Pattern Recognition 132 (2022) 108965

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

JSL3D: Joint subspace learning with implicit structure supervision for 3D pose estimation

Mengxi Jiang[a,b], Shihao Zhou[b], Cuihua Li[b], Yunqi Lei[b,*]

[a] School of Advanced Manufacturing, Fuzhou University, Jinjiang 362251, China
[b] Department of Computer Science, Xiamen University, Xiamen 361005, China

ARTICLE INFO

Article history:
Received 19 November 2019
Revised 2 July 2022
Accepted 9 August 2022
Available online 10 August 2022

Keywords:
3D pose estimation
Sparse representation model
Implicit structure supervision
Joint subspace learning

ABSTRACT

Estimating 3D human poses from a single image is an important task in computer graphics. Most model-based estimation methods represent the labeled/detected 2D poses and the projection of approximated 3D poses using vector representations of body joints. However, such lower-dimensional vector representations fail to maintain the spatial relations of original body joints, because the representations do not consider the inherent structure of body joints. In this paper, we propose JSL3D, a novel joint subspace learning approach with implicit structure supervision based on Sparse Representation (SR) model, capturing the latent spatial relations of 2D body joints by an end-to-end autoencoder network. JSL3Djointly combines the learned latent spatial relations and 2D joints as inputs for the standard SR inference frame. The optimization is simultaneously processed via geometric priors in both latent and original feature spaces. We have evaluated JSL3Dusing four large-scale and well-recognized benchmarks, including `Human3.6M`, `HumanEva-I`, `CMU MoCap` and `MPII`. The experiment results demonstrate the effectiveness of JSL3D.

1. Introduction

Estimating 3D human poses from a single RGB image containing the human activity is a fundamental yet challenging task in many applications [1], such as virtual reality [2]. Most existing approaches first use robust 2D detectors [3] to obtain 2D body joints from the image, and then design a post-processing mapping procedure to reconstruct 3D pose from these detected 2D joints [4,5]. Since similar 2D body joints may be projected from various 3D human activities, these approaches are often suffered from the inherent ambiguity of 2D-to-3D mapping.

To alleviate such inference ambiguity, researchers have explored various solutions, which can be roughly divided into learning-based and model-based methods. The former directly learns an end-to-end mapping from 2D to 3D joints, such as deep learning based approaches [6] [5,7]. Generally, the estimation capability of learning-based approaches depends on a large number of 2D-3D paired data for supervised learning. Such weakness can be eliminated by a model-based method called sparse representation model [8], which fits a parametric body model to 2D joints based on geometry prior [9–11].

During the past few years, several sparse representation (SR)-based approaches have been proposed to infer the 3D human pose [4,11,12]. In the standard procedure of SR-based approaches, the parametric body model is used for the 3D pose estimation. By integrating given (i.e., manual annotations) or detected 2D joints and 3D geometry priors, the body model parameters are solved by minimizing the Euclidean Distance in Cartesian space between the 2D pose and the projection of the approximated 3D pose. In particular, the 2D and 3D poses are described as $2 * N$ and $3 * N$ dimensional vector representations of the N body joints, respectively. The inference procedure in the standard SR-based approach is shown in Fig. 1(a). Although the effectiveness of the SR-based paradigm in a constrained environment (i.e., without paired training data), there are still limitations that hinder better performance in 3D human pose estimation.

1.1. Limitations and insights

Firstly, vector representations of human pose lose the inherent structure of the original data [13], resulting in the degradation of estimation performance. Since the human body is highly structured, the spatial relation of body joints is crucial when reconstructing the 3D pose from a 2D pose, as shown in Fig. 1(b). In the figure, each joint has spatial relations with all the other joints in the human body. Such relations should be consistently preserved between 2D and 3D spaces. For example, the location of hand joint

[*] Corresponding author.
E-mail addresses: jiangmengxi@stu.xmu.edu.cn (M. Jiang), shzhou@stu.xmu.edu.cn (S. Zhou), chli@xmu.edu.cn (C. Li), yqlei@xmu.edu.cn (Y. Lei).

https://doi.org/10.1016/j.patcog.2022.108965
0031-3203/© 2022 Elsevier Ltd. All rights reserved.

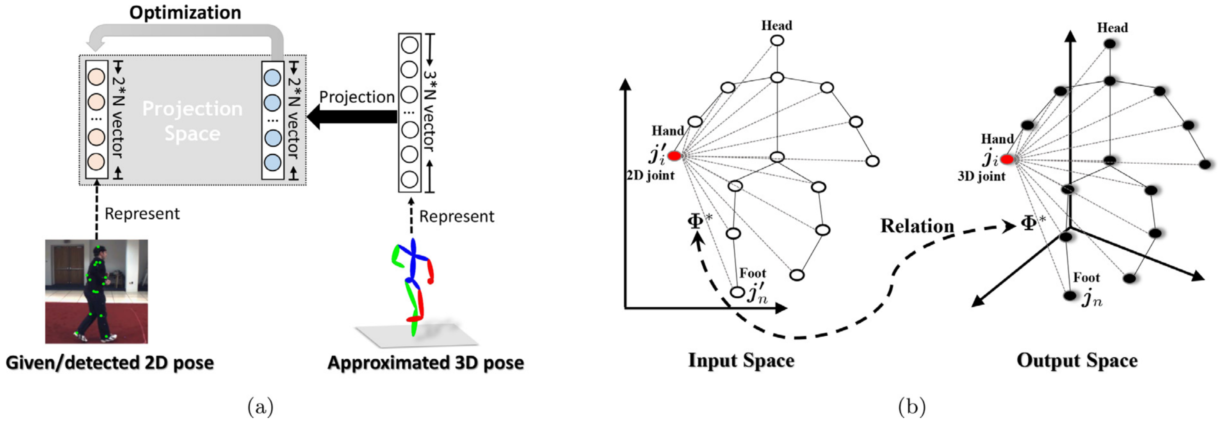


Fig. 1. (a) The inference procedure in the standard SR-based approach. (b) The illustration of maintaining relation consistency between input 2D and output 3D poses. Note that a known skeleton configuration connects the body joints of the human pose.

j_i' is between the head and foot joints in the input space, which also should be precisely preserved in the output space. Previous SR-based approaches have attempted to keep the original structure of the input 2D pose by imposing additional configuration constraints, such as the proportions of limbs [12] and joint angle limits [14]. However, these strategies only preserve the partial information of the body joints based on anthropometric knowledge. To the best of our knowledge, exploiting the deep implicit topological relations for human pose joints is an unexplored yet important problem to the current SR-based 3D human pose estimation approach.

Secondly, as shown in Fig. 1(a), the standard SR-based approach implements the optimization procedure in 2D projection space. The 3D human pose is not inferred directly in the original 3D space. Since the real projection matrix is unknown, such a minimization procedure based on a single projection space introduces estimation bias for the resulting 3D inference.

1.2. Our solution

In this paper, we propose a Joint Subspace Learning with Implicit Structure Supervision for 3D Pose Estimation, namely JSL3D, to alleviate the above issues. Firstly, we propose to capture implicit structure representations from all input 2D joints by leveraging an unsupervised learning-based approach called autoencoder. Secondly, to keep the consistent implicit spatial relations of body joints between 2D and 3D spaces, we propose a new joint subspace optimization frame based on latent and projection spaces by integrating the implicit spatial structure, 2D body joints, and 3D geometric prior.

Figure 2 gives an overview of our solution. Concretely, Firstly, given an input image, the 2D pose is obtained by a set of body joints that are labeled manually or detected by a deep convolutional neural network (CNN), as depicted in Fig. 2(a). Secondly, to estimate the 3D pose from the inputted 2D pose, we firstly obtain an original 3D pose by a sparse linear combination of a set of basis poses. Then, we project this estimated 3D pose (i.e., approximated 3D pose by a linear combination) into the 2D space, as shown in Fig. 2(b). Thirdly, the inputted and projected 2D poses are fed into an autoencoder network to generate the latent representation, as shown in Fig. 2(c). Finally, based on the 2D pose and its latent representations, we propose a joint subspace learning strategy that is built upon both the 2D and the latent spaces. By minimizing loss functions of the joint subspace learning, we optimize the estimated 3D pose by updating linear combination coefficients.

In summary, the main contributions of this paper are as follows.

- We propose a Joint Subspace Learning with Implicit Structure Supervision for 3D Pose Estimation, JSL3D, to estimate 3D human pose from a 2D pose. It is the first attempt to impose implicit structure representations captured by the learning-based approach into the inference frame of the SR model, which boosts the SR model capturing complex structure relations of the input 2D pose.
- This paper introduces a novel optimization procedure built on projection and latent spaces for the 3D human pose estimation, in which the implicit spatial relation consistency of body joints between 2D and 3D spaces is enforced.
- Compared to existing model-based approaches, JSL3Dachieves superior overall performance across all quantitative experiments on the well-recognized benchmarks. JSL3Deven shows competitive results compared with several representative learning-based approaches.

2. Related work

The task of 3D human pose estimation from 2D observations has been extensively studied in the literature. The early works attempt to estimate 3D pose from 2D observations, such as silhouettes [15] and edges [16]. With the release of Motion Capture (MoCap) datasets [17,18], a large number of 2D-3D human body joint annotations are available. Based on these MoCap datasets, considerable effort has been devoted to inferring 3D human poses from the 2D body joints. A typical solution makes use of depth information [19] and multi-view images [20] captured in some highly sensored environment. Since sensor devices are difficult to deploy in real scenarios, it is more realistic to estimate the 3D human pose simply from a single image [7,9]. In the following, we focus on the literature of 3D pose estimation from the single view. The related works to the estimation techniques used in JSL3Dare introduced, especially model-based approaches using SR and the methods capturing human pose structures.

Model-based Approaches

In model-based approaches, 3D human pose estimation is modeled as a parameter prediction problem of a given model, such as a known articulated skeleton. Recently, sparse representation (SR) model [8] has been applied for the 3D human pose estimation [9,11,12]. It is an effective tool in capturing complex signal variability, which has been successfully used in various computer vision tasks, such as image classification [21]. Ramakrishna et al. [12] first apply SR model to estimate 3D human pose and propose a matching pursuit algorithm for the prediction of model

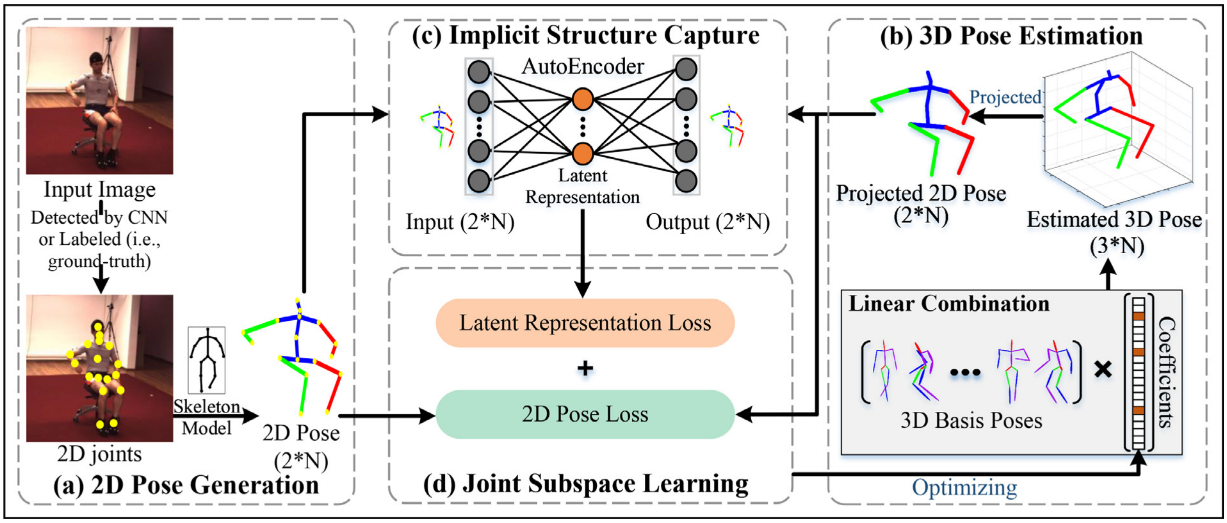


Fig. 2. The overview of the proposed approach. (a) Given an image, we obtain a set of 2D joints which are connected into a 2D pose by a known skeleton configuration. (b) To estimate the 3D pose for the inputted 2D pose, we firstly represent a 3D pose as a sparse linear combination of learned basis poses. This approximated 3D pose is then projected into 2D space to produce a projected 2D pose. (c) The inputted and projected 2D poses are fed into an autoencoder network to generate the latent representations, respectively. (d) Based on 2D poses and their latent representations, we design a joint subspace learning built upon the 2D and the latent spaces. By minimizing loss functions of the joint subspace learning, linear combination coefficients of the estimated 3D pose are optimized. Note that the figure indicates an iterative process.

parameters. To enhance the work [12], Zhou et al. [9] propose a convex relaxation algorithm for the solution of model parameters by imposing the orthogonal constraints. Instead of using ℓ_2-norm to measure two 2D pose vectors, Wang et al. [10] apply ℓ_1-norm to alleviate the influence of 2D outliers. It is known that using the projection error alone is not enough to ensure the most desired 3D pose. Jiang et al. [22] adjust the estimated 3D poses directly in the 3D space. In the work [11], more geometric priors are learned from the limited diversities of the training set for the 3D human pose estimation. The above model-based approaches only use the vector coordinates of 2D joints as input and perform minimization in a projections space. In contrast to these approaches, we integrate the 2D joint locations and its implicit structure representations into the optimization frame, in which a minimization objective function is built on both 2D projection space and other latent space.

In addition to the model-based techniques, several deep learning-based approaches have been proposed and achieved great success by devising specific network architectures [5,23]. We suggest the interested readers refer to [1] on this topic. Compared to most of learning-based approaches, JSL3Ddoes not require paired 2D-3D training samples.

Human Pose Structure Capture

For capturing human pose structure, physiological knowledge is often used as regularizing constraints which are imposed in 3D pose inference, such as joint angle limits [14] and limbs lengths [10]. However, such physiological constraints can only capture partial relations between body joints. In addition to using physiological constraints, ordinal depth relations are introduced to model the structure between 2D joints by using additional 2D depth annotations [6]. Recently, there are some deep learning-based works apply transformer architecture to capture the relationship among human body joints [7]. The most relevant work to us is [23], which uses autoencoder to learn latent representations for 3D poses, then directly map input 2D joints to the latent representations based on a large number of paired 2D-3D training data. The implicit relations in the input pose are not carefully explored in such a data-driven strategy.

Most existing approaches either use body physiological constraints or additional data information (i.e., depth annotations or paired 2D-3D data) for the consistency of human pose structure. Unlike these strategies, in this paper, we propose to capture the implicit spatial structure relations for input 2D joints, which are imposed into a model-based frame to guide the 3D human pose inference. Notice that additional data are not required in our approach.

3. Background

This section introduces the preliminary knowledge of this work, including the weak perspective camera model and the Sparse Representation (SR) in 3D pose estimation.

3.1. Weak perspective camera model

In this work, a human body is represented as a skeleton with N joints, in which 2D and 3D poses are denoted as $\boldsymbol{X} = \{\boldsymbol{j'}_i\}_{i=1}^{N} \in \mathbb{R}^{2N\times 1}$ and $\boldsymbol{Y} = \{\boldsymbol{j}_i\}_{i=1}^{N} \in \mathbb{R}^{3N\times 1}$, respectively, where $\boldsymbol{j'}_i$ and $\boldsymbol{j}_i$ are its corresponding 2D and 3D coordinates of joint i, respectively. By applying a weak perspective camera model, the dependence between the 2D pose and its corresponding 3D pose is described as:

$$\boldsymbol{X} = (\boldsymbol{I}_{N\times N} \otimes \boldsymbol{M})\boldsymbol{Y} + \boldsymbol{t} \otimes 1_N, \tag{1}$$

where $\boldsymbol{M} \in \mathbb{R}^{2\times 3}$ is the camera projection which contains both scaling and rotation parameters. $\otimes$ is the Kronecker product and I is the identity matrix. $\boldsymbol{t} \in \mathbb{R}^{2\times 1}$ is the camera translation vector. Based on Eq. (1), our aim is to find the 3D pose $\boldsymbol{Y}$ whose 2D projection is required to be consistent with the given 2D pose $\boldsymbol{X}$ as much as possible.

3.2. Sparse representation in 3D pose estimation

It is an ill-posed problem to obtain 3D pose $\boldsymbol{Y}$ by solving Eq. (1) since there may be many feasible solutions in the mathematical sense. To alleviate this issue, SR model is introduced to infer 3D pose $\boldsymbol{Y}$ from 2D projections $\boldsymbol{X}$. With the SR model, an original 3D pose $\boldsymbol{Y}$ is approximated as a sparse linear combination of a set of basis poses:

$$\boldsymbol{Y} = \sum_{j=1}^{k} c_j \boldsymbol{b}_j, \tag{2}$$

where k is the number of basis poses. $\boldsymbol{b}_j \in \mathbb{R}^{3N}$ represents a basis pose, and c_j is the corresponding coefficient. For the problem formulation, we eliminate translation vector $\boldsymbol{t}$ in Eq. (1) by centralizing the data and plug Eq. (2) into Eq. (1), we have

$$\boldsymbol{X} = (\boldsymbol{I}_{N\times N} \otimes \boldsymbol{M})(\boldsymbol{Bc}), \tag{3}$$

where $\boldsymbol{c} = [c_1, \dots c_k]^T$ is a coefficient vector. Basis pose set $\boldsymbol{B} = \{\boldsymbol{b}_1, \boldsymbol{b}_2, \dots, \boldsymbol{b}_k\}$ is named as overcomplete dictionary in the SR model and learned from the training poses in the MoCap dataset. Since the dictionary $\boldsymbol{B}$ can be pre-learned by existing dictionary learning algorithms. The estimation problem of 3D pose from 2D joints is converted to the calculation of camera parameters $\boldsymbol{M}$ and coefficient vector $\boldsymbol{c}$.

Note that $\boldsymbol{c}$ is expected to be only a few nonzero entries under SR assumption. To enforce sparsity on coefficient $\boldsymbol{c}$, ℓ_1-norm is introduced. Thus, the problem is modeled by the following optimization formulation:

$$\arg\min_{\boldsymbol{M},\boldsymbol{c}} \|\boldsymbol{c}\|_1 \quad \textbf{s.t.} \quad \boldsymbol{X} = (\boldsymbol{I}_{N\times N} \otimes \boldsymbol{M})(\boldsymbol{Bc}). \tag{4}$$

Considering the observation noise, we relax the equation constraint in Eq. (4) by adopting lagrangian multiplier as follows

$$\arg\min_{\boldsymbol{M},\boldsymbol{c}} \frac{1}{2}\|\boldsymbol{X} - (\boldsymbol{I}_{N\times N} \otimes \boldsymbol{M})(\boldsymbol{Bc})\|_F^2 + \lambda\|\boldsymbol{c}\|_1, \tag{5}$$

where $\|\cdot\|_1$ denotes the ℓ_1-norm and $\|\cdot\|_F$ represents the Frobenius norm of a matrix. In Eq. (5), the first term is the reconstruction error in the projection space, and the second term is the sparsity regularization to induce a sparse 3D representation. $\lambda > 0$ is the balance parameter for two terms.

4. Our approach

This section gives the detailed design of our JSL3D. We first explore the representation learning with the implied structural information for the input 2D poses using an end-to-end autoencoder. Then, we introduce joint subspace learning for our 3D pose estimation using a latent representation capturing the spatial relations between body joints and the 2D joint coordinates. Last, the optimization procedure is described.

4.1. Autoencoder for implicit structure capture

To capture the spatial relations of body joints, we use an end-to-end autoencoder to learn a robust representation by mapping input 2D joints to a latent space. Autoencoder has shown promising performance in unsupervised learning [24]. In this work, we encode input 2D joints into a latent representation using an autoencoder with one hidden layer. Formally, given a training set of 2D poses $\boldsymbol{X} = [\boldsymbol{X}_1, \boldsymbol{X}_2, \dots, \boldsymbol{X}_M]$, the standard training procedure of an autoencoder is to learn the parameter set $\boldsymbol{\Phi} = \{\boldsymbol{W}_{enc}, \boldsymbol{b}_{enc}, \boldsymbol{W}_{dec}, \boldsymbol{b}_{dec}\}$ by minimizing the following square loss:

$$\boldsymbol{\Phi}^* = \arg\min_{\boldsymbol{\Phi}} \sum_{m=1}^{M} \|\boldsymbol{X}_m - \hat{\boldsymbol{X}}_m\|_2^2, \tag{6}$$

where M is the number of the training 2D poses. $\boldsymbol{W}_{enc} \in \mathbb{R}^{N_l \times 2N}$ and $\boldsymbol{W}_{dec} \in \mathbb{R}^{2N \times N_l}$ denote the weight matrices for encoding and decoding. $\boldsymbol{b}_{enc} \in \mathbb{R}^{N_l}$ and $\boldsymbol{b}_{dec} \in \mathbb{R}^{2N}$ are corresponding bias terms. N_l is the number of hidden layer nodes. $\hat{\boldsymbol{X}}_m = f(\boldsymbol{X}_m, \boldsymbol{\Phi})$ represents the reconstructed inputs for the mth sample, where $f(\cdot)$ describes a mapping function of a complete autoencoder. For an autoencoder with only one hidden layer, $f(\cdot)$ comprises an encoding and a decoding processes as

$$\begin{cases} \boldsymbol{L_m} = E(\boldsymbol{X}_m), \\ \hat{\boldsymbol{X}}_{\boldsymbol{m}} = D(\boldsymbol{L}_m), \end{cases} \tag{7}$$

where $\boldsymbol{L}_m \in \mathbb{R}^{N_l}$ denotes the latent representation for the mth sample, termed as the implicit structure representation, which encodes the input 2D joints. $E(\boldsymbol{X}_m)$ and $D(\boldsymbol{L}_m)$ are encode and decode functions respectively, which are given as

$$\begin{cases} E(\boldsymbol{X}_m) = g(\boldsymbol{W}_{enc}\boldsymbol{X}_m + \boldsymbol{b}_{enc}), \\ D(\boldsymbol{L}_m) = g(\boldsymbol{W}_{dec}\boldsymbol{L}_m + \boldsymbol{b}_{dec}), \end{cases} \tag{8}$$

where $g(\cdot)$ represents a nonlinear activation function. Autoencoder can induce robust representation directly from input data while holding the original information. All joints of the original input pose $\boldsymbol{X}$ are encoded and preserved in each element of $\boldsymbol{L_m}$ which reflects implicit correlations between body 2D joints, instead of simply the coordinates.

4.2. Joint subspace learning for 3D pose estimation

After training the autoencoder, we obtain the encode function $E(\cdot)$, which can map a 2D pose $\boldsymbol{X}$ to the latent space and generate implicit structure representation $\boldsymbol{L}$. Based on the standard SR model formulated by (5), we map the given and the projected 2D pose into the latent space to produce implicit structure representations. Note that the projected 2D pose is projected from the estimated 3D pose (i.e., the 3D pose approximated by a linear combination. Then, we introduce an equality constraint to enforce implicit structure representations of 2D poses in the latent space. As a result, Eq. (5) is reformulated as

$$\begin{aligned} &\arg\min\nolimits_{\boldsymbol{M},\boldsymbol{c}} \tfrac{1}{2}\|\boldsymbol{X} - (\boldsymbol{I}_{N\times N} \otimes \boldsymbol{M})(\boldsymbol{Bc})\|_F^2 + \lambda\|\boldsymbol{c}\|_1 \\ &\quad \textbf{s.t.} \quad E(\boldsymbol{X}) = E((\boldsymbol{I}_{N\times N} \otimes \boldsymbol{M})(\boldsymbol{Bc})), \end{aligned} \tag{9}$$

where $E(\boldsymbol{X}) = \boldsymbol{L}$ and the result of $E((\boldsymbol{I}_{N\times N} \otimes \boldsymbol{M})(\boldsymbol{Bc}))$ are the implicit structure representations for the input 2D pose and the estimated 3D pose respectively. The model (9) states that, in addition to the minimization requirement in the 2D projection space between the given 2D pose and the projected 2D pose, the implicit structure representation in the latent space is also required to be fitted. The model (9) is the final objective function built upon the joint 2D space and the latent space to enable our optimization procedure.

4.3. Optimization

We present an algorithm to implement our model (9). The Augmented Lagrangian Methods (ALMs) is applied to solve the equality constraint problem in model (9). By introducing a dual variable $\boldsymbol{Z}$, the Augmented Lagrangian function of (9) is denoted as

$$\begin{aligned} \mathcal{L}(\boldsymbol{M}, \boldsymbol{c}, \boldsymbol{Z}) &= \tfrac{1}{2}\|\boldsymbol{X} - (\boldsymbol{I}_{N\times N} \otimes \boldsymbol{M})(\boldsymbol{Bc})\|_F^2 + \lambda\|\boldsymbol{c}\|_1 \\ &+ \tfrac{\mu}{2}\|E(\boldsymbol{X}) - E((\boldsymbol{I}_{N\times N} \otimes \boldsymbol{M})(\boldsymbol{Bc}))\|_F^2 \\ &+ < \boldsymbol{Z}, E(\boldsymbol{X}) - E((\boldsymbol{I}_{N\times N} \otimes \boldsymbol{M})(\boldsymbol{Bc})) >, \end{aligned} \tag{10}$$

where $\mu > 0$ is a penalty parameter. Then, the alternating direction method of multipliers (ADMM) is used to update the values of variables $\boldsymbol{M}, \boldsymbol{c}, \boldsymbol{Z}$ by the minimizing following sub-problems until convergence:

$$\boldsymbol{M}^{t+1} = \arg\min_{\boldsymbol{M}} \mathcal{L}(\boldsymbol{M}^t, \boldsymbol{c}^t, \boldsymbol{Z}^t), \tag{11}$$

$$\boldsymbol{c}^{t+1} = \arg\min_{\boldsymbol{c}} \mathcal{L}(\boldsymbol{M}^{t+1}, \boldsymbol{c}^t, \boldsymbol{Z}^t), \tag{12}$$

$$\boldsymbol{Z}^{t+1} = \boldsymbol{Z}^t + \mu(E(\boldsymbol{X}) - E((\boldsymbol{I}_{N\times N} \otimes \boldsymbol{M}^{t+1})(\boldsymbol{Bc}^{t+1}))), \tag{13}$$

where t denotes the tth iteration. The sub-problems (11) and (12) can be solved by using Accelerated Proximal Gradient (APG) and the manifold optimization solver in the Manopt toolbox respectively. When $\boldsymbol{c}$ reaches the optimum, we can obtain an estimation result of the 3D pose by calculating Eq. (2). The algorithm with a maximum number $\ell_{\max}$ of iterations is summarized in Algorithm 1.

Algorithm 1 ADMM to solve the problem (10).

```
Input: $\boldsymbol{X}, \boldsymbol{B}, \boldsymbol{W}_{enc}, \boldsymbol{b}_{enc}$   //The input 2D joints, the 3D basis dictionaries, encode matrix and bias.
Parameter: $\tau, \lambda, \mu$   //The convergence tolerance, the hyper-parameters.
Output: $\boldsymbol{Y}$   //The 3D human pose.
initialize $\boldsymbol{c}, \boldsymbol{M}, \boldsymbol{Z}, \tau, \lambda, \mu$.
while $\|\boldsymbol{r}\|_2 > \tau$ or $\ell < \ell_{\max}$ do
    update $\boldsymbol{M}$ by (11).
    update $\boldsymbol{c}$ by (12).
    update $\boldsymbol{Z}$ by (13).
    calculate $\boldsymbol{r} = \boldsymbol{X} - (\boldsymbol{I}_{N \times N} \otimes \boldsymbol{M})(\boldsymbol{Bc})$. // the estimation residual.
    update $\ell = \ell + 1$.   // iteration count.
end while
calculate $\boldsymbol{Y}$ by (2);
```

Table 1
The brief summary of four evaluation datasets.

Dataset	Size	# of Actions	# of Subjects
Human3.6M [18]	1376 videos	15	11
HumanEva-I [17]	56 videos	6	4
CMU MoCap [25]	2605 videos	23	109
MPII [26]	25,000 images	N.A.	N.A.

5. Experimental setup and evaluation

5.1. Evaluation datasets and protocols

The extensive evaluations of JSL3Dare performed on three public datasets, i.e., Human3.6M [18], HumanEva-I [17], CMU MoCap [25], and MPII [26]. The brief information about evaluation datasets is described in Table 1. We conduct the quantitative experiments on the first three datasets and qualitative experiments on the last.

Human3.6M contains millions of paired (2D, 3D) poses with corresponding RGB images. It includes 15 activities performed by seven actors in a configured indoor environment. The standard evaluation protocol uses five subjects (S1, S5, S6, S7, and S8) and two subjects (S9 and S11) for the training and testing, respectively [4,22]. This standard evaluation protocol is named protocol #1. In some literature, six subjects (S1, S5, S6, S7, S8, and S9) are used for training and only S11 for testing [27,28]. This protocol is named protocol #2. These two different protocols are found in the existing literature. Moreover, for the training, there are also two different ways. The one trains a universal model for all actions of the testing set [4]. In comparison, another trains special models for each testing class [23]. Two different evaluation protocols and training ways are also considered in this paper. For protocol #1, we train a universal dictionary only using 3D poses from the training data [4]. For protocol #2, following the same training protocol in [29], we learn special dictionaries for each testing activity.

HumanEva-I also contains images with corresponding poses (2D, 3D) captured in indoor. It includes six actions performed by four actors. Following the standard evaluation protocol [4,10], we train our dictionary by using the training set of HumanEva-I, and test on the walking and jogging performed by three subjects (S1, S2, and S3) from the validation set. Following the same training protocol in [4,22], we learn action-specific dictionaries for each subject separately.

CMU MoCap contains more than 3 million annotated 3D human poses with corresponding synchronized videos performed by 144 subjects. Following the standard evaluation protocol used in related model-based approaches [9], we conduct our experiments on eight categories (i.e., walk, run, jump, climb, box, dance, sit, and basketball). For each category, we randomly select six video sequences and corresponding human pose annotations, in which three sequences are used for dictionary learning and the remaining three for testing. The 2D human poses of CMU MoCap dataset are projected from 3D human poses by simulating an orthographic camera motion with 360-degree rotation [9]. Following the same training protocol in [9], we learn a single dictionary for all testing examples.

MPII includes over 410 activities of 40K people around 25K Internet images in various outdoor scenes. For each image, only corresponding 2D annotations are provided.

5.2. Implementation details

In our implementation, similar to previous work [4,10], we use the stacked hourglass model [3] for the 2D joints detection, which is pre-trained on the MPII and fine-tuned on Human3.6M. In our model, the hyper-parameter μ controls the optimization step, which is empirically set $\mu = 0.0001$. The maximum iteration $\ell_{\max}$ and convergence tolerance τ are set to 1000 and 0.0001, respectively. For the hyper-parameter λ, we conduct the ablation analysis in the next section. For camera parameter $\boldsymbol{M}$ and coefficient vector $\boldsymbol{c}$ in Eq. (11), we initialize them as an identity matrix and zero vectors, respectively.

The dictionary $\boldsymbol{B}$ is learned by the algorithm proposed in [9]. Considering the different amount of training examples, we set different dictionary sizes. Specifically, the dictionary sizes are set to 400 and 150 for the universal and action-specific dictionaries on Human3.6M, respectively. For HumanEva-I and CMU MoCap, k is set to 358 and 128, respectively. The dictionary learned from CMU MoCap is used for the qualitative experiments of MPII [9].

To obtain the latent representation $\boldsymbol{L_m}$ in Eq. (7) for input 2D poses, we train a simple autoencoder with one hidden layer in Pytorch. The hidden layer contains the same number of nodes as the input layer, that is, $N_l = 2 * N$. We use PReLU as the activation function $g(\cdot)$ across our autoencoder architecture. The autoencoder model is trained in an end-to-end pattern by using only 2D annotations in MPII.

5.3. Evaluation settings

5.3.1. Evaluation metric

To evaluate the estimation quality of the 3D human pose, we follow the two standard metrics, i.e., mean per joint 3D error and mean estimation error. The first one calculates the average Euclidean distance between the estimated 3D pose and ground-truth 3D pose over all the joints. The second is defined as the per joint 3D error up to a similarity transformation for two 3D poses. Following the standard evaluation protocol, we use the mean per joint position error and mean estimation error on Human3.6M. For HumanEva-I, only mean estimation error metric is considered.

5.3.2. Comparison approaches

We compare the performances of JSL3Dwith several existing methods on Human3.6M and HumanEva-I. In this paper, we study the ability of the SR model with unsupervised features learning for the 3D human pose estimation. Thus, for a fair comparison, we mainly focus on comparing the competing approaches that do not need paired 2D-3D data. To verify the effectiveness of JSL3D, we use the standard SR-based model (as formulated in Eq. (5)) as the comparison baseline, denoted as "Base". The same dictionary is used for both "Base" and JSL3Dfor a fair comparison. Moreover, to validate the effectiveness of JSL3D, we also compare our performance with several representative learning-based approaches. Recently, some learning-based approaches have been devoted to exploring a unsupervised [30,32] or weakly supervised strategy [35,36] for the 3D human pose estimation. These approaches are also considered in the comparison. All compared approaches may be using labeled (i.e., ground-truth) or detected 2D as input. Since the reconstruction performance depends on the accuracy of the input 2D annotations, we partition the results into ground-truth 2D and detected 2D groups.

5.4. Results and analysis

5.4.1. Human3.6M

The first comparative study is conducted on the Human3.6M dataset. We use the per joint 3D pose error and the mean estimation error metrics [18] in two evaluation protocols (i.e., protocol #1 and protocol #2) in this part. The first metric is widely applied for the evaluation of Human3.6M. All compared results are taken from the corresponding literature if there is no special explanation.

Under protocol #1, the per joint 3D pose errors of the compared approaches are summarized in Table 2. From the table, we observe that the performance of JSL3Doutperforms the Base in all cases and obtains better accuracy by more than 24% (ground-truth 2D) and 12% (detected 2D) improvements on average since the implicit correlations of input 2D joints are captured by JSL3D.

Moreover, compared to several approaches (including recent learning-based approaches) using the ground-truth 2D as their inputs, JSL3Doutperforms all of them in 12 out of 15 categories by more than 5% improvement on average. Although the detected 2D pose affects the performance due to inaccurate 2D estimations, the lowest error is still achieved on average by JSL3D. In particular, JSL3Dis superior to recent the model-based approaches [4,11] (TPAMI 19', NCAA 21') and the learning-based approach [30] (ICCV 21') in the average performance of the test categories.

The mean estimation errors of JSL3Dand representative works are reported in Table 3. As expected, JSL3Dstill consistently performs better than Base in most cases and achieves a better improvement by more than 13% (ground-truth 2D) and 4% (detected 2D) on average. The improvement is marginal when using the detected 2D annotations as our inputs. The reason is that the inaccurate inputs lead to a misleading implicit representation. Compared to recent representative approaches, Table 3 shows none of the approaches dominate across all cases, while JSL3Dyields a lower overall error. For example, JSL3Dachieves superior performance in 9 out of 15 motions when using the ground-truth 2D as the inputs. On average, JSL3Dachieves the best performance with more than 4% (ground-truth 2D) and 5% (detected 2D) improvements. Moreover, compared to several unsupervised learning-based approaches [32,35,36], JSL3Dstill outperforms these approaches in most categories and achieves a better performance improvement on average.

Under protocol #2, the mean estimation errors are reported in Table 4. When using ground-truth 2D joints as input, we observe that JSL3Doutperforms Base and comparison approaches [27,28] in most testing motions and more than 6% and 24% improvements on average. Note that work [27,28] uses paired 2D-3D training data. When using detected 2D joints as input, none of the comparison approaches dominate across all categories. JSL3Dstill performs better than them in 8 out of 15 categories and yields the lower estimation errors on average. Specifically, JSL3Dattains a better reconstruction performance by more than 5% and 2% improvements than Base and other competitive approaches on average.

5.4.2. HumanEva-I

We have also compared JSL3Dwith Base and several approaches on HumanEva-I. The mean 3D pose errors are reported in

Table 2
Per joint 3D errors (mm) on Human3.6M under protocol #1. “P” denotes using paired 2D-3D training data, while “NP” denotes not using. Best results are marked in bold.

Methods	Direct.	Discuss	Eat	Greet	Phone	Photo	Pose	Buy	Sit	SitDown	Smoke	Wait	WalkDog	Walk	WalkPair	Avg.
							Ground-Truth 2D									
(NP) Ramakrishna et al. [12]†	127.6	134.8	127.6	144.7	129.7	138.	137.0	146.2	149.8	161.5	132.0	156.7	120.2	159.3	159.8	141.4
(NP) Zhou et al. [9]†	94.8	89.7	83.3	99.1	85.2	109.2	95.3	84.9	80.3	97.2	82.2	94.4	83.4	**81.4**	92.3	90.2
(NP) Zhou et al. [4]	92.8	88.2	82.3	97.4	83.3	107.1	93.2	83.7	79.0	96.8	80.9	92.4	81.6	81.5	**91.3**	88.7
(NP) Jiang et al. [22]	97.7	86.7	74.0	100.1	80.0	96.5	97.0	83.1	83.5	97.8	77.8	92.3	84.0	86.4	96.5	87.7
(NP) SDM3d [11]	94.2	84.1	73.8	98.3	76.1	105.1	**86.7**	84.3	74.9	107.9	76.1	88.9	88.1	86.0	93.9	86.3
(NP) Yu et al. [30]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	85.3
(P) Zhou et al. [28]	88.1	64.3	73.0	62.1	84.4	128.5	77.1	70.8	96.3	89.9	68.8	62.7	64.8	65.9	67.5	79.8
(NP) **Base**	123.9	114.6	75.6	124.8	96.5	119.9	110.5	128.7	98.5	147.4	92.8	107.3	102.6	99.3	110.0	108.2
(NP) **JSL3D**	**89.9**	**79.5**	**67.2**	**92.6**	**75.7**	**93.9**	88.5	**78.1**	**69.0**	**81.4**	**72.6**	**86.3**	**79.1**	95.3	95.3	**81.4**
							Detected 2D									
(NP) Ionescu et al. [18]	132.7	183.6	132.4	164.4	162.1	205.9	150.6	171.3	151.6	243.0	162.1	170.7	177.1	96.6	127.9	162.1
(NP) Novotny et al. [31]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	153.0
(NP) Rhodin et al. [32]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	131.7
(NP) Wang et al. [10]	90.3	117.6	86.0	111.0	123.5	154.9	100.5	97.3	130.6	200.7	130.6	110.3	124.0	64.9	88.0	115.3
(NP) Chen and Ramanan [33]	89.9	97.6	90.0	107.9	107.3	139.2	93.6	136.1	133.1	240.1	106.7	106.2	114.1	87.0	90.6	114.2
(NP) SDM3d [11]	113.1	101.1	88.5	111.4	**97.9**	121.4	107.6	99.1	115.1	151.7	97.8	117.8	103.4	105.7	117.4	108.6
(NP) Zhou et al. [4]	**82.8**	**88.2**	93.3	**93.0**	111.7	**115.9**	**85.4**	131.4	126.8	226.8	97.6	**91.7**	99.7	83.5	88.4	107.8
(P) Zhou et al. [34]	91.8	102.4	96.7	98.8	113.4	125.2	90.0	**93.8**	132.2	159.0	107.0	94.4	126.0	79.0	99.0	107.3
(NP) **Base**	137.5	122.7	89.5	134.4	106.6	129.0	123.5	140.9	118.7	171.0	105.2	124.9	116.2	112.0	125.9	121.5
(NP) **JSL3D**	101.9	100.8	**83.9**	111.5	98.8	118.6	109.9	94.7	**110.3**	**142.1**	**94.8**	118.5	**96.9**	111.4	117.1	**106.2**

Note: The literature with marker † denotes that corresponding results are obtain from Zhou et al. [4].

Table 3

Mean estimation errors (mm) on `Human3.6M` under protocol #1. “P” denotes using paired 2D-3D training data, while “NP” denotes not using. Best results are marked in bold.

Methods	Direct.	Discuss	Eat	Greet	Phone	Photo	Pose	Buy	Sit	SitDown	Smoke	Wait	WalkDog	Walk	WalkPair	Avg.
Ground-Truth 2D																
(NP) 3DInterpreter [36] *	56.3	77.5	96.2	71.6	96.3	106.7	59.1	109.2	111.9	111.9	124.2	93.3	-	58.0	-	88.6
(NP) AIGN [35]	53.7	71.5	82.3	58.6	86.9	98.4	57.6	104.2	100.0	112.5	83.3	68.9	-	57.0	-	79.0
(NP) Zhou et al. [4]	52.0	54.0	59.1	61.7	74.2	70.7	51.5	60.3	83.9	119.9	66.9	54.8	64.5	55.6	59.1	65.9
(P) Morenonoguer et al. [29]	53.5	50.5	65.8	62.5	56.9	60.6	50.8	56.0	79.6	**63.7**	80.8	61.8	59.4	68.5	62.1	62.2
(NP) Wang et al. [37]	48.4	57.1	49.8	54.8	57.2	**50.9**	51.6	76.0	109.8	55.3	74.5	57.0	**40.2**	61.3	**47.2**	59.4
(NP) Jiang et al. [22]	51.2	48.1	**46.1**	57.4	51.2	58.2	50.1	**47.6**	61.7	82.1	48.6	53.5	54.4	50.3	54.5	53.8
(NP) **Base**	52.2	53.9	49.6	58.3	58.6	71.9	51.1	65.0	63.9	91.2	54.5	56.2	58.5	50.9	58.5	58.9
(NP) **JSL3D**	**47.8**	**45.6**	46.6	**54.2**	**49.6**	61.3	**48.6**	48.1	**53.4**	65.9	**46.9**	**50.3**	54.4	**49.5**	56.1	**51.2**
Detected 2D																
(NP) Akhter and Black [14] †	199.2	177.6	161.8	197.8	176.2	86.5	195.4	167.3	160.7	173.7	177.8	181.9	176.2	198.6	192.7	181.1
(NP) Zhou et al. [9] †	99.7	95.8	87.9	116.8	108.3	107.3	93.5	95.3	109.1	137.5	106.0	102.2	106.5	110.4	115.2	106.7
(NP) 3DInterpreter [36] *	78.6	90.8	92.5	89.4	108.9	112.4	77.1	106.7	127.4	139.0	103.4	91.4	-	79.1	-	98.4
(NP) Rhodin et al. [32]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	98.2
(NP) AIGN [35]	77.6	91.4	89.9	88	107.3	110.1	75.9	107.5	124.2	137.8	102.2	90.3	-	78.6	-	97.2
(P) Morenonoguer [29]	69.5	80.2	78.2	87.0	100.8	**76.0**	69.7	104.7	113.9	**89.7**	102.7	98.5	79.2	82.4	77.2	87.3
(NP) Bogo et al. [38]	62.0	60.2	67.8	76.5	92.1	77.0	73.0	75.3	100.3	137.3	83.4	77.3	86.8	79.7	87.7	82.3
(P) Lin et al. [39]	58.0	68.2	63.3	**65.8**	75.3	93.1	61.2	65.7	98.7	127.7	70.4	68.2	72.9	**50.6**	**57.7**	73.1
(NP) **Base**	63.5	64.9	62.0	68.9	68.8	82.6	**60.9**	73.1	88.5	111.6	67.0	69.4	72.4	62.3	68.9	71.8
(NP) **JSL3D**	59.4	63.9	**57.7**	68.1	**66.5**	79.4	62.9	**57.6**	**82.1**	102.9	**65.8**	72.4	**68.4**	60.6	67.2	**68.9**

Note: The literature with marker † and * denotes that corresponding results are obtain from Bogo et al. [38] and Tung et al. [35], respectively.

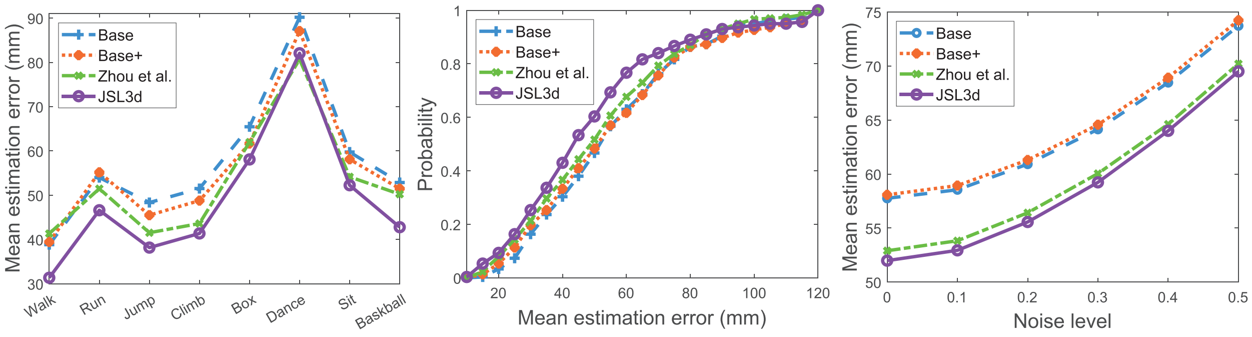
Table 4

Mean estimation errors (mm) on `Human3.6M` under protocol #2. “P” denotes using paired 2D-3D training data, while “NP” denotes not using. Best results are marked in bold.

Methods	Direct.	Discuss	Eat	Greet	Phone	Photo	Pose	Buy	Sit	SitDown	Smoke	Wait	WalkDog	Walk	WalkPair	Avg.
Ground-Truth 2D																
(P) Yasin et al. [27]	60.0	54.7	71.6	67.5	63.8	96.9	61.9	55.7	73.9	110.8	78.9	67.9	67.9	89.3	47.5	70.5
(P) Zhou et al. [28]	59.1	63.3	70.6	65.1	61.2	68.4	73.2	83.7	84.9	72.7	84.3	81.9	75.1	57.9	**49.6**	70.0
(NP) **Base**	51.2	50.3	51.8	61.1	47.8	73.3	59.3	43.9	**62.1**	75.6	58.9	62.2	54.1	40.3	57.4	56.6
(NP) **JSL3D**	**45.6**	**42.9**	**52.0**	**55.70**	**46.5**	**67.5**	**48.4**	**61.6**	62.5	**68.3**	**55.2**	**50.2**	**49.4**	**39.0**	52.9	**53.2**
Detected 2D																
(P) Yasin et al. [27]	88.4	72.5	108.5	110.2	97.1	81.6	107.2	119.0	170.8	108.2	142.5	86.9	92.1	165.7	102.0	108.3
(P) Zhou et al. [28]	67.9	65.4	77.7	69.3	68.9	75.9	86.5	105.3	81.5	86.3	**73.6**	102.3	59.1	69.8	52.6	76.1
(NP) Chen and Ramanan [33]	71.6	66.6	74.7	79.1	70.1	67.6	89.3	90.7	195.6	83.5	93.3	71.2	**55.8**	85.8	62.5	82.7
(P) Nie et al. [40]	62.8	69.2	79.6	78.8	80.8	72.5	73.9	96.1	106.9	88.0	86.9	**70.7**	71.9	76.5	73.2	79.5
(P) Morenonoguer [29]	66.1	61.7	84.5	73.7	65.2	67.2	**60.9**	**67.3**	103.5	**74.8**	92.6	69.6	71.5	78.0	73.2	74.0
(NP) **Base**	56.1	56.6	60.1	69.1	56.9	74.0	103.2	147.5	**61.8**	96.0	90.6	69.5	60.3	70.1	65.8	75.8
(NP) **JSL3D**	**50.8**	**51.6**	**58.0**	**62.6**	**53.2**	**63.6**	109.7	144.5	65.6	82.3	89.6	59.4	57.2	**68.7**	**62.4**	**71.9**

Table 5
Mean estimation errors (mm) on HumanEva-I. "P" denotes using paired 2D-3D training data, while "NP" denotes not using. Best results are marked in bold.

Method	Walking(C1)			Jogging(C1)			
	S1	S2	S3	S1	S2	S3	Avg.
				Ground-truth 2D			
(P) Morenonoguer [29]	28.4	27.8	31.7	47.8	27.8	30.2	37.1
(NP) SDM3d [11]	30.0	30.0	32.3	27.9	27.1	31.9	29.9
(NP) Jiang et al. [22]	21.5	20.4	**20.0**	**21.5**	21.5	21.9	21.1
(NP) **Base**	46.2	32.5	46.7	22.0	15.4	24.5	31.2
(NP) **JSL3D**	**16.1**	**13.0**	33.6	21.7	**14.4**	**17.2**	**19.3**
				Detected 2D			
(NP) Yasin et al. [27]	35.8	32.4	41.6	46.6	41.4	35.4	38.9
(NP) Wang et al. [10]	40.3	37.6	**37.4**	39.7	36.2	38.4	38.3
(NP) Zhou et al. [4]	34.3	31.6	49.3	48.6	34.0	**30.0**	37.9
(P) Katircioglu et al. [23]	29.3	**17.9**	59.5	-	-	-	**35.6**
(NP) **Base**	51.0	55.0	68.2	35.6	37.1	52.2	49.9
(NP) **JSL3D**	**23.9**	28.2	55.8	**32.1**	**32.0**	45.6	36.7



(a) (b) (c)

Fig. 3. Quantitative results on CMU MoCap dataset. (a) Mean 3D estimation errors on different testing motions. (b) The distribution of mean 3D estimation errors. The y-axis is the percentage of the testing examples whose estimation errors are less than the corresponding x-axis value. (c) Mean 3D estimation errors on various standard deviations of Gaussian noise.

Table 5. The figures in the table are cited from the original papers. Similar to the evaluation on Human3.6M, we divide all compared results into two groups, i.e., ground-truth 2D and detected 2D. Similar to the evaluation results on Human3.6M, JSL3Dboosts the performance of Base in all categories of HumanEva-I. Such observations confirm that the effectiveness of JSL3Dis supported by a learned implicit representation of the input 2D pose. In particular, compared to Base, JSL3Dachieves better performance with more than 38% (ground-truth 2D) and 26% (detected 2D) performance improvements, respectively.

In addition, JSL3Dachieves the best performance on average compared to all model-based and several representative learning-based approaches. Specifically, when using the ground-truth 2D as the inputs, JSL3Doutperforms all comparison algorithms in 4 out of 6 categories and obtains more than 8% improvement on average. Using the detected 2D poses as the inputs, JSL3Dstill achieves the best performance in 4 out of 6 categories with more than 3% improvement on average compared to the existing model-based approach [4] (TPAMI 19'). When comparing with deep learning-based approaches [23] (IJCV 18'), which also aims to maintain the implicit structure of poses, JSL3Dstill outperforms this approach in 2 out of 3 categories. Note that the paired 2D-3D data are required in the work [23], which are not needed in JSL3D.

There are exceptions in a few cases, such as the "Walking" category of S3. The performance of JSL3Dis not very high. We found that some annotations with outliers exist in this category. Since JSL3Dlearns a robust representation from the input 2D joints directly, the outliers impact the accuracy of implicit representation.

5.4.3. *CMU MoCap*

Although the few works evaluate their approach on CMU MoCap, the work [9]) still reports the experiment results on this dataset. Since the work [9] is closely related to ours, we also evaluate our approach on this dataset. In the following comparisons, "Base+" is an alternative optimization solution of the standard SR model, which is also reported in literature [9]. All comparison approaches in this dataset do not use paired 2D-3D training data.

The first experiment is conducted on different motions, as shown in Fig. 3(a). JSL3Doutperforms baseline approaches and [9] across most motions. We observe that in categories with less depth ambiguity (e.g., walk), JSL3Dshows more significant improvements due to capturing more accurate joint implicit structure. We further present the percentage of mean 3D estimation error ranges of all testing examples, as shown in Fig. 3(b). As seen from the figure, JSL3Dperforms better estimation on most testing examples than comparison approaches. Significantly, the percentage of JSL3Dreaches 60% when the mean estimation errors are smaller than 50 (mm). This percentage is reduced to around 42% of comparison approaches. To analyze the robustness of JSL3Dagainst noise, we evaluate JSL3Dby adding Gaussian noises with different standard deviations on inputted 2D joints. The results are reported in Fig. 3(c). The 3D estimation errors of JSL3Dare consistently lower than the comparison approaches across all noise levels.

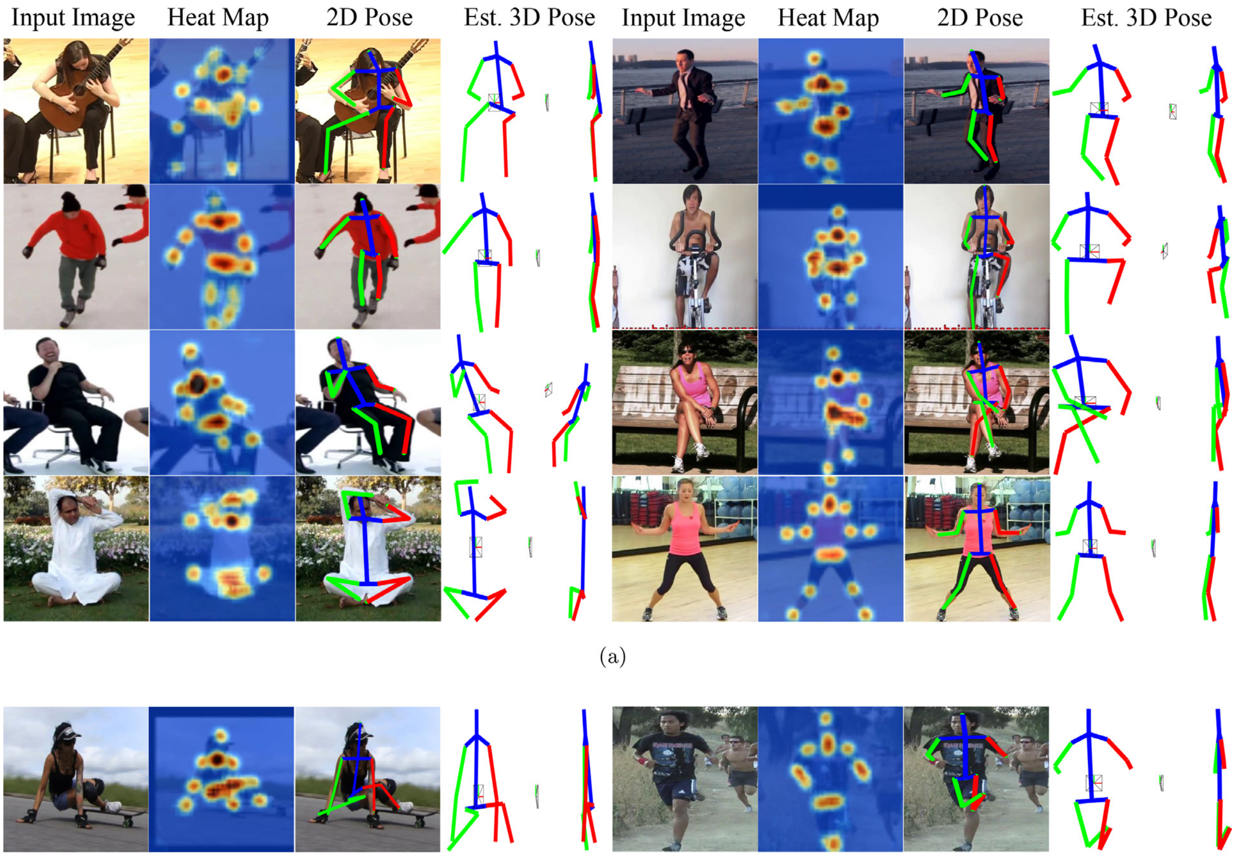
Considering that the deformation of body joints leads to different degrees of depth ambiguity, we present the estimation performance of approaches on different joints. The mean estimation errors on different body joints are reported in Table 6. We observe that JSL3Dconsistently yields lower estimation errors across

Table 6
Mean 3D estimation errors (mm) of different joints on CMU MoCap dataset. Note that RH, RK, RA, LH, LK, RE, RW, LS, LE, and LW denote right hip, right knee, right ankle, left hip, left knee, right elbow, right wrist, left shoulder, left elbow, and left wrist, respectively.

Methods	Pelvis	RH	RK	RA	LH	LK	LA	Neck	Head	Spine	RE	RW	LS	LE	LW	Avg.
Base	40.0	48.4	59.8	69.5	48.2	61.9	74.9	29.0	46.6	41.6	56.6	89.8	46.8	64.3	88.8	57.8
Base+	41.4	48.9	62.4	70.1	49.1	66.0	77.8	26.8	47.7	39.2	60.0	86.2	45.9	66.1	83.9	58.1
Zhou et al.	**35.3**	45.5	53.3	**64.2**	48.3	**53.8**	**68.1**	25.6	44.5	39.9	58.0	**76.7**	43.7	61.0	75.7	52.9
JSL3D	36.7	**45.4**	**52.4**	67.7	**44.2**	55.3	72.2	**24.6**	**40.7**	**38.5**	**54.5**	77.1	**43.1**	**54.2**	**75.2**	**52.1**

Table 7
Mean running time (ms) of different motions on CMU MoCap dataset.

Methods	Walk	Run	Jump	Climb	Box	Dance	Sit	Basketball	Avg.
Base	20.52	20.38	19.80	18.23	27.16	23.89	33.29	19.28	22.82
Base+	37.62	40.0	40.23	54.64	48.03	54.54	60.65	48.72	48.05
Zhou et al.	1017.56	752.97	955.42	839.22	771.54	801.78	801.78	838.95	855.09
JSL3D	410.75	517.61	591.95	608.02	594.06	761.80	638.48	561.76	585.93



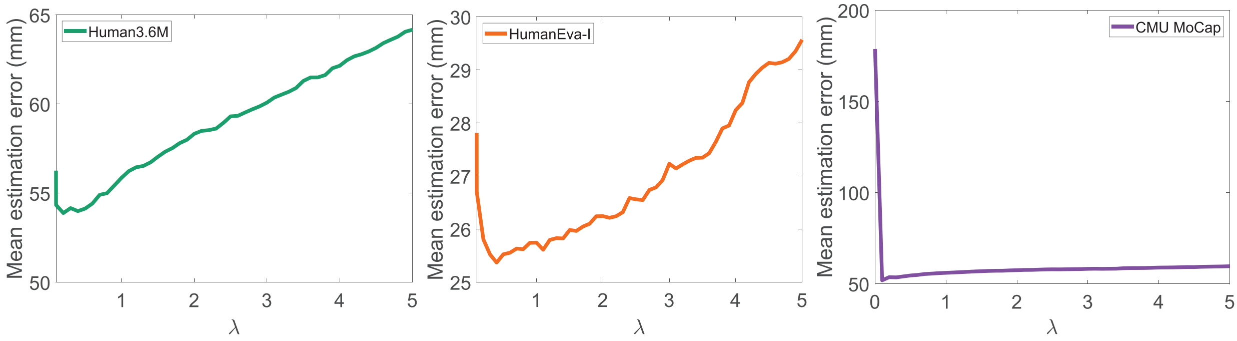
(a)

(b)

Fig. 4. Qualitative experiments on MPII. (a) The successful estimation results. (b) The failure estimation results.

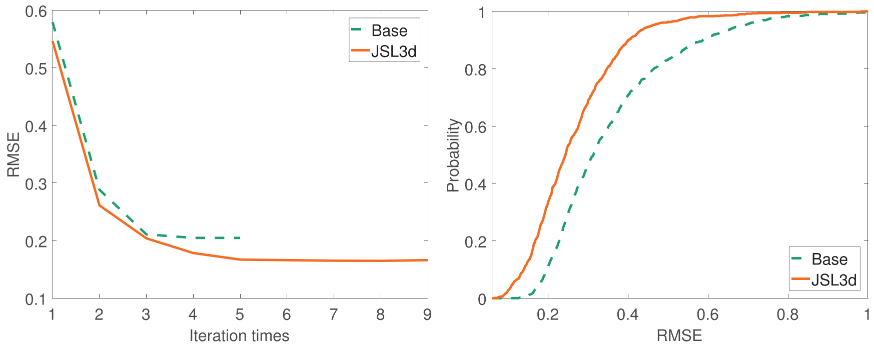
all joints than Base and Base+, and more than 9% improvement on average. Compared to [9], JSL3Dachieves better performance in 10 out of 15 joints and the average error. Since self-occlusion and severe deformation seldom occur in some human body joints (e.g., head and elbow), the depth ambiguity of these joints is relatively not serious. As we expected,JSL3Dshows more significant improvements in these joints since the more accurate joint implicit structure is learned. Specifically, the JSL3Dachieves a better reconstruction performance by more than 8% improvement on the “head” joint.

It seems that the margin between our work and [9] is narrow in a few cases, such as the estimation error of “Dance” in Fig. 3(a) and the average error in Table 6. However, the execution speed of JSL3Dis faster than [9] under the same running environment configuration. The mean running times on different testing motions are presented in Table 7. The experiments are implemented in MATLAB on a laptop with an Intel i7 2.30 GHz CPU, an Nvidia RTX 3060 GPU, and 32 GB RAM. It is not surprising that “Base” runs the fastest since it is a standard SR-based model. However, the estimation performances of “Base” are often unsatisfactory. Running time increases when more complex optimization strategies (i.e., “Base+” and [9]) are used. The work [9] handles complex body variability using a convex optimization strategy during the model inference. In this paper, we use a pre-trained autoencoder to obtain the implicit structure of the human body, which relieves the stress of the optimization process, resulting in reduced running time. Specifi-



(a) (b) (c)

Fig. 5. Sensitivity of the hyper-parameter λ on Humen3.6M, HumanEva-I, and CMU MoCap datasets, respectively.



(a) (b)

Fig. 6. Comparison of 2D pose RMSE between Base and JSL3D. (a) RSME values during one iteration. (b) RSME values distribution of all testing examples.

cally, as seen from Table 7, compared to [9], JSL3Dhas less running time in all motions, with an average running time reduction of more than 31%.

5.4.4. MPII

In the following, we explore the applicability of JSL3Don the MPII dataset. For each input 2D image, we locate its 2D joints heat map using the stacked hourglass 2D detector [3]. Since MPII does not provide any 3D annotation, a dictionary learned from Human3.6M is adopted for the inference of 3D poses. The estimation results with various activities are presented in Fig. 4, where each row includes two examples. We can see that JSL3Dis able to produce reasonable 3D human poses from an image for a wide variety of viewpoints and activities, as the successful examples shown in Fig. 4(a). The failed examples shown in the Fig. 4(b) are mainly because heavy occlusions cause incorrect 2D detection and some extreme activities.

5.4.5. Ablation study

To explore the impacts of λ in Eq. (10), we have further conducted ablation experiments using different datasets. In addition, the effectiveness of JSL3Dduring optimization is also studied.

For the ablation analysis, using ground-truth 2D poses of testing samples as inputs, we tested the value of the parameter λ in the range of [0, 5]. The ratios of the mean estimation errors to λ for Human3.6M, HumanEva-I, and CMU MoCap are presented in Fig. 5(a), (b), and (c), respectively. The variation in λ leads to the precision fluctuation of 3D estimation, while the mean estimation errors reach small values when λ is in the range of [0, 1] both for the three datasets. Thus, we fix $\lambda = 0.3$ for Human3.6M and HumanEva-I, $\lambda = 0.1$ for CMU MoCap in the experiments.

In addition, to verify the effectiveness of the proposed scheme, we have conducted convergence experiments. Compared to the Base built upon on the standard SR model (5), our model (9) imposes the implicit structure constraint of a latent space on the standard model. It is observed that after the convergence of the Base, JSL3Dcontinues to iterate to find a smaller Root Mean Square Errors (RSME) under the same tolerance value, as shown in Fig. 6(a). Moreover, the RMSE distribution of all testing examples after algorithms convergence is shown in Fig. 6(b). Note that the y-axis is the percentage of the testing cases whose RMSE is less than the x-axis value. As expected, JSL3Dachieves lower RMSE values than Base.

6. Conclusion

In this paper, we presented JSL3D, a novel joint subspace learning approach with implicit structure supervision based on the SR model, for precisely estimating human 3D poses. Instead of imposing a pose structure learning module on the optimization procedure, JSL3Ddirectly obtains the spatial relations of body joints through an autoencoder pre-trained on 2D joints of the human body. Then, JSL3Dcombines original input 2D joints, and the learned implicit representation capturing the spatial relations of body joints as supervisions for the SR model, in which the optimization is processed on both the 2D and the latent spaces. Such strategies enable the standard SR model to capture the implicit structure for input signal without introducing additional computational cost, which may be able to extend to other SR-based signal processing applications. We have evaluated JSL3Don four large-scale datasets (i.e., Human3.6M, HumanEva-I, CMU MoCap and MPII) with the comparison of several well-recognized benchmarks. The experiment results demonstrate that JSL3Dshows superior overall performance across all quantitative evaluations compared with the state-of-art model-based approaches and achieves competitive performance compared with several representative learning-based approaches. In our approach, the projection ambiguity is a critical factor that affects the structure capture accuracy of the 2D human pose. To alleviate this issue, in the future, we may leverage context information of the input image to explore joint depth that may be a useful cue for the pose structure capture.

Declaration of Competing Interest

We wish to confirm that there are no known conflicts of interest associated with this publication and there has been no significant financial support for this work that could have influenced its outcome.

We confirm that the manuscript has been read and approved by all named authors and that there are no other persons who satisfied the criteria for authorship but are not listed. We further confirm that the order of authors listed in the manuscript has been approved by all of us.

We confirm that we have given due consideration to the protection of intellectual property associated with this work and that there are no impediments to publication, including the timing of publication, with respect to intellectual property. In so doing we confirm that we have followed the regulations of our institutions concerning intellectual property.

We understand that the Corresponding Author is the sole contact for the Editorial process (including Editorial Manager and direct communications with the office). He is responsible for communicating with the other authors about progress, submissions of revisions and final approval of proofs. We confirm that we have provided a current, correct email address which is accessible by the Corresponding Author and which has been configured to accept email from (yqlei@xmu.edu.cn).

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research was supported by the National Nature Science Foundation of China (grant no 61671397). We gratefully acknowledge Dr. Guanping Xiao from Nanjing University of Aeronautics and Astronautics for helpful discussions.

References

[1] J. Yan, M. Zhou, J. Pan, M. Yin, B. Fang, Recent advances in 3D human pose estimation: from optimization to implementation and beyond, Int. J. Pattern Recognit. Artif. Intell. 36 (02) (2022) 2255003.
[2] M.-Y. Wu, P.-W. Ting, Y.-H. Tang, E.-T. Chou, L.-C. Fu, Hand pose estimation in object-interaction based on deep learning for virtual reality applications, J. Vis. Commun. Image Represent. 70 (2020) 102802.
[3] A. Newell, K. Yang, J. Deng, Stacked hourglass networks for human pose estimation, in: European Conference on Computer Vision, 2016, pp. 483–499.
[4] X. Zhou, M. Zhu, G. Pavlakos, S. Leonardos, K.G. Derpanis, K. Daniilidis, MonoCap: monocular human motion capture using a CNN coupled with a geometric prior, IEEE Trans. Pattern Anal. Mach. Intell. 41 (4) (2019) 901–914.
[5] W. Hu, C. Zhang, F. Zhan, L. Zhang, T.-T. Wong, Conditional directed graph convolution for 3D human pose estimation, in: Proceedings of Conference on International Conference on Multimedia, 2021, pp. 602–611.
[6] G. Pavlakos, X. Zhou, K. Daniilidis, Ordinal depth supervision for 3D human pose estimation, in: Computer Vision and Pattern Recognition, 2018, pp. 7307–7316.
[7] C. Zheng, S. Zhu, M. Mendieta, T. Yang, C. Chen, Z. Ding, 3D human pose estimation with spatial and temporal transformers, in: International Conference on Computer Vision, 2021, pp. 11656–11665.
[8] B.A. Olshausen, D.J. Field, Sparse coding with an overcomplete basis set: a strategy employed by v1? Vision Res. 37 (23) (1997) 3311–3325.
[9] X. Zhou, M. Zhu, S. Leonardos, K. Daniilidis, Sparse representation for 3D shape estimation: aconvex relaxation approach, IEEE Trans. Pattern Anal. Mach. Intell. 39 (8) (2017) 1648–1661.
[10] C. Wang, Y. Wang, Z. Lin, A. Yuille, Robust 3D human pose estimation from single images or video sequences, IEEE Trans. Pattern Anal. Mach. Intell. 41 (5) (2019) 1227–1241.
[11] M. Jiang, Z. Yu, C. Li, Y. Lei, SDM3d: shape decomposition of multiple geometric priors for 3D pose estimation, Neural Comput. Appl. 33 (7) (2021) 2165–2181.
[12] V. Ramakrishna, T. Kanade, Y. Sheikh, Reconstructing 3D human pose from 2D image landmarks, in: European Conference on Computer Vision, 2012, pp. 573–586.
[13] M. Pelillo, E.R. Hancock, X. Li, V. Murino, Guest editorial special section on learning in non-(geo) metric spaces, IEEE Trans. Neural Netw. Learn. Syst. 27 (6) (2016) 1290–1293.
[14] I. Akhter, M.J. Black, Pose-conditioned joint angle limits for 3D human pose reconstruction, in: Computer Vision and Pattern Recognition, 2015, pp. 1446–1455.
[15] L. Sigal, M. Isard, H. Haussecker, M.J. Black, Loose-limbed people: estimating 3D human pose and motion using non-parametric belief propagation, Int. J. Comput. Vis. 98 (1) (2012) 15–48.
[16] M. Hofmann, D.M. Gavrila, Multi-view 3D human pose estimation combining single-frame recovery, temporal integration and model adaptation, in: IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 2214–2221.
[17] L. Sigal, M.J. Black, HumanEva: synchronized video and motion capture dataset for evaluation of articulated human motion, Int. J. Comput. Vis. 87 (1–2) (2006) 4–27.
[18] C. Ionescu, D. Papava, V. Olaru, C. Sminchisescu, Human3.6M: large scale datasets and predictive methods for 3D human sensing in natural environments, IEEE Trans. Pattern Anal. Mach. Intell. 36 (7) (2014) 1325–1339.
[19] Y. Kim, D. Kim, A CNN-based 3D human pose estimation based on projection of depth and ridge data, Pattern Recognit. 106 (2020) 107462.
[20] D. Sun, C. Zhang, A multi-view 3D human pose estimation algorithm based on positional attention, in: Intelligent Computing and Signal Processing, 2022, pp. 125–128.
[21] L. Cui, L. Bai, Y. Wang, S.Y. Philip, E.R. Hancock, Fused lasso for feature selection using structural information, Pattern Recognit. 119 (2021) 108058.
[22] M. Jiang, Z. Yu, Y. Zhang, Q. Wang, C. Li, Y. Lei, Reweighted sparse representation with residual compensation for 3D human pose estimation from a single RGB image, Neurocomputing 358 (C) (2019) 332–343.
[23] I. Katircioglu, B. Tekin, M. Salzmann, V. Lepetit, P. Fua, Learning latent representations of 3D human pose with deep neural networks, Int. J. Comput. Vis. 126 (12) (2018) 1–16.
[24] Y. Wang, J. Song, K. Zhou, Y. Liu, Unsupervised deep hashing with node representation for image retrieval, Pattern Recognit. 112 (2021) 107785.
[25] Mocap: Carnegie mellon university motion capture database, http://mocap.cs.cmu.edu/.
[26] M. Andriluka, L. Pishchulin, P. Gehler, B. Schiele, 2D human pose estimation: new benchmark and state of the art analysis, in: Computer Vision and Pattern Recognition, 2014, pp. 3686–3693.
[27] H. Yasin, U. Iqbal, B. Krüger, A. Weber, J. Gall, A dual-source approach for 3D pose estimation from a single image, in: Computer Vision and Pattern Recognition, 2016, pp. 4948–4956.
[28] S. Zhou, M. Jiang, Q. Wang, Y. Lei, Towards locality similarity preserving to 3D human pose estimation, in: Asian Conference on Computer Vision, 2020.
[29] F. Morenonoguer, 3D human pose estimation from a single image via distance matrix regression, in: Computer Vision and Pattern Recognition, 2017, pp. 1561–1570.
[30] Z. Yu, B. Ni, J. Xu, J. Wang, C. Zhao, W. Zhang, Towards alleviating the modeling ambiguity of unsupervised monocular 3D human pose estimation, in: International Conference on Computer Vision, 2021, pp. 8651–8660.
[31] D. Novotny, N. Ravi, B. Graham, N. Neverova, A. Vedaldi, C3DPO: Canonical 3D pose networks for non-rigid structure from motion, in: Proceedings of IEEE International Conference on Computer Vision, 2019, pp. 7688–7697.
[32] H. Rhodin, M. Salzmann, P. Fua, Unsupervised geometry-aware representation for 3D human pose estimation, in: European Conference on Computer Vision, 2018, pp. 750–767.
[33] C.-H. Chen, D. Ramanan, 3D human pose estimation = 2D pose estimation+ matching, in: Computer Vision and Pattern Recognition, 2017, pp. 5759–5767.
[34] X. Zhou, X. Sun, W. Zhang, S. Liang, Y. Wei, Deep kinematic pose regression, in: European Conference on Computer Vision, 2016, pp. 186–201.
[35] H.-Y.F. Tung, A.W. Harley, W. Seto, K. Fragkiadaki, Adversarial inverse graphics networks: learning 2D-to-3D lifting and image-to-image translation from unpaired supervision, in: International Conference on Computer Vision, 2017, pp. 4364–4372.
[36] J. Wu, T. Xue, J.J. Lim, Y. Tian, J.B. Tenenbaum, A. Torralba, W.T. Freeman, Single image 3D interpreter network, in: European Conference on Computer Vision, 2016, pp. 365–382.
[37] K. Wang, L. Lin, C. Jiang, C. Qian, P. Wei, 3D human pose machines with self–supervised learning, IEEE Trans. Pattern Anal. Mach. Intell. (2019) 1.
[38] F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero, M.J. Black, Keep it SMPL: automatic estimation of 3D human pose and shape from a single image, in: European Conference on Computer Vision, 2016, pp. 561–578.
[39] M. Lin, L. Liang, X. Liang, K. Wang, C. Hui, M. Lin, L. Liang, X. Liang, K. Wang, C. Hui, Recurrent 3D pose sequence machines, in: Computer Vision and Pattern Recognition, 2017, pp. 5543–5552.
[40] B.X. Nie, P. Wei, S.C. Zhu, Monocular 3D human pose estimation by predicting depth on joints, in: International Conference on Computer Vision, 2017, pp. 3467–3475.

Mengxi Jiang received her BS degree in Computer Science and Technology from Fuzhou University, China, in 2013, MS degree in Computer Technology from Xi'an Polytechnic University, China, in 2016, and PhD. degree in Computer Science and Technology at Xiamen University, China, in 2021. She is interested in the research of computer vision, machine learning, and 3D pose estimation and recognition, etc.

Shihao Zhou received the BS degree in Computer Science and Technology from Zhejiang Gongshang University, China, in 2018. From 2018 to now, he is studying at Xiamen University for MS degree in Computer Science and Technology. He is interested in the research of computer vision, machine learning, and 3D human pose estimation.

Cuihua Li received the BS degree in computational mathematics from Shandong University, China, in 1983, MS degree in computational mathematics and PhD in automatic control theory and engineering from Xi'an Jiaotong University, China, in 1989 and 1999, respectively. He was an Associate Professor with the School of Science, Xi'an Jiaotong University before 1999. He is currently with the Department of Computer Science, Xiamen University, Xiamen, China. His current research interests include computer vision, video and image processing, and super-resolution image reconstruction algorithms.

Yunqi Lei received his BS degree in electronics from University of Science and Technology of China, MS degree in marine electric engineering from University of The Navy Engineering, China, and the PhD degree in automation from National University of Defense Technology, China, in 1982, 1984 and 1988, respectively. His current research interests include deep learning, computer vision and image processing, big data, cloud computing and computer networks. As the group leader, he is now conducting a project of China NSF regarding manifold learning and 3D object recognition.