that training the networks using $\mathcal{L}_{\text{contrast}}$ and $\mathcal{L}_{\text{consist}}$ alternatively produces better results than using them jointly, while $\mathcal{L}_{\text{reproj}}$ is always used.

5. Experiments

We evaluate our method on a large-scale human motion dataset, a categorical objects dataset, a facial landmark dataset and a mesh dataset, which are representative deformable shapes. We first introduce the datasets and the experimental setups. Next the reconstruction results are reported. Finally, we analyze the proposed model in detail.

5.1. Datasets and Setups

CMU MOCAP. The CMU Motion Capture dataset[1] consists of 144 subjects, and most subjects contain tens of human activity sequences. In each activity, ground truth 3D coordinates of 31 keypoints are recorded in the world coordinate system. CMU MOCAP is diverse and large enough for verifying the PR-RRN. We select 9 subjects from CMU MOCAP. For fair comparison with previous methods, we build the training and testing set following [20]: the first 80% the activity sequences in a subject are concatenated as the training set and the remaining 20% are used as testing set (Unseen). Random orthogonal projections are applied to 3D shapes to obtain 2D observations. The coordinates of the 3D shapes are centered to zero in each frame to cancel the transition term in camera projection. Note that in training deep networks, the data will be shuffled every epoch, so that the input frames are *orderless*.

PASCAL3D+. PASCAL3D+ datasets [45] contains 12 categories of objects with 3D annotations from around 80 CAD models. Each category contains about 3000 objects on average. For fair comparison with previous works, we follow [20] to use the categories with at least 8 points of annotation, and do not split the dataset into training and testing set. The ground truth 3D shapes and 2D observations are also zero-centered.

MUCT Face. The MUCT Face dataset [29] consists of 3755 faces with 76 facial landmarks annotations. The dataset is diverse in lighting, age and races. The face images are collected with five cameras from different viewpoints. In our experiments, we use all the 76 keypoints. As there is no 3D ground-truth of the points, we use the MUCT for qualitative evaluation.

TWO CLOTHS. The TWO CLOTHS dataset [38] is a popular dataset for mesh reconstruction, which contains 163 frames of two fast deforming cloths. The dataset provides the 2D trajectory of a 525-point grid mesh.

Handling Missing Points. MUCT contains some missing points caused by occlusion. In the experiment on MUCT dataset, the input coordinates of missing 2D points are simply set to zero, and the normal points are subtracted by their mean to become zero-centered. In training, the msr are calculated only with common visible points of two observations, and the losses are masked with the visibility.

Evaluation Metrics. Following previous works [6, 7], we use the normalized mean 3D error to evaluate the shape recovery accuracy. Before evaluation, the predicted 3D shape is aligned to the ground truth using Procrustes algorithm. The metric is calculated as:

$$e_{3\mathrm{D}} = \frac{1}{F} \sum_{i=1}^{F} \frac{\left\| \mathrm{S}_i^{\mathrm{gt}} - \hat{\mathrm{S}}_i \right\|_2}{\left\| \mathrm{S}_i^{\mathrm{gt}} \right\|_2}, \tag{19}$$

where $\mathrm{S}_i^{\mathrm{gt}}$ is the ground-truth 3D shape of i-th frame.

Training Details. In all experiments, we set $\lambda_1 = 0.1$, $\lambda_2 = 0.2$, $\tau = 0.02$, $\xi = 0.04$, $N_{mem} = 1024$. For the RRN, the number of RR modules n is set to 5, the channels of the modules are $128, 64, 32, 16, 8$, so that the dimension of shape representation h_i is 8. In the rotation estimation network, the sizes of linear layers are $128, 32, 8$. The recursive time T is set to 3 for CMU MOCAP dataset and 2 for PASCAL3D+. The network is trained with Adam optimizer with a learning rate of 0.001 and an exponential decay rate

[1] http://mocap.cs.cmu.edu/

	CSF	KSTA	BMM	CNS	NLO	RIKS	SPS	SFC	MUS	URN	C3D	DNR	Ours
Aeroplane	0.363	0.175	1.459	0.416	0.876	0.132	0.930	0.504	0.261	0.121	0.272	**0.024**	0.031
Bicycle	0.424	0.245	1.376	0.356	0.269	0.136	1.322	0.372	0.178	0.328	0.585	**0.003**	0.005
Bus	0.217	0.199	1.023	0.250	0.140	0.160	0.604	0.251	0.113	0.097	0.271	**0.004**	0.008
Car	0.195	0.186	1.278	0.258	0.104	0.097	0.872	0.282	0.078	0.104	0.276	0.009	**0.005**
Chair	0.398	0.399	1.297	0.170	0.146	0.192	1.046	0.226	0.210	0.115	0.658	**0.007**	0.025
Diningtable	0.406	0.267	1.000	0.170	0.109	0.207	1.050	0.221	0.264	0.115	0.441	0.060	**0.015**
Motorbike	0.278	0.255	0.857	0.457	0.432	0.118	0.986	0.361	0.222	0.287	0.492	**0.002**	0.006
Sofa	0.409	0.307	1.126	0.250	0.149	0.228	1.328	0.302	0.167	0.181	0.343	**0.004**	0.007
Average	0.336	0.223	1.178	0.291	0.278	0.159	1.017	0.315	0.186	0.168	0.417	0.014	**0.013**

Table 2. The reconstruction error e_{3D} on PASCAL3D+. The performances of compared methods are quoted from [2, 8, 20].

of 0.95 for 700 epochs. The pairwise consistency loss and the contrastive loss are used alternatively for 100 epochs.

5.2. NRSfM Results

CMU MOCAP. PR-RRN is compared with several strong methods. As the number of frames is large, classic methods like [16, 11, 17] fail, except CSF [15] and CNS [6]. CSF and CNS assume temporal smooth trajectories of points, so that the sequential frames of CMU MOCAP datasets will put these two methods in advantage. URN [7], C3DPO [30], DNRSFM [20] and ours PR-RRN are deep models which can deal with large-scale reconstruction and do not assume temporal smoothness. Tab. 1 shows the results on the 9 subjects of CMU MOCAP. For PR-RRN, the results of unseen shapes (test set) are reported. One can see that PR-RRN outperforms all four competing methods in 9 subjects. On Subject 20, 33 and 43, PR-RRN surpasses the state-of-the-art approaches by a large margin. It is also worth noting that PR-RRN achieves high accuracy when tested with Unseen shapes on Subject 07, 38, 39 and 43. This may be from a small domain gap between the training set and the testing set. In short, the results validate the capability of PR-RRN for accurate recovery of non-rigid shapes.

PASCAL3D+. For PASCAL3D+ we consider more methods for comparison, including CSF [15], KSTA [16], BMM [11], CNS [6], NLO [5], RIKS [17], SPS [19], SFC [21], MUS [2], URN [8], C3DPO [30] and DNRSFM [20]. The results are shown in Tab. 2. PR-RRN and DNRSFM both achieve higher accuracy on all 8 selected categories of PASCAL3D+ than other methods, while the PR-RRN performs better than DNR on average and especially on `Diningtable` class. The results show that PR-RRN also performs well on categorical objects reconstruction tasks.

MUCT. The reconstruction results on MUCT dataset are visualized in Fig. 6. From the results one can verify that the recovery of non-rigid facial landmarks is successful under realistic camera motions of MUCT dataset.

TWO CLOTHS. We test our method on the TWO CLOTHS dataset to validate mesh reconstruction. As there is no ground-truth, we visualize the qualitative result in Fig. 5, where our model produces plausible deformation and clear segmentation of the two cloths.

5.3. Model Analysis

Structure of RRN. We give an ablation study to validate the residual-recursive design in RRN. We set up a **Vanilla** baseline where the shape network $\mathcal{F}$ contains standard convolution layers with same number of parameters as RRN, and the rotation network $\mathcal{G}$ remains the same as RRN. Note that the only difference between Vanilla and RRN is the residual-recursive structure. We compare the two models together with DNRSFM in Tab. 3. As shown, in Subject 20, 23, 33 and 43, RRN outperforms Vanilla by a margin, which verifies the effectiveness of the structure. The RRN structure does not work well on the difficult Subject 93, however it can be improved by the pairwise regularizations. It is also worth noting that Vanilla model outperforms DNRSFM in Subject 33 and 43 and is competitive in Subject 20 and 93.

Effectiveness of Constrast and Consistent Losses. To understand the effectiveness of proposed Residual-Recursive Networks and two novel losses, we conduct experiments with three variations of the PR-RRN: 1) **RRN.** The Residual-Recursive Networks trained with reprojection loss only. 2) **RRN-Contrast.** The RRN trained with reprojection loss and Pairwise Constrastive Loss. 3) **RRN-Consist.** The RRN trained with reprojection loss and Pairwise Consistency Loss. The reconstruction results are reported in Tab. 4, together with the full model PR-RRN for comparison. From the table, one can see that the proposed RRN achieves high accuracy on CMU MOCAP Subject 20, 23, 33 and 43, and it is further improved by Contrast Loss and Consistency Loss. For Subject 93, the performance of RRN is significantly enhanced by the regularization losses.

Limitations. In our experiments, our method can address points from 8 to more than 500. However, the SVD in Contrastive Loss becomes a bottleneck for the entire model when handling a large scale of points, *e.g.* 5000 points. When a shape contains this amount of points, training PR-RRN will become computationally prohibitive.

Robustness. We analyze the robustness of PR-RRN un-

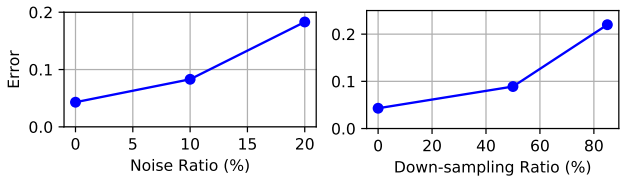

Figure 4. Performance with noisy or down-sampled data.

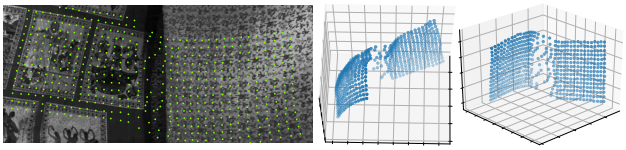

Figure 5. Results on TWO CLOTHS Dataset.

der noisy and small-scale data. (1) We add Gaussian noise to Subject 33 of CMU MOCAP dataset. We follow [20] to calculate the noise ratio: $\|\text{noise}\|_F \,/\, \|\mathbf{W}\|_F$. (2) We train our model on down-sampled Subject 33 and test it on full dataset. In Fig. 4, one can see that the proposed method is capable to achieve reasonable accuracy on corrupted data.

Model	20	23	33	43	93
DNRSFM [20]	0.137	0.053	0.137	0.125	0.214
Vanilla	0.147	0.352	0.060	0.072	0.213
RRN	0.041	0.050	0.051	0.047	0.305

Table 3. Analysis on RRN structure.

Model	20	23	33	43	93
RRN	0.041	0.050	0.051	0.047	0.305
RRN-Contrast	0.039	0.043	0.046	0.033	0.255
RRN-Consist	0.038	0.045	0.044	0.034	0.160
PR-RRN (full)	0.034	0.039	0.043	0.028	0.152

Table 4. Analysis on pairwise regularizations.

6. Conclusion

We present PR-RRN, a novel deep-networks based approach to NRSfM. We introduce a novel Residual-Recursive Network, which can estimate the 3D shape and camera rotation from 2D inputs. We propose a rigidity-based pairwise contrastive loss and a pairwise consistency loss for regularizing the shape representation learning without assuming global distribution or manifold. Experiments on CMU MOCAP and PASCAL3D+ datasets show that the proposed method achieves state-of-the-art shape recovery accuracy for large-scale human motion and categorical objects reconstruction. PR-RRN is also capable to reconstruct facial landmarks and meshes.

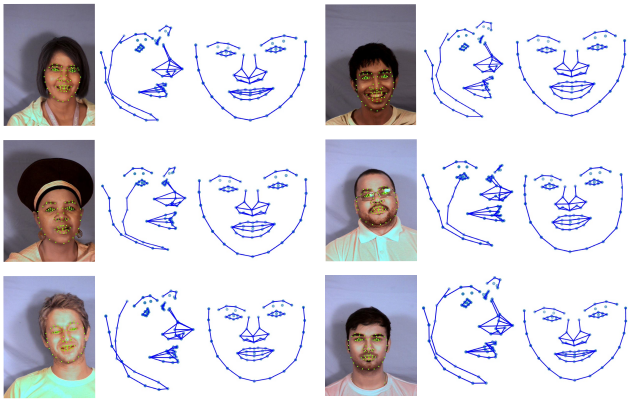

Figure 6. Visualization of some reconstruction results on MUCT dataset [29]. Left: Origin pictures of different people. Center: Side views of the reconstructed shapes. Right: Front views of reconstructions.

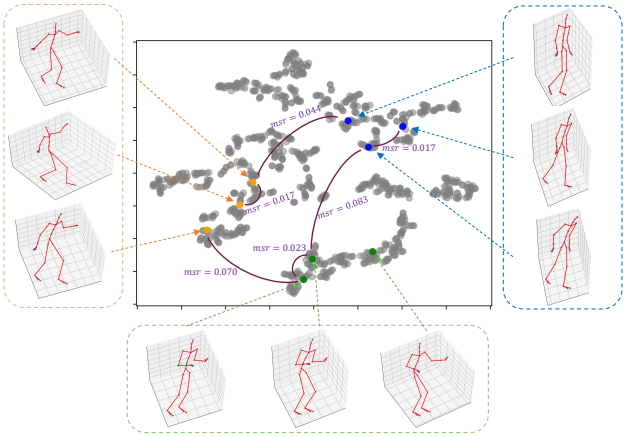

Figure 7. t-SNE [43] visualization of the shape representation learned by PR-RRN on CMU MOCAP Subject 20. The grey points are 1000 randomly selected frames out of a total of 4183 frames. We show 9 reconstructed shapes which can be coarsely divided into three groups. One can see that shape representations are spatially closer to the shapes in the same group than shapes in other groups. Additionally, we mark out some pairwise rigidity measure msr, colored in purple. Qualitatively, the msr correctly reflects the similarity of different 3D shapes, and generally agrees with the distance of representation. Best viewed in color.

Acknowledgments

This work was done when Haitian Zeng interned at Baidu Research. Yuchao Dai was supported in part by National Natural Science Foundation of China (61871325) and National Key Research and Development Program of China (2018AAA0102803). We would like to thank the anonymous reviewers and the area chairs for their useful feedback.

References

[1] Antonio Agudo. Unsupervised 3d reconstruction and grouping of rigid and non-rigid categories. *IEEE TPAMI*, 2020.

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

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

[4] Christoph Bregler, Aaron Hertzmann, and Henning Biermann. Recovering non-rigid 3d shape from image streams. In *CVPR*, pages 2690–2696, 2000.

[5] Alessio Del Bue, Fabrizio Smeraldi, and Lourdes Agapito. Non-rigid structure from motion using ranklet-based tracking and non-linear optimization. *Image Vis. Comput.*, 25(3):297–310, 2007.

[6] Geonho Cha, Minsik Lee, and Songhwai Oh. Reconstruct as far as you can: Consensus of non-rigid reconstruction from feasible regions. *IEEE TPAMI*, pages 1–1, 2019.

[7] Geonho Cha, Minsik Lee, and Songhwai Oh. Unsupervised 3d reconstruction networks. In *ICCV*, pages 3848–3857, 2019.

[8] Ching-Hang Chen, Ambrish Tyagi, Amit Agrawal, Dylan Drover, Rohith MV, Stefan Stojanov, and James M. Rehg. Unsupervised 3d pose estimation with geometric self-supervision. In *CVPR*, pages 5707–5717, 2019.

[9] Ajad Chhatkuli, Daniel Pizarro, Toby Collins, and Adrien Bartoli. Inextensible non-rigid shape-from-motion by second-order cone programming. In *CVPR*, pages 1719–1727, 2016.

[10] Ryan Dahl, Mohammad Norouzi, and Jonathon Shlens. Pixel recursive super resolution. In *ICCV*, pages 5449–5458, 2017.

[11] Yuchao Dai, Hongdong Li, and Mingyi He. A simple prior-free method for non-rigid structure-from-motion factorization. In *CVPR*, 2012.

[12] Joao Fayad, Lourdes Agapito, and Alessio Del Bue. Piecewise quadratic reconstruction of non-rigid surfaces from monocular sequences. In *ECCV*, pages 297–310, 2010.

[13] Joao Fayad, Alessio Del Bue, Lourdes Agapito, and Pedro Aguiar. Non-rigid structure from motion using quadratic deformation models. In *BMVC*, pages 1–11, 2009.

[14] Katerina Fragkiadaki, Marta Salas, Pablo Andrés Arbeláez, and Jitendra Malik. Grouping-based low-rank trajectory completion and 3d reconstruction. In *NeurIPS*, pages 55–63, 2014.

[15] Paulo F. U. Gotardo and Aleix M. Martinez. Computing smooth time trajectories for camera and deformable shape in structure from motion with occlusion. *IEEE TPAMI*, 33(10):2051–2065, 2011.

[16] Paulo F. U. Gotardo and Aleix M. Martinez. Kernel non-rigid structure from motion. In *ICCV*, pages 802–809, 2011.

[17] Onur C. Hamsici, Paulo F. U. Gotardo, and Aleix M. Martinez. Learning spatially-smooth mappings in non-rigid structure from motion. In *ECCV*, pages 260–273, 2012.

[18] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, pages 9726–9735, 2020.

[19] Chen Kong and Simon Lucey. Prior-less compressible structure from motion. In *CVPR*, pages 4123–4131, 2016.

[20] Chen Kong and Simon Lucey. Deep non-rigid structure from motion with missing data. *IEEE TPAMI*, pages 1–1, 2020.

[21] Chen Kong, Rui Zhu, Hamed Kiani, and Simon Lucey. Structure from category: A generic and prior-less approach. In *3DV*, pages 296–304, 2016.

[22] Suryansh Kumar. Jumping manifolds: Geometry aware dense non-rigid structure from motion. In *CVPR*, pages 5346–5355, 2019.

[23] Suryansh Kumar, Anoop Cherian, Yuchao Dai, and Hongdong Li. Scalable dense non-rigid structure-from-motion: A grassmannian perspective. In *CVPR*, pages 254–263, 2018.

[24] Suryansh Kumar, Yuchao Dai, and Hongdong Li. Spatio-temporal union of subspaces for multi-body non-rigid structure-from-motion. *Pattern Recognition*, 71:428–443, 2017.

[25] Minsik Lee, Jungchan Cho, Chong-Ho Choi, and Songhwai Oh. Procrustean normal distribution for non-rigid structure from motion. *IEEE TPAMI*, 39(7):1388–1400, 2017.

[26] Jake Levinson, Carlos Esteves, Kefan Chen, Noah Snavely, Angjoo Kanazawa, Afshin Rostamizadeh, and Ameesh Makadia. An analysis of SVD for deep rotation estimation. In *NeurIPS*, 2020.

[27] Xiu Li, Hongdong Li, Hanbyul Joo, Yebin Liu, and Yaser Sheikh. Structure from recurrent motion: From rigidity to recurrency. In *CVPR*, pages 3032–3040, 2018.

[28] Guangcan Liu, Zhouchen Lin, and Yong Yu. Robust subspace segmentation by low-rank representation. In *ICML*, pages 663–670, 2010.

[29] Stephen Milborrow, John Morkel, and Fred Nicolls. The MUCT Landmarked Face Database. *Pattern Recognition Association of South Africa*, 2010.

[30] David Novotny, Nikhila Ravi, Ben Graham, Natalia Neverova, and Andrea Vedaldi. C3dpo: Canonical 3d pose networks for non-rigid structure from motion. In *ICCV*, 2019.

[31] Marco Paladini, Adrien Bartoli, and Lourdes Agapito. Sequential non-rigid structure-from-motion with the 3d-implicit low-rank shape model. In *ECCV*, 2010.

[32] Marco Paladini, Alessio Del Bue, Marko Stosic, Marija Dodig, Joao Xavier, and Lourdes Agapito. Factorization for non-rigid and articulated structure using metric projections. In *CVPR*, 2009.

[33] Hyun Soo Park and Yaser Sheikh. 3d reconstruction of a smooth articulated trajectory from a monocular image sequence. In *ICCV*, pages 201–208, 2011.

[34] Sungheon Park, Minsik Lee, and Nojun Kwak. Procrustean regression networks: Learning 3d structure of non-rigid objects from 2d annotations. In *ECCV*, pages 1–18, 2020.

[35] Aditya Sanghi. Info3d: Representation learning on 3d objects using mutual information maximization and contrastive learning. In *ECCV*, pages 626–642, 2020.

[36] Yujiao Shi, Hongdong Li, and Xin Yu. Self-supervised visibility learning for novel view synthesis. In *CVPR*, pages 9675–9684, 2021.

[37] Vikramjit Sidhu, Edgar Tretschk, Vladislav Golyanik, Antonio Agudo, and Christian Theobalt. Neural dense non-rigid structure from motion with latent space constraints. In *ECCV*, pages 204–222, 2020.

[38] Jonathan Taylor, Allan D. Jepson, and Kiriakos N. Kutulakos. Non-rigid structure from locally-rigid motion. *CVPR*, pages 2761–2768, 2010.

[39] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *ECCV*, pages 776–794, 2020.

[40] Carlo Tomasi and Takeo Kanade. Shape and motion from image streams under orthography: a factorization method. *IJCV*, 9(2):137–154, 1992.

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

[42] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *CoRR*, abs/1807.03748, 2018.

[43] Laurens van der Maaten and Geoffrey Hinton. Visualizing high-dimensional data using t-sne. *JMLR*, 9:2579–2605, 2008.

[44] Sara Vicente and Lourdes Agapito. Soft inextensibility constraints for template-free non-rigid reconstruction. In *ECCV*, pages 426–440, 2012.

[45] Yu Xiang, Roozbeh Mottaghi, and Silvio Savarese. Beyond PASCAL: A benchmark for 3d object detection in the wild. In *WACV*, pages 75–82, 2014.

[46] Jing Xiao, Jinxiang Chai, and Takeo Kanade. A closed-form solution to non-rigid shape and motion recovery. *IJCV*, 67(2):233–246, 2006.

[47] Jing Xiao and Takeo Kanade. Non-rigid shape and motion recovery: Degenerate deformations. In *CVPR*, pages 668–675, 2004.

[48] Xin Yu, Yurun Tian, Fatih Porikli, Richard Hartley, Hongdong Li, Huub Heijnen, and Vassileios Balntas. Unsupervised extraction of local image descriptors via relative distance ranking loss. In *CVPRW*, 2019.

[49] Yang Zhang, Ivor W. Tsang, Yawei Luo, Changhui Hu, Xiaobo Lu, and Xin Yu. Recursive copy and paste gan: Face hallucination from shaded thumbnails. *IEEE TPAMI*, 2021.

[50] Yupei Zheng, Xin Yu, Miaomiao Liu, and Shunli Zhang. Single-image deraining via recurrent residual multiscale networks. *IEEE TNNLS*, 2020.

[51] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *CVPR*, pages 5745–5753, 2019.

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