Non-rigid Structure-from-Motion: Temporally-smooth Procrustean Alignment and Spatially-variant Deformation Modeling

Jiawei Shi, Hui Deng, Yuchao Dai *
School of Electronics and Information, Northwestern Polytechnical University
Shaanxi Key Laboratory of Information Acquisition and Processing
{sjw2018, denghui986}@mail.nwpu.edu.cn, daiyuchao@nwpu.edu.cn

Abstract

Even though Non-rigid Structure-from-Motion (NRSfM) has been extensively studied and great progress has been made, there are still key challenges that hinder their broad real-world applications: 1) the inherent motion/rotation ambiguity requires either explicit camera motion recovery with extra constraint or complex Procrustean Alignment; 2) existing low-rank modeling of the global shape can over-penalize drastic deformations in the 3D shape sequence. This paper proposes to resolve the above issues from a spatial-temporal modeling *perspective. First, we propose a novel Temporally-smooth Procrustean Alignment module that estimates 3D deforming shapes and adjusts the camera motion by aligning the 3D shape sequence consecutively. Our new alignment module remedies the requirement of complex reference 3D shape during alignment, which is more conductive to non-isotropic deformation modeling. Second, we propose a spatial-weighted approach to enforce the* low-rank constraint adaptively *at different locations to accommodate drastic spatially-variant deformation reconstruction better. Our modeling outperform existing low-rank based methods, and extensive experiments across different datasets validate the effectiveness of our method*[1].

1. Introduction

Non-rigid structure-from-motion (NRSfM) targets at jointly estimating the deforming 3D shapes and the camera motion from the 2D observation sequence, which is a classical and long-lasting geometric vision problem [8]. In contrast to the recently booming deep learning based solutions [14, 24, 25, 34, 42, 48], the optimization based solutions do not rely on training data and can be directly applied to arbitrary classes such as human, animal, and deforming surface. Even though considerable progress has been made [11] and a series of milestones have been achieved [1, 27, 30, 49], the performance of NRSfM is still far from satisfactory. The challenges mainly lie in 1) the inherent motion/rotation ambiguity requires either explicit camera motion recovery with extra constraint or complex Procrustean Alignment; 2) existing global modeling (*i.e.*, low-rank [11, 26, 27], union-of-subspace [28, 49]) of the deforming 3D shape cannot handle drastic spatially-variant 3D deformations.

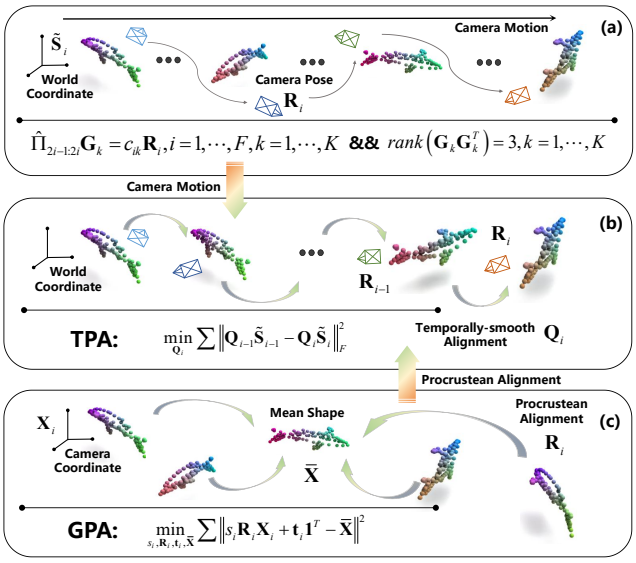


Figure 1. Overview of our proposed TPA Module. (a) Camera motion can be recovered by orthographic and rank-3 constraints within the matrix factorization framework[11]. (c) The Procrustean alignment framework uses GPA to resolve the rotation ambiguity. (b) We propose the TPA module, which aligns the 3D shapes of consecutive frames and corrects the camera motion by the temporal smoothing property.

Existing NRSfM statistical prior frameworks can be roughly classified into two categories: 1) explicit motion estimation-based approaches [2, 8, 11, 15, 26, 27] and 2) motion estimation-free approaches [30, 31, 41]. The former usually employs the matrix factorization framework

*Yuchao Dai is the corresponding author.
[1]Project page: https://npucvr.github.io/TSM-NRSfM.

and orthographic constraint to solve for camera motion and 3D shapes separately. Dai *et al*. [11] introduced the low-rank constraint into the recovery of camera motion. Kumar *et al*. dig into the correction matrix to improve the motion estimation results by using the smoothing prior of camera motions [26] or the fusion of different camera motion sequences [27]. However, these camera motion estimation methods have difficulties in tackling the inherent rotation/motion ambiguity problem in NRSfM [47].

The latter solves the shapes directly and remedies the requirement of motion estimation. Lee *et al*. [30, 31, 41] argued that non-rigid shapes can be recovered unambiguously under the coordinate with almost no rigid relative motion. This type of framework resolves the rotation ambiguity through General Procrustean Analysis (GPA) [18] to achieve alignment of the 3D shapes. However, the high dependence on the alignment reference shape makes the model extremely complex [30, 41] and tends to over-penalize non-isotropic deformations.

Explicit deforming 3D sequence modeling is another difficulty in NRSfM. The seminal work by Dai *et al*. [11] recovers the 3D shape sequence using only the low-rank constraint. Although the low-rank constraint is valid in most cases, it is insufficient to model the objects with severe spatially-variant deformation [27]. Kumar *et al*. [26, 27] improve the estimation of low-rank structures but still use global low-rank modeling. Some methods [19, 48] tried to introduce additional combinations of linear basis to model drastic deformation, but the lack of effective constraints makes the recovery of deformation basis difficult.

In this paper, we propose to resolve the above issues in camera motion and deforming shape recovery from a *spatial-temporal modeling perspective*. The object of our study is the *2D observation sequences*, and this is a common setup for traditional NRSfM [2, 15, 17, 31, 41]. Existing methods do not utilize sequence information well, which can help us deal with the above problems effectively. First, by digging into the temporal smoothing property of non-rigid sequences, we propose a novel *Temporally-smooth Procrustean Alignment (TPA)* module that corrects camera motion through 3D shape alignment. The TPA module can reduce the effects of rotation ambiguity as GPA by exploiting the temporal similarity between consecutive 3D shapes, and it also eliminates the need for additional modeling of reference shapes as the GPA-based methods [30, 41](*cf*. Fig. 1). Second, to tackle the drastic spatially-variant deformations in real-world 3D shape sequences, we propose to use the spatial information obtained by analyzing the trajectory space to segment the regions with different levels of non-rigid deformation. Afterward, we propose a *Spatial-Weighted Nuclear Norm (SWNN)* optimization to improve the adaptation performance of low-rank constraint for severe spatially-variant deformation.

Our main contributions are summarized as:

- We propose a Temporally-smooth Procrustean alignment (TPA) module to estimate 3D deforming shapes and adjust camera motion by aligning the 3D shape sequence consecutively. The TPA module is more conductive to non-isotropic deformation modeling by remedying the requirement of mean shape as alignment reference.
- We contribute a spatial-weighted approach to enforce the low-rank constraint adaptively at different locations, which can better accommodate severe spatially-variant deformation than global low-rank modeling.
- We develop a unified spatial-temporal modeling framework for NRSfM. Extensive experiments across datasets demonstrate the superiority of our approach. Ablation studies show the effectiveness of each algorithm module.

2. Related Work

2.1. Matrix Factorization based NRSfM

Since Bregler *et al*. [8] first applied the factorization framework in non-rigid reconstruction, researchers begun to shift their focus from pre-learned modeling methods [4, 5] to the optimization methods, such as using Metric Projection [36, 37], deformable surface modeling [43, 46], and some priors, *e.g*., motion states [20], shape priors [12, 13], DCT trajectory basis [2, 17], *etc*. However, Xiao *et al*. [47] pointed out that, unlike the rigid version of factorization framework [44], there is an inherent ambiguity in the solution process using only the orthographic constraint, which would lead to non-unique shape basis and corresponding coefficients. But Akhter *et al*. [3] later proved that inherent ambiguity does not affect obtaining a unique result. Based on this, Dai *et al*. [11] proposed a new pipeline that employs low-rank constraint on camera motion and shape estimation, combined with the solution technique of Brand [7], this method successfully achieves an unambiguous shape estimation using the factorization framework without introducing an overly strong prior.

Subsequent researchers have derived a number of different solution methods based on the application of low-rank constraints in the factorization framework. Kumar *et al*. [26, 27] noted that the reconstruction accuracy of 3D shapes can be improved by preserving the main algebraic features of the matrix in the nuclear norm optimization process. In addition, [1, 28, 49] model 3D shapes using union of subspace constraint, [10, 32] model spatial structures using consensus assumptions, and some work represent non-rigid 3D surfaces using manifolds[29, 39, 40]. The factorization framework has been shown to be effective in NRSfM, but a number of issues remain. The accuracy of the unique shape obtained under the existing pipeline is compromised by the inherent ambiguity, and global low-rank modeling does not well tackle the reconstruction of severely deformed objects. And this paper intends to fill those gaps.

2.2. Procrustean Alignment for NRSfM

Parallel to the development of the factorization framework, the alignment framework chose to reduce the impact of the inherent ambiguity on reconstruction by avoiding estimating motion. Torresani *et al*. [45] used Gaussian distribution as a prior to represent non-rigid shapes and proposed an Expectation Maximization (EM) solving framework. To separate rigid motion and non-rigid deformation, Lee *et al*. [30] introduced the General Procrustean Analysis (GPA) algorithm to construct a special Procrustean Normalized Distribution (PND) that effectively represents the data properties of non-rigid shapes. Lee *et al*. [31] added the hidden Markov process to PND, strengthening the temporal dependence in the shape sequence. Park *et al*. [41] designed a more extensible regression framework based on PND and applied it to deep learning method [42]. The separation operation of the alignment framework has a significant effect on mitigating the rotation ambiguity than most of factorization methods, but on the other hand, the pure alignment framework fails to capitalize on the advantages of the factorization framework as well. Therefore, this paper aims to combine the advantages of both to obtain better results.

3. Method

In this section, we introduce our spatial-temporal modeling solution for NRSfM. We first recap the definition of two classical models in NRSfM. Then we improve the modeling of non-rigid deformations from both temporal and spatial perspectives, *i.e*., temporally-smooth Procrustean alignment and spatial-weighted non-rigid deformation modeling.

3.1. Problem Formulation and Optimization

NRSfM aims to recover 3D deforming shape sequence $\mathbf{S}=[\mathbf{S}_1;\cdots;\mathbf{S}_F]\in\mathbb{R}^{3F\times P}$ in the world coordinate and camera motion $\mathbf{R}=diag(\mathbf{R}_i)\in\mathbb{R}^{2F\times 3F}$ from 2D measurements $\mathbf{W}=[\mathbf{W}_1;\cdots;\mathbf{W}_F]\in\mathbb{R}^{2F\times P}$, *i.e*., $\mathbf{W}=\mathbf{RS}$ [8], F and P denote the number of frames and points. The matrix factorization model [8, 11, 44] assumes the non-rigid 3D shapes can be expressed by the linear combination of K shape basis $\mathbf{B}_j$, *i.e*., $\mathbf{S}_i=\sum_{j=1}^{K}c_j\mathbf{B}_j$, where c_j is the shape basis coefficients. Following this assumption, $\mathbf{R}$ and $\mathbf{S}$ can be estimated by the Singular Value Decomposition (SVD) of $\mathbf{W}$ coupled with orthographic constraint. Since $\forall\ \mathbf{Q}\in SO(3), \mathbf{W}_i=\mathbf{R}_i\mathbf{Q}^T\mathbf{Q}\mathbf{S}_i$ always holds, there is an inherent rotation ambiguity in the absence of additional constraints, which leads to incorrect estimation of camera motion and 3D shapes. [11, 26, 27] have proposed a number of effective prior-free methods to recover camera motion and shape sequence, but these methods can not resolve the rotation ambiguity problem theoretically.

Different from the above methods, another class of methods[30, 41] estimates alignment rotations by Procrustean Alignment to substitute the recovery of camera motion. The General Procrustean Alignment (GPA) model is defined as:

$$\tilde{\mathbf{R}}_i=\arg\min_{\mathbf{R}_i}\left\|\mathbf{R}_i\mathbf{S}_i\mathbf{T}-\bar{\mathbf{S}}\right\|,\mathbf{R}_i\in SO(3), \tag{1}$$

where $\mathbf{T}=\mathbf{I}-\frac{1}{P}\mathbf{1}\mathbf{1}^T$ is the translation elimination matrix, $\bar{\mathbf{S}}$ is the mean shape used as the reference shape during alignment, and $\mathbf{S}_i$ is the 3D shape in camera coordinate. Such methods effectively resolve the rotation ambiguity by separating rigid motion from non-rigid deformation. But for non-isotropic deformations, variations along certain directions may markedly influence the mean shape, leading to an exaggerated penalization of shapes during optimization.

In addition to the reprojection information $\mathbf{W}$ in camera coordinate, both the factorization framework and Procrustean alignment framework are looking for a canonical coordinate (*i.e*., world or alignment coordinate) to enforce extra regularization, such as low-rank. We use $\mathbf{S}\in\mathbb{R}^{3F\times P}$ to denote the sequence of 3D shapes in the camera coordinate, and $\hat{\mathbf{S}}\in\mathbb{R}^{3F\times P}$ is the shape sequence in the canonical coordinate. We rethink these two types of frameworks and provide a new unified framework:

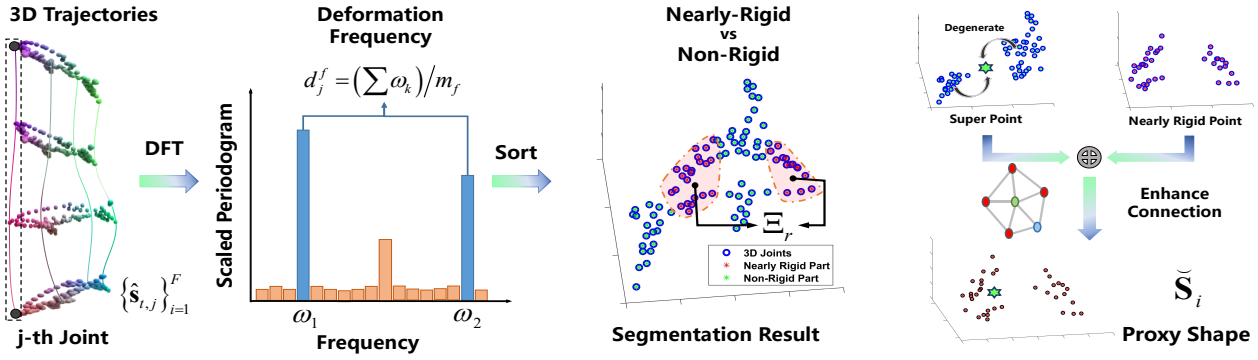
$$\begin{aligned}&\min_{\mathbf{S},\hat{\mathbf{S}},\mathbf{R}}\ \mathcal{F}(\mathbf{S})+\mathcal{G}(\hat{\mathbf{S}})+\mathcal{P}(\mathbf{R},\Phi),\\&\text{s.t. }\hat{\mathbf{S}}=\mathbf{RS},\end{aligned} \tag{2}$$

where $\mathcal{F}(\cdot)$ denotes the data term, such as reprojection error, and $\mathcal{G}(\cdot)$ denotes the regularization term, such as low-rank or smoothing. $\mathcal{P}(\cdot)$ represents the transformation refinement module, $\mathbf{R}$ is the transformation between the camera and canonical coordinates, Φ is the optimization parameters. $\mathbf{R}$ denotes pure 3D rotation transformation, while we have removed scale and translation from measurement matrix $\mathbf{W}$ by regularization and centralization as [11]. When $\mathbf{R}$ is computed in advance and fixed during optimization, the model (2) degenerates to prior-free methods, such as [11, 26, 27]. If the optimization goal $\mathcal{P}(\cdot)$ is set to GPA (1) and Φ is defined as $\{\mathbf{S},\bar{\mathbf{S}}\}$, model (2) degenerates into the framework in [41].

3.2. Temporally-smooth Procrustean Alignment

Under the matrix factorization framework, the 3D reconstruction results will be affected by camera motion estimation as incorrect camera motion estimation may invalidate the underlying low-rank assumption. The Procrustean alignment methods make the aligned sequence satisfy the low-rank constraint by minimizing the rigid motion between 3D shapes. Therefore, such methods can obtain improved results when they begin with good initialization.

In the real world, the nonrigid objects deform continuously and smoothly. Therefore, the low rank as well as



(a) Process for nearly rigid region segmentation.

(b) Construction for proxy Shape.

Figure 2. Overview of region segmentation and proxy shape construction. (a) We use DFT to analyze the 3D trajectories, dividing them by comparing the frequency components contained in each trajectory. The figure shows a segmentation result. (b) The geometric center of the non-rigid region is set as the super point, and it is linearly combined with nearly rigid points to construct the proxy shape.

other deformation regularization constraints should be satisfied in the smooth motion sequence. We propose a camera motion refinement module called **Temporally-smooth Procrustean Alignment (TPA)**, which aligns the shape sequence under the smoothness constraint as follows:

$$\mathcal{L}_{tpa}\left(\mathbf{R},\mathbf{S}\right)=\frac{1}{2}\sum_{i=1}^{F-1}\left\|\mathbf{R}_i\mathbf{S}_i-\mathbf{R}_{i+1}\mathbf{S}_{i+1}\right\|_F^2, \quad (3)$$

where $\mathbf{R}=diag\left(\mathbf{R}_i\right),\mathbf{R}_i\in SO(3)$. The TPA module (3) estimates alignment matrix $\mathbf{R}$ by the principle of minimizing the difference between the 3D shapes of consecutive frames after rotation transformation, as in Fig. 1 (b). Different from GPA (1), TPA does not need to estimate the mean shape of the sequence, which greatly simplifies the optimization and avoids over-penalizing non-isotropic deformations. In addition, we have experimentally verified that sequences aligned by the TPA module can have more similar low-rank and smoothing properties to the real sequences compared with GPA, see Sec. 4.3 and Fig. 4.

Under the NRSfM problem setting, the TPA module can gradually put the shape sequence in a smooth state by adjusting the camera motion. Assume that the estimated camera poses are $\tilde{\mathbf{R}}_i, i=1,\cdots,F$ and define the initialization of the alignment transformation as $\mathbf{R}_{pi}=\tilde{\mathbf{R}}_i^T$. We denote $\mathbf{R}_{pi}\mathbf{S}_i$ as $\tilde{\mathbf{S}}_i$, where $\mathbf{S}_i$ is in the camera coordinate. Then we use the TPA module to estimate the correction rotation $\mathbf{Q}=diag(\mathbf{Q}_i)$ as follows:

$$\begin{aligned}\mathcal{L}_{tpa}^{Q}\left(\mathbf{Q},\tilde{\mathbf{S}}\right)&=\frac{1}{2}\sum_{i=1}^{F-1}\left\|\mathbf{Q}_i\tilde{\mathbf{S}}_i-\mathbf{Q}_{i+1}\tilde{\mathbf{S}}_{i+1}\right\|_F^2\\&=\frac{1}{2}\sum_{i=1}^{F-1}\sum_{j=1}^{P}\left\|\mathbf{Q}_i\tilde{\mathbf{s}}_{i,j}-\mathbf{Q}_{i+1}\tilde{\mathbf{s}}_{i+1,j}\right\|_2^2,\end{aligned} \quad (4)$$

where $\tilde{\mathbf{s}}_{i,j}\in\mathbb{R}^{3\times1}$ is the j-th column of $\tilde{\mathbf{S}}_i$. We take the Lie algebra $\phi_i\in so(3)$ of $\mathbf{Q}_i$ as the optimization variable, and the derivative of the optimization objective $\mathcal{L}_{tpa}^{Q}$ (4) with respect to ϕ_i can be derived as:

$$\frac{\partial\mathcal{L}_{tpa}^{Q}}{\partial\phi_i}=\sum_{j=1}^{P}\left[\left(\mathbf{Q}_i\tilde{\mathbf{s}}_{i,j}\right)^{\wedge}\right]^T\left(r_{i,j}^{(1)}-r_{i,j}^{(0)}\right), \quad (5)$$

where $r_{i,j}^{(m)}=\mathbf{Q}_{i-m}\tilde{\mathbf{s}}_{i-m,j}-\mathbf{Q}_{i-m+1}\tilde{\mathbf{s}}_{i-m+1,j}$, and $(\mathbf{a})^{\wedge}$ represents the skew-symmetric matrix of vector $\mathbf{a}$. Once the gradient is calculated, we can use traditional numerical optimization algorithms to solve model (4), *e.g.*, BFGS, LMF, *etc.* [33], see more derivation details and discussions in the Supplementary Material (Supp.) Sec 1.

TPA largely separates the rigid motion between consecutive frames in the sequence $\tilde{\mathbf{S}}$ and makes it gradually converge to a smooth state. After reducing the interference of rigid motion, we can better utilize the low-rank constraint to recover non-rigid deformation of the aligned sequence $\mathbf{Q}\tilde{\mathbf{S}}$. In the following, we introduce the shape estimation method and unify them afterward.

3.3. Spatial-Weighted Non-rigid Structure Model

3.3.1 Spatial-Weighted Nuclear Norm

In real-world scenarios, low-rank regularization can handle many non-rigid deformation situations but always over-penalize drastic deformations. In the absence of priors to reconstruct the deformation, we can give higher degrees of freedom for shape recovery by relaxing the low-rank constraint. An object whose overall or local deformation is very drastic can lead to its 3D shape sequence $\hat{\mathbf{S}}$ in the world coordinate not satisfying the low-rank constraint. But we can transform its spatial structure by linearly combining the trajectories of $\hat{\mathbf{S}}$ to degenerate the object to a low-rank state. Given the affinity matrix $\mathbf{\Lambda}\in\mathbb{R}^{P\times P}$, we define the sequence of **proxy shape** as $\breve{\mathbf{S}}=\hat{\mathbf{S}}\mathbf{\Lambda}$. We can then utilize low-rank constraint to recover the proxy shapes, which in turn reduces the strength of the constraint in the region of severe deformation to better recover $\hat{\mathbf{S}}$.

We use the nuclear norm $\|\cdot\|_*$ with the rearrangement operator $g(\cdot)$ [11] to build the low-rank constraint, and denote the reshuffled sequence as $\hat{\mathbf{S}}^\sharp = g(\hat{\mathbf{S}}) \in \mathbb{R}^{F\times 3P}$. Different from existing methods, we propose a **Spatial-Weighted Nuclear Norm (SWNN)** optimization model that imposes low-rank constraint on the reshuffle of the proxy shapes $\check{\mathbf{S}}^\sharp = g(\hat{\mathbf{S}}\mathbf{\Lambda})$ instead of $\hat{\mathbf{S}}^\sharp$, so the new low-rank optimization model can be written as:

$$\min_{\hat{\mathbf{S}}} \|\check{\mathbf{S}}^\sharp\|_*. \tag{6}$$

The key to the model (6) is how to design the spatial weight matrix $\mathbf{\Lambda}$. Next, we present a method of constructing proxy shapes $\check{\mathbf{S}}_i, i = 1, \cdots, F$ to relax the low-rank constraint on severe deformation regions. We first introduce an algorithm for segmenting objects according to the level of deformation in different regions. Based on this, a **kernel based** method for the construction of proxy shapes is proposed.

3.3.2 Nearly Rigid Region Segmentation

In reality, not all regions of non-rigid objects undergo the same deformations, but are spatially-variant. For example, a talking mouth on a human face will undergo more severe deformation than the nose. Thus different regions should fit the low-rank regularization differently. In this section, we introduce the partition algorithm for different regions of non-rigid objects, see Fig. 2a for the process.

We define $\mathcal{T}_j = \{\hat{\mathbf{s}}_{t,j} \in \mathbb{R}^{3\times 1}\}_{t=1}^F$ as the j-th column of shape sequence $\hat{\mathbf{S}}$, which represents a 3D trajectory. The x-y-z coordinates of a 3D trajectory can be viewed as time domain signals, and can be transformed into the frequency domain by Discrete Fourier Transform (DFT):

$$d(k/F) = F^{-1/2} \sum_{t=1}^{F} \hat{\mathbf{s}}_{t,j} e^{-2\pi i t k/F}, k = 0, \cdots, F-1, \tag{7}$$

we define $\omega_k = k/F$, which symbolize the different frequency components. Intuitively, when the time domain signal changes drastically, its corresponding frequency domain signal will contain more high-frequency components. We can compare the severity of different time domain signal changes by calculating the intensity of different frequency components of the frequency domain signal. Here we use the **scaled periodogram** to calculate the intensity of the signal in the frequency domain. The formula is as follows:

$$P_g(\omega_k) = \frac{4}{F} |d(\omega_k)|^2, k = 0, \cdots, F-1. \tag{8}$$

Then, we compare $\{P_g(\omega_k)\}_k$ and filter out the frequency components ω_k corresponding to the first $m_f(=2)$ maxima. Their average value $d_j^f = \sum \omega_k / m_f$ is noted as the **deformation frequency** of the trajectory. We sort all the points according to the magnitude of the deformation frequencies $\{d_j^f\}_{j=1}^P$ from low to high, defining the points with low frequency of $\alpha_r \times 100\%$ as **nearly rigid points**, and denote the set of their subscripts as $\mathbb{A}_r$. Here, $\alpha_r \in [0, 1]$ is a hyperparameter that depends on the deformation properties.

In this way, the whole object is divided into a nearly rigid region and a non-rigid region, as the example shown in Fig. 2a. We summarize the operation above as:

$$\mathbb{A}_r = \mathcal{A}_{nrs}\left(\hat{\mathbf{S}}, \alpha_r, m_f\right). \tag{9}$$

3.3.3 Kernel Based Proxy Shape

After completing the segmentation, we compute the proxy shape $\check{\mathbf{S}}_i$ of $\hat{\mathbf{S}}_i$. We model the weight matrix $\mathbf{\Lambda}$ by the kernel function $k(\mathcal{T}_i, \mathcal{T}_j)$, where $\mathcal{T}_j$ is the 3D trajectory of the j-th point. The weight matrix can be expressed as follows:

$$\mathbf{\Lambda} = \begin{bmatrix} \langle\phi(\mathcal{T}_1), \phi(\mathcal{T}_1)\rangle & \cdots & \langle\phi(\mathcal{T}_1), \phi(\mathcal{T}_P)\rangle \\ \vdots & \ddots & \vdots \\ \langle\phi(\mathcal{T}_P), \phi(\mathcal{T}_1)\rangle & \cdots & \langle\phi(\mathcal{T}_P), \phi(\mathcal{T}_P)\rangle \end{bmatrix}, \tag{10}$$

where $\phi(\cdot)$ represents a feature map, $\langle\cdot,\cdot\rangle$ is the vector inner product. We use the inner product to compute the correlation between the vectors after the feature mapping as the combination weights between different points in the space. In order to increase the difference between the low-rank constraint effect on the nearly rigid and non-rigid regions, we set up the feature mapping $\phi(\cdot)$ in the following form:

$$\phi(\mathcal{T}_i) = \begin{cases} \sqrt{1-\delta_r^2}\mathbf{1}_i + \delta_r \mathbf{1}_{P+1}, & i \in \mathbb{A}_r \\ \delta_{nr}\mathbf{1}_{P+1}, & i \notin \mathbb{A}_r \end{cases} \tag{11}$$

where $\mathbf{1}_i$ is a 0-1 vector in $P+1$ dimensions, taking 1 only in the i-th dimension. δ_r is a constant, and δ_{nr} is set to $1/\sqrt{(1-\alpha_r)P}$. Here we provide a brief discussion of model (10). When $\alpha_r = 1$, *i.e.*, all points are in set $\mathbb{A}_r$, $\check{\mathbf{S}}_i$ is just a shrink state of the original shape. If $\alpha_r < 1$, all points in the non-rigid region degenerate into the same **super point** (as shown in Fig. 2b). The super point prevents the low-rank constraint from acting directly on each point in the non-rigid region, which can help reduce the intensity of the low-rank constraint on the non-rigid region. In addition, the proxy shape calculation model (10) ensures that the super point and nearly rigid points remain a unified whole[2].

3.4. Complete Model and Solution

In this section, we integrate the complete model with the unified framework (2) and perform the solution. For the data term $\mathcal{F}$, we use the reprojection constraint as the optimization target. The SWNN model (6) is set to the regularization term $\mathcal{G}$ and constrains the shape sequence under canonical coordinate. The link between the camera coordinate and the canonical coordinate is constructed through the

[2]Refer to more analysis on proxy shape and its role in the Supp. Sec 3.

Algorithm 1 Overview of the Optimization Algorithm

Input: Initialize $\{\mathbf{S}, \tilde{\mathbf{S}}, \hat{\mathbf{S}}, \breve{\mathbf{S}}^{\sharp}\}$ via Pseudo inverse shape in [11], $\mathbf{R}_p$, $\mathbf{Q} = \mathbf{I}$, $\mathbf{\Lambda} = \mathbf{I}$, $\beta = 1e^{-4}$ and $\epsilon = 1e^{-6}$
repeat
  Updating variables $\{\breve{\mathbf{S}}^{\sharp}, \hat{\mathbf{S}}, \tilde{\mathbf{S}}, \mathbf{S}\}$ alternately according to model (13) yields $\{\breve{\mathbf{S}}_{-}^{\sharp}, \hat{\mathbf{S}}_{-}, \tilde{\mathbf{S}}_{-}, \mathbf{S}_{-}\}$
until $\|\mathbf{S}_{-} - \mathbf{S}\|_{\infty} < \epsilon$
Initialization: Calculate $\mathbb{A}_r$ by Eq. (9), build weight matrix $\mathbf{\Lambda}$ by (10), reset $\beta = \beta_d$
repeat
  Updating variables Ω alternately according to model (13) yields $\{\breve{\mathbf{S}}_{-}^{\sharp}, \hat{\mathbf{S}}_{-}, \tilde{\mathbf{S}}_{-}, \mathbf{S}_{-}, \mathbf{Q}_{-}\}$
until $\|\mathbf{S}_{-} - \mathbf{S}\|_{\infty} < \epsilon$
Output: $\mathbf{S}, \mathbf{Q}$

TPA module (4). The complete model is as follows:

$$\min_{\mathbf{S},\mathbf{Q}} \frac{\mu_1}{2}\|\mathbf{W}-\mathbf{\Pi}\mathbf{S}\|_F^2 + \mu_2\left\|\breve{\mathbf{S}}^{\sharp}\right\|_* + \frac{\mu_3}{2}\sum_{i=1}^{F-1}\left\|\mathbf{Q}_i\tilde{\mathbf{S}}_i - \mathbf{Q}_{i+1}\tilde{\mathbf{S}}_{i+1}\right\|_F^2$$
$$s.t.\begin{cases}\breve{\mathbf{S}}^{\sharp} = g(\hat{\mathbf{S}}\mathbf{\Lambda}) \\ \hat{\mathbf{S}}_i = \mathbf{Q}_i\tilde{\mathbf{S}}_i, i = 1, \cdots, F \\ \tilde{\mathbf{S}}_i = \mathbf{R}_{pi}\mathbf{S}_i, i = 1, \cdots, F\end{cases} \tag{12}$$

where $\mathbf{\Pi} = diag(\Pi_i), \Pi_i \in \mathbb{R}^{2\times 3}$ is the camera projection matrix, and we use orthogonal projection as [11, 26]. $\mathbf{S} \in \mathbb{R}^{3F\times P}$ is the 3D shape sequence under the camera coordinate. Then we improve the initialized camera motion $\mathbf{R}_p$ by estimating $\mathbf{Q}_i \in SO(3)$ and compute the 3D shape sequence $\hat{\mathbf{S}} \in \mathbb{R}^{3F\times P}$ in the canonical coordinate. Finally, we introduce the kernel based weight matrix $\mathbf{\Lambda}$ to construct proxy shapes that more satisfy the low-rank constraint.

We optimize the model (12) by ADMM [6], and the Lagrange multiplier of the whole model is as follows:

$$\begin{aligned}\min_{\Omega} \mathcal{L} = &\frac{\mu_1}{2}\|\mathbf{W}-\mathbf{\Pi}\mathbf{S}\|_F^2 + \mu_2\left\|\breve{\mathbf{S}}^{\sharp}\right\|_* + \frac{\mu_3}{2}\sum_{i=1}^{F-1}\left\|\mathbf{Q}_i\tilde{\mathbf{S}}_i - \mathbf{Q}_{i+1}\tilde{\mathbf{S}}_{i+1}\right\|_F^2 \\ &+ \frac{\beta}{2}\left\|\breve{\mathbf{S}}^{\sharp} - g(\hat{\mathbf{S}}\mathbf{\Lambda})\right\|_F^2 + \left\langle \mathbf{Y}_1, \breve{\mathbf{S}}^{\sharp} - g(\hat{\mathbf{S}}\mathbf{\Lambda})\right\rangle \\ &+ \frac{\beta}{2}\left\|\hat{\mathbf{S}} - \mathbf{Q}\tilde{\mathbf{S}}\right\|_F^2 + \left\langle \mathbf{Y}_2, \hat{\mathbf{S}} - \mathbf{Q}\tilde{\mathbf{S}}\right\rangle \\ &+ \frac{\beta}{2}\left\|\tilde{\mathbf{S}} - \mathbf{R}_p\mathbf{S}\right\|_F^2 + \left\langle \mathbf{Y}_3, \tilde{\mathbf{S}} - \mathbf{R}_p\mathbf{S}\right\rangle\end{aligned} \tag{13}$$

where $\Omega = \{\mathbf{S}, \tilde{\mathbf{S}}, \hat{\mathbf{S}}, \breve{\mathbf{S}}^{\sharp}, \mathbf{Q}\}$ denotes the variables to be updated, $\{\mathbf{Y}_n\}_{n=1}^3$ are Lagrange multipliers. For the update of $\breve{\mathbf{S}}^{\sharp}$, we use the rectification algorithm in [26]. The difference is that when calculating the weights Θ, we truncate the singular values according to the shape basis dimension K_s and normalize the weights. In real scenes, the captured images are often obscured and it is difficult to observe all the keypoints in each frame. Our proposed framework can handle the problem of missing points by simply adding visible information to the data term. More details and formulas are provided in the Supp. Sec 2.

4. Experiments

4.1. Implementation Details and Evaluation Metric

Implementation Details. The parameter settings in the ADMM optimization algorithm are the same as in the Organic Priors Method (OPM) [27]. The model (12) is a non-convex optimization that requires the initialization of camera motion and 3D shapes. We use the camera motion estimation algorithm in BMM [11] to initialize $\mathbf{R}_p$. To build the weight matrix $\mathbf{\Lambda}$, a good initialization of the shape sequence $\hat{\mathbf{S}}$ is needed to calculate the segmentation of the non-rigid region. Since our model is a unified framework, there is no need to use other methods, which can be accomplished using only model (12). As shown in Algorithm 1, we first fix the correction rotation $\mathbf{Q}$ and set $\mathbf{\Lambda}$ to the Identity matrix. After convergence, the weight matrix is calculated and all parameters Ω are well initialized. In addition, β_d in Algorithm 1 is generally $1e^{-2}$ or $1e^{0}$, and $\Psi = \{\mu_1, \mu_2, \mu_3, \alpha_r, \delta_r, K_s\}$ is adjusted to the dataset (see Supp. Sec 4 for more settings).

Evaluation Metric. We follow the setup in [27] using the mean normalized 3D reconstruction error metric to evaluate the shape reconstruction results on the motion capture benchmark (MoCap), semi-dense, and H3WB dataset. The metric is defined as $e_{3d} = \frac{1}{F}\sum_{i=1}^{F}\left\|\mathbf{S}_i^{est} - \mathbf{S}_i^{gt}\right\|_F / \left\|\mathbf{S}_i^{gt}\right\|_F$ and $\mathbf{S}_i^{est}$, $\mathbf{S}_i^{gt}$ denote the estimated 3D shape and the corresponding ground-truth (GT) value respectively. We remove the global ambiguity [2, 23] as in [27] before computing the 3D reconstruction error. To evaluate our approach on the NRSfM benchmark dataset [23], we use the officially supplied metric script.

4.2. Datasets and Results

MoCap Benchmark Dataset. This dataset is a standard benchmark for NRSfM consisting of 8 real sequences. Akhter *et al.* [2] introduced five sequences: Drink, Pickup, Yoga, Stretch, and Dance. And the other three, Face, Walking, and Shark, were presented by Torresani *et al.* [45]. Tab. 1 and Fig. 3a demonstrate the reconstruction errors e_{3d} of our method compared to other methods and some visual results, respectively. As shown in Tab. 1, our method performs best or second-best across multiple sequences, indicating that our method is able to accommodate diverse types of deformation. Our method also achieves comparable results in sequences such as Shark and Walking, outperforming the pure low-rank constraint methods [11, 26, 27].

NRSfM Challenge Dataset. Jensen *et al.* [23] recently proposed a new challenging benchmark. This dataset contains five types of non-rigid deformation: Articulated, Balloon, Paper, Stretch, and Tearing. Each subject contains six observation sequences captured by different types of camera motion, *i.e.*, circle, flyby, line, semi-circle, tricky, and zigzag. For each subject, we calculate the reconstruction

Table 1. 3D reconstruction errors on MoCap dataset. Our method shows advantages over many matrix factorization methods and Procrustean alignment methods. The second-best results are underlined, and the shape basis dimension K_s is shown in brackets.

Data	CSF1 [15]	CSF2 [17]	KSTA [16]	PND [30]	PMP [31]	CNS [32]	PR [41]	BMM [11]	R-BMM [26]	OPM [27]	Ours
Drink	0.0223	0.0223	0.0156	0.0037	**0.0018**	0.0431	0.0063	0.0152	0.0119	0.0071	0.0031(13)
Pickup	0.2301	0.2277	0.2322	0.0372	0.0127	0.1281	0.0157	0.0315	0.0198	0.0152	**0.0126**(12)
Yoga	0.1467	0.1464	0.1476	0.0140	0.0128	0.1845	0.0175	0.0225	0.0129	0.0122	**0.0109**(10)
Stretch	0.0710	0.0685	0.0674	0.0156	0.0124	0.0939	0.0156	0.0247	0.0144	0.0124	**0.0114**(12)
Dance	0.2705	0.1983	0.2504	0.1454	0.1278	**0.0759**	0.1266	0.1445	0.1491	0.01209	0.0921(13)
Face	0.0363	0.0314	0.0339	0.0165	0.0166	0.0248	0.0164	0.0206	0.0179	0.0145	**0.0144**(5)
Walking	0.1893	0.1035	0.1029	0.0465	0.0424	**0.0396**	0.0544	0.0908	0.0882	0.0816	0.0710(4)
Shark	**0.0081**	0.0444	0.0160	0.0135	0.0099	0.0832	0.0272	0.2311	0.0551	0.0550	0.0258(6)

Table 2. Reconstruction error comparison with state-of-the-art on NRSfM Challenge dataset. We report the results in millimeters.

Data	CSF2 [17]	BMM [11]	R-BMM [26]	AOW [21]	BP [35]	OPM [27]	Ours
Articul.	11.52	18.49	16.00	15.03	16.10	12.18	**10.69**
Balloon	10.14	10.39	7.84	8.05	8.29	**6.29**	7.28
Paper	9.72	8.94	10.69	10.45	**6.70**	8.86	7.91
Stretch	8.65	10.02	7.53	9.01	7.66	6.36	**5.43**
Tearing	12.04	14.23	16.34	16.20	11.26	10.91	**10.77**

Table 3. Mean normalized 3D reconstruction errors on Semi-dense dataset. ' - ' indicates the estimation failed due to excessive computational overhead.

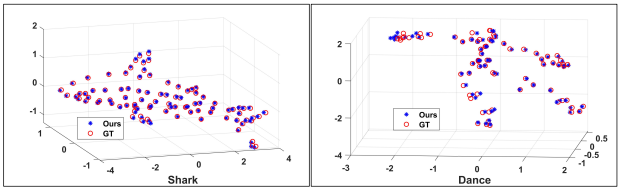
Data	CSF2 [17]	BMM [11]	CNS [32]	R-BMM [26]	Ours	Ours(I)
Kinect	0.0232	0.1212	0.0453	0.0199	0.0356	**0.0161**
Rug	0.0189	0.0109	-	0.0135	**0.0088**	**0.0088**
Mat	0.0620	0.1088	0.0197	0.0285	0.0238	**0.0182**

Table 4. Comparison of mean normalized 3D reconstruction error with well-known sparse NRSfM methods on H3WB dataset.

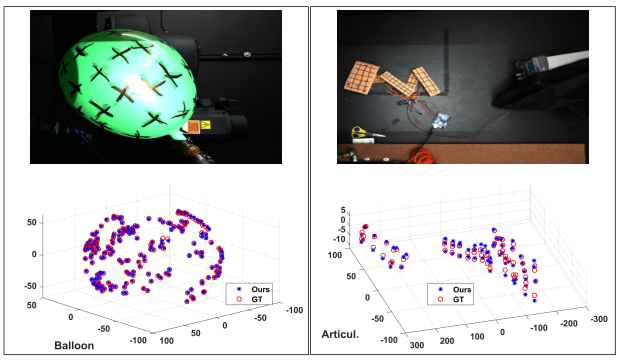
Data	CSF2 [17]	BMM [11]	PND [30]	CNS [32]	R-BMM [26]	Ours
Eating2	0.2662	0.2084	0.2309	0.2250	0.1779	**0.1767**
Smoking1	0.3534	0.2287	0.3397	0.2982	0.2290	**0.2040**
Directions	0.2953	0.2872	0.3002	**0.2674**	0.2964	0.2688
Smoking	0.1980	0.1759	0.4349	0.1968	0.1760	**0.1693**
Waiting2	0.2125	0.1712	0.3306	0.1848	0.1478	**0.1097**
Mean	0.2651	0.2143	0.3272	0.2344	0.2054	**0.1857**

errors under the six camera motions and take the average as the final error for that subject. The quantitative comparison with other methods and the qualitative visualization are shown in Tab. 2 and Fig. 3b. Our method achieves the best results on Articul., Stretch, and Tearing, and also has high reconstruction accuracy on Balloon and Paper. The numerical results further show the advantages of our method on multi-category deformation reconstruction.

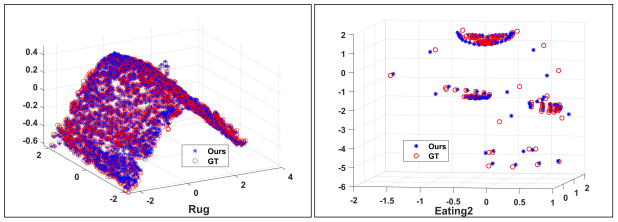
Semi-dense Dataset. We evaluated our method on the kinect paper, rug, table mat datasets [38](191, 159, 60 frames and 1503, 3912, 1500 keypoints respectively). We compared our method with other well-known sparse NRSfM methods (PND [30] fails on this dataset due to excessive computational overhead caused by the large number of points), see the results in Tab. 3 and Fig. 3c. We denote our method that initializes $\mathbf{R}_p$ using the modified algorithm in [26] as Our(I). Our proposed model is also effective in reducing the reconstruction error on semi-dense dataset. Our method can significantly reduce the effect of poor camera motion on shape estimation and also further improve reconstruction accuracy with better motion initialization.



(a) Visualization on MoCap Dataset



(b) Visualization on NRSfM Challenge Dataset



(c) Visualization on Semi-dense and H3WB Dataset

Figure 3. (a) Visualization of Shark and Dance results. (b) Visualization of Balloon and Articulated results. Top row is the image in dataset, and bottom row is the 3D reconstruction shape. (c) Visualization of Rug and Eating2 results.

H3WB Dataset. Human3.6M 3D WholeBody (H3WB) dataset [50] is an entire human body 3D dataset, including face, hands, body, and feet. H3WB is an extension of Human3.6M [22] and contains multiple categories of common life actions such as Directions, Eating, and Smoking. The human body in H3WB is annotated with 133 keypoints, 17 for body, 6 for feet, 68 for face, and 42 for hands. We selected five sequences in H3WB and processed them to obtain a new evaluation dataset: Eating2, Smoking1, Directions, Smoking, Waiting2 (185, 265, 245, 180 and 335 frames, respectively), see Supp. Sec 4 for more details.

Table 5. Comparison of reconstruction errors with and without SWNN module. The table shows the comparison results on the MoCap dataset and the Articul.(circle), Articul.(flyby), Stretch(zigzag) sequences in NRSfM Challenge dataset.

Method	Drink	Pickup	Yoga	Stretch	Dance	Face	Articul.(c)	Articul.(f)	Stretch(z)
Ours(w/)	**0.0031**	**0.0126**	**0.0109**	**0.0114**	**0.0921**	**0.0144**	**2.3442**	**3.8229**	**0.6460**
Ours(w/o)	0.0050	0.0137	0.0120	0.0132	0.0932	0.0161	3.5626	4.8042	0.7341

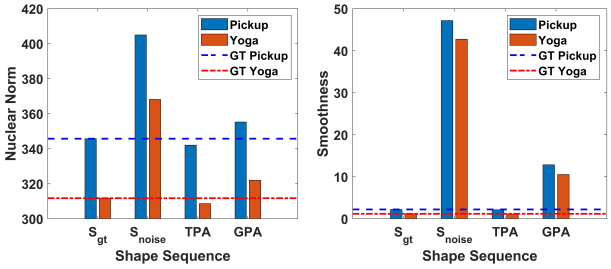


Figure 4. Shape alignment test for the TPA module. We compare the low-rank and smoothing properties of GT non-rigid sequences, randomly rotated disrupted sequences, and sequences aligned by TPA or GPA. The results show that the TPA-aligned sequences have more similar properties to the GT sequences.

As manifested in Tab. 4, our method shows better reconstruction results than other prior-free and procrustean-based methods (Visualization illustrated in Fig. 3c).

4.3. Ablation Study

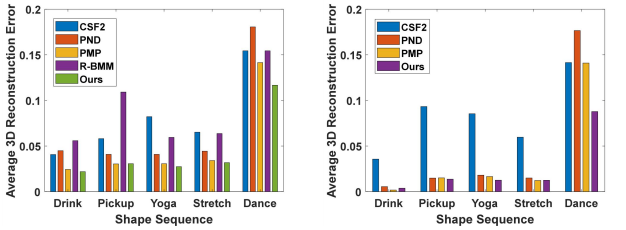
Role of TPA Module. First, we verify the alignment function of TPA module (4) on Pickup and Yoga sequences. We randomly sample Lie algebra $p_n \in \mathbb{R}^{3\times F}$ from Gaussian noise $\mathcal{N}(0, 0.1)$, and then map it to a 3D rotation matrix $\mathbf{R}_d \in \mathbb{R}^{3F\times 3F}$ using Rodrigues' rotation formula. We define the shape sequence disrupted by random rotations as $\mathbf{S}_{noise} = \mathbf{R}_d\mathbf{S}_{gt}$, where $\mathbf{S}_{gt} \in \mathbb{R}^{3F\times P}$ is the GT sequence in the world coordinate. We realigned the sequence with the TPA module and compared the result with GPA [30]. Fig. 4 compares the nuclear norm (*i.e.*, $\|g(\cdot)\|_*$) and the first-order smoothness [11] of $\mathbf{S}_{gt}, \mathbf{S}_{noise}$ and the sequences aligned by the two algorithms, showing that the TPA-aligned sequence has the most similar properties to the GT sequence.

We also test the stability of the TPA module for camera motion initialization. We used the same approach to sample perturbed rotations $\mathbf{R}_d$ with standard deviation σ of 0.1, 0.2, and 0.5. Adding ambiguity $\mathbf{R}_d$ to the camera motion initialized by BMM [11], we then compare the four methods BMM, R-BMM, BMM+Smooth (Add first-order smoothness constraint), BMM+TPA. For each σ, we experimented five times and calculated the mean, the results are shown in Tab. 6. Compared to other methods, the TPA module maintains its effectiveness even when a large deviation has been applied to the camera motion. This shows that our method can effectively deal with the rotation ambiguity in camera motion estimation and can improve the effectiveness and stability of the smoothing constraint.

Effectiveness of SWNN. We compared the two cases of using and not using (with $\mathbf{\Lambda} = \mathbf{I}$) the SWNN module in model (12). The comparison results are reported in Tab. 5. SWNN improves the accuracy of 3D shape recovery in most deformation settings. The accuracy improvement mainly depends on the segmentation result and the smoothing property of the sequence, for more analysis see Supp. Sec 3.

Table 6. Experiment for the effectiveness of the TPA module in correcting camera motion. Adjustment for weight μ_3 in brackets.

SD	BMM [11]	R-BMM [26]	BMM+Smooth	BMM+TPA
$\sigma = 0.1$	0.0493	0.0507	0.0845	**0.0237**($1e^{-1}$)
$\sigma = 0.2$	0.0845	0.0769	0.1149	**0.0592**($1e^{-2}$)
$\sigma = 0.5$	0.1743	0.1681	0.1905	**0.1665**($1e^{-3}$)



(a) e_{3d} with noise (b) e_{3d} with occlusion

Figure 5. (a) Shape errors on noisy sequences. (b) Experiment for 3D reconstruction on missing data.

Performance on Noisy and Missing Data. We follow the approaches in [30] to add noise and occlusion to the data. For missing data, we first complete the measurement matrix by [9] and then run our method. Fig. 5 shows that our method performs better than others in most cases[3].

5. Conclusion

In this paper, we have proposed a spatial-temporal modeling framework for NRSfM. We proposed a Temporally-smooth Procrustean Alignment module to tackle the rotation ambiguity. Furthermore, we introduced a spatial-weighted nuclear norm model with the shape proxy strategy to ensure the effectiveness of low-rank constraint in coping with severe spatially-variant deformations. Our method achieves superior 3D reconstruction performance across a wide range of deformation sequences. Moreover, there is still much research to do in the future on more accurate trajectory segmentation and explicit modeling of complex deformation.

Acknowledgements: This research was supported in part by the National Natural Science Foundation of China (62271410), and the Fundamental Research Funds for the Central Universities.

[3] See Supp. for more tests on TPA, SWNN and Missing Data.

References

[1] Antonio Agudo. Unsupervised 3d reconstruction and grouping of rigid and non-rigid categories. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(1):519–532, 2020. 1, 2

[2] Ijaz Akhter, Yaser Sheikh, Sohaib Khan, and Takeo Kanade. Nonrigid structure from motion in trajectory space. In *Adv. Neural Inform. Process. Syst.*, pages 41–48, 2008. 1, 2, 6

[3] Ijaz Akhter, Yaser Sheikh, and Sohaib Khan. In defense of orthonormality constraints for nonrigid structure from motion. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1534–1541, 2009. 2

[4] Benedicte Bascle and Andrew Blake. Separability of pose and expression in facial tracking and animation. In *Int. Conf. Comput. Vis.*, pages 323–328, 1998. 2

[5] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pages 187–194, 1999. 2

[6] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, Jonathan Eckstein, et al. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning*, 3 (1):1–122, 2011. 6

[7] Matthew Brand. A direct method for 3d factorization of nonrigid motion observed in 2d. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 122–128, 2005. 2

[8] Christoph Bregler, Aaron Hertzmann, and Henning Biermann. Recovering non-rigid 3d shape from image streams. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 690–696, 2000. 1, 2, 3

[9] Ricardo Cabral, Fernando De la Torre, João P Costeira, and Alexandre Bernardino. Unifying nuclear norm and bilinear factorization approaches for low-rank matrix decomposition. In *Int. Conf. Comput. Vis.*, pages 2488–2495, 2013. 8

[10] Geonho Cha, Minsik Lee, Jungchan Cho, and Songhwai Oh. Reconstruct as far as you can: Consensus of non-rigid reconstruction from feasible regions. *IEEE Trans. Pattern Anal. Mach. Intell.*, 43(2):623–637, 2019. 2

[11] Yuchao Dai, Hongdong Li, and Mingyi He. A simple prior-free method for non-rigid structure-from-motion factorization. *Int. J. Comput. Vis.*, 107(2):101–122, 2014. 1, 2, 3, 5, 6, 7, 8

[12] Alessio Del Bue. A factorization approach to structure from motion with shape priors. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1–8, 2008. 2

[13] Alessio Del Bue, Xavier Llad, and Lourdes Agapito. Non-rigid metric shape and motion recovery from uncalibrated images using priors. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1191–1198, 2006. 2

[14] Hui Deng, Tong Zhang, Yuchao Dai, Jiawei Shi, Yiran Zhong, and Hongdong Li. Deep non-rigid structure-from-motion: A sequence-to-sequence translation perspective. *arXiv preprint arXiv:2204.04730*, 2022. 1

[15] Paulo FU Gotardo and Aleix M Martinez. Computing smooth time trajectories for camera and deformable shape in structure from motion with occlusion. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(10):2051–2065, 2011. 1, 2, 7

[16] Paulo FU Gotardo and Aleix M Martinez. Kernel non-rigid structure from motion. In *Int. Conf. Comput. Vis.*, page 802, 2011. 7

[17] Paulo FU Gotardo and Aleix M Martinez. Non-rigid structure from motion with complementary rank-3 spaces. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3065–3072, 2011. 2, 7

[18] John C Gower. Generalized procrustes analysis. *Psychometrika*, 40:33–51, 1975. 2

[19] Stella Graßhof and Sami Sebastian Brandt. Tensor-based non-rigid structure from motion. In *IEEE Winter Conference on Applications of Computer Vision(WACV)*, pages 3011–3020, 2022. 2

[20] Mei Han and Takeo Kanade. Reconstruction of a scene with multiple linearly moving objects. *Int. J. Comput. Vis.*, 59: 285–300, 2004. 2

[21] José Pedro Iglesias, Carl Olsson, and Marcus Valtonen Örnhag. Accurate optimization of weighted nuclear norm for non-rigid structure from motion. In *Eur. Conf. Comput. Vis.*, pages 21–37, 2020. 7

[22] 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 Trans. Pattern Anal. Mach. Intell.*, 36(7):1325–1339, 2013. 7

[23] Sebastian Hoppe Nesgaard Jensen, Mads Emil Brix Doest, Henrik Aanæs, and Alessio Del Bue. A benchmark and evaluation of non-rigid structure from motion. *IEEE Trans. Pattern Anal. Mach. Intell.*, 129(4):882–899, 2021. 6

[24] Haorui Ji, Hui Deng, Yuchao Dai, and Hongdong Li. Unsupervised 3d pose estimation with non-rigid structure-from-motion modeling. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 3314–3323, 2024. 1

[25] Chen Kong and Simon Lucey. Deep non-rigid structure from motion with missing data. *IEEE Trans. Pattern Anal. Mach. Intell.*, 43(12):4365–4377, 2021. 1

[26] Suryansh Kumar. Non-rigid structure from motion: Prior-free factorization method revisited. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 51–60, 2020. 1, 2, 3, 6, 7, 8

[27] Suryansh Kumar and Luc Van Gool. Organic priors in non-rigid structure from motion. In *Eur. Conf. Comput. Vis.*, pages 71–88, 2022. 1, 2, 3, 6, 7

[28] 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. 1, 2

[29] Suryansh Kumar, Anoop Cherian, Yuchao Dai, and Hongdong Li. Scalable dense non-rigid structure-from-motion: A grassmannian perspective. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 254–263, 2018. 2

[30] Minsik Lee, Jungchan Cho, Chong-Ho Choi, and Songhwai Oh. Procrustean normal distribution for non-rigid structure from motion. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1280–1287, 2013. 1, 2, 3, 7, 8

[31] Minsik Lee, Chong-Ho Choi, and Songhwai Oh. A procrustean markov process for non-rigid structure recovery. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2014. 1, 2, 3, 7

[32] Minsik Lee, Jungchan Cho, and Songhwai Oh. Consensus of non-rigid reconstructions. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 4670–4678, 2016. 2, 7

[33] Jorge Nocedal and Stephen J Wright. *Numerical Optimization*. Springer, New York, NY, USA, 2e edition, 2006. 4

[34] David Novotny, Nikhila Ravi, Benjamin Graham, Natalia Neverova, and Andrea Vedaldi. C3dpo: Canonical 3d pose networks for non-rigid structure from motion. In *Int. Conf. Comput. Vis.*, pages 7688–7697, 2019. 1

[35] Marcus Valtonen Ornhag, José Pedro Iglesias, and Carl Olsson. Bilinear parameterization for non-separable singular value penalties. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3897–3906, 2021. 7

[36] 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 *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2898–2905, 2009. 2

[37] Marco Paladini, Alessio Del Bue, Joao Xavier, Lourdes Agapito, Marko Stošić, and Marija Dodig. Optimal metric projections for deformable and articulated structure-from-motion. *Int. J. Comput. Vis.*, 96:252–276, 2012. 2

[38] Shaifali Parashar, Daniel Pizarro, and Adrien Bartoli. Isometric non-rigid shape-from-motion with riemannian geometry solved in linear time. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(10):2442–2454, 2017. 7

[39] Shaifali Parashar, Daniel Pizarro, and Adrien Bartoli. Local deformable 3d reconstruction with cartan's connections. *IEEE Trans. Pattern Anal. Mach. Intell.*, 42(12):3011–3026, 2019. 2

[40] Shaifali Parashar, Daniel Pizarro, and Adrien Bartoli. Robust isometric non-rigid structure-from-motion. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(10):6409–6423, 2021. 2

[41] Sungheon Park, Minsik Lee, and Nojun Kwak. Procrustean regression: A flexible alignment-based framework for non-rigid structure estimation. *IEEE Trans. Image Process.*, 27 (1):249–264, 2017. 1, 2, 3, 7

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

[43] Mathieu Salzmann, Julien Pilet, Slobodan Ilic, and Pascal Fua. Surface deformation models for nonrigid 3d shape recovery. *IEEE Trans. Pattern Anal. Mach. Intell.*, 29(8):1481–1487, 2007. 2

[44] Carlo Tomasi and Takeo Kanade. Shape and motion from image streams under orthography: a factorization method. *Int. J. Comput. Vis.*, 9(2):137–154, 1992. 2, 3

[45] Lorenzo Torresani, Aaron Hertzmann, and Chris Bregler. Nonrigid structure-from-motion: Estimating shape and motion with hierarchical priors. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30(5):878–892, 2008. 3, 6

[46] Aydin Varol, Mathieu Salzmann, Engin Tola, and Pascal Fua. Template-free monocular reconstruction of deformable surfaces. In *Int. Conf. Comput. Vis.*, pages 1811–1818, 2009. 2

[47] Jing Xiao, Jin-xiang Chai, and Takeo Kanade. A closed-form solution to non-rigid shape and motion recovery. In *Eur. Conf. Comput. Vis.*, pages 573–587, 2004. 2

[48] Haitian Zeng, Xin Yu, Jiaxu Miao, and Yi Yang. Mhr-net: Multiple-hypothesis reconstruction of non-rigid shapes from 2d views. In *Eur. Conf. Comput. Vis.*, pages 1–17, 2022. 1, 2

[49] Yingying Zhu, Dong Huang, Fernando De La Torre, and Simon Lucey. Complex non-rigid motion 3d reconstruction by union of subspaces. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1542–1549, 2014. 1, 2

[50] Yue Zhu, Nermin Samet, and David Picard. H3wb: Human3.6m 3d wholebody dataset and benchmark. In *Int. Conf. Comput. Vis.*, pages 20166–20177, 2023. 7