Towards Locality Similarity Preserving to 3D Human Pose Estimation

Shihao Zhou[0000−0002−9202−9761], Mengxi Jiang, Qicong Wang, and Yunqi Lei✉

Department of Computer Science, Xiamen University, Xiamen 361005, China
{shzhou,jiangmengxi}@stu.xmu.edu.cn, {qcwang,yqlei}@xmu.edu.cn

Abstract. Estimating 3D human pose from an annotated or detected 2D pose in a single RGB image is a challenging problem. A successful way to address this problem is the example-based approach. The existing example-based approaches often calculate a global pose error to search a single match 3D pose from the source library. This way fails to capture the local deformations of human pose and highly dependent on a large training set. To alleviate these issues, we propose a simple example-based approach with locality similarity preserving to estimate 3D human pose. Specifically, first of all, we split an annotated or detected 2D pose into 2D body parts with kinematic priors. Then, to recover the 3D pose from these 2D body parts, we recombine a 3D pose by using 3D body parts that are split from the 3D pose candidates. Note that joints in the combined 3D parts are refined by a weighted searching strategy during the inference. Moreover, to increase the search speed, we propose a candidate selecting mechanism to narrow the original source data. We evaluate our approach on three well-design benchmarks, including Human3.6M, HumanEva-I, and MPII. The extensive experimental results show the effectiveness of our approach. Specifically, our approach achieves better performance than compared approaches while using fewer training samples.

Keywords: 3D human pose estimation · Locality Similarity

1 Introduction

3D human pose estimation from a single RGB image is quite an important task in the field of computer vision with a variety of practical applications, such as human-robot interaction, virtual reality, activity recognition, and abnormal behavior detection [1–6]. Estimating 3D human pose from a single image is a typical ill-posed problem since similar projections in low dimension may be derived from different 3D poses. To alleviate this problem, a wide range of approaches with different strategies have been introduced in recent years.

Most of the existing literature apply discriminative strategy, with its representative work (*i.e.*, neural networks model) [7–11]. Generally, these learning-based approaches intend to learn a mapping between images and 3D human poses with plenty of 2D-3D paired examples. As a result, a large number of training data are required to learn a satisfactory mapping function.

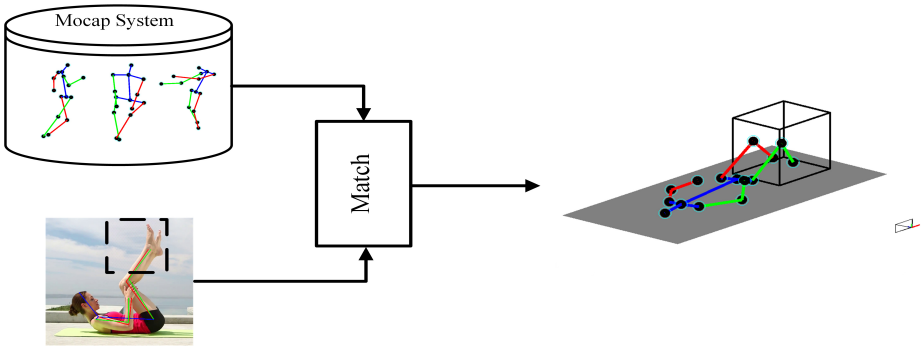

Fig. 1. Impact of global matching strategy to standard example-based approach.

The other branch of discriminative strategy is example-based approaches [12–14]. These works usually search for a 3D human pose from the source data by calculating the global pose error rather than learning a mapping function. Such a global matching strategy may fail to capture the deformation of the local part in a human pose, as shown in Fig.1. In this figure, the traditional example-based approach searches for the best match with minimum joints position error among the library. Moreover, this issue can be aggravated if there are insufficient source data. Intuitively, these approaches obtain a pose with a high global score, but fails in capturing the local deformations. Of course, this issue can be alleviated by augmenting more diverse samples (*e.g.*, different scenes, subjects, and actions) to the source library. However, since the deployment of capture equipment is constrained by the outdoor environment, it is difficult to obtain 3D annotations of real scenarios. Existing widely used datasets (*e.g.*, Human3.6M [15] and HumanEva-I [16]) collect 2D and 3D annotated poses performed by a few of subjects with specific actions indoors. Thus, available rich source samples are limited. Though some data augmentation strategies are proposed [17, 18] for synthesizing the training examples, there is still a gap between the synthesized and complex real scenario data on the diversity [19]. As a result, in order to alleviate these issues, we propose a novel and simple example-based model to estimate the 3D pose.

Considering the fact that a 3D pose obtained by minimizing global pose error may mismatch in local parts (*i.e.*, small global pose error and relatively large error in pose local parts, as shown in Fig.1), while two poses with huge global pose error may perfect match in a few of local parts. Inspired by these observations, we argue that combine similar local parts from different poses could obtain pose with similar parts preserving and further boost the performance. However, the traditional example-based approach usually search a single pose match using the global error, which often fails to capture similar local part. Therefore, in this paper, instead of searching for a single best match as traditional example-based ways, we proposed to search for multiple candidates with similar local parts. More specifically, we split a pose into different parts and search the matched

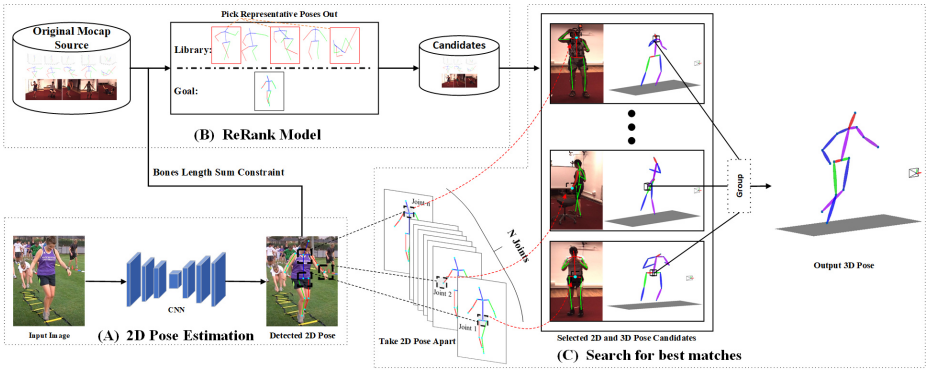

Fig. 2. Overview. Our approach can be mainly divided into three parts. The first part is about 2D pose estimation. We train a 2D detector Hourglass Network for 2D poses estimation. After that, we take both the detected 2D pose and the original mocap source into the reRank model. With a bone length sum constraint, we filter a lot of poses out, and narrow the training source into a quite small candidate set. Finally, we take the estimated 2D pose into different parts and search the best match for every single point with a weight mechanism. Then, we group different key points from all searched best matches into a whole 3D pose. Since we get a final pose from different candidates, we can nearly avoid generating predictions in Fig.1.

poses for each part. Then, we use key points extracted from different parts (*e.g.*, a limb or a trunk) to group a pose, as illustrated in Fig.2.

We evaluate our approach on several datasets, including indoor 3D datasets Human3.6M [15] and HumanEva-I [16] for quantitative evaluation and outdoor 2D dataset MPII [20] for qualitative evaluation. On all the evaluation datasets, we achieve competitive results to example-based approaches and our approach even outperforms some learning-based ways.

The main contributions of our work include:

(1) In this study, to alleviate the pose local structure match issue, we propose a simple example-based approach with locality similarity preserving. This work, for the first time, introduces a novel body parts match of splitting a 2D and 3D human pose into the body parts with kinematic priors.

(2) In our approach, a source library narrowing strategy is designed to increase the searching speed, while remaining the most representative candidates.

(3) Extensive evaluated experiments are conducted on three public benchmarks. Our approach achieves superior estimation performance than considered comparison approaches. Especially, comparing with several approaches, our approach uses fewer training samples while obtaining better estimation accuracy.

2 Related Work

3D human pose estimation has been a well-studied problem these years. Traditionally, approaches proposed to solve this problem can be simply divided

into two classes, generative ways [21–26] and discriminative ones [14, 13, 27, 28]. The greatest strength for generative approaches is that they need quite a small size of the training set. While discriminative approaches generally take as input plenty of 2D-3D training samples. Moreover, discriminative ways can be classified into two subtypes, learning-based approachess [29–33] and example-based ones [34–36]. Prior to these traditional approaches, recent works raise interest in weak or unsupervised learning for wild human pose estimation [37–40]. Some other approaches combine tasks for both shape and pose estimation together [41–43].Considering the graph nature of human body, [28, 44] make a great breakthrough by introducing graph operation. Although numerous methods have boosted the interest for 3D human pose estimation, we will focus our review on example-based pose estimation.

2.1 Example-based Approaches

Exampled-based approaches, also called pose retrieval methods, intended to search for the best match from the library with the goal one. Such methods benefit in fitness on anthropology from searching in training source, composed completely of the realistic human 3D pose. While other kinds of approaches (*i.e.* learning-based approaches and generative ways) are more likely to predict human poses as an outlier with less mean per joint position errors (MPJPE) [45, 46]. Noted the performance of the example-based ways may not as good as learning-based approaches, but the obtained poses based on sample retrieval are usually more in line with the physical constraints. Moreover, example-based approaches never generate unreasonable pose, which is the main flaw of learning-based ways. So this topic is still worth exploring.

Recently, exampled-based works have attempted to acquire better results via introducing networks or other priors. Yasin et al. [14] utilized a dual-source approach to do 3D pose estimation from a single image and became the-state-of-art at that time. Chen and Ramanan [13] proposed to split the task 3D poses estimation apart into 2D pose estimation and matching to the 3D poses in Library. Li et al. [34] introduced a deep-network to do pose prediction auxiliary task and transferred the prediction problem into maximum-margin structured learning. However, these traditional retrieval approaches usually deal with the whole human pose, which hardly captures local deformations. Therefore, we propose to solve this by introducing a local similarity preserving strategy.

2.2 Locality Similarity Preserving for Human Pose Estimation

Many works have made efforts to keep local part preserving. For preserving local structure in the original space, Tian et al. [47] introduced Latent Variable Models to learn latent spaces for both image features and 3D poses. Fan et al. [48] developed a block-structural pose dictionary to explicitly encourage pose locality in human-body modeling. Rogez and Schmid [12] selected different images patches for a 3D pose in the library via keeping their 2D pose similar. Yasin [35] learned a 3D local pose model in low Principle Component Analysis space via

retrieving nearest neighbors. Zhou et al. [7] embedded kinematic function as part of the network for fully exploited geometric validity. Varolet al. [17] addressed segmentation by training a pixel-wise classifier. Tang and Wu [30] trained a part-based branching network for leaning specific feature. ke et al. [49] proposed a local refinement part for more compact limbs.

However, they are either greedy for plenty of training data, which without considering unaffordable computational cost, or aiming for local compact structure in the 2D dimension [50]. Hence, our approach designs a new model for simple and fast pose estimation with keeping local similarity. Instead of simply implementing constraints or synthesizing data for learning-based ways, we redesign the search rule from the whole body into local parts.

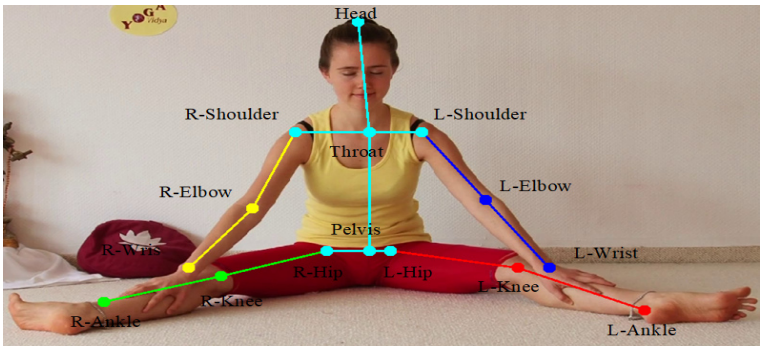

Fig. 3. Skeleton. As shown in the figure, a human pose can be presented with 15 joints and divided into various five groups. Note that, joints belong to the same group are painted with the same color.

3 Background

In this section, we introduce the background knowledge of this paper, including the problem definition and the standard example-based approaches [14, 13].

In this paper, we represent a human body as a skeleton with N joints. The 2D and 3D pose corresponding to the human body are denoted as $\boldsymbol{X} = \{x_i\}_{i=1}^{N} \in \mathbb{R}^{2\times N}$ and $\boldsymbol{Y} = \{y_i\}_{i=1}^{N} \in \mathbb{R}^{3\times N}$ respectively.

To estimate a 3D human pose from an annotated or detected 2D pose $\boldsymbol{X}$, traditional example-based ways firstly search for the best match through minimizing the distance between the annotated or detected 2D pose and the retrieved 2D pose $\hat{\boldsymbol{X}} = \{\hat{x_i}\}_{i=1}^{N} \in \mathbb{R}^{2\times N}$. Normally, the distance function in 2D space can be formulated as:

$$\arg\min_{\hat{\boldsymbol{X}}} \|x_i - \hat{x_i}\|_2 \tag{1}$$

where x_i and $\hat{x_i}$ denote the i-th joint position of the annotated or detected 2D pose and the retrieved pose, respectively. Then, by minimizing $Eq.(1)$, the

traditional example-based approach obtains a best global match 2D pose whose corresponding 3D pose is the final 3D prediction. However, two poses with the global minimum error may have large local differences in some body parts as Fig.1 shown, and the estimation performance of the traditional examples-based approach is affected since they fail to capture the local deformation.

4 Proposed Model

To alleviate the issue discussed above, we propose our solution in this paper. Specifically, firstly, we propose to represent a whole 2D human pose as different parts that satisfy human body kinematic constraints. Then, instead of searching a single match 3D pose from the source library as most of the example-based approaches, we retrieve body parts that come from different source poses. In this work, our aim is to enable our model to preserve the locality similarity of the local parts within human bodies for pose estimation.

4.1 Human Pose Split Using Kinematic Priors

Kinematic priors [51, 7] interpret the interrelationship between the body components, which capture the inherent connectivity of human pose. In this paper, in order to split the 2D human pose while preserving correlation within the local structure, we provide local structural descriptions of an annotated or detected 2D human pose in terms of the kinematic priors. Specifically, considering that the arms and feet of the human body belong to different local parts, a human pose can be divided into various groups. For example, a pose with 15 joints can be divided into five sub-groups in total, including: (1) right elbow and right wrist, (2) left elbow and left wrist, (3) right knee and right ankle, (4) left knee and left ankle, and (5) head, throat, right shoulder, left shoulder, pelvis, right hip and left rip, as shown in Fig.3. As a result, to formulate such local structure, a 2D human posture with N joint nodes can be rewritten as $\boldsymbol{X} = \{\boldsymbol{P}_k\}_{k=1}^{K}$, where K indicates the number of the subgroups, $\boldsymbol{P}_k$ is the specific part. Such a split strategy allows us to search for the best match through part by part.

4.2 The Locality Similarity Preserving for Human Pose

In our retrieval strategy, instead of searching the whole pose as traditional example-based approaches, we retrieve parts $\hat{\boldsymbol{P}}_k$ for locality similarity preserving.

$$\arg\min_{\hat{\boldsymbol{P}}_k} \|\boldsymbol{P}_k - \hat{\boldsymbol{P}}_k\|_2 \tag{2}$$

By minimizing $Eq.(2)$, we obtain the best match part $\hat{\boldsymbol{P}}_k$ for each part $\boldsymbol{P}_k$. In practice, $Eq.(2)$ is too strict to be satisfied due to the annotated or detected 2D pose is usually wild (*i.e.* the annotated or detected 2D pose is not included in the indoor source library). Thus, it may leads to a inferior retrieval part.

To further improve the parts retrieval performance, we aim to search each joint in the part carefully rather than a whole part. To obtain each joint in the part, we future reformulate $Eq.(2)$ as:

$$\arg\min_{\hat{x}_a} \boldsymbol{W}_k^a * \|\boldsymbol{P}_k - \hat{\boldsymbol{P}}_k\|_2, \ \forall a \in \{1, 2, ..., M_k\} \tag{3}$$

where $\hat{x}_a \in \hat{\boldsymbol{P}}_k$ denotes the query joint in the k-th part. $\boldsymbol{W}_k^a = \{\omega_k^{an}\}_{n=1}^{M_k}$ is used to increase the impact of adjacent joints of $\hat{x}_a$, which helps to find a better match [12]. M_k is the number of joints in the k-th part. Moreover, the weight ω_k^{an} is calculated by

$$\omega_k^{an} = \begin{cases} \dfrac{1}{\|x_a - x_n\|_2} + \dfrac{1}{\|\hat{x}_a - \hat{x}_n\|_2}, & a \neq n \\ 0, & a = n \end{cases} \tag{4}$$

where $\hat{x}_n$ is the n-th joint in k-th part, $n \in \{1, 2, ..., M_k\}$. x_a, x_n are the corresponding joints of $\hat{x}_a, \hat{x}_n$ in the given 2D parts. ω_k^{an} is the sum of inversely proportional to the distance between joint n-th and the a-th joint. $Eq.(4)$ implies that the joint closer to the a-th joint is given a higher weight.

By minimizing the $Eq.(3)$ M_k times, we obtain a list of M_k match joints $\{\hat{x}_a\}_{a=1}^{M_k}$. Since parts are paired (2D-3D) samples in source libary, we can acquire another list of M_k match joints in 3D space at the same time and denote as $\{\hat{y}_a\}_{a=1}^{M_k}$, which is illustrated in (B) of Fig.2. By combining these joints $\{\hat{x}_a\}_{a=1}^{M_k}$, we obtain the final k-th part as

$$\hat{\boldsymbol{P}}_k = \{\hat{x}_a\}_{a=1}^{M_k} \tag{5}$$

Similarly, we can obtain the final k-th part in according 3D space as:

$$\hat{\boldsymbol{Q}}_k = \{\hat{y}_a\}_{a=1}^{M_k} \tag{6}$$

where $\hat{\boldsymbol{Q}}_k$ is the k-th part in 3D space, and $\hat{y}_a \in \hat{\boldsymbol{Q}}_k$ denotes the 3D coordinate of query joint in the k-th part. After retrieving the all K parts $\hat{\boldsymbol{Q}}_k$, we assemble them into one pose as the final prediction Y.

Given a source pose library $\boldsymbol{\mathcal{S}}$, we can search the joints in $\hat{\boldsymbol{P}}_k$ and $\hat{\boldsymbol{Q}}_k$. Therefore, the speed of the workflow is largely influenced by the size of the source pose library $\boldsymbol{\mathcal{S}}$. In order to increase the search speed, we propose a candidate mechanism to narrow the source library $\boldsymbol{\mathcal{S}}$. By introducing a weak physical constraint, we improve the retrieval speed by selecting the representative source poses from the library $\boldsymbol{\mathcal{S}}$. Especially, our strategy can be formulated as:

$$\boldsymbol{C} = \arg\min_{\boldsymbol{S}}(L(\boldsymbol{X}), L(\boldsymbol{S})), \ \boldsymbol{S} \in \boldsymbol{\mathcal{S}} \tag{7}$$

Algorithm 1 The locality similarity preserving algorithm

Input: $\boldsymbol{X}$, $\boldsymbol{\mathcal{S}}$ //2D human pose,source pose libary
Output: $\boldsymbol{Y}$ //Predicted 3D human pose
Parameter: N, K, M_k //joints of a human pose, parts of the split strategy, joints in each part

```
Calculate C by Eq.(7).
Split X into K parts with kinematic priors.
for k = 1 to K do
    for a = 1 to M_k do
        Calculate x̂_a by Eq.(3).
    end for
    Calculate P̂_k by Eq.(5).
    Calculate Q̂_k by Eq.(6).
end for
Calculate Y = {Q̂_k}_{k=1}^K.
```

where $\boldsymbol{C} = \{\boldsymbol{S}_1^c, \boldsymbol{S}_2^c, ..., \boldsymbol{S}_{N_c}^c\}$ represents the candidate set including a set of representative source poses, $\boldsymbol{S}_1^c$ is one of the poses, N_c is the number of candidate pose in the $\boldsymbol{C}$. $L(\boldsymbol{X})$, $L(\boldsymbol{S})$ are the sum of bones length to the annotated or detected 2D pose and the retrieval 2D pose, and $L(\boldsymbol{X}) = \sum_{i=1}^{N-1} \|x_i - parent(x_i)\|_2,\ x_i \in \boldsymbol{X}$. $\|x_i - parent(x_i)\|_2$ is the length of the joint x_i to its parent in 2D pose $\boldsymbol{X}$. By minimizing $Eq.(7)$, we pick up poses with similar bones length sum from $\boldsymbol{\mathcal{S}}$, and these poses are combined as candidate set $\boldsymbol{C}$. As a result, we narrow source pose library $\boldsymbol{\mathcal{S}}$ into a limited set, which accelerates the searching speed.

The complete workflow of the proposed approach is described in Algorithm1.

5 Experiment

Since various datasets provide different human pose representation, we unify all these original poses into a single one, with N = 15 points. We quantitatively evaluate our approach on two different datasets: **Human3.6M** dataset [15], one of the largest indoor human pose datasets, and **HumanEva-I** dataset [16]. We qualitatively verify our approach on **MPII** dataset [20]. In this paper, to verity the effectiveness of the proposed approach, we conduct the experiments using annotated and detected 2D pose as input, respectively. Following the same setting of previous works [10, 37, 39], we apply the stacked hourglass network (SH) [52] to obtain the detected 2D pose for the fair comparison. Note that SH is pre-trained on the MPII dataset at first and then fine-tuned on the Human3.6M dataset to be in line with the literature [10, 37, 39].

5.1 Datasets

Human3.6M. It is a large-scale indoor dataset with 3D annotations. For there are many protocols proposed these years and it is quite difficult to perform

Table 1. Mean reconstruction errors (mm) under Protocol #1 and mean per joint errors (mm) under Protocol #2 of Human3.6M. − indicates that the result for the specific action is not reported. 'GT' means ground truth (annotated) 2D input. 'DT' means detected 2D input. 'IM' means image input. † indicates learning-based approach. Best result in bold.

Protocol #1	Dir.	Disc.	Eat	Greet	Phone	Pose	Purch.	Sit	SitD.	Smoke	Photo	Wait	Walk	WalkD.	WalkT.	AVG
Rogez (IM) (NIPS'16) [12]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	88.1
†Nie et al. (DT) (ICCV'17) [8]	**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
Chen (DT) (CVPR'17) [13]	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
†Moreno-Noguer (DT) (CVPR'17) [9]	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**
ADSA (DT) (CVPR'16) [14]	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
Ours (DT)	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
ADSA (GT)(CVPR'16) [14]	60.0	**54.7**	71.6	67.5	63.8	**61.9**	**55.7**	**73.9**	110.8	**78.9**	96.9	**67.9**	**47.5**	89.3	53.4	70.5
Ours (GT)	**59.1**	63.3	**70.6**	**65.1**	**61.2**	73.2	83.7	84.9	**72.7**	84.3	**68.4**	81.9	57.5	**75.1**	**49.6**	**70.0**
Protocol #2	Dir.	Disc.	Eat	Greet	Phone	Pose	Purch.	Sit	SitD.	Smoke	Photo	Wait	Walk	WalkD.	WalkT.	AVG
Li et al. (IM) (ICCV'15) [34]	-	149.1	109.9	136.9	-	-	-	-	-	-	179.9	-	83.6	147.2	-	135.6
†Du et al. (IM) (ECCV'16) [32]	85.1	112.7	104.9	122.1	139.1	105.9	166.2	117.5	226.9	120.0	135.9	117.7	99.3	137.4	106.5	126.5
Rogez (IM) (NIPS'16) [12]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	121.2
†Zhou et al. (IM) (ECCVW'16) [7]	91.8	102.4	97.0	98.8	113.4	90.0	**93.9**	132.2	159.0	106.9	125.2	94.4	79.0	126.0	99.0	107.3
Chen (DT) (CVPR'17) [13]	89.9	97.6	90.0	107.9	107.3	93.6	136.1	133.1	240.1	106.7	139.2	106.2	87.0	114.1	90.6	114.2
†Kudo et al. (DT) (arXiv'18) [37]	161.3	174.3	143.1	169.2	161.7	180.7	178.0	170.6	191.4	157.4	174.1	182.3	180.3	180.7	193.4	173.2
†Novotnỹ et al. (DT) (ICCV'19) [10]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	153.0
†Wandt et al. (DT) (CVPR'19) [39]	77.5	85.2	82.7	93.8	93.9	82.9	102.6	**100.5**	125.8	88.0	**101.0**	84.8	72.6	**78.8**	79.0	89.9
†Li et al. (IM) (ICCV'19) [11]	**70.4**	83.6	76.6	77.9	85.4	**72.3**	102.9	115.8	165.0	82.4	106.1	**74.3**	**60.2**	94.6	**70.7**	**88.8**
ADSA (DT) (CVPR'16) [14]	97.3	103.2	97.2	110.4	115.1	127.3	90.7	104.6	160.2	173.8	103.0	117.2	99.7	93.1	94.9	112.5
Ours (DT)	75.5	**80.0**	**75.3**	**71.8**	**77.0**	84.3	97.2	105.4	**101.0**	**78.1**	132.3	96.5	92.8	88.8	79.2	89.5
†Kudo et al. (GT) (arXiv'18) [37]	125.0	**44.4**	107.2	65.1	115.1	147.7	128.7	134.7	139.8	114.5	127.3	147.1	125.6	130.8	151.1	130.9
†Novotnỹ et al. (GT) (ICCV'19) [10]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	101.8
ADSA (GT)(CVPR'16) [14]	**80.5**	77.0	**72.1**	90.4	92.1	103.1	84.7	**72.0**	103.9	107.1	**87.5**	83.1	84.6	79.8	67.6	85.7
Ours (GT)	88.1	64.3	73.0	**62.1**	**84.4**	**77.1**	**70.8**	96.3	**89.9**	**68.8**	128.5	**62.7**	**65.9**	**64.8**	**67.5**	**79.8**

a comprehensive comparison to all the existing experiments. We followed the standard protocol according to [14] noted as Protocol #1, and Protocol #2 from [33]. Under Protocol #1, training is performed on subjects (1,5,6,7,8,9), and the valid set consists of subject (11). While the train set is made up of subjects (1,5,6,7,8) and test on the subjects (9,11) is Protocol #2.

HumanEva-I. It is a common used benchmark with annotated 3D pose. Following the same setting of pervious works [14, 53, 23], we take all the training sequences as input while validate on the "walking" and "jogging" actions.

MPII. It is an in-the-wild dataset with 2D annotations. For only 2D annotation is provided, we perform qualitatively validation on it.

5.2 Comparison Approaches

A dual-source approach (**ADSA**) proposed by Yasin et al. [14], which used images with annotated 2D poses and accurate 3D motion capture data to do 3D pose estimation from a single image. ADSA is a standard example-based approach, in which a single 3D pose in the source library is selected as the final best match by minimizing the global pose error. In our approach, in order to preserve the local part similarity, we combine a 3D pose by using multiple candidates that are searched via a weighted similarity mechanism. Moreover, recent three other example-based approaches [13, 34, 12] and several representative works [7–11, 23, 24, 32, 37, 39, 53, 54], including generative and discriminative ways, are also considered in the comparison.

Table 2. Mean reconstruction errors (mm) under Protocol #1 of Human3.6M. Comparison to example-based approaches with different size of training data. Best result in bold.

Method	2D source	3D source	AVG
Rogez (NIPS'16) [12]	207k	190k	88.1
Chen (CVPR'17) [13]	180k	180k	82.4
ADSA (CVPR'16) [14]	64,000k	380k	108.3
Ours	375k	375k	**76.1**

5.3 Evaluation Protocols

There are two popular criterions to evaluate the pose estimation accuracy, the **per joint error** and the **reconstruction error** [21]. The **per joint error** calculates the average Euclidean distance of the estimated joints to ground truth. While the **reconstruction error** makes the same calculation but with a rigid transformation. Following the same evaluation protocols in most literature [8–14, 24, 54], we take per joint error as the evaluation metric for Human3.6M Protocol #1 while reconstruction error for HumanEva-I and Human3.6M Protocol #2.

5.4 Quantitative Evaluation on Human3.6M

We first report the results of our approach and the representative works both under Protocol #1 and #2 in Table 1. It is easy to find out that there is a huge promotion between our approach and the baseline (ADSA). More specifically, under Protocol #1, we demonstrate that our pipeline on the standard benchmark with a relative error reduction greater than **30%** to the baseline (ADSA) and 13% on average to other compared approaches. Under Protocol #2, our approach also outperforms the baseline (ADSA) by 20%. Moreover, as expected our proposed approach outperforms all example-based approaches [12–14, 34]. It should be noted that our approach can even beyond many learning-based approaches [7, 8, 32, 37], and comparative to some recent representative works [10, 39]. Even though our proposed approach performs slightly worse than learning-based approaches [9, 39] on average, we outperform both of the two approaches on more than half categories. While [39] trains subject-wise models relying on multiple priors and [9] use more training data (400k), comparing to ours. Moreover, both [9, 39] report their results after a fine-tuning process while our approach needs no hyperparameter and training stage. Noted that, all compared results are taken from original papers except for [14] under Protocol #2, which we implement with their publicly available code.

As mentioned above, we achieve quite competitive results on both two protocols. Under Protocol #1, our approach can achieve the best results for only several actions, while the average results show that our approach can be competitive with the learning-based ways [9, 8]. We attribute this to effective of our split strategy, which is easier to handle challenges of intricate actions comparing to the previous approaches, as shown in Fig.2. In general, intricate actions, such as

Table 3. Mean reconstruction errors (mm) on the HumanEva-I. 'GT' means ground truth (annotated) 2D input. 'DT' means detected 2D input. † indicates learning-based approach. Best result in bold.

Approaches	Walking			Jogging			AVG
	S1	S2	S3	S1	S2	S3	
Radwan et al. (DT) (ICCV'13) [24]	75.1	99.8	93.8	79.2	89.8	99.4	89.5
Wang et al. (DT) (CVPR'14) [54]	71.9	75.7	85.3	62.6	77.7	54.4	71.3
ADSA (DT)(CVPR'16) [14]	59.5	43.9	**63.4**	61.0	51.2	**55.7**	55.8
Ours (DT)	**39.1**	**21.2**	87.5	**39.1**	**48.2**	64.5	**53.2**
†Simo-Serra et al. (GT) (CVPR'13) [53]	65.1	48.6	73.5	74.2	46.6	**32.2**	56.7
Kostrikov et al. (GT) (BMVC'14) [23]	44.0	30.9	41.7	57.2	35.0	33.0	39.6
ADSA (GT)(CVPR'16) [14]	41.1	39.9	48.4	53.4	36.0	43.1	43.6
Ours (GT)	**27.3**	**13.2**	**37.6**	**44.0**	**34.1**	50.2	**34.4**

"SitDown", can be harder to make predictions for heavy self-occlusion than the simple ones like "Direction" or "Discuss". However, our proposed pipeline works for such challenging actions, which is difficult to traditional example-based ways [14, 13]. Under Protocol #2, beyond greater scores in most categories, we also found out that our approach meets slight degradation in some specific actions, such as "Sit" and "Photo". We argue that this may be attributed to the unrepresentative candidates. In our pipeline, we simply generate candidates with similarity bones sum, which is work for most cases. While we do not take other poses with similar shape but different scale into consideration. We believe the proposed approach could boost the performance by a margin through considering translate poses into a similar scale. Moreover, we found that actions like "Purchase" and "WalkDog" can not achieve the best results. This may due to strong occlusions of these specific categories.

Moreover, we can notice that there is no great gap after replacing the estimated 2D pose with ground truth. We argue that estimated 2D ground truth rather than reprojected one can explain this phenomenon, and also this can show that our approach still works with a slightly worse 2D estimated pose. Noted that, some recent works [31, 28] tend to take as input 2D pose reprojected via 3D pose with responding camera parameter, while this is impractical in the real world. Therefore we regard as input 2D pose the dataset provided, which is estimated and leads to less accuracy at first as ground truth and harder for making an accurate prediction. However, this input can be more close to a practical case.

We compare the results with different approaches take as input various size of the training set in Table 2. For our approach can be roughly classified into example-based ways, we are supposed to do the comparison with the same group, *e.g.*, [12–14, 34]. We can observe that our approach achieves the best scores with quite a small size of data as the training source. For this, we argue the strategy to take pose into pieces does help enlarge the search space. We believe this module improves performance, but in the same way, could increase the time to do the search process. Therefore, we conduct a relevant experiment to verify our thoughts in Sec.5.7.

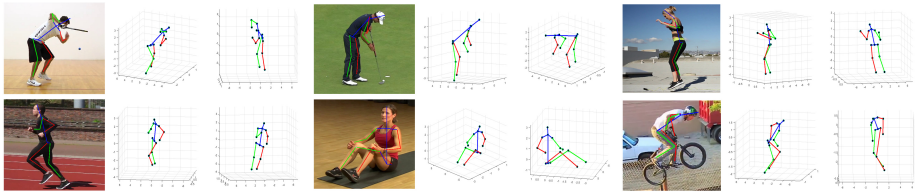

Fig. 4. Examples successes on MPII. For each example, the first column is the input image with its 2D pose, the second and the third columns are estimated 3D poses from different views.

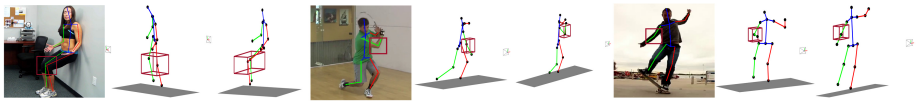

Fig. 5. Qualitative comparison. For each example, the images from left-to-right correspond to input the image with its 2D pose, estimated 3D poses from ADSA [14] and ours, respectively. Note that the reconstruction results of the body local parts are highlighted by red cubes.

5.5 Quantitative Evaluation on HumanEva-I

In this section, the results of HumanEva-I are presented. Since this dataset is quite small, few recent learning-based approaches treat it as significant as Human3.6M and conduct experiments on it [31]. Therefore, we compare representative learning-based ways (*e.g.*, [53]) with lightweight architecture to avoid overfitting. As shown in Table 3, we can outperform most works and achieve best for several sequences, except for the "Walking" and "Jogging" sequence of S3 subject. It is easy to find out many works degrade on the "Walking" sequence of subject S3, and this may due to inaccurate annotations that exist in the testing data [55]. Similar to human3.6M, due to the heavy self-occlusion, the performance of the proposed approach has also been affected ("Jogging" sequence of subject S3). Visual inspection of these results suggests that the extremely rare 3D poses are beyond the representational capacity of the model. Noted that all compared results are taken from original papers except for [14], which we implement with their publicly available code. For a fair comparison, we report the result with the same "CMU" skeleton [14].

5.6 Qualitative Evaluation on MPII

We also implement qualitative validation on MPII dataset. The successful results can be viewed in Fig.4. In Fig.5, the qualitative results from ADSA [14] and our approach are provided. It is clear observer that our approach achieve the better local part reconstruction than ADSA. Moreover, there are some failure cases are presented in Fig.6. This may due to great depth ambiguity and severe occlusions.

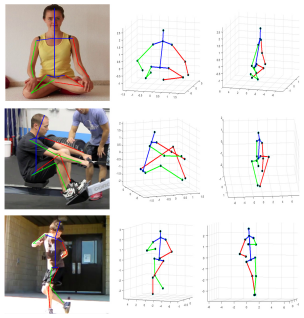

Fig. 6. Example fails on MPII. For this example, the first column is the input image with its 2D pose, the second and the third column are estimated 3D poses from different views.

Table 4. Per joint errors (mm) under Protocol #1 of Human3.6M. Comparison with various strategies for local similarity preserving. 'SH' means detected 2D input with Stacked Hourglass Networks. 'GT' means ground truth (annotated) 2D input. Best result in bold.

Strategy	MPJPE (SH)	MPJPE (GT)
Pose-Pose	110.0	101.4
Part-Part	92.5	75.2
Part-Joint	**76.1**	**70.0**

Table 5. Per joint errors (mm) under Protocol #1 of Human3.6M. Comparison with various strategies for adjacent joints weight calculation. 'SH' means detected 2D input with Stacked Hourglass Networks. 'GT' means ground truth (annotated) 2D input. Best result in bold.

Strategy	MPJPE (SH)	MPJPE (GT)
Pose	93.7	86.4
Part	**76.1**	**70.0**

5.7 Ablation Study

Different size of Candidate Set. Our approach introduces a module called reRank to narrow the searching space $\boldsymbol{\mathcal{S}}$ into a relatively small one C, as shown in (B) of Fig.2. Various settings of the size of the candidate N_c for C will generate different results and take diverse seconds to complete the search process. Therefore, we do a series of experiments to find out the best parameter for the whole process. As shown in Fig.7, the MPJPE decreases with the growth of N_c, while the time cost surge at the same time. Thus, we are not simply increasing N_c, but choose a proper value for good performance with sustainable time cost, and in this experiment, we set N_c as 35,000 for comparison in Table 1.

Different local similarity preserving strategies. We compare different strategies to estimate 3D human pose, including: **(i)** searching the whole pose via $Eq.(1)$ as the prediction (denoted as Pose-Pose); **(ii)** searching pre-defined body parts via $Eq.(2)$ and assemble the 3D pose (denoted as Part-Part); and **(iii)** searching joints via $Eq.(5)$ from different parts to reconstruct the 3D pose (denoted as Part-Joint). Our proposed strategy searching joints **(iii)** achieves **31%** improvement than searching the whole pose **(i)** , and **7%** improvement than searching parts **(ii)** , which is shown in Table 4.

Different adjacent joints weights W_k^a. We take different ways to calculate the weight of adjacent joints, including: **(i)** treating joints in the whole pose as neighbors in $Eq.(4)$ to calculate the similarity (denoted as Pose); **(ii)** treating joints in the pre-defined body part as neighbors to do the calculation in $Eq.(4)$ (denoted as Part). As shown in Table 5, the results demonstrate proposed adjacent joints weights calculation can pick up the similar local parts effectively.

Table 6. Cross-dataset validation. Mean reconstruction error (mm) on HumanEva-I given training source from Human3.6M (under Protocol #1). Best result in bold.

Size	Strategy	Walking			Jogging			AVG
		S1	S2	S3	S1	S2	S3	
30k	Pose-Pose	104.1	108.1	138.9	125.3	129.2	135.7	123.6
	Part-Joint	65.6	74.4	94.8	85.3	87.5	**88.6**	82.7
375k	Pose-Pose	77.5	76.4	107.3	91.7	96.4	105.7	92.5
	Part-Joint	**65.4**	**69.7**	**91.9**	**80.5**	**84.0**	89.2	**80.1**

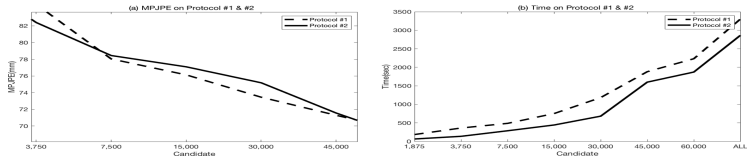

Fig. 7. Different size of Candidate Set. Fig.7 (a) and (b) show the different parameter C can impact the whole workflow on accuracy and time cost.

Cross-dataset validation. To further verify the generalization of the proposed approach, we quantitatively evaluate the cross-dataset ability, in which we perform accuracy evaluation on HumanEva-I given training source from Human3.6M (under Protocol #1). We take the ground-truth (annotated) 2D pose as input. Various sizes of training sources (30k and 375k) and different local similarity preserving strategies (Pose-Pose and Part-Joint) are taken into consideration. As shown in Table 6, increasing the number of training sources would boost the accuracy generally. However, different from the Pose-Pose strategy, our strategy (Part-Joint) is insensitive to the training source. More specifically, though both two strategies obtain promotion with more training sources, there is a huger gap for the Pose-Pose strategy. Moreover, with fewer training sources, our pipeline still performs superior. Noted that our strategy performs better with 30k training size on "Jogging" of S3, and this may due to the representative candidates.

6 Conclusion

In our work, we propose a simple yet effective approach to estimate 3D human pose by taking the 2D pose apart and searching a group of 3D poses to assemble a new one. Noted, our approach has dramatically reduced reliance on massive samples and improve the performance. Extensive experiments demonstrate the effectiveness of our approach. For example, under Protocol #1 of Human3.6M, we achieve a relative error reduction greater than **30%** to ADSA and **13%** on average to other compared approaches. More interesting, our approach yields competitive results with only 1% training data of ADSA.

Acknowledgement. This research was supported by the National Natural Science Foundation of China (Grant no 61671397).

References

1. Wang, Y., Xiao, Y., Xiong, F., Jiang, W., Cao, Z., Zhou, J.T., Yuan, J.: 3dv: 3d dynamic voxel for action recognition in depth video. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2020) 508–517
2. Wang, Z., Yu, P., Yang, Z., Zhang, R., Zhou, Y., Yuan, J., Chen, C.: Learning diverse stochastic human-action generators by learning smooth latent transitions. In: Proceedings of the AAAI Conference on Artificial Intelligence. Volume 34. (2020) 12281–12288
3. Weng, J., Liu, M., Jiang, X., Yuan, J.: Deformable pose traversal convolution for 3d action and gesture recognition. In: European Conference on Computer Vision (ECCV). (2018) 142–157
4. Tu, Z., Li, H., Zhang, D., Dauwels, J., Li, B., Yuan, J.: Action-stage emphasized spatiotemporal vlad for video action recognition. IEEE Transactions on Image Processing **28** (2019) 2799–2812
5. Tu, Z., Xie, W., Qin, Q., Poppe, R., Veltkamp, R.C., Li, B., Yuan, J.: Multi-stream cnn: Learning representations based on human-related regions for action recognition. Pattern Recognition **79** (2018) 32–43
6. Tu, Z., Xie, W., Dauwels, J., Li, B., Yuan, J.: Semantic cues enhanced multi-modality multistream cnn for action recognition. IEEE Transactions on Circuits and Systems for Video Technology **29** (2019) 1423–1437
7. Zhou, X., Sun, X., Zhang, W., Liang, S., Wei, Y.: Deep kinematic pose regression. In: European Conference on Computer Vision Workshops (ECCVW). (2016) 186–201
8. Nie, B.X., Wei, P., Zhu, S.: Monocular 3d human pose estimation by predicting depth on joints. In: Proceedings of IEEE International Conference on Computer Vision (ICCV). (2017) 3447–3455
9. Moreno-Noguer, F.: 3d human pose estimation from a single image via distance matrix regression. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2017) 1561–1570
10. Novotny, D., Ravi, N., Graham, B., Neverova, N., Vedaldi, A.: C3dpo: Canonical 3d pose networks for non-rigid structure from motion. In: Proceedings of IEEE International Conference on Computer Vision (ICCV). (2019) 7688–7697
11. Li, Z., Wang, X., Wang, F., Jiang, P.: On boosting single-frame 3d human pose estimation via monocular videos. In: Proceedings of IEEE International Conference on Computer Vision (ICCV). (2019) 2192–2201
12. Rogez, G., Schmid, C.: MoCap-guided Data Augmentation for 3D Pose Estimation in the Wild. In: Advances in Neural Information Processing Systems (NIPS). (2016) 3108–3116
13. Chen, C., Ramanan, D.: 3d human pose estimation = 2d pose estimation + matching. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2017) 5759–5767
14. Hashim, Y., Umar, I., Björn, K., Andreas, W., Juergen, G.: A dual-source approach for 3d pose estimation from a single image. In: Proceeding of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2016) 4948–4956
15. Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C.: Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. IEEE Transactions on Pattern Analysis and Machine Intelligence **36** (2014) 1325–1339

16. Sigal, L., Balan, A., Black, M.J.: HumanEva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. International Journal of Computer Vision **87** (2010) 4–27
17. Varol, G., Romero, J., Martin, X., Mahmood, N., Black, M.J., Laptev, I., Schmid, C.: Learning from synthetic humans. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2017) 109–117
18. Chen, W., Wang, H., Li, Y., Su, H., Wang, Z., Tu, C., Lischinski, D., Cohen-Or, D., Chen, B.: Synthesizing training images for boosting human 3d pose estimation. In: Proceedings of International Conference on 3D Vision (3DV). (2016) 479–488
19. Wu, J., Xue, T., Lim, J.J., Tian, Y., Tenenbaum, J.B., Torralba, A., Freeman, W.T.: Single Image 3D Interpreter Network. In: European Conference on Computer Vision (ECCV). (2016) 365–382
20. Andriluka, M., Pishchulin, L., Gehler, P., Schiele, B.: 2d human pose estimation: New benchmark and state of the art analysis. In: Proceeding of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2014) 3686–3693
21. Zhou, X., Zhu, M., Pavlakos, G., Leonardos, S., Derpanis, K.G., Daniilidis, K.: Monocap: Monocular human motion capture using a cnn coupled with a geometric prior. IEEE Transactions on Pattern Analysis and Machine Intelligence **41** (2019) 901–914
22. Jiang, M., Yu, Z.L., Zhang, Y., Wang, Q., Li, C., Lei, Y.: Reweighted sparse representation with residual compensation for 3d human pose estimation from a single rgb image. Neurocomputing **358** (2019) 332–343
23. Kostrikov, I., Gall, J.: Depth sweep regression forests for estimating 3d human pose from images. In: Proceedings of the British Machine Vision Conference (BMVC). Volume 1. (2014) page 5
24. Radwan, I., Dhall, A., Goecke, R.: Monocular image 3d human pose estimation under self-occlusion. In: Proceedings of IEEE International Conference on Computer Vision (ICCV). (2013) 1888–1895
25. Zhou, X., Zhu, M., Leonardos, S., Daniilidis, K.: Sparse representation for 3d shape estimation: A convex relaxation approach. IEEE Transactions on Pattern Analysis and Machine Intelligence **39** (2017) 1648–1661
26. Jiang, M., Yu, Z., Li, C., Lei, Y.: Sdm3d: shape decomposition of multiple geometric priors for 3d pose estimation. Neural Computing and Applications (2020)
27. Sarafianos, N., Boteanu, B., Ionescu, B., Kakadiaris, I.A.: 3d human pose estimation. Computer Vision and Image Understanding **152** (2016) 1–20
28. Zhao, L., Peng, X., Tian, Y., Kapadia, M., Metaxas, D.N.: Semantic graph convolutional networks for 3d human pose regression. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2019) 3425–3435
29. Li, S., Chan, A.B.: 3d human pose estimation from monocular images with deep convolutional neural network. In: Proceedings of Asian Conference on Computer Vision (ACCV). (2014) 332–347
30. Tang, W., Wu, Y.: Does learning specific features for related parts help human pose estimation? In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2019) 1107–1116
31. Martinez, J., Hossain, R., Romero, J., Little, J.J.: A simple yet effective baseline for 3d human pose estimation. In: Proceedings of IEEE International Conference on Computer Vision (ICCV). (2017) 2659–2668
32. Du, Y., Wong, Y., Liu, Y., Han, F., Gui, Y., Wang, Z., Kankanhalli, M., Geng, W.: Marker-less 3D human motion capture with monocular image sequence and height-maps. In: European Conference on Computer Vision (ECCV). (2016) 20–36

33. Luo, C., Chu, X., Yuille, A.L.: Orinet: A fully convolutional network for 3d human pose estimation. arXiv:1811.04989 (2018)
34. Li, S., Zhang, W., Chan, A.B.: Maximum-margin structured learning with deep networks for 3d human pose estimation. In: Proceedings of IEEE International Conference on Computer Vision (ICCV). (2015) 2848–2856
35. Yasin, H.: Towards efficient 3d pose retrieval and reconstruction from 2d landmarks. In: Proceedings of international symposium on multimedia (ISM). (2017) 169–176
36. Yu, J., Hong, C.: Exemplar-based 3d human pose estimation with sparse spectral embedding. Neurocomputing **269** (2017) 82–89
37. Kudo, Y., Ogaki, K., Matsui, Y., Odagiri, Y.: Unsupervised adversarial learning of 3d human pose from 2d joint locations. arXiv: 1803.08244 (2018)
38. Tung, H.F., Harley, A.W., Seto, W., Fragkiadaki, K.: Adversarial inverse graphics networks: Learning 2d-to-3d lifting and image-to-image translation from unpaired supervision. In: Proceedings of IEEE International Conference on Computer Vision (ICCV). (2017) 4364–4372
39. Wandt, B., Rosenhahn, B.: Repnet: Weakly supervised training of an adversarial reprojection network for 3d human pose estimation. In: Proceedings of the Computer Vision and Pattern Recognition (CVPR). (2019) 7782–7791
40. Dong, J., Jiang, W., Huang, Q., Bao, H., Zhou, X.: Fast and robust multi-person 3d pose estimation from multiple views. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2019) 7792–7801
41. Kanazawa, A., Black, M.J., Jacobs, D.W., Malik, J.: End-to-end recovery of human shape and pose. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2018) 7122–7131
42. Pavlakos, G., Zhu, L., Zhou, X., Daniilidis, K.: Learning to estimate 3D human pose and shape from a single color image. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2018) 459–468
43. Xu, Y., Zhu, S., Tung, T.: Denserac: Joint 3d pose and shape estimation by dense render-and-compare. In: Proceedings of IEEE International Conference on Computer Vision (ICCV). (2019) 7760–7770
44. Cai, Y., Ge, L., Liu, J., Cai, J., Cham, T., Yuan, J., Thalmann, N.M.: Exploiting spatial-temporal relationships for 3d pose estimation via graph convolutional networks. In: Proceedings of IEEE International Conference on Computer Vision (ICCV). (2019) 2272–2281
45. Kocabas, M., Karagoz, S., Akbas, E.: Self-supervised learning of 3d human pose using multi-view geometry. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2019) 1077–1086
46. Bogo, F., Kanazawa, A., Lassner, C., Gehler, P., Romero, J., Black, M.J.: Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In: European Conference on Computer Vision (ECCV). (2016) 561–578
47. Tian, Y., Sigal, L., La Torre, F.D., Jia, Y.: Canonical locality preserving latent variable model for discriminative pose inference. Image and Vision Computing **31** (2013) 223–230
48. Fan, X., Zheng, K., Zhou, Y., Wang, S.: Pose locality constrained representation for 3d human pose reconstruction. In: European Conference on Computer Vision (ECCV). (2014) 174–188
49. Sun, K., Lan, C., Xing, J., Zeng, W., Liu, D., Wang, J.: Human pose estimation using global and local normalization. In: Proceedings of IEEE International Conference on Computer Vision (ICCV). (2017) 5600–5608

50. Luo, Y., Xu, Z., Liu, P., Du, Y., Guo, J.: Combining fractal hourglass network and skeleton joints pairwise affinity for multi-person pose estimation. Multimedia Tools and Applications **78** (2019) 7341–7363
51. Isack, H., Haene, C., Keskin, C., Bouaziz, S., Boykov, Y., Izadi, S., Khamis, S.: Repose: Learning deep kinematic priors for fast human pose estimation. arXiv:2002.03933 (2020)
52. Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. In: European Conference on Computer Vision (ECCV). (2016) 483–499
53. Simo-Serra, E., Quattoni, A., Torras, C., Moreno-Noguer, F.: A joint model for 2d and 3d pose estimation from a single image. In: Proceeding of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2013) 3634–3641
54. Wang, C., Wang, Y., Lin, Z., Yuille, A.L., Gao, W.: Robust estimation of 3d human poses from a single image. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2014) 2369–2376
55. Pavllo, D., Feichtenhofer, C., Grangier, D., Auli, M.: 3d human pose estimation in video with temporal convolutions and semi-supervised training. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2019) 7745–7754