RMPE: Regional Multi-Person Pose Estimation

Hao-Shu Fang[1]*, Shuqin Xie[1], Yu-Wing Tai[2], Cewu Lu[1]§
[1]Shanghai Jiao Tong University, China [2] Tencent YouTu
fhaoshu@gmail.com qweasdshu@sjtu.edu.cn yuwingtai@tencent.com lucewu@sjtu.edu.cn

Abstract

Multi-person pose estimation in the wild is challenging. Although state-of-the-art human detectors have demonstrated good performance, small errors in localization and recognition are inevitable. These errors can cause failures for a single-person pose estimator (SPPE), especially for methods that solely depend on human detection results. In this paper, we propose a novel regional multi-person pose estimation (RMPE) framework to facilitate pose estimation in the presence of inaccurate human bounding boxes. Our framework consists of three components: Symmetric Spatial Transformer Network (SSTN), Parametric Pose Non-Maximum-Suppression (NMS), and Pose-Guided Proposals Generator (PGPG). Our method is able to handle inaccurate bounding boxes and redundant detections, allowing it to achieve **76.7** *mAP on the MPII (multi person) dataset[3]. Our model and source codes are made* ***publicly available****.*†.

1. Introduction

Human pose estimation is a fundamental challenge for computer vision. In practice, recognizing the pose of multiple persons in the wild is a lot more challenging than recognizing the pose of a single person in an image [30, 31, 21, 23, 38]. Recent attempts approach this problem by using either a two-step framework [28, 12] or a part-based framework [7, 27, 17]. The two-step framework first detects human bounding boxes and then estimates the pose within each box independently. The part-based framework first detects body parts independently and then assembles the detected body parts to form multiple human poses. Both frameworks have their advantages and disadvantages. In the two-step framework, the accuracy of pose estimation highly depends on the quality of the detected bounding boxes. In the part-based framework, the assembled human poses are ambiguous when two or more persons are too close together. Also, part-based framework loses the capability to recognize body parts from a global pose view due to the mere utilization of second-order body parts dependence.

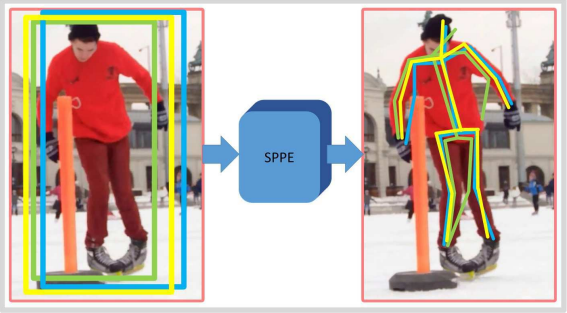

Figure 2. Problem of redundant human detections. The left image shows the detected bounding boxes; the right image shows the estimated human poses. Because each bounding box is operated on independently, multiple poses are detected for a single person.

Our approach follows the two-step framework. We aim to detect accurate human poses even when given inaccurate bounding boxes. To illustrate the problems of previous approaches, we applied the state-of-the-art object detector Faster-RCNN [29] and the SPPE Stacked Hourglass model [23]. Figure 1 and Figure 2 show two major problems: the localization error problem and the redundant detection problem. In fact, SPPE is rather vulnerable to bounding box errors. Even for the cases when the bounding boxes are considered as correct with $IoU > 0.5$, the detected human poses can still be wrong. Since SPPE produces a pose for each given bounding box, redundant detections result in redundant poses.

To address the above problems, a regional multi-person pose estimation (RMPE) framework is proposed. Our framework improves the performance of SPPE-based human pose estimation algorithms. We have designed a new symmetric spatial transformer network (SSTN) which is attached to the SPPE to extract a high-quality single person region from an inaccurate bounding box. A novel parallel SPPE branch is introduced to optimize this network. To address the problem of redundant detection, a parametric

*part of this work was done when Hao-Shu Fang was an student intern in Tencent
§corresponding author is Cewu Lu
†https://cvsjtu.wordpress.com/rmpe-regional-multi-person-pose-estimation/

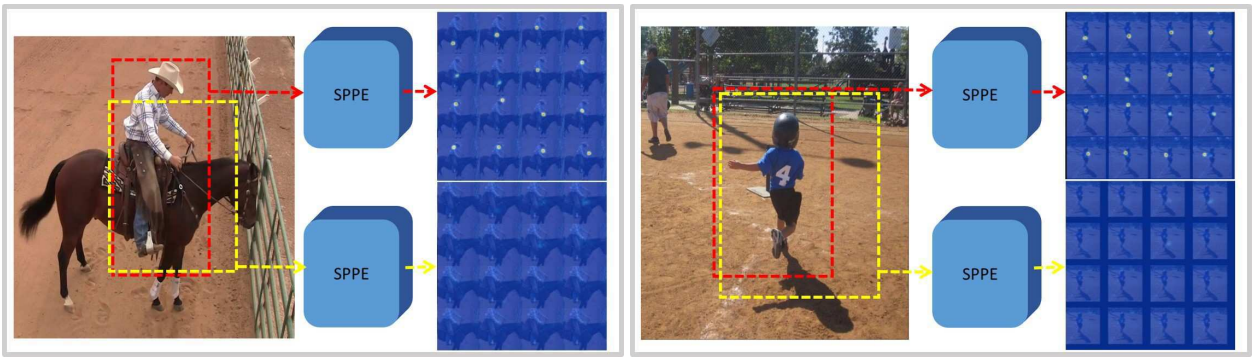

Figure 1. Problem of bounding box localization errors. The red boxes are the ground truth bounding boxes, and the yellow boxes are detected bounding boxes with $IoU > 0.5$. The heatmaps are the outputs of SPPE [23] corresponding to the two types of boxes. The corresponding body parts are not detected in the heatmaps of the yellow boxes. Note that with $IoU > 0.5$, the yellow boxes are considered as "correct" detections. However, human poses are not detected even with the "correct" bounding boxes.

pose NMS is introduced. Our parametric pose NMS eliminates redundant poses by using a novel pose distance metric to compare pose similarity. A data-driven approach is applied to optimize the pose distance parameters. Lastly, we propose a novel pose-guided human proposal generator (PGPG) to augment training samples. By learning the output distribution of a human detector for different poses, we can simulate the generation of human bounding boxes, producing a large sample of training data.

Our RMPE framework is general and is applicable to different human detectors and single person pose estimators. We applied our framework on the MPII (multi-person) dataset [3], where it outperforms the state-of-the-art methods and achieves 76.7 mAP. We have also conducted ablation studies to validate the effectiveness of each proposed component of our framework. Our model and source codes are made publicly available to support reproducible research.

2. Related Work

2.1. Single Person Pose Estimation

In single person pose estimation, the pose estimation problem is simplified by only attempting to estimate the pose of a single person, and the person is assumed to dominate the image content. Conventional methods considered pictorial structure models. For example, tree models [37, 30, 40, 36] and random forest models [31, 8] have demonstrated to be very efficient in human pose estimation. Graph based models such as random field models [20] and dependency graph models [14] have also been widely investigated in the literature [13, 32, 21, 26].

More recently, deep learning has become a promising technique in object/face recognition, and human pose estimation is of no exception. Representative works include DeepPose (Toshev *et al*) [34], DNN based models [24, 11] and various CNN based models [19, 33, 23, 4, 38]. Apart from simply estimating a human pose, some studies [9, 25] consider human parsing and pose estimation simultaneously. For single person pose estimation, these methods could perform well only when the person has been correctly located. However, this assumption is not always satisfied.

2.2. Multi Person Pose Estimation

Part-based Framework Representative works on part-based framework [7, 12, 35, 27, 17] are reviewed. Chen *et al*. presented an approach to parse largely occluded people by graphical model which models humans as flexible compositions of body parts [7]. Gkiox *et al* used k-poselets to jointly detect people and predict locations of human poses [12]. The final pose localization is predicted by a weighted average of all activated poselets. Pishchulin *et al*. proposed DeepCut to first detect all body parts, and then label and assemble these parts via integral linear programming[27]. A stronger part detector based on ResNet[15] and a better incremental optimization strategy is proposed by Insafutdinov *et al* [17]. While part-based methods have demonstrated good performance, their body-part detectors can be vulnerable since only small local regions are considered.

Two-step Framework Our work follows the two-step framework [28, 12]. In our work, we use a CNN based SPPE method to estimate poses, while Pishchulin *et al*. [28] used conventional pictorial structure models for pose estimation. In particular, Insafutdinov *et al* [17] propose a similar two-step pipeline which uses the Faster R-CNN as their human detector and a unary DeeperCut as their pose estimator. Their method can only achieve 51.0 in mAP on MPII dataset, while ours can achieve 76.7 mAP. With the development of object detection and single person pose estimation, the two-step framework can achieve further advances in its performance. Our paper aims to solve the problem of imperfect human detection in the two-step framework in order to maximize the power of SPPE.

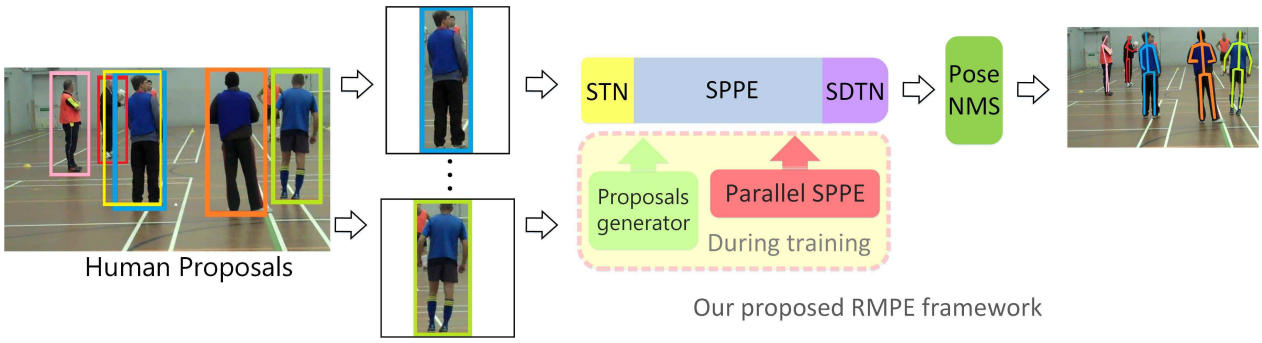

Figure 3. Pipeline of our RMPE framework. Our **Symmetric STN** consists of **STN** and **SDTN** which are attached before and after the SPPE. The **STN** receives human proposals and the **SDTN** generates pose proposals. The **Parallel SPPE** acts as an extra regularizer during the training phase. Finally, the **parametric Pose NMS (p-Pose NMS)** is carried out to eliminate redundant pose estimations. Unlike traditional training, we train the SSTN+SPPE module with images generated by **PGPG**.

3. Regional Multi-person Pose Estimation

The pipeline of our proposed RMPE is illustrated in Figure 3. The human bounding boxes obtained by the human detector are fed into the "**Symmetric STN** + SPPE" module, and the pose proposals are generated automatically. The generated pose proposals are refined by **parametric Pose NMS** to obtain the estimated human poses. During the training, we introduce "**Parallel SPPE**" in order to avoid local minimums and further leverage the power of SSTN. To augment the existing training samples, a **pose-guided proposals generator (PGPG)** is designed. In the following sections, we present the three major components of our framework.

3.1. Symmetric STN and Parallel SPPE

Human proposals provided by human detectors are not well-suited to SPPE. This is because SPPE is specifically trained on single person images and is very sensitive to localisation errors. It has been shown that small translation or cropping of human proposals can significantly affect performance of SPPE [23]. Our symmetric STN + parallel SPPE was introduced to enhance SPPE when given imperfect human proposals. The module of our SSTN and parallel SPPE is shown in Figure 4.

STN and SDTN The spatial transformer network [18](STN) has demonstrated excellent performance in selecting region of interests automatically. In this paper, we use the STN to extract high quality dominant human proposals. Mathematically, the STN performs a 2D affine transformation which can be expressed as

$$\begin{pmatrix} x_i^s \\ y_i^s \end{pmatrix} = \begin{bmatrix} \boldsymbol{\theta_1} & \boldsymbol{\theta_2} & \boldsymbol{\theta_3} \end{bmatrix} \begin{pmatrix} x_i^t \\ y_i^t \\ 1 \end{pmatrix}, \tag{1}$$

where $\boldsymbol{\theta_1}$, $\boldsymbol{\theta_2}$ and $\boldsymbol{\theta_3}$ are vectors in $\mathbb{R}^2$. $\{x_i^s, y_i^s\}$ and $\{x_i^t, y_i^t\}$ are the coordinates before and after transformation, respectively. After SPPE, the resulting pose is mapped into the original human proposal image. Naturally, a spatial de-transformer network (SDTN) is required to remap the estimated human pose back to the original image coordinate. The SDTN computes the $\boldsymbol{\gamma}$ for de-transformation and generates grids based on $\boldsymbol{\gamma}$:

$$\begin{pmatrix} x_i^t \\ y_i^t \end{pmatrix} = \begin{bmatrix} \boldsymbol{\gamma_1} & \boldsymbol{\gamma_2} & \boldsymbol{\gamma_3} \end{bmatrix} \begin{pmatrix} x_i^s \\ y_i^s \\ 1 \end{pmatrix} \tag{2}$$

Since SDTN is an inverse procedure of STN, we can obtain the following:

$$\begin{bmatrix} \boldsymbol{\gamma_1} & \boldsymbol{\gamma_2} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\theta_1} & \boldsymbol{\theta_2} \end{bmatrix}^{-1} \tag{3}$$

$$\boldsymbol{\gamma_3} = -1 \times \begin{bmatrix} \boldsymbol{\gamma_1} & \boldsymbol{\gamma_2} \end{bmatrix} \boldsymbol{\theta_3} \tag{4}$$

To back propagate through SDTN, $\frac{\partial J(W,b)}{\partial \theta}$ can be derived as

$$\begin{aligned} \frac{\partial J(W,b)}{\partial \begin{bmatrix} \boldsymbol{\theta_1} & \boldsymbol{\theta_2} \end{bmatrix}} &= \frac{\partial J(W,b)}{\partial \begin{bmatrix} \boldsymbol{\gamma_1} & \boldsymbol{\gamma_2} \end{bmatrix}} \times \frac{\partial \begin{bmatrix} \boldsymbol{\gamma_1} & \boldsymbol{\gamma_2} \end{bmatrix}}{\partial \begin{bmatrix} \boldsymbol{\theta_1} & \boldsymbol{\theta_2} \end{bmatrix}} \\ &+ \frac{\partial J(W,b)}{\partial \boldsymbol{\gamma_3}} \times \frac{\partial \boldsymbol{\gamma_3}}{\partial \begin{bmatrix} \boldsymbol{\gamma_1} & \boldsymbol{\gamma_2} \end{bmatrix}} \times \frac{\partial \begin{bmatrix} \boldsymbol{\gamma_1} & \boldsymbol{\gamma_2} \end{bmatrix}}{\partial \begin{bmatrix} \boldsymbol{\theta_1} & \boldsymbol{\theta_2} \end{bmatrix}} \end{aligned} \tag{5}$$

with respect to θ_1 and θ_2, and

$$\frac{\partial J(W,b)}{\partial \boldsymbol{\theta_3}} = \frac{\partial J(W,b)}{\partial \boldsymbol{\gamma_3}} \times \frac{\partial \boldsymbol{\gamma_3}}{\partial \boldsymbol{\theta_3}} \tag{6}$$

with respect to θ_3. $\frac{\partial \begin{bmatrix} \boldsymbol{\gamma_1} & \boldsymbol{\gamma_2} \end{bmatrix}}{\partial \begin{bmatrix} \boldsymbol{\theta_1} & \boldsymbol{\theta_2} \end{bmatrix}}$ and $\frac{\partial \boldsymbol{\gamma_3}}{\partial \boldsymbol{\theta_3}}$ can be derived from Eqn. (3) and (4) respectively.

After extracting high quality dominant human proposal regions, we can utilize off-the-shelf SPPE for accurate pose estimation. In our training, the SSTN is fine-tuned together with our SPPE.

Parallel SPPE To further help STN extract good human-dominant regions, we add a parallel SPPE branch in the training phrase. This branch shares the same STN with

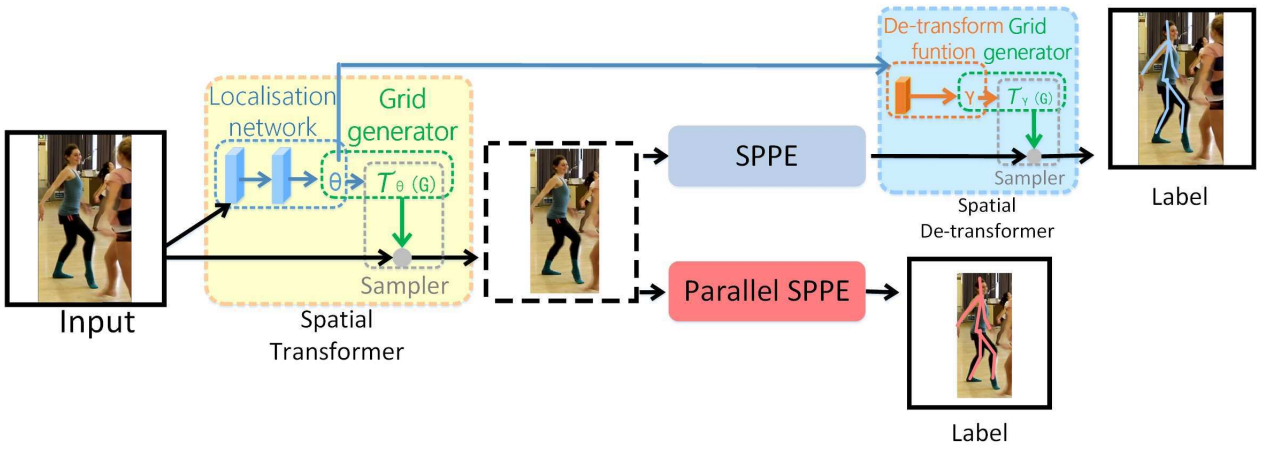

Figure 4. An illustration of our symmetric STN architecture and our training strategy with parallel SPPE. The STN used was developed by Jaderberg *et al.* [18]. Our SDTN takes a parameter $\boldsymbol{\theta}$, generated by the localization net and computes the $\boldsymbol{\gamma}$ for de-transformation. We follow the grid generator and sampler [18] to extract a human-dominant region. For our parallel SPPE branch, a center-located pose label is specified. We freeze the weights of all layers of the parallel SPPE to encourage the STN to extract a dominant single person proposal.

the original SPPE, but the spatial de-transformer (SDTN) is omitted. The human pose label of this branch is specified to be centered. To be more specific, the output of this SPPE branch is directly compared to labels of center-located ground truth poses. We freeze all the layers of this parallel SPPE during the training phase. The weights of this branch are fixed and its purpose is to back-propagate center-located pose errors to the STN module. If the extracted pose of the STN is not center-located, the parallel branch will back-propagate large errors. In this way, we can help the STN focus on the correct area and extract high quality human-dominant regions. In the testing phase, the parallel SPPE is discarded. The effectiveness of our parallel SPPE will be verified in our experiments.

Discussions The parallel SPPE can be regarded as a regularizer during the training phase. It helps to avoid a poor solution (local minimum) where the STN does not transform the pose to the center of extracted human regions. The likelihood of reaching a local minimum is increased because compensation from the SDTN will make the network generate fewer errors. These errors are necessary to train the STN. With the parallel SPPE, the STN is trained to move the human to the center of the extracted region to facilitate accurate pose estimation by SPPE.

It may seem intuitive to replace parallel SPPE with a center-located poses regression loss in the output of SPPE (before SDTN). However, this approach will degrade the performance of our system. Although STN can partly transform the input, it is impossible to perfectly place the person at the same location as the label. The difference in coordinate space between the input and label of SPPE will largely impair its ability to learn pose estimation. This will cause the performance of our main branch SPPE to decrease. Thus, to ensure that both STN and SPPE can fully leverage their own power, a parallel SPPE with frozen weights is indispensable for our framework. The parallel SPPE always produces large errors for non-center poses to push the STN to produce a center-located pose, without affecting the performance of the main branch SPPE.

3.2. Parametric Pose NMS

Human detectors inevitably generate redundant detections, which in turn produce redundant pose estimations. Therefore, pose non-maximum suppression (NMS) is required to eliminate the redundancies. Previous methods [5, 7] are either not efficient or not accurate enough. In this paper, we propose a parametric pose NMS method. Similar to the previous subsection, the pose P_i, with m joints is denoted as $\{\langle k_i^1, c_i^1\rangle, \ldots, \langle k_i^m, c_i^m\rangle\}$, where k_i^j and c_i^j are the j^{th} location and confidence score of joints respectively.

NMS scheme We revisit pose NMS as follows: firstly, the most confident pose is selected as reference, and some poses close to it are subject to elimination by applying *elimination criterion*. This process is repeated on the remaining poses set until redundant poses are eliminated and only unique poses are reported.

Elimination Criterion We need to define pose similarity in order to eliminate the poses which are too close and too similar to each others. We define a pose distance metric $d(P_i, P_j|\Lambda)$ to measure the pose similarity, and a threshold η as elimination criterion, where Λ is a parameter set of function $d(\cdot)$. Our elimination criterion can be written as follows:

$$f(\ P_i, P_j|\Lambda, \eta) = 1[d(P_i, P_j|\Lambda, \lambda) \leq \eta] \qquad (7)$$

If $d(\cdot)$ is smaller than η, the output of $f(\cdot)$ should be 1,

which indicates that pose P_i should be eliminated due to redundancy with reference pose P_j.

Pose Distance Now, we present the distance function $d_{pose}(P_i, P_j)$. We assume that the box for P_i is B_i. Then we define a soft matching function

$$K_{Sim}(P_i, P_j|\sigma_1) = \begin{cases} \sum_n \tanh \frac{c_i^n}{\sigma_1} \cdot \tanh \frac{c_j^n}{\sigma_1}, & \text{if } k_j^n \text{ is within } \mathcal{B}(k_i^n) \\ 0 & \text{otherwise} \end{cases} \quad (8)$$

where $\mathcal{B}(k_i^n)$ is a box center at k_i^n, and each dimension of $\mathcal{B}(k_i^n)$ is $1/10$ of the original box B_i. The $\tanh$ operation filters out poses with low-confidence scores. When two corresponding joints both have high confidence scores, the output will be close to 1. This distance softly counts the number of joints matching between poses.

The spatial distance between parts is also considered, which can be written as

$$H_{Sim}(P_i, P_j|\sigma_2) = \sum_n \exp[-\frac{(k_i^n - k_j^n)^2}{\sigma_2}] \quad (9)$$

By combining Eqn (8) and (9), the final distance function can be written as

$$d(P_i, P_j|\Lambda) = K_{Sim}(P_i, P_j|\sigma_1) + \lambda H_{Sim}(P_i, P_j|\sigma_2) \quad (10)$$

where λ is a weight balancing the two distances and $\Lambda = \{\sigma_1, \sigma_2, \lambda\}$. Note that the previous pose NMS [7] set pose distance parameters and thresholds manually. In contrast, our parameters can be determined in a data-driven manner.

Optimization Given the detected redundant poses, the four parameters in the eliminate criterion $f(P_i, P_j|\Lambda, \eta)$ are optimized to achieve the maximal mAP for the validation set. Since exhaustive search in a 4D space is intractable, we optimize two parameters at a time by fixing the other two parameters in an iterative manner. Once convergence is achieved, the parameters are fixed and will be used in the testing phase.

3.3. Pose-guided Proposals Generator

Data Augmentation For the two-stage pose estimation, proper data augmentation is necessary to make the SSTN+SPPE module adapt to the 'imperfect' human proposals generated by the human detector. Otherwise, the module may not work properly in the testing phase for the human detector. An intuitive approach is to directly use bounding boxes generated by the human detector during the training phase. However, the human detector can only produce one bounding box for each person. By using the proposals generator, this quantity can be greatly increased. Since we already have the ground truth pose and an object detection bounding box for each person, we can generate a large sample of training proposals with the same distribution as the output of the human detector. With this technique, we are able to further boost the performance of our system.

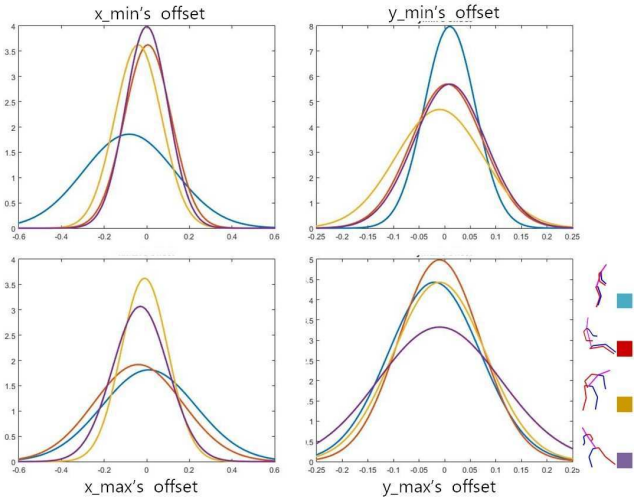

Figure 5. Gaussian distributions of bounding box offsets for several different atomic poses. More results are available in supplementary materials. Best viewed in color.

Insight We find that the distribution of the relative offset between the detected bounding box and the ground truth bounding box varies across different poses. To be more specific, there exists a distribution $P(\delta B|P)$, where δB is the offset between the coordinates of a bounding box generated by human detector and the coordinates of the ground truth bounding box, and P is the ground truth pose of a person. If we can model this distribution, we are able to generate many training samples that are similar to human proposals generated by the human detector.

Implementation To directly learn the distribution $P(\delta B|P)$ is difficult due to the variation of human poses. So instead, we attempt to learn the distribution $P(\delta B|atom(P))$, where $atom(P)$ denotes the atomic pose [39] of P. We follow the method used by Andriluka *et al* [3] to learn the atomic poses. To derive the atomic poses from annotations of human poses, we first align all poses so that their torsos have the same length. Then we use the k-means algorithm to cluster our aligned poses, and the computed cluster centers form our atomic poses. Now for each person instance sharing the same atomic pose a, we calculate the offsets between its ground truth bounding box and detected bounding box. The offsets are then normalized by the corresponding side-length of ground truth bounding box in that direction. After these processes, the offsets form a frequency distribution, and we fit our data to a Gaussian mixture distribution. For

different atomic poses, we have different Gaussian mixture parameters. We visualize some of the distributions and their corresponding clustered human poses in Figure 5.

Proposals Generation During the training phase of the SSTN+SPPE, for each annotated pose in the training sample we first look up the corresponding atomic pose a. Then we generate additional offsets by dense sampling according to $P(\delta B|a)$ to produce augmented training proposals.

4. Experiments

The proposed method is qualitatively and quantitatively evaluated on two standard multi-person datasets with large occlusion cases: MPII [3] and MSCOCO 2016 Keypoints Challenge dataset[1].

4.1. Evaluation datasets

MPII Multi-Person Dataset The challenging benchmark MPII Human Pose (multi-person)[3] consists of 3,844 training and 1,758 testing groups with both occluded and overlapped people. Moreover, it contains more than 28,000 training samples for single person pose estimation. We use all the training data in the single person dataset and 90% of the multi-person training set to fine-tune the SPPE, leaving 10% for validation.

MSCOCO Keypoints Challenge We also evaluate our method on the MSCOCO Keypoints Challenge dataset[1]. This dataset requires localization of person keypoints in challenging, uncontrolled conditions. It consists of 105,698 training and around 80,000 testing human instances. The training set contains over 1 million total labeled keypoints. The testing set are divided into four roughly equally sized splits: test-challenge, test-dev, test-standard, and test-reserve.

4.2. Implementation details in testing

In this paper, we use the VGG-based SSD-512 [22] as our human detector, as it performs object detection effectively and efficiently. In order to guarantee that the entire person region will be extracted, detected human proposals are extended by 30% along both the height and width directions. We use the stacked hourglass model [23] as the single person pose estimator because of its superior performance. For the STN network, we adopt the ResNet-18 [15] as our localization network. Considering the memory efficiency, we use a smaller 4-stack hourglass network as the parallel SPPE.

4.3. Results

Results on MPII dataset. We evaluated our method on full MPII multi-person test set. Quantitative results on the full testing set are given in Table 1. Notably, we achieve an average accuracy of 72 mAP on identifying difficult joints such as wrists, elbows, ankles, and knees, which is 3.3 mAP higher than the previous state-of-the-art result. We reach a final accuracy of 70.4 mAP for the wrist and an accuracy of 73 mAP for the knee. We present some of our results in Figure 6. These results show that our method can accurately predict pose in multi-person images. More results are presented in supplementary materials.

Results on MSCOCO Keypoints dataset. We fine-tuned the SPPE on the MSCOCO Keypoints training + validating sets and leave 5,000 images for validation. Quantitative results on the test-dev set are given in Table 2. Our method achieves the state-of-the-art performance. Note that when using a similar human detector and SPPE, our method outperforms *R4D* and *Caltech* by a great margin. Even when using an inferior human detector (SSD, 28.8 mAP VS G-RMI, 41.6 mAP on MSCOCO Detections test-dev set), without extra in-house training data and ensembling, our two step method outperforms the performance of *G-RMI* by 1.3 mAP, which demonstrates the effectiveness of our proposed framework.

Team	**AP**	AP^{50}	AP^{75}	AP^{M}	AP^{L}
CMU-Pose[6]	61.8	**84.9**	67.5	57.1	**68.2**
G-RMI	60.5	82.2	66.2	57.6	66.6
DL-61	54.4	75.3	50.9	58.3	54.3
R4D	51.4	75	55.9	47.4	56.7
umich_vl	46	74.6	48.4	38.8	55.6
Caltech	40.2	65.2	41.9	34.9	49.2
ours	**61.8**	83.7	**69.8**	**58.6**	67.6

Table 2. Results on the MSCOCO Keypoint Challenge (AP) dataset [2]. The MSCOCO website provides a technical overview only. Our result is obtained without ensembling.

4.4. Ablation studies

We evaluate the effectiveness of the three proposed components, i.e., symmetric STN, pose-guided proposals generator and parametric pose NMS. The ablative studies have been conducted by removing the proposed components from the pipeline or replacing the proposed components with conventional solvers. The straightforward two-step method without the three components and the upper-bound of our framework are tested for comparison. We conducted these experiments on the MPII validation set. In addition, we replace our human detection module to prove the generality of our framework.

Symmetric STN and Parallel SPPE To validate the importance of symmetric STN and parallel SPPE, two experiments were conducted. In the first experiment, we removed the SSTN, including the parallel SPPE, from our pipeline.

	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Total
full testing set								
Iqbal&Gall, ECCVw16 [35]	58.4	53.9	44.5	35.0	42.2	36.7	31.1	43.1
DeeperCut, ECCV16 [17]	78.4	72.5	60.2	51.0	57.2	52.0	45.4	59.5
Levinkov *et al.*, CVPR17[10]	89.8	85.2	71.8	59.6	71.1	63.0	53.5	70.6
Insafutdinov *et al.*, CVPR17[16]	88.8	87.0	75.9	64.9	74.2	68.8	60.5	74.3
Cao *et al.*, CVPR17[6]	**91.2**	**87.6**	77.7	66.8	**75.4**	68.9	61.7	75.6
ours	88.4	86.5	**78.6**	**70.4**	74.4	**73.0**	**65.8**	**76.7**

Table 1. Results on the MPII multi-person test set (mAP).

Figure 6. Some results of our model's predictions.

In the second experiment, we only removed the parallel SPPE and kept the symmetric STN structure. Both of these results are shown in Table 3(a). We can observe performance degradation when removing parallel SPPE, which implies that parallel SPPE with single person image labels strongly encourages the STN to extract single person regions to minimize the total losses.

Pose-guided Proposals Generator In Table 3(b), we demonstrate that our pose-guided proposals generator also plays an important role in our system. In this experiment, we first remove the data augmentation from our training phase. The final mAP drops to 73.0%. Then we compare our data augmentation technique with a simple baseline. The baseline is formed by jittering the locations and aspect ratios of the bounding boxes produced by person detector to generate a large number of additional proposals. We choose those that have IoU>0.5 with ground truth boxes. From our result in Table 3(b), we can see that our technique is better than the baseline method. Generating training proposals according to the distribution can be regarded as a kind of data re-sampling, which can help the model to better fit human proposals.

Parametric Pose NMS Since pose NMS is an independent module, we can directly remove it from our final model. The experimental results are shown in Table 3(c). As we can see, the mAP drops significantly if the parametric pose NMS is removed. This is because the increase in the number of redundant poses will ultimately decrease our precision. We note that the previous pose NMS can also eliminate redundant detection to some extent. The state-of-

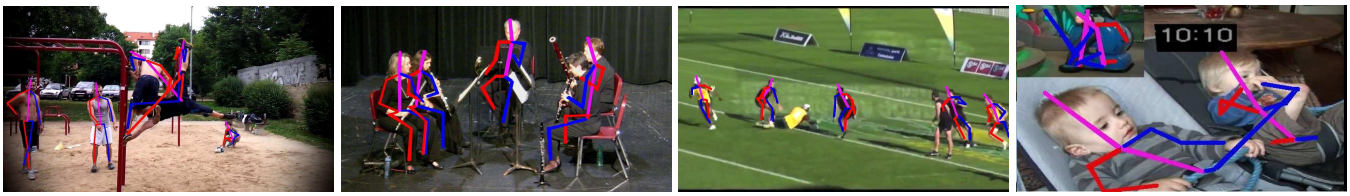

Figure 7. Example failure cases of our model.

	Methods	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Total
	RMPE, full	**90.7**	**89.7**	**84.1**	**75.4**	**80.4**	**75.5**	**67.3**	**80.8**
a)	w/o SSTN+parallel SPPE	89.0	86.9	82.8	73.5	77.1	73.3	65.0	78.2
	w/o parallel SPPE only	89.9	88.0	83.4	74.7	77.8	74.0	65.8	79.1
b)	w/o PGPG	82.8	81.0	77.5	68.2	74.6	66.8	60.1	73.0
	random jittering*	89.3	87.8	82.3	70.4	78.4	73.3	63.8	77.9
c)	w/o PoseNMS	85.1	83.6	79.2	69.8	76.4	72.2	63.6	75.7
	PoseNMS [7]	88.9	87.8	83.0	73.8	78.7	74.6	66.3	79.1
	PoseNMS [5]	90.0	88.6	83.7	74.6	79.7	75.1	67.0	79.9
d)	straight forward two-steps	81.9	80.4	74.1	68.5	69.0	66.1	62.2	71.7
e)	oracle human detection	94.3	93.4	87.7	80.2	84.3	78.9	70.6	84.2

Table 3. Results of the ablation experiments on our validation set. "w/o X" means without X module in our pipeline. "random jittering*" means generating training proposals by jittering locations and aspect ratios of the detected human bounding boxes. "PoseNMS [x]" reports the result when using the pose NMS algorithm developed in paper [x].

the-art pose NMS algorithms [5, 7] are used to replace our parametric pose NMS, with the results given in Table 3(c). These schemes perform less effectively than ours, since the parameter learning is missing. In terms of efficiency, on our validation set which contains 1300 images, the publicly available implementation of [5][‡]takes 62.2 seconds to perform pose NMS while using our algorithm takes only 1.8 seconds.

Upper Bound of Our Framework The upper bound of our framework is tested, where we use the ground truth bounding boxes as human proposals. As shown in Table 3(e), this setting could yield 84.2% mAP. It verifies that our system is already close to the upper-bound of two-step framework.

Human Detector Modules By replacing the human detector with ResNet50 based Faster-RCNN[29] (32.0 mAP on MSCOCO Detections test-dev set), the final result on our MPII valid set achieves 81.4 mAP and the result on MSCOCO Keypoints test-dev set achieves 63.3 mAP. It shows that with stronger human detector, our framework can have better performance, which proves that our RMPE framework is general and is applicable to different human detectors.

4.5. Failure cases

We present some failure cases in Figure 7. It can be seen that the SPPE can not handle poses which are rarely occurred (e.g. the person performing the 'Human Flag' in the first image). When two persons are highly overlapped, our system get confused and can not separate them apart (e.g. the two persons in the left of the second image). The misses of person detector will also cause the missing detection of human poses (e.g. the person who has laid down in the third image). Finally, erroneous pose may still be detected when an object looks very similar to human which can fool both human detector and SPPE (e.g. the background object in the forth image).

[‡]http://www.vision.caltech.edu/ dhall/projects/MergingPoseEstimates/

5. Conclusion

In this paper, a novel regional multi-person pose estimation (RMPE) framework is proposed, which significantly outperforms the state-of-the-art methods for multi-person human pose estimation in terms of accuracy and efficiency. It validates the potential of two-step frameworks, i.e., human detector + SPPE, when SPPE is adapted to a human detector. Our RMPE framework consists of three novel components: symmetric STN with parallel SPPE, parametric pose NMS, and pose-guided proposals generator (PGPG). In particular, PGPG is used to greatly argument the training data by learning the conditional distribution of bounding box proposals for a given human pose. The SPPE becomes adept at handling human localization errors due to the utilization of symmetric STN and parallel SPPE. Finally, the parametric pose NMS can be used to reduce redundant detections. In our future work, it would be interesting to explore the possibility of training our framework together with the human detector in an end-to-end manner.

References

[1] MSCOCO keypoint challenge 2016. http://mscoco.org/dataset/keypoints-challenge2016. 4326

[2] http://mscoco.org/dataset/#keypoints-leaderboard, 2016. 4326

[3] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. 4321, 4322, 4325, 4326

[4] V. Belagiannis and A. Zisserman. Recurrent human pose estimation. In *arXiv preprint arXiv:1605.02914*, 2016. 4322

[5] X. Burgos-Artizzu, D. Hall, P. Perona, and P. Dollar. Merging pose estimates across space and time. In *British Machine Vision Conference (BMVC)*, 2013. 4324, 4328

[6] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 4326, 4327

[7] X. Chen and A. L. Yuille. Parsing occluded people by flexible compositions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3945–3954, 2015. 4321, 4322, 4324, 4325, 4328

[8] M. Dantone, J. Gall, C. Leistner, and L. Van Gool. Human pose estimation using body parts dependent joint regressors. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3041–3048, 2013. 4322

[9] J. Dong, Q. Chen, X. Shen, J. Yang, and S. Yan. Towards unified human parsing and pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 843–850, 2014. 4322

[10] S. T. M. O. E. I. A. K. C. R. T. B. B. S. B. A. Evgeny Levinkov, Jonas Uhrig. Joint graph decomposition and node labeling: Problem, algorithms, applications. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 4327

[11] X. Fan, K. Zheng, Y. Lin, and S. Wang. Combining local appearance and holistic view: Dual-source deep neural networks for human pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1347–1355. IEEE, 2015. 4322

[12] G. Gkioxari, B. Hariharan, R. Girshick, and J. Malik. Using k-poselets for detecting people and localizing their keypoints. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3582–3589, 2014. 4321, 4322

[13] A. Gupta, T. Chen, F. Chen, D. Kimber, and L. S. Davis. Context and observation driven latent variable model for human pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8. IEEE, 2008. 4322

[14] K. Hara and R. Chellappa. Computationally efficient regression on a dependency graph for human pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3390–3397, 2013. 4322

[15] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. 2016. 4322, 4326

[16] E. Insafutdinov, M. Andriluka, L. Pishchulin, S. Tang, E. Levinkov, B. Andres, and B. Schiele. Arttrack: Articulated multi-person tracking in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 4327

[17] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, and B. Schiele. DeeperCut: A Deeper, Stronger, and Faster Multi-Person Pose Estimation Model. In *European Conference on Computer Vision (ECCV)*, May 2016. 4321, 4322, 4327

[18] M. Jaderberg, K. Simonyan, A. Zisserman, et al. Spatial transformer networks. In *Conference on Neural Information Processing Systems (NIPS)*, pages 2017–2025, 2015. 4323, 4324

[19] A. Jain, J. Tompson, M. Andriluka, G. W. Taylor, and C. Bregler. Learning human pose estimation features with convolutional networks. In *arXiv preprint arXiv:1312.7302*, 2013. 4322

[20] M. Kiefel and P. V. Gehler. Human pose estimation with fields of parts. In *European Conference on Computer Vision (ECCV)*, pages 331–346. Springer, 2014. 4322

[21] L. Ladicky, P. H. Torr, and A. Zisserman. Human pose estimation using a joint pixel-wise and part-wise formulation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3578–3585, 2013. 4321, 4322

[22] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, and S. Reed. SSD: Single Shot MultiBox Detector. In *European Conference on Computer Vision (ECCV)*, 2016. 4326

[23] A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. In *arXiv preprint arXiv:1603.06937*, 2016. 4321, 4322, 4323, 4326

[24] W. Ouyang, X. Chu, and X. Wang. Multi-source deep learning for human pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014. 4322

[25] S. Park and S.-C. Zhu. Attributed grammars for joint estimation of human attributes, part and pose. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2372–2380, 2015. 4322

[26] L. Pishchulin, M. Andriluka, P. Gehler, and B. Schiele. Strong appearance and expressive spatial models for human pose estimation. In *IEEE International Conference on Computer Vision (ICCV)*, pages 3487–3494, 2013. 4322

[27] 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 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 4321, 4322

[28] L. Pishchulin, A. Jain, M. Andriluka, T. Thormählen, and B. Schiele. Articulated people detection and pose estimation: Reshaping the future. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3178–3185, 2012. 4321, 4322

[29] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Conference on Neural Information Processing Systems (NIPS)*, pages 91–99, 2015. 4321, 4328

[30] B. Sapp, A. Toshev, and B. Taskar. Cascaded models for articulated pose estimation. In *European Conference on Computer Vision (ECCV)*, pages 406–420. Springer, 2010. 4321, 4322

[31] M. Sun, P. Kohli, and J. Shotton. Conditional regression forests for human pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3394–3401. IEEE, 2012. 4321, 4322

[32] M. Sun, M. Telaprolu, H. Lee, and S. Savarese. An efficient branch-and-bound algorithm for optimal human pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1616–1623. IEEE, 2012. 4322

[33] J. J. Tompson, A. Jain, Y. LeCun, and C. Bregler. Joint training of a convolutional network and a graphical model for human pose estimation. In *Conference on Neural Information Processing Systems (NIPS)*, pages 1799–1807, 2014. 4322

[34] A. Toshev and C. Szegedy. Deeppose: Human pose estimation via deep neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. 4322

[35] J. G. Umar Iqbal. Multi-person pose estimation with local joint-to-person associations. In *European Conference on Computer Vision Workshops 2016 (ECCVW'16) - Workshop on Crowd Understanding (CUW'16)*, 2016. 4322, 4327

[36] F. Wang and Y. Li. Beyond physical connections: Tree models in human pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 596–603, 2013. 4322

[37] Y. Wang and G. Mori. Multiple tree models for occlusion and spatial constraints in human pose estimation. In *European Conference on Computer Vision (ECCV)*, pages 710–724. Springer, 2008. 4322

[38] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh. Convolutional pose machines. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4724–4732, 2016. 4321, 4322

[39] B. Yao and L. Fei-Fei. Recognizing human-object interactions in still images by modeling the mutual context of objects and human poses. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 34(9):1691–1703, 2012. 4325

[40] X. Zhang, C. Li, X. Tong, W. Hu, S. Maybank, and Y. Zhang. Efficient human pose estimation via parsing a tree structure based human model. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1349–1356. IEEE, 2009. 4322