Single-Stage Multi-Person Pose Machines

Xuecheng Nie[1] Jiashi Feng[1] Jianfeng Zhang[1] Shuicheng Yan[1,2]
[1]Department of Electrical and Computer Engineering, National University of Singapore, Singapore
[2]Yitu Technology
niexuecheng@u.nus.edu elefjia@nus.edu.sg elezji@nus.edu.sg shuicheng.yan@yitu-inc.cn

Abstract

Multi-person pose estimation is a challenging problem. Existing methods are mostly two-stage based—one stage for proposal generation and the other for allocating poses to corresponding persons. However, such two-stage methods generally suffer low efficiency. In this work, we present the first single-stage *model, Single-stage multi-person Pose Machine (SPM), to simplify the pipeline and lift the efficiency for multi-person pose estimation. To achieve this, we propose a novel Structured Pose Representation (SPR) that unifies person instance and body joint position representations. Based on SPR, we develop the SPM model that can directly predict structured poses for multiple persons in a single stage, and thus offer a more compact pipeline and attractive efficiency advantage over two-stage methods. In particular, SPR introduces the root joints to indicate different person instances and human body joint positions are encoded into their displacements w.r.t. the roots. To better predict long-range displacements for some joints, SPR is further extended to hierarchical representations. Based on SPR, SPM can efficiently perform multi-person poses estimation by simultaneously predicting root joints (location of instances) and body joint displacements via CNNs. Moreover, to demonstrate the generality of SPM, we also apply it to multi-person 3D pose estimation. Comprehensive experiments on benchmarks MPII, extended PASCAL-Person-Part, MSCOCO and CMU Panoptic clearly demonstrate the state-of-the-art efficiency of SPM for multi-person 2D/3D pose estimation, together with outstanding accuracy.*

1. Introduction

Multi-person pose estimation from a single monocular RGB image aims to simultaneously isolate and locate body joints of multiple person instances. It is a fundamental yet challenging task with broad applications in action recognition [7], person Re-ID [32], pedestrian tracking [2], etc.

Existing methods typically adopt two-stage solutions. As shown in Figure 1 (b), they either follow the top-down strategy [12, 35, 17, 9, 8, 34] that employs off-the-shelf detectors to localize person instances at first and then locates their joints individually; or the bottom-up strategy [3, 16, 26, 31, 25] that locates all the body joints at first and then assigns them to the corresponding person. Though with high accuracy, these methods are not efficient as they require two-stage processing to predict human poses with computational redundancy. We observe that such a requirement mainly comes from the conventional pose representation they adopt. As shown in Figure 2 (b), absolute positions of allocated body joints separate the position information w.r.t. person instances and body joints, each of which requires a stage to process, leading to low efficiency.

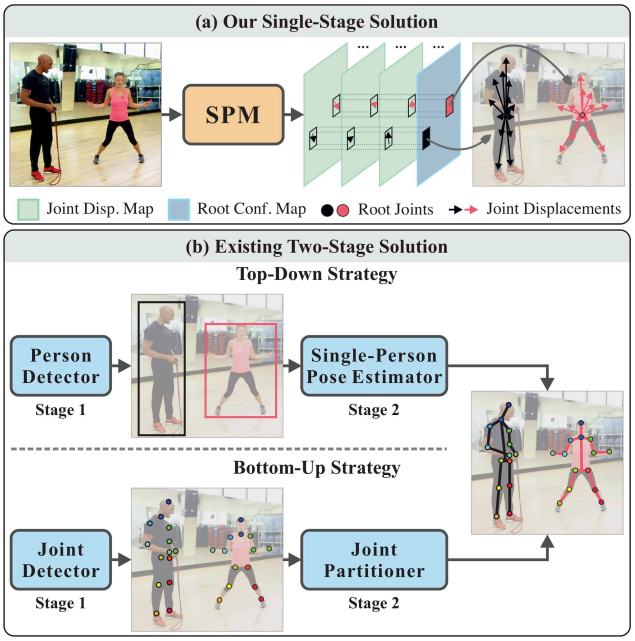


Figure 1. Comparison between (a) our single-stage solution and (b) existing two-stage solution to multi-person pose estimation. The proposed SPM model directly predicts structured poses of multiple persons in a single stage, offering a more compact pipeline and attractive efficiency advantages over two-stage based top-down or bottom-up strategies. See more details in the main text.

To overcome such an intrinsic limitation, we propose a new Structured Pose Representation (SPR) to unify position information of person instances and body joints. SPR allows to simplify the pipeline for person separation and body joint localization and thus enables a much more efficient *single-stage* solution to multi-person pose estimation. In particular, SPR defines a unique identity joint, the *root joint*, for each person instance to indicate its position in the image. Then, the positions of body joints are encoded by their displacements w.r.t. the root joints. In this way, the pose of a person instance is represented together with its location, as shown in Figure 2 (c), making a single-stage solution feasible. To tackle the long-range displacements (*e.g.* limb joints), we further extend SPR to a hierarchical one by dividing body joints into hierarchies induced from articulated kinematics [20]. Such a Hierarchical Structured Pose Representation is shown in Figure 2 (d).

Based on SPR, we propose a *Single-stage multi-person Pose Machine* (SPM) model to solve multi-person pose estimation with compact pipeline and high efficiency. As aforementioned, existing two-stage models isolate different instances and estimate their poses separately. Different from them, SPM maps a given image to multiple human poses represented by SPR in a single-stage manner. As shown in Figure 1 (a), it simultaneously regresses the root joint positions and body joint displacements, predicting multi-person poses within one stage. We implement SPM with Convolutional Neural Networks (CNNs) based on the state-of-the-art Hourglass architecture [27] for learning and inferring root joint position and body joint displacement simultaneously and end-to-end.

Comprehensive experiments on benchmarks MPII [1], extended PASCAL-Person-Part [38], MSCOCO [23] and CMU Panoptic [19] evidently demonstrate the high efficiency of the proposed SPM model. In addition, it achieves new state-of-the-art on MPII and extended PASCAL-Person-Part datasets, and competitive performance on MSCOCO dataset. Moreover, it also achieves promising results on CMU Panoptic dataset for multi-person 3D pose estimation. Our contributions is summarized as: 1) We propose the first *single-stage* solution to multi-person 2D/3D pose estimation. 2) We propose novel structured pose representations to unify position information of person instances and body joints. 3) Our model achieves outperforming efficiency with competitive accuracy on multiple benchmarks.

2. Background

In this section, we review the state-of-the-art multi-person pose estimation methods based on conventional pose representation. Given an image I, multi-person pose estimation targets at estimating human poses $\bar{\mathcal{P}}$ of all the person instances in I via inferring coordinates of their body joints. Conventionally, poses are represented as

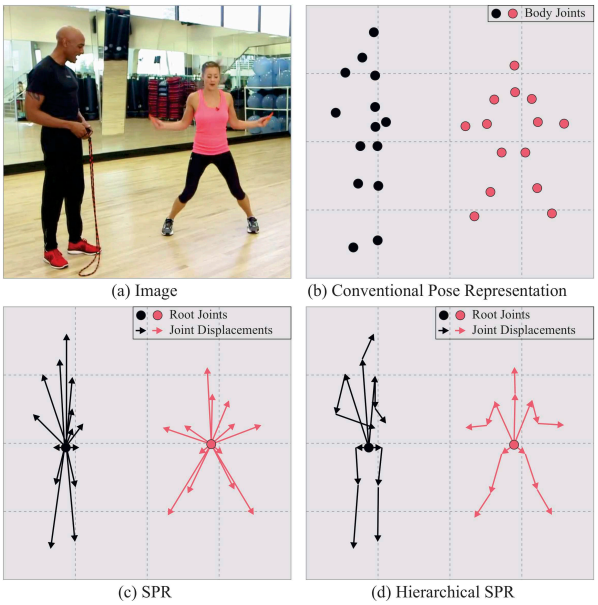
$$\bar{\mathcal{P}} = \{\mathbf{P}_i^1, \mathbf{P}_i^2, \dots, \mathbf{P}_i^K\}_{i=1}^N, \tag{1}$$

where N is the number of persons in I, K is the number of joint categories, and $\mathbf{P}_i^j$ denotes coordinates of the jth body joint from person i, where $\mathbf{P}_i^j{=}(x_i^j, y_i^j)$ for 2D case while $\mathbf{P}_i^j{=}(x_i^j, y_i^j, z_i^j)$ for 3D case. To obtain $\bar{\mathcal{P}}$, existing methods typically exploit two-stage solutions, *i.e.* separately predicting positions of person instances and their body joints. Based on the processing order, they can be divided into two categories: the top-down methods and the bottom-up ones.

A top-down method generates multiple human poses $\bar{\mathcal{P}}$ as follows. It first uses a person detector f to localize and separate person instances, and then conducts single-person pose estimation using a single-person model g to individually locate body joints for each person instance. Formally, the process can be summarized as

$$\begin{aligned} f &: I \rightarrow \mathcal{B}, \\ g &: \mathcal{B}, I \rightarrow \bar{\mathcal{P}}. \end{aligned} \tag{2}$$

Here $\mathcal{B}$ denotes person instance localization results that are usually represented by a set of bounding boxes. Following this strategy, for 2D case, Gkioxari *et al.* [12] exploited a Generalized Hough Transform framework to detect person instances and then localize body joints via classifying poselets—the tightly clustered body parts with similar appearances and configurations. Iqbal and Gall [17] improved the person detector and single-person model via exploiting deep



(a) Image (b) Conventional Pose Representation (c) SPR (d) Hierarchical SPR

Figure 2. Different pose representations for multiple person instances in image (a). (b) Conventional pose representation with each joint represented by absolute coordinates. (c) Proposed structured pose representation w.r.t. root joints. (d) Proposed hierarchical structured pose representation. See more details in main text.

learning based techniques, including Faster-RCNN [33] and convolutional pose machine [37], to acquire more accurate human poses. Similarly, Fang *et al.* [9] proposed to incorporate spatial transformer network [18] and Hourglass network [27] to further improve person instance and body joint detections. Papandreou *et al.* [29] further improved the top-down strategy via location refinement with predictions of 2D offset vector from a pixel to the corresponding joint. For 3D case, Rogez [34] first utilized region proposal network to detect persons of interest and found 3D anchor pose for each detection, then exploit iterative regression for refinement. Dong [8] performed top-down multi-person 2D pose estimation for images from multiple views and reconstructed 3D pose for each person from multi-view 2D poses.

In contrast, to obtain poses $\bar{\mathcal{P}}$, a bottom-up method first utilizes a body joint estimator g' to localize body joints for all instances, and then estimates the position of each instance and the joint allocation by solving a graph partition problem with the model f', formulated as

$$\begin{aligned} g' &: I \rightarrow \mathcal{J}, \mathcal{C} \\ f' &: \mathcal{J}, \mathcal{C} \rightarrow \bar{\mathcal{P}}, \end{aligned} \tag{3}$$

where $\mathcal{J}$ denotes the set of joint candidates and $\mathcal{C}$ the affinities for assigning joint candidates to person instances. In [16], Insafutdinov *et al.* exploited Residual networks [14] as the joint detector and defined geometric correlations for allocating body joints, and then performed Integer Linear Programming to partition joint candidates. Cao *et al.* [3] proposed a real-time model with improved joint correlations via introducing part affinity fields to encode location and orientation of limbs and allocate joint candidates via solving a maximum weight bipartite graph matching problem. Later, Mehta [25] extended [3] to multi-person 3D pose estimation. Newell and Deng [26] introduced the associative embedding model followed by a greedy algorithm for allocating body joints. Papandreou *et al.* [28] presented the bottom-up PersonLab model by defining different levels of offsets to calculate association scores and adjust joint positions for grouping joint candidates into person instance and refining pose estimations.

Different from all the previous methods relying on a two-stage pipeline, we present a new pose representation method that unifies positions of person instances and body joints, enabling a compact and efficient single-stage solution to multi-person 2D/3D pose estimation, as explained below.

3. Structured pose representation

In this section, we elaborate on the proposed Structured Pose Representations (SPR) for multi-person pose estimation. Different from the conventional pose representation in Eqn. (1), SPR aims to unify the position information of person instance and body joint to deliver a single-stage solution for multi-person pose estimation. In particular, SPR introduces an auxiliary joint, the root joint, to denote the person instance position. It is a unique identity joint for a specific person instance. In the following, we illustrate the formulations of SPR in 2D case for simplification, which can be directly extended to 3D case via replacing 2D coordinates with 3D ones. Specifically, we use $(x_i^{\mathrm{r}}, y_i^{\mathrm{r}})$ to denote the root joint position of the ith person. Then the position of the jth joint of person i can be defined as

$$(x_i^j, y_i^j) = (x_i^{\mathrm{r}}, y_i^{\mathrm{r}}) + (\delta x_i^j, \delta y_i^j), \tag{4}$$

where $(\delta x_i^j, \delta y_i^j)$ represents the displacement of the jth body joint position w.r.t. the root joint. Eqn. (4) directly establishes the structured relationship between person instance position and body joint position. Thus, we use the Structured Pose Representations to represent human poses with the root joint position and body joint displacements, formulated as

$$\mathcal{P}=\left\{(x_i^{\mathrm{r}}, y_i^{\mathrm{r}}), (\delta x_i^1, \delta y_i^1), (\delta x_i^2, \delta y_i^2), \ldots, (\delta x_i^K, \delta y_i^K)\right\}_{i=1}^{N}. \tag{5}$$

By the definition in Eqn. (5), SPR unifies position information of the person instance and the body joint and can be obtained in an efficient single-stage prediction. In addition, SPR can be effortlessly converted back to the conventional pose representation based on Eqn. (4). Here, we exploit the person centroid as the root joint of the person instance, due to its stability and robustness in discriminating person instances even with extreme poses. An example of SPR representing multiple human poses is shown in Figure 2 (c).

Hierarchical SPR SPR in Eqn. (5) may involve long-range displacements between body joints and the root joint due to possible large pose deformation, *e.g.*, wrists and ankles relative to the person centroid, bringing difficulty to displacement estimation by mapping from image representation to the vector domain. Thereby, we propose to factorize long-range displacements into accumulative shorter ones to further improve SPR. Specifically, we divide the root joint and body joints into four hierarchies based on articulated kinematics [20] by their degrees of freedom and extent of deformation. Here, the root joint is placed in the first hierarchy; torso joints including neck, shoulders and hips are in the second one; head, elbows and knees are put in the third; wrists and ankles are put in the fourth. Then we can identify joint positions via shorter-range displacements between joints in adjacent hierarchies. For example, the wrist position can be encoded by its displacement relative to the elbow. Modeling short-range displacements can alleviate the learning difficulty of mapping from image representation to the vector domain and better utilize appearance cues along limbs. Formally, for the jth joint in the lth layer (*e.g.*, wrist in the 4th layer) and its corresponding j'th

joint in the $(l-1)$th layer (*e.g.*, elbow in the 3rd layer), the relation between their positions (x_i^j, y_i^j) and $(x_i^{j'}, y_i^{j'})$ can be formulated as

$$(x_i^j, y_i^j) = (x_i^{j'}, y_i^{j'}) + (\delta\tilde{x}_i^j, \delta\tilde{y}_i^j), \tag{6}$$

where $(\delta\tilde{x}_i^j, \delta\tilde{y}_i^j)$ denotes the displacement between joints in adjacent hierarchies. According to the articulated kinematics, we can define an articulated path (a set of ordered joints) connecting the root joint to any body joint. Then, the body joint can be identified via the root joint position and accumulation of short-range displacements along the articulated path. Namely,

$$(x_i^j, y_i^j) = (x_i^{\mathrm{r}}, y_i^{\mathrm{r}}) + \sum_{h \in \mathcal{H}^j \setminus \{\mathrm{r}\}} (\delta\tilde{x}_i^h, \delta\tilde{y}_i^h), \tag{7}$$

where $\mathcal{H}^j = \{\mathrm{r}, a^{(1)}, \ldots, a^{(m)}, j\}$ represents the articulated path between the root joint and the jth body joint and $a^{(n)}$ denotes the nth articulated joint on the path. In this way, we propose the Hierarchical Structured Pose Representations to denote a human pose with the root joint position, the short-range body joint displacements between neighboring hierarchies, and the articulated path set $\mathcal{H}$ as

$$\mathcal{P}=\left\{(x_i^{\mathrm{r}}, y_i^{\mathrm{r}}), (\delta\tilde{x}_i^1, \delta\tilde{y}_i^1), (\delta\tilde{x}_i^2, \delta\tilde{y}_i^2), \ldots, (\delta\tilde{x}_i^K, \delta\tilde{y}_i^K)\right\}_{i=1}^N, \text{ given } \mathcal{H}. \tag{8}$$

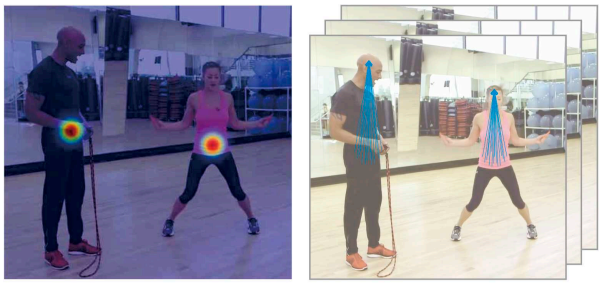
Similar to SPR, hierarchical SPR defined in Eqn. (8) also unifies representations of person instance position and body joint position, leading to a single-stage solution to multi-person pose estimation as well. Moreover, hierarchical SPR factorizes displacements between the root joint and long-range body joints, benefiting estimation results for the cases with large body joint displacements. Hierarchical SPR can also be easily converted to SPR and conventional pose representation via Eqn. (7). Figure 2 (d) gives an example of Hierarchical SPR for multi-person pose representation.

4. Single-stage multi-person pose machine

With SPR, we propose to construct a regression model, termed as Single-stage multi-person Pose Machine (SPM), to map an input image I to the poses of multiple persons $\mathcal{P}$:

$$\mathrm{SPM}: I \rightarrow \mathcal{P}, \tag{9}$$

which tackles the multi-person pose estimation problem in a single-stage manner. Different from two-stage solutions in Eqn. (2) and (3), SPM only needs to learn a single mapping function. Motivated by recent success of Convolutional Neural Networks (CNNs) in computer vision tasks [14, 22, 24], we implement SPM with a CNN model. Below we will describe regression targets, network architecture, and training and inference details of SPM in 2D case for simplification. For 3D case[1], the same scheme can be exploited with 3D coordinates.



(a) Root Joint Conf. Map (b) Dense Disp. Maps

Figure 3. Regression targets of the proposed SPM. (a) Confidence map for root joint. (b) Dense displacement maps for body joints.

4.1. Regression targets

Since the root joint $(x_i^{\mathrm{r}}, y_i^{\mathrm{r}})$ and body joint displacements $\{(\delta x_i^1, \delta y_i^1), (\delta x_i^2, \delta y_i^2), \ldots, (\delta x_i^K, \delta y_i^K)\}$ are respectively in the coordinate and vector domains, we construct different regression targets for the proposed SPM to learn to predict these two kinds of information.

Regression target for root joint position According to previous works [4, 31], it is difficult to directly regress the absolute joint coordinates in an image. To reliably detect root joint positions, we exploit a confidence map to encode probabilities of the root joint of a person instance at each location in the image. The root joint confidence map is constructed by modeling the root joint position as Gaussian peaks. We use C^{r} to denote the root joint confidence map and $\mathrm{C}_i^{\mathrm{r}}$ the root joint map of the ith person. For a position (x, y) in the given image I, $\mathrm{C}_i^{\mathrm{r}}(x, y)$ is calculated by

$$\mathrm{C}_i^{\mathrm{r}}(x, y) = \exp(-\|(x, y) - (x_i^{\mathrm{r}}, y_i^{\mathrm{r}})\|_2^2 / \sigma^2),$$

where $(x_i^{\mathrm{r}}, y_i^{\mathrm{r}})$ is the groundtruth root joint position of the ith person instance and σ is an empirically chosen constant to control the variance of Gaussian distribution, set as σ=7 in our experiments. The root joint confidence map C^{r} is an aggregation of peaks of all persons in a single map. Here, we choose to take the maximum of confidence maps rather than their average to maintain distinctions between close-by peaks [3], *i.e.*, $\mathrm{C}^{\mathrm{r}}(x, y) = \max_i \mathrm{C}_i^{\mathrm{r}}(x, y)$. An example of the root joint confidence map is shown in Figure 3 (a).

Regression target for body joint displacement We construct a dense displacement map for each joint. We use D^j to denote it for joint j and D_i^j to denote the one for joint j of person i. For a location (x, y) in image I, $\mathrm{D}_i^j(x, y)$ is calculated by

$$\mathrm{D}_i^j(x, y) = \begin{cases} \frac{(\delta x, \delta y)}{Z} & \text{if } (x, y) \in \mathcal{N}_i^{\mathrm{r}} \\ 0 & \text{otherwise} \end{cases},$$

[1] We set the camera position as the origin of the 3D coordinate system.

$$(\delta x, \delta y) = (x_i^j, y_i^j) - (x, y),$$

where $\mathcal{N}_i^{\mathrm{r}} = \left\{(x,y) | \|(x,y) - (x_i^{\mathrm{r}}, y_i^{\mathrm{r}})\|_2^2 \leq \tau\right\}$ denotes the neighboring positions of the root joint of person i, $Z{=}\sqrt{H^2+W^2}$ is the normalization factor, with H and W denoting the height and width of I, and τ is a constant controlling the neighborhood size, set as 7 in our experiments. Then, we define the dense displacement map D^j for the jth joint to be the average for all persons:

$$\mathrm{D}^j(x,y) = \frac{1}{M^j} \sum_i \mathrm{D}_i^j(x,y),$$

where M^j is the number of non-zero vectors at position (x,y) across all persons. Figure 3 (b) shows examples for the constructed dense displacement maps. For hierarchical SPR, D^j is constructed in a similar way, just replacing the root joint with the one in the neighbor hierarchy.

4.2. Network architecture

We use the Hourglass network [26], the state-of-the-art architecture for human pose estimation, as the backbone of SPM. It is a fully convolutional network composed of multiple stacked Hourglass modules. Each Hourglass module, as shown in Figure 4, adopts a U-Shape structure that first decreases feature map resolution to learn abstract semantic representations and then upsamples the feature maps for body joint localization. Additionally, skip connections are added between feature maps with the same resolution for reusing low-level spatial information to refine high-level semantic information. In the original design, the Hourglass network utilizes a single branch to predict body joint confidence maps for single-person pose estimation. In this paper, SPM exploits the confidence regression branch of the Hourglass network to regress confidence maps for the root joint. In addition, SPM extends the Hourglass network via adding a displacement regression branch, to estimate body joint displacement maps. In this way, SPM can produce (Hierarchical) SPR in a single forward pass.

4.3. Training and inference

For training SPM, we adopt ℓ_2 loss $\mathcal{L}^{\mathrm{C}}$ and smooth ℓ_1 loss [11] $\mathcal{L}^{\mathrm{D}}$ for root joint confidence and dense displacement map regression respectively. Intermediate supervision is applied at all Hourglass modules to avoid gradient vanishing. The total loss $\mathcal{L}$ is the accumulation of weighted sum of $\mathcal{L}^{\mathrm{C}}$ and $\mathcal{L}^{\mathrm{D}}$ across all hourglass modules:

$$\mathcal{L} = \sum_{t=1}^{T} \left(\mathcal{L}^{\mathrm{C}}(\hat{\mathrm{C}}_{(t)}^{\mathrm{r}}, \mathrm{C}^{\mathrm{r}}) + \beta \mathcal{L}^{\mathrm{D}}(\hat{\mathrm{D}}_{(t)}, \mathrm{D})\right),$$

where T is the number of Hourglass modules, set as $T{=}8$, $\hat{\mathrm{C}}_{(t)}^{\mathrm{r}}$ and $\hat{\mathrm{D}}_{(t)}$ denote the predicted root joint confidence map and dense displacement maps at the tth stage, and β is a constant weight factor to balance two kinds of losses, set as $\beta{=}0.01$ in our experiments. The overall framework of SPM is end-to-end trainable via gradient backpropagation.

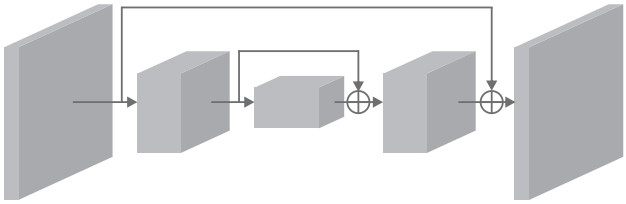


Figure 4. The backbone of SPM: Hourglass network.

The overall inference procedure for SPM to predict SPR is illustrated in Figure 1 (a). Given an image, SPM first produces root joint confidence map $\hat{\mathrm{C}}^{\mathrm{r}}$ and displacement maps $\hat{\mathrm{D}}$ via a CNN. Then, it performs NMS on $\hat{\mathrm{C}}^{\mathrm{r}}$ to generate root joint positions $\left\{(\hat{x}_i^{\mathrm{r}}, \hat{y}_i^{\mathrm{r}})\right\}_{i=1}^{\hat{N}}$, with $\hat{N}$ denoting the estimated number of persons. After that, SPM gets the displacement of the body joint j of person i by $Z{\cdot}\mathrm{D}^j(\hat{x}_i^{\mathrm{r}}, \hat{y}_i^{\mathrm{r}})$. Finally, SPM outputs human poses represented by SPRs via combining root joint positions and body joint displacements. For predicting hierarchical SPRs, SPM follows the above procedure to sequentially get joint displacements according to the joint hierarchies in Eqn. (7).

5. Experiments

5.1. Experiment setup

Datasets We evaluate the proposed SPM model for multi-person pose estimation on three widely adopted 2D benchmarks: MPII [1] dataset, extended PASCAL-Person-Part [38] dataset and MSCOCO [23] dataset, and one 3D benchmark CMU Panoptic dataset [19].

MPII dataset contains 5,602 groups of images of multiple persons, which are split into 3,844 for training and 1,758 for testing. It also provides over 28,000 annotated single-person pose samples. Each person is annotated with 16 body joints. We use the official mean Average Precision (mAP) for evaluation on this dataset. The extended PASCAL-Person-Part dataset consists of 1,716 training and 1,817 testing images collected from the original PASCAL-Person-Part dataset [5], and provides 14 body joint annotations for each person. Similar to MPII, this dataset also adopts mAP as the evaluation metric. MSCOCO dataset contains about 60,000 training images with 17 annotated body joints per person. Evaluations are conducted on the test-dev subset, including roughly 20,000 images, with the official Average Precision (AP) as metric.

CMU Panoptic is a large scale dataset providing 3D pose annotations for multiple people engaging social activities. It totally includes 65 videos with multi-view annotations, but only 17 of them are in multi-person scenario and given the

camera parameters. We use the front-view captures of these 17 videos in our experiments, which contains 75,552 images in total and are randomly split into 65,552 for training and 10,000 for testing. We following conventions [25, 34] to utilize 3D-PCK@150mm as metric.

Data augmentation We follow the conventional data augmentation strategies for multi-person pose estimation via cropping original images centered at person centroid to 384×384 input samples to SPM. For MPII and extended PASCAL-Person-Part datasets, we augment training samples with rotation degrees in $[-40°, 40°]$, scaling factors in $[0.7, 1.3]$, translation offset in $[-40\text{px}, 40\text{px}]$ and horizontally flipping. For MSCOCO dataset, scaling factors are sampled in $[0.5, 1.5]$ and other augmentation parameters are set the same as MPII and extended PASCAL-Person-Part datasets. For CMU Panoptic dataset, we conduct data augmentation with scale factors in $[0.9, 1.5]$ and set the other augmentation parameters the same as 2D case.

Implementation For MPII dataset, we randomly select 350 groups of multi-person training samples as the validation dataset and use the remaining training samples and all single-person pose images to learn SPM. For MSCOCO dataset, we use the standard training split for training the model. Following conventions [3, 37] for multi-person pose estimation, we normalize the input image to CNN with mean 0.5 and standard deviation 1.0 for RGB channels. We implement SPM with Pytorch [30] and utilize RMSprop [36] as the optimizer with an initial learning rate of 0.003. For MPII dataset, we train SPM for 250 epochs and decrease learning rate by a factor of 2 at the 150th, 170th, 200th, 230th epoch. For extended PASCAL-Person-Part dataset, we fine-tune the model pre-trained on MPII for 30 epochs. For MSCOCO dataset, SPM is trained for 100 epochs and learning rate is decreased at the 30th, 60th, and 80th epoch by a factor of 2. For CMU Panoptic dataset, we adopt the same training strategy as MPII. Testing is performed on six-scale image pyramids with flipping for both datasets. Specially, we follow previous works [3, 26] to refine estimation results with a single-person model trained on the same dataset on MPII and MSCOCO.

5.2. Results on MPII dataset

Comparison with state-of-the-arts In Table 1, we compare our SPM model with hierarchical SPR to state-of-the-arts on the full test split of MPII dataset[2]. We can see that our SPM model only requires 0.058s to process an image, about $5\times$ faster than the bottom-up model [26] with state-of-the-art speed, verifying the efficiency advantage of the proposed single-stage solution over existing two-stage ones for multi-person pose estimation. In addition, our SPM model achieves new state-of-the-art 78.5% mAP on MPII dataset and improves accuracies for most kinds of body joints, which demonstrates its superior performance for estimating human poses of multiple persons in a single stage.

[2]For our SPM model, the time is counted with single-scale testing on GPU TITAN X and CPU Intel I7-5820K 3.3GHz, excluding the refinement time by single-person pose estimation. For time evaluation on [26], we report the runtime with the code provided by authors in the link: https://github.com/umich-vl/pose-ae-train. For runtime on [3], we refer to its speed for single-scale inference setting on MPII testing set, which can be found in Table 1 of 1st version of [3].

Table 1. Comparison with state-of-the-arts on the full testing set of MPII dataset (mAP).

Method	Head	Sho.	Elb.	Wri.	Hip	Knee	Ank.	Total	Time[s]
Iqbal and Gall [17]	58.4	53.9	44.5	35.0	42.2	36.7	31.1	43.1	10
Insafutdinov *et al.* [16]	78.4	72.5	60.2	51.0	57.2	52.0	45.4	59.5	485
Levinkov *et al.* [21]	89.8	85.2	71.8	59.6	71.1	63.0	53.5	70.6	-
Insafutdinov *et al.* [15]	88.8	87.0	75.9	64.9	74.2	68.8	60.5	74.3	-
Cao *et al.* [3]	91.2	87.6	77.7	66.8	75.4	68.9	61.7	75.6	0.6
Fang *et al.* [9]	88.4	86.5	78.6	70.4	74.4	73.0	65.8	76.7	0.4
Newell and Deng [26]	**92.1**	89.3	78.9	69.8	76.2	71.6	64.7	77.5	0.25
Fieraru *et al.* [10]	91.8	**89.5**	80.4	69.6	**77.3**	71.7	65.5	78.0	-
SPM (Ours)	89.7	87.4	**80.4**	**72.4**	76.7	**74.9**	**68.3**	**78.5**	**0.058**

Table 2. Ablation experiments on MPII validation dataset (mAP).

Method	Head	Sho.	Elb.	Wri.	Hip	Knee	Ank.	Total	Time[s]
SPM-Vanilla	91.7	87.5	76.1	65.2	75.2	71.4	60.3	75.3	0.058
SPM-Hierar	92.0	88.5	78.6	69.4	77.7	73.8	63.9	77.7	0.058

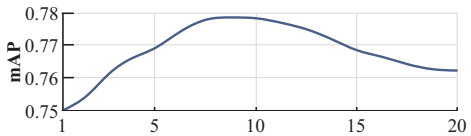


Figure 5. Analysis on hyper-parameter τ, the neighborhood size for constructing regression target for body joint displacement.

Ablation analysis We conduct ablation analysis on MPII validation dataset. We first evaluate the impact of the hierarchical division to SPR on the proposed SPM model. Results are shown in Table 2. We use SPM-Vanilla and SPM-Hierar to denote the models for predicting SPR and Hierarchical SPR, respectively.

We can see SPM-Vanilla achieves 75.3% mAP with 0.058s per image. By introducing joint hierarchies, SPM-Hierar improves the performance to 77.7% mAP without increasing time cost as SPR and hierarchical SPR have the same complexity and both of them are generated by SPM in a single-stage manner. In addition, we can see SPR-Hierar improves the accuracy of all joints. Moreover, we can also see that improvements by SPM-Hierar on long-range body joints wrists and ankles are significant, from 65.2% to 69.4% mAP and 60.3% to 63.9% mAP, respectively, verifying the effectiveness of shortening long-range displacements with Hierarchical SPR that divides body joints to dif-

Table 3. Comparison with state-of-the-arts on the testing set of the extended PASCAL-Person-Part dataset (mAP)

Method	Head	Sho.	Elb.	Wri.	Hip	Knee	Ank.	Total
Chen and Yuille [6]	45.3	34.6	24.8	21.7	9.8	8.6	7.7	21.8
Insafutdinov et al. [16]	41.5	39.3	34.0	27.5	16.3	21.3	20.6	28.6
Xia et at. [38]	58.0	52.1	43.1	37.2	22.1	30.8	31.1	39.2
SPM (Ours)	**65.4**	**60.8**	**50.2**	**47.7**	**29.0**	**35.3**	**34.6**	**46.1**

ferent hierarchies. These results clearly show the efficacy of incorporating hierarchical SPR to improve performance and efficiency of multi-person pose estimation.

We then conduct experiments to analyze the impact of important hyper-parameter τ, the neighborhood size in constructing regression targets for body joint displacements in Section 4.1, on the proposed SPM model. We range τ from 1 to 20 and results are given in Figure 5. From Figure 5, we can see increasing τ from 1 to 7 gradually improves the performance, mainly because with the increase of positive samples, more variations of body joints can be covered for displacement regression in training. Further increasing τ from 7 to 10 cannot achieve performance improvement. However, when τ>10, we observe performance drop. This is because noise from background is taken as positive samples and the overlap of displacement fields among multiple persons degrades the performance. Hence, we set τ=7 in our experiments for the trade-off of efficiency and accuracy.

Qualitative results Qualitative results on MPII dataset are shown in the top row of Figure 6. We can see that the proposed SPM is effective and robust for estimating human poses represented by Hierarchical SPRs even in challenging scenarios, *e.g.*, large pose deformation (1st example), blurred and cluttered background (2nd example), occlusion and person overlapping (3rd example), and illumination variations (4th example). These results further validate the efficacy of SPM.

5.3. Results on PASCAL-Person-Part dataset

Table 3 shows the comparison results with state-of-the-arts on the extended PASCAL-Person-Part dataset. We can see that the proposed SPM model achieves 46.1% mAP and provides new state-of-the-art. Besides, SPM outperforms previous models for all body joints, demonstrating the effectiveness of the proposed single-stage model for tackling the multi-person pose estimation problem.

Qualitative results are shown in the middle row of Figure 6. We observe SPM can deal with person scale variations (1st example), occlusion (2nd to 4th examples) and person overlapping (the last example), showing the efficacy of SPM on producing robust pose estimation in various challenging scenes.

Table 4. Comparison with state-of-the-arts on the MSCOCO test-dev (AP).

Method	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}	Time[s]
CMU-Pose [3]	0.618	0.849	0.675	0.571	0.682	0.6
RMPE [9]	0.618	0.837	0.698	0.586	0.676	0.4
Mask-RCNN [13]	0.627	0.870	0.684	0.574	0.711	0.2
G-RMI [29]	0.649	0.855	0.713	0.623	0.700	-
AssocEmbedding [26]	0.655	0.868	0.723	0.606	0.726	0.25
PersonLab [28]	0.687	0.890	0.754	0.641	0.755	0.464
SPM (Ours)	0.669	0.885	0.729	0.626	0.731	**0.058**

5.4. Results on MSCOCO dataset

Table 4 shows experimental results on MSCOCO test-dev. We can see that the proposed SPM model achieves overall 0.669 AP, which is slightly lower than the state-of-the-art [28]. However, our SPM achieves superior speed, 8× faster than [28]. These results further confirm the superior efficiency of our single-stage solution over existing two-stage top-down or bottom-up strategies, while achieving very competitive performance, for addressing the multi-person pose estimation tasks.

Qualitative results on MSCOCO dataset are shown in the bottom row of Figure 6. We can see that our SPM model is effective in challenging scenes, *e.g.*, appearance variations (1st example) and occlusion (2nd to 4th examples).

5.5. Results on CMU Panoptic dataset

We evaluate the proposed SPM model for multi-person 3D pose estimation on the CMU Panoptic dataset, which provides large-scale data with accurate 3D pose annotations and thus is suitable to be an evaluation benchmark. Since previous works [19, 8] only conduct qualitative evaluation on this dataset, there are no reported quantitative results for comparison. For better understanding the model performance, we present the first quantitative evaluation here. We separate 10,000 images from the dataset to form the testing split and use the remaining for training as mentioned in Section 5.1. In particular, our SPM model achieves 77.8% 3D-PCK, a promising result for multi-person 3D pose estimation. The effectiveness of our SPM model can be also verified through the qualitative results in Figure 7. We can see our SPM model is robust for pose variations (1st and 2nd examples), self occlusions (3rd example), scale and depth changes (4th and 5th examples).

In addition, the proposed SPM model achieves attractive efficiency with speed of about 20 FPS. Moreover, its single-stage design also significantly simplifies the pipeline for multi-person 3D pose estimation from a single monocular RGB image, alleviating the requirements of intermediate 2D pose estimations [25] or 3D pose reconstructions from multiple views [8].

Figure 6. Qualitative results on MPII dataset (top), extended PASCAL-Person-Part dataset (middle) and MSCOCO dataset (bottom).

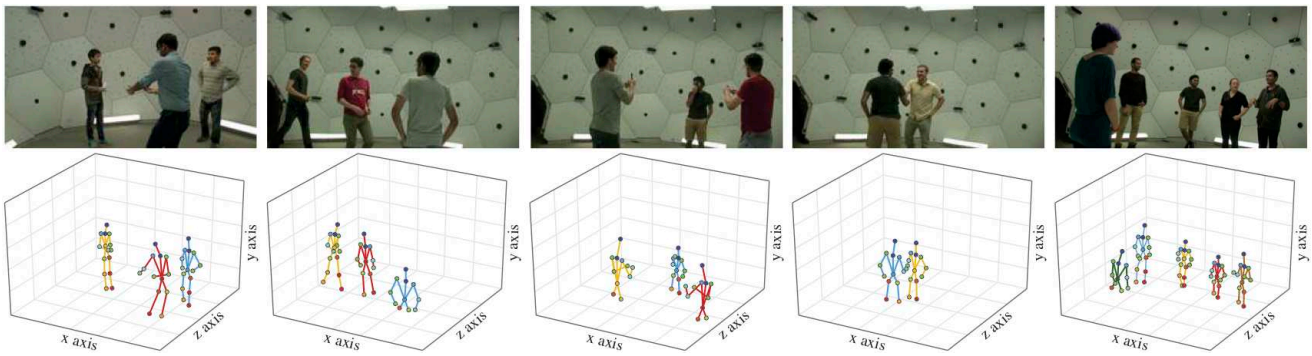


Figure 7. Qualitative results on CMU Panoptic dataset. 1st row is the input image and 2nd row is the corresponding multi-person 3D pose estimation with the proposed SPM. Best viewed in color and 2× zoom.

6. Conclusion

In this paper, we present the first single-stage model, Single-stage multi-person Pose Machine (SPM), for multi-person pose estimation. The SPM model offers a more compact pipeline and attractive efficiency advantage over existing two-stage based solutions. The superiority of SPM mainly comes from a novel Structured Pose Representation (SPR) that unifies the person instance and body joint position information and overcomes the intrinsic limitations of conventional pose representations. In addition, we present a hierarchical extension of SPR to effectively factorize long-range displacements into accumulative short-range ones between adjacent articulated joints, without introducing extra complexity to SPR. With SPR, SPM can estimate poses of multiple persons in a single-stage feed-forward manner. We implement SPM with CNNs, which can perform end-to-end learning and inference. Moreover, SPM can be flexibly adopted in both 2D and 3D scenarios. Extensive experiments on 2D benchmarks demonstrate the state-of-the-art speed of the proposed SPM model also with superior performance for predicting poses of multiple persons. Results on 3D benchmark also show the promising performance of our SPM model with attractive efficiency.

Acknowledgement

Jiashi Feng was partially supported by NUS IDS R-263-000-C67-646, ECRA R-263-000-C87-133 and MOE Tier-II R-263-000-D17-112.

References

[1] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *CVPR*, 2014.

[2] Mykhaylo Andriluka, Stefan Roth, and Bernt Schiele. Monocular 3d pose estimation and tracking by detection. In *CVPR*, 2010.

[3] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017.

[4] Joao Carreira, Pulkit Agrawal, Katerina Fragkiadaki, and Jitendra Malik. Human pose estimation with iterative error feedback. In *CVPR*, 2016.

[5] Xianjie Chen, Roozbeh Mottaghi, Xiaobai Liu, Sanja Fidler, Raquel Urtasun, and Alan L Yuille. Detect what you can: Detecting and representing objects using holistic models and body parts. In *CVPR*, 2014.

[6] Xianjie Chen and Alan L Yuille. Parsing occluded people by flexible compositions. In *CVPR*, 2015.

[7] Guilhem Chéron, Ivan Laptev, and Cordelia Schmid. P-cnn: Pose-based cnn features for action recognition. In *ICCV*, 2015.

[8] Junting Dong, Wen Jiang, Qixing Huang, Hujun Bao, and Xiaowei Zhou. Fast and robust multi-person 3d pose estimation from multiple views. *arXiv*, 2019.

[9] Haoshu Fang, Shuqin Xie, Yuwing Tai, and Cewu Lu. RMPE: Regional multi-person pose estimation. In *ICCV*, 2017.

[10] Mihai Fieraru, Anna Khoreva, Leonid Pishchulin, and Bernt Schiele. Learning to refine human pose estimation. In *CVPRw*, 2018.

[11] Ross Girshick. Fast r-cnn. In *ICCV*, 2015.

[12] Georgia Gkioxari, Bharath Hariharan, Ross Girshick, and Jitendra Malik. Using k-poselets for detecting people and localizing their keypoints. In *CVPR*, 2014.

[13] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, 2017.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[15] Eldar Insafutdinov, Mykhaylo Andriluka, Leonid Pishchulin, Siyu Tang, Bjoern Andres, and Bernt Schiele. Articulated multi-person tracking in the wild. In *CVPR*, 2017.

[16] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, and B. Schiele. Deepercut: A deeper, stronger, and faster multi-person pose estimation model. In *ECCV*, 2016.

[17] Umar Iqbal and Juergen Gall. Multi-person pose estimation with local joint-to-person associations. In *ECCV*, 2016.

[18] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. In *NIPS*, 2015.

[19] Hanbyul Joo, Tomas Simon, Xulong Li, Hao Liu, Lei Tan, Lin Gui, Sean Banerjee, Timothy Scott Godisart, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. Panoptic studio: A massively multiview system for social interaction capture. *IEEE Trans. on Pattern Anal. and Mach. Intell.*, 2017.

[20] Kathleen M Knutzen. *Kinematics of human motion*. Wiley Online Library, 1998.

[21] Evgeny Levinkov, Jonas Uhrig, Siyu Tang, Mohamed Omran, Eldar Insafutdinov, Alexander Kirillov, Carsten Rother, Thomas Brox, Bernt Schiele, and Bjoern Andres. Joint graph decomposition & node labeling: Problem, algorithms, applications. In *CVPR*, 2017.

[22] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, 2017.

[23] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014.

[24] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015.

[25] Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Srinath Sridhar, Gerard Pons-Moll, and Christian Theobalt. Single-shot multi-person 3d pose estimation from monocular rgb. In *3DV*, 2018.

[26] Alejandro Newell and Jia Deng. Associative embedding: End-to-end learning for joint detection and grouping. In *NIPS*, 2017.

[27] A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. In *ECCV*, 2016.

[28] George Papandreou, Tyler Zhu, Liang-Chieh Chen, Spyros Gidaris, Jonathan Tompson, and Kevin Murphy. Personlab: Person pose estimation and instance segmentation with a bottom-up, part-based, geometric embedding model. In *ECCV*, 2018.

[29] George Papandreou, Tyler Zhu, Nori Kanazawa, Alexander Toshev, Jonathan Tompson, Chris Bregler, and Kevin Murphy. Towards accurate multi-person pose estimation in the wild. In *CVPR*, 2017.

[30] Adam Paszke, Sam Gross, and Soumith Chintala. Pytorch, 2017.

[31] L. Pishchulin, E. Insafutdinov, S. Tang, B. Andres, M. Andriluka, P. Gehler, and B. Schiele. Deepcut: Joint subset partition and labeling for multi person pose estimation. In *CVPR*, 2016.

[32] Xuelin Qian, Yanwei Fu, Wenxuan Wang, Tao Xiang, Yang Wu, Yu-Gang Jiang, and Xiangyang Xue. Pose-normalized image generation for person re-identification. In *CVPR*, 2018.

[33] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, 2015.

[34] Gregory Rogez, Philippe Weinzaepfel, and Cordelia Schmid. Lcr-net++: Multi-person 2d and 3d pose detection in natural images. *IEEE Trans. on Pattern Anal. and Mach. Intell.*, 2019.

[35] Min Sun and Silvio Savarese. Articulated part-based model for joint object detection and pose estimation. In *ICCV*, 2011.

[36] T. Tieleman and G. Hinton. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA*, 2012.

[37] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh. Convolutional pose machines. In *CVPR*, 2016.

[38] Fangting Xia, Peng Wang, Xianjie Chen, and Alan L Yuille. Joint multi-person pose estimation and semantic part segmentation. In *CVPR*, 2017.