Spatial-Aware Regression for Keypoint Localization

Dongkai Wang Shiliang Zhang
National Key Laboratory for Multimedia Information Processing,
School of Computer Science, Peking University
{dongkai.wang, slzhang.jdl}@pku.edu.cn

Abstract

Regression-based keypoint localization shows advantages of high efficiency and better robustness to quantization errors than heatmap-based methods. However, existing regression-based methods discard the spatial location prior in input image with a global pooling, leading to inferior accuracy and are limited to single instance localization tasks. We study the regression-based keypoint localization from a new perspective by leveraging the spatial location prior. Instead of regressing on the pooled feature, the proposed Spatial-Aware Regression (SAR) maintains the spatial location map and outputs spatial coordinates and confidence score for each grid, which are optimized with a unified objective. Benefited by the location prior, these spatial-aware outputs can be efficiently optimized, resulting in better localization performance. Moreover, incorporating spatial prior makes SAR more general and can be applied into various keypoint localization tasks. We test the proposed method in 4 keypoint localization tasks including single/multi-person 2D/3D pose estimation, and the whole-body pose estimation. Extensive experiments demonstrate its promising performance, e.g., consistently outperforming recent regressions-based methods[†].

1. Introduction

Keypoint localization aims to locate target keypoints from an input image and is a fundamental task in the field of computer vision. It has a wide range of applications in human pose estimation [21, 26–28] and facial landmark detection [19], *et al*. Existing methods for keypoint localization can be summarized into two categories: heatmap-based [21, 29, 31] and regression-based [10, 23, 25], respectively. Regression-based method directly adopts neural network to learn the mapping from input RGB image to keypoint coordinates. Heatmap-based method uses a probability map (also referred as heatmap) to encode the likelihood of the target location and retrieves it by selecting location with the highest probability.

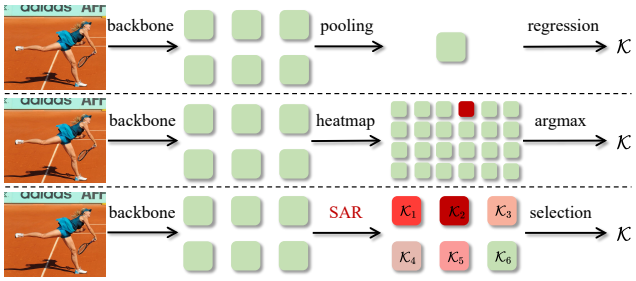

Figure 1. Illustration of (Top) regression-based method, (Middle) standard heatmap-based method, and (Bottom) the proposed SAR for keypoint localization.

As illustrated in Fig. 1, heatmap-based method selects pre-defined points on heatmaps as localization results, which are easy to optimize. However, low-resolution heatmap leads to high quantization error and high-resolution heatmap enlarges the computation and storage cost. Regression-based method is more efficient and robust to quantization error, but is hard to optimize and commonly achieve inferior performance. One reason is that conventional regression destroys the spatial location information of the feature map by a global pooling, thus cannot provide a good initialization for the following regression. This design also limits the application of regression to differentiate and locate multiple keypoints of the same type, *e.g.*, multi-person pose estimation.

This work is motivated to facilitate the regression-based localization by embedding spatial location prior into regression. Grids in the extracted feature map provide different starting points for regression, making them fitted to locate different keypoints. Regressing from different starting points also introduces duplicate predictions, and we do not know which grids produce the best localization results. Prior works [5, 18] assume the results of grids near the target location are accurate and select them via a separate classification branch. We argue that this heuristic design is not optimal in all cases, *e.g.*, for occluded or truncated

[†]Project page: https://github.com/kennethwdk/SAR

keypoints. Moreover, this multi-task learning pipeline introduces optimization inconsistency between classification and regression [2] and is sensitive to many hyperparameters.

This work presents the Spatial-Aware Regression (SAR), a novel regression method that effectively utilizes the spatial location prior in input image to generate spatial-aware outputs and automatically select the best prediction. As shown in Fig. 1, SAR regresses coordinates on each grid to utilize its spatial location cues. Benefited by the prior, SAR is able to leverage better starting points. Selecting different starting points also helps to differentiate similar keypoints, making SAR applicable to more challenging tasks like multi-person pose estimation.

SAR performs localization on a set of grids, which can be extracted by deep neural networks like CNN [4] or Vision Transformer [31]. To utilize the spatial location prior, SAR introduces a spatial-aware regressor to locate target locations based on spatial location of each grid. To handle duplicate predictions, we propose a spatial-aware selector to evaluate the quality of each regression as confidence score, and select the best prediction. The selector is jointly optimized with the regressor with a unified objective, leading to automatic regression and selection without heuristic design and complex hyperparameters. The introduced selector also depresses the influence of inaccurate predictions. SAR can work well on a low-resolution feature map, thus introduces marginal computational overheads and maintains similar efficiency with existing regression-based methods.

SAR shares all merits of conventional regression and surpasses it in many aspects. We test its effectiveness on various keypoint localization tasks including single/multi-person 2D/3D pose estimation and whole-body pose estimation. Extensive experiments on 7 keypoint localization benchmarks demonstrate its superior performance in keypoint localization. For example, SAR obtains 72.5% AP on COCO Keypoint dataset [14], which is higher than conventional regression and heatmap by 16.5% and 1.8%. SAR is robust to various input size and output stride, making it more general to deal with complex scenarios. SAR can also generalize well to detect various types of keypoints, arbitrary number of keypoints, as well as 2D/3D keypoints.

2. Related Work

Heatmap-based Keypoint Localization encodes keypoint location with a probability map, which is introduced by [24]. This type of methods estimates heatmaps and retrieves keypoint coordinates with a post-processing operation. Heatmap-based methods dominate the field of keypoint localization because heatmap is easy to learn with CNN. Pioneer works [16, 21, 29] design powerful CNN models to estimate high resolution heatmaps for human pose estimation and facial landmark detection, then the target keypoint can be simply obtained by a post-processing shifting [16, 33]. Due to the limitation of feature map size, some works [18] combine regression and add an offset branch to avoid quantization error. These methods improve the performance of heatmap. However, they rely on high resolution heatmap to locate keypoints, which results in high computation and storage cost.

Regression-based Keypoint Localization directly learns the mapping from input image to output coordinates via a neural network, which is adopted by several classical methods [1, 25]. Researchers have proposed many methods to improve the performance of direct regression. The first kind of methods changes the way of regression. Integral pose regression [23] leverages the soft-argmax operation to regress keypoint locations by integrating a latent heatmap, which is proved to be superior to direct regression. Sampling-argmax [11] further improves soft-argmax by minimizing the error between samples drawing from a distribution with groundtruth, avoiding unconstrained probability map in previous method. Some work improves regression by proposing new loss functions. RLE [10] changes the predefined Gaussian or Laplace distribution in commonly used regression loss with a learned distribution via normalizing flow. Recently, researchers also try to improve direct regression by proposing more powerful backbones in Transformer architecture [31, 32], such as TokenPose [12] and PETR [20].

Although many regression-based works have been proposed, they ignore the spatial location prior, leading to inferior performance and cannot be applied to multiple keypoints localization tasks. This work shows that embedding the spatial prior into regression significantly improves its performance and generalization capability on various human keypoint localization tasks. A more detailed comparison with heatmap-based and existing regression-based methods is presented in the Sec. 3.3.

3. Method

3.1. Overview

The goal of keypoint localization is to estimate the coordinates of target keypoints from input images, which can be denoted as,

$$\{\mathcal{K}_s\}_{s=1}^{m} = \text{locate}(\mathcal{I}), \tag{1}$$

where $\mathcal{K}_s$ denotes the s-th keypoint coordinates and m is the number of keypoints in this image, which is equal to 1 for single-keypoint localization and larger than 1 for multi-keypoint localization.

Given an input image $\mathcal{I} \in \mathbb{R}^{3\times H\times W}$, existing methods first adopt a backbone $\Phi(\cdot)$, *e.g.*, CNN-based [21] or Transformer-based [31] network to extract feature map $\mathbf{F} \in \mathbb{R}^{c\times h\times w}$, *i.e.*,

$$\mathbf{F} = \{\mathcal{F}_{i,j}, \mathcal{P}_{i,j}\}_{i=1\ldots h}^{j=1\ldots w} = \Phi(\mathcal{I}), \tag{2}$$

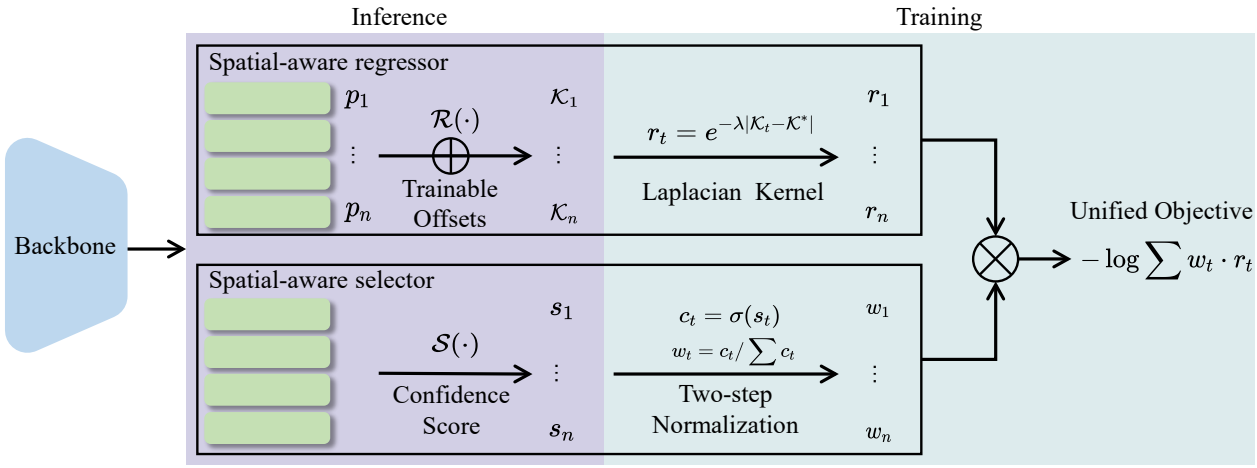

Figure 2. The pipeline of the proposed Spatial-Aware Regression (SAR) for keypoint localization. SAR first adopts a backbone to extract a set of features, which are passed to spatial-aware regressor and selector to generate corresponding keypoint coordinates and confidence scores. During training, SAR applies Laplacian kernel function and normalization to convert the outputs into corresponding score and calculate the unified objective to train the whole model. During inference, SAR selects regressed outputs with large confidence score.

where $f = \mathcal{F}_{i,j}$ is a visual feature on the feature map $\mathbf{F}$, $p = \mathcal{P}_{i,j}$ is the spatial location of this grid, h, w denotes the spatial size of $\mathbf{F}$. $\mathcal{P} = \{(i,j)\}_{i=1\ldots h}^{j=1\ldots w}$ is named as the spatial location prior, which is crucial for keypoint localization.

Conventional regression will apply a pooling on $\mathbf{F}$ to get a global feature f^g, then adopt a regressor to obtain final localization results,

$$\mathcal{K} = \mathcal{R}(f^g). \tag{3}$$

A ℓ_1 or ℓ_2 loss is applied on the regressed output to train model, *e.g.*, $\mathcal{L} = \ell_1(\mathcal{K}, \mathcal{K}^*)$.

The above paradigm has several drawbacks. First, pooling only preserves visual information f^g and discards the spatial location prior $\mathcal{P}$, making Eq. (3) hard to optimize. This pooling operation also makes regression sensitive to instance scale. Second, it is difficult to differentiate multiple keypoints of the same category only with f^g, because those keypoints present similar visual appearances and different locations. Although extra operation such as box cropping can remedy this issue, it involves extra object detection process and can be sensitive to detection error.

We present the Spatial-Aware Regression (SAR) method to integrate the spatial location prior $\mathcal{P}$ into regression to pursue better localization performance. SAR adopts both spatial location prior $\mathcal{P}$ and visual feature f^g or $\mathcal{F}$ to perform localization, which can be denoted as,

$$\mathcal{K} = \mathrm{SAR}(\mathcal{P}, f^g | \mathcal{F}). \tag{4}$$

SAR shares all the merits of regression-based methods, which are efficient and robust to quantization error. Benefited by the spatial location prior, SAR is easier to optimize and achieves superior keypoint localization performance. It also works well in multi-keypoints localization tasks to detect arbitrary number of keypoints from an input image. All these make SAR a better and robust keypoint localization method. In the following we will present the detailed implementation of SAR.

3.2. Spatial-Aware Regression

Given the spatial location prior $\mathcal{P}$ and visual feature $\mathcal{F}$ or f^g, SAR aims to locate the target location $\mathcal{K}^*$. We first consider the single keypoint localization that only one target exists in $\mathcal{I}$. Multiple keypoints localizations can be estimated by repeating the same process for each target in the input image. The core components of SAR are spatial-aware regressor to generate multiple outputs and spatial-aware selector to select the optimal output, respectively. Both components are jointly optimized by a unified training objective.

Spatial-aware regressor aims to get the target location by regressing the coordinates from each grid in $\mathcal{P}$. To relieve the difficulty of direct regression, we introduce the spatial location prior in original $\mathbf{F}$ into regression process, which can be denoted as,

$$\mathcal{K}_t = \mathcal{R}(f_t) + p_t, \tag{5}$$

where f_t denotes the visual feature of t-th grid and p_t is its location prior. $\mathcal{K}_t$ is the regressed output by t-th grid. f_t can be obtained by directly taking the feature at corresponding location of $\mathcal{F}_t$. It can also be generated from the pooled feature f^g with a grid-wise FC layer, which produces a comparable performance as shown in our experiments.

Compared with Eq. (3), Eq. (5) involves more detailed feature grids and each grid has different spatial location prior p_t to regress the target. Intuitively, for some grids

near the target $\mathcal{K}^*$, they will produce more reliable predictions, thus relieving the difficulty of direct regression in Eq. (3). As each grid generates a prediction, we further propose spatial-aware selector to handle duplicate predictions and select the accurate one.

Spatial-aware selector. It is not reasonable to optimize every $\{\mathcal{K}_t\}$ *w.r.t.* the target, because not all grids are suited to predict a specific keypoint. Equally optimizing all $\{\mathcal{K}_t\}$ will cause the learning degenerate to conventional regression as the model tends to focus on learning hard samples for localization. To reduce the influence of inaccurate predictions and improve the training efficiency, we introduce a spatial-aware selector $\mathcal{S}(\cdot)$ to automatically evaluate the importance of each grid in locating target $\mathcal{K}^*$. We predict a confidence score for each regressed output based on the feature f_t, *i.e.*,

$$s_t = \mathcal{S}(f_t), \tag{6}$$

where s_t denotes the score of selecting the t-th output as the final result. Adopting a heuristic predefined strategy, *e.g.*, selecting grids near the target, is unreasonable to optimize $\mathcal{S}(\cdot)$ and may introduce optimization conflict with regressor. Instead, we jointly optimize the selector with regressor under a unified objective, letting model to select proper points by itself.

Unified objective. The goal of SAR is to optimize $\{\mathcal{K}_t, s_t\}$ to accurately locate target $\mathcal{K}^*$. We propose a unified objective to jointly optimize the regressor and selector to avoid optimization contradiction in naive multi-task learning pipeline. We rely on regression to achieve the goal of localization and regard selection as an auxiliary task to reduce the difficulty of regression. This intuition leads to a unified objective function that aims to maximize the overall regression quality score weighted by the regression confidence, *i.e.*,

$$\mathcal{L} = -\log \sum w_t \cdot r_t, \tag{7}$$

where $r_t \in [0, 1]$ denotes the regression quality score of the t-th output and w_t denotes the corresponding regression weight. During training, we measure the regression quality score of each grid with the groundtruth $\mathcal{K}^*$ by a Laplacian kernel to convert the estimated location $\mathcal{K}_t$ into a score r_t, that is

$$r_t = e^{-\lambda \cdot |\mathcal{K}_t - \mathcal{K}^*|}, \tag{8}$$

where $r_t \in [0, 1]$ indicates the accuracy of the regressed output and λ is the only hyperparameter of SAR. We set λ to 16 for all experiments and normalize the coordinates by the feature map size.

The weight w_t generated by the t-th grid should be correlated with other grids. In other words, once some outputs are selected, the remaining should be suppressed. To achieve this goal, we adopt a two-step normalization operation on the confidence score $\{s_t\}$ to generate w_t, *i.e.*,

$$c_t = \sigma(s_t), w_t = \frac{c_t}{\sum c_t}, \tag{9}$$

where $\sigma(\cdot)$ denotes the `sigmoid` function to convert score into $c_t \in [0, 1]$. Eq. (9) first normalizes each score into $[0, 1]$ individually then utilizes element-sum to correlate them. Therefore, optimizing a high score s_t will decrease others. One advantage of the two-step normalization is the introduced extra score c_t, which is not influenced by the target number and can be used to select outputs by a fixed threshold during inference. Another way to generate weight w_t is to apply `softmax` on s_t. However, this one-step normalization cannot generate reliable high confidence score in multiple keypoints localization task during inference, because the value of generated w_t depends on the number of targets in the image. Therefore, it is hard to select final outputs by a fixed threshold γ.

The unified objective weights each output by the confidence score and is adopted to maximize the overall regression quality. We then give a detailed analysis on Eq. (7) and show how it optimizes both coordinates and confidence scores with a unified loss function. Specifically, the goal of SAR is to minimize the following objective,

$$\arg\min_{w_t, r_t} -\log \sum w_t \cdot r_t, \text{ s.t. } \sum w_t = 1. \tag{10}$$

Due to the monotonic increase property of $\log(\cdot)$ function, the optimal w_t, r_t satisfy that $\sum w_i \cdot r_t = 1$. With the normalization in selection process, we also can get $\sum w_t = 1$. Therefore, optimizing Eq. (10) is equal to maximizing the regression score r_t and will result in two cases:

- $\forall t \in [1, n], r_t \to 1$. This means that all grids can accurately regress the target, *i.e.*, $\mathcal{K}_t = \mathcal{K}^*$. In this case, w_t is not important because we can select any output as final result. However, this case rarely happens, especially for multi-keypoints localization task where the same regressed output $\mathcal{K}_t$ will be used to compute r_t for different groundtruth $\mathcal{K}^*$.
- $\exists t \in [1, n], r_t \to 1, \forall s \in [1, n] \backslash t, w_s \to 0$. This means that some grids can accurately regress the target location ($r_t \to 1$). Meanwhile, the model tends to output low confidence scores $w_s \to 0$ to relieve the influence of inaccurate regression outputs. This case effectively leverages spatial location priors at some grids to generate accurate predictions, which are hence selected by the selector as final results.

The above analysis shows that, our proposed training objective trains regressor and selector to boost the accuracy of keypoint localization. It also relieves need of careful multi-task design and hyperparameter tuning in different tasks.

Inference. Similar to regression-based methods, SAR only needs a simple decoding process for inference. It first

Figure 3. Visualization of heatmap [29] and confidence score map predicted by SAR in single/multi-keypoint localization. Target keypoints are denoted by star. SAR can locate keypoints that are out of range of the image and effectively avoids missed detection in crowded scenes.

forwards input to get the regressed output $\{\mathcal{K}_t\}$ and confidence score $\{c_t\}$. It then selects the top m predictions with scores larger than γ as the final result,

$$\{\mathcal{K}\} = \{\mathcal{K}_t\}_{t\in\Delta}, \Delta = \underset{t}{\text{top}}(m, \{c_t \geq \gamma\}). \quad (11)$$

The value m is decided by different tasks, *e.g.*, in top-down task we set $m = 1$ and in bottom-up task we set $m = 30$. Compared with previous heatmap-based methods, SAR does not require complex post-processing operation like shifting [29] or DARK [33] and can be implemented on GPU parallelly. It also works well on low-resolution feature maps, thus is more efficient. SAR is also faster than previous detection-based methods that adopt a Hough voting [18] or accumulation operation [8].

3.3. Discussions

Comparison to regression. SAR can be regarded as a generalized regression to process feature map. It degenerates to conventional regression when the size of feature map is 1×1. SAR extends the regression-based method by enhancing its performance and capability in handling multiple keypoint localization by exploring the spatial prior. SAR shares all merits of regression, *e.g.*, low computation and storage complexity and produces a continuous output to relieve the quantization error.

Comparison to heatmap. SAR does not suffer from the quantization error in heatmap and removes the necessity of complex post-processing operation. SAR also differs from methods that combines heatmap with offset regression in both motivation and implementation. Those methods treat heatmap generation and offset regression with equal importance, and their performance is limited by the quality of heatmap. SAR does not assume fixed anchor point (e.g., person center) for regression, thus is general and robust to occlusion and truncation. This property enables SAR to locate keypoints out of input image as shown in Fig. 3.

Visualization. In Fig. 3, we visualize the confidence score map predicted by SAR and heatmap generated by [29]. Compared with heatmap, SAR is more effective in locating multiple keypoints and is more robust to crowded scenes, *e.g.*, keypoints missed by the heatmap can be reliably detected by SAR. More extensive experiment and visualization results will be presented in following section.

4. Experiments

In this section we validate the effectiveness of the proposed SAR on various keypoint localization tasks. We first test SAR on the widely studied *2D human pose estimation* task to demonstrate its basic keypoint localization ability. Moreover, we further verify SAR is more general than conventional regression in three aspects, *i.e.*, generalization to various type of keypoints on *whole-body pose estimation* task, generalization to arbitrary number of keypoints on *multi-person pose estimation* task and generalization to 3D keypoint localization on *3D human pose estimation* task. Descriptions of datasets, evaluation metrics, detailed implementation of experiments and qualitative results are provided in supplemental material.

4.1. 2D Human Pose Estimation

2D human pose estimation is a classical localization task that aims to locate the human body keypoints such as knee or shoulder and researchers have proposed many methods to boost its performance. We first conduct experiments on this task to verify the effectiveness of the proposed method, including commonly used large-scale in-the-wild benchmarks COCO Keypoint [14] and MPII [25]. Following previous works, we report the OKS-based AP and PCKh@0.5/0.1 for COCO Keypoint and MPII evaluation respectively. Ex-

Method	GFLOPs	Deconvs	AP	AP^{50}	AP^{75}	Kpt.Error(px)
Regression	4.0		53.8	82.4	57.7	8.5
RLE	4.0		69.6	89.3	76.0	7.8
Heatmap	9.7	✓	70.7	91.2	77.6	7.6
Heatmap+Offset	9.7	✓	70.9	91.3	77.8	7.6
SAR	4.0		71.3	91.5	79.1	7.3
SAR	9.7	✓	**72.5**	**92.3**	**79.9**	**7.1**
SAR (f^g+FC)	5.3		71.1	91.4	78.6	7.3

Table 1. Comparison with baselines on COCO Keypoint `val` set based on SimplePose [29]. Input size is 256×192.

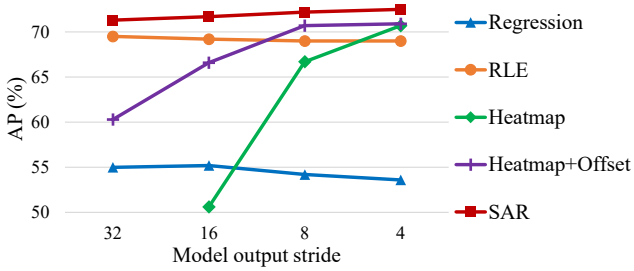

Figure 4. Comparison with baselines under different output stride.

periments on MPII are shown in supplemental material.

Comparison with baselines. We first evaluate the proposed SAR with two types of commonly used baselines under the same setting. Experiments are conducted based on SimplePose [29], ResNet-50 [4] is adopted as the backbone. The results are summarized in Table 1. Compared with conventional regression that adopts pooled feature to regress target, SAR obtains superior performance without carefully designed loss or modules. It improves the regression baseline from 53.8 to 72.5. SAR is also better than heatmap-based methods that also utilize the spatial location prior to locate target. We also compare a simple multi-task method that also combine heatmap and regression, which is denoted as heatmap+offset. It can be observed that simply combining two tasks cannot get better performance because it introduces contradiction and more hyperparameters into training, and their performance is also limited by heatmap. We also show that the improvement of SAR mainly comes from the spatial location prior $\mathcal{P}$, rather than the visual feature $\mathcal{F}$. We keep the same input f^g of SAR and conventional regression, thus the only difference is the introduced location prior $\mathcal{P}$, which is denoted as f^g+FC. f^g+FC explicitly compensates the missed spatial information in f^g and obtains 71.1 AP, which is higher than most baselines.

Analysis on spatial size. SAR shares the merits of regression that is free of quantization error, we conduct experiments on various network output strides and input sizes to demonstrate the superior merits of the proposed SAR. In Fig. 4 we investigate the effect of different output stride to localization performance. We can observe that heatmap-based methods are affected by the output stride greatly due

Method	Input size	AP	AP^{50}	AP^{75}	AR
Regression	64×64	22.7	55.0	14.6	37.5
Heatmap	64×64	33.2	70.6	26.8	40.2
SimCC [13]	64×64	37.5	71.9	34.7	43.1
SAR	64×64	**39.2**	**74.9**	**36.3**	**45.6**
Regression	128×128	43.1	75.2	45.2	57.9
Heatmap	128×128	60.0	87.8	68.1	65.0
SimCC [13]	128×128	61.9	87.8	68.7	66.3
SAR	128×128	**62.5**	**87.8**	**70.4**	**67.4**
Regression	256×192	53.8	82.4	57.7	67.6
Heatmap	256×192	70.7	91.2	77.6	74.3
SimCC [13]	256×192	71.2	91.2	78.7	74.9
SAR	256×192	**72.5**	**92.3**	**79.9**	**76.1**

Table 2. Comparison with baselines under different input size.

Figure 5. Illustration of sampled keypoint localization results of (Top) 2D human pose estimation, (Middle) whole-body pose estimation, and (Bottom) 3D human pose estimation.

to the quantization error in low resolution heatmap. However, SAR obtains stable performance on various output strides and is constantly superior to heatmap-based methods. Therefore, SAR can be integrated to different backbone without any modification. We also test SAR with other methods on different input size to show its robustness to different input resolution. The results are shown in Table 2. SAR can handle different input size well and achieves superior performance, *e.g.*, higher than previous SimCC [13] under the same setting.

Comparison with other methods. Finally, we give a comprehensive comparison with other methods on COCO Keypoint test-dev set. Results are shown in Table 3. Equipped with different backbones, SAR outperforms previous methods and achieves superior performance.

4.2. Whole-Body Pose Estimation

We further verify the generalization of the proposed SAR to other type of keypoints on whole-body pose estimation task, which aims to locate keypoints of human body, foot, face and hand. Compared with previous task that only focuses

Method	Backbone	Input size	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}	AR
Heatmap-based								
SimplePose [29]	ResNet-50	384×288	71.5	91.1	78.7	67.8	78.0	76.9
HRNet [21]	HRNet-W48	384×288	75.5	92.5	83.3	71.9	81.5	80.5
SimCC [13]	HRNet-W48	384×288	76.0	92.4	83.5	72.5	81.9	81.1
Regression-based								
DeepPose [25]	ResNet-152	256×192	59.3	87.6	66.7	56.8	64.9	-
Integral Pose [23]	ResNet-101	256×256	67.8	88.2	74.8	63.9	74.0	-
RLE [10]	HRNet-W48	384×288	75.7	92.3	82.9	72.3	81.3	-
Ours (SAR)	ResNet-50	384×288	73.5	91.9	80.9	69.6	79.7	78.8
Ours (SAR)	HRNet-W48	384×288	**76.3**	**92.5**	**83.6**	**72.6**	**82.4**	**81.2**

Table 3. Comparison with other methods on COCO Keypoint `test-dev` set.

Method	Input size	Backbone	body		foot		face		hand		whole-body	
			AP	AR	AP	AR	AP	AR	AP	AR	AP	AR
Baselines												
Heatmap [29]	256×192	ResNet-50	65.2	73.8	61.5	74.9	60.6	71.5	46.0	58.4	52.1	63.3
Heatmap [29]	256×192	HRNet-W48	70.1	77.6	67.5	78.7	65.6	74.3	53.5	63.9	57.9	68.1
SAR	256×192	ResNet-50	67.3	74.9	61.5	75.2	83.2	88.7	48.3	60.8	59.1	68.0
SAR	256×192	HRNet-W48	**71.0**	**78.9**	**69.1**	**81.8**	**88.1**	**92.3**	**58.4**	**69.2**	**65.1**	**74.0**
SoTA methods												
DeepPose [25]	384×288	ResNet-101	44.4	56.8	36.8	53.7	49.3	66.3	23.5	41.0	33.5	48.4
SimplePose [29]	384×288	ResNet-50	66.6	74.7	63.5	76.3	73.2	81.2	53.7	64.7	57.3	67.1
HRNet [21]	384×288	HRNet-W48	72.2	79.1	69.6	80.1	77.6	83.4	58.7	67.8	63.2	71.7
ZoomNet [7]	384×288	HR32+HR18	**74.5**	**81.0**	60.9	70.8	88.0	92.4	57.9	73.4	63.0	74.2
ZoomNAS [30]	384×288	-	74.0	80.7	61.7	71.8	88.9	93.0	**62.5**	**74.0**	65.4	74.4
Ours (SAR)	384×288	HRNet-W48	71.2	79.6	**69.3**	**82.6**	**90.3**	**93.3**	61.3	72.1	**66.6**	**75.6**

Table 4. Comparison with other methods on COCO Whole-Body benchmark.

on human body, whole-body pose estimation is more challenging due to the large scale variance of different type of keypoints. Therefore, this task can be used to test the ability of localization method on handling scale variance. Following previous works, we conduct experiments on large-scale in-the-wild benchmark COCO Whole-Body [7] and report the OKS-based AP on each subset.

Table 4 shows results. We first compare SAR with heatmap baseline using ResNet-50 and HRNet-W48 as backbone. SAR outperforms heatmap by a large margin, especially on small scale keypoints of face and hand. This demonstrates the benefit of continuous output of SAR, which is free of quantization error in heatmap. We also compare SAR with top-down based methods SimplePose, HRNet, ZoomNet and ZoomNAS. We implement SAR with top-down HRNet under the same setting, and Table 4 shows that our method achieves the best performance on most subsets, without carefully designed multi-branch model in ZoomNet and ZoomNAS. It demonstrates that SAR can generalize to locate various types of keypoints.

4.3. Multi-Person Pose Estimation

One advantage of SAR is that it can be applied to multiple keypoints localization tasks, while most of previous methods, *e.g.*, regression-based methods and SimCC [13] cannot be used to locate multiple keypoints of the same type simultaneously. Multiple keypoints localization task such as multi-person pose estimation (MPPE) is common because it eliminates the assumption that only one instance exists in the input image in single-keypoint localization. Benefited by embedding spatial prior, the proposed method can be applied to MPPE tasks such as bottom-up MPPE and single-stage MPPE. Following previous works, we conduct experiments on COCO [14], OCHuman [34] and CrowdPose [9] and report OKS-based AP metrics.

For bottom-up MPPE, we adapt SAR with AE [17] to locate multiple keypoints simultaneously because it separately implements localization and grouping. For single-stage MPPE, we conduct experiments on two widely used methods, *i.e.*, CenterNet [35] and DEKR [3], to locate multiple person center points.

All experiments are conducted under the same setting. As shown in Table 5, SAR can successfully detect arbitrary number of keypoints and produce more reliable results than heatmap on all benchmarks. From Fig. 3 we can observe that SAR outputs higher confidence score for detected keypoints, which is important in MPPE because it avoids missed localization.

Method	COCO				OCHuman				CrowdPose			
	AP	AP^M	AP^L	AR	AP	AP^M	AP^L	AR	AP	AP^E	AP^M	AP^L
Bottom-up MPPE based on AE [17]												
Heatmap	48.4	44.1	54.8	54.1	-	-	-	-	46.7	59.9	46.6	36.7
SAR	**50.0**	**45.0**	**56.9**	**56.2**	-	-	-	-	**48.2**	**61.1**	**48.0**	**38.9**
Single-stage Regression MPPE												
CenterNet [35]	57.6	50.5	68.4	63.1	34.1	15.2	36.2	69.3	60.9	68.9	61.7	51.7
+SAR	**60.3**	**52.6**	**71.1**	**64.7**	**39.6**	**15.5**	**40.5**	**71.5**	**62.2**	**69.8**	**63.0**	**53.7**
DEKR [3]	61.8	54.7	72.8	67.5	35.6	9.6	37.5	71.3	65.2	73.0	66.0	55.9
+SAR	**64.5**	**56.8**	**75.5**	**69.1**	**40.9**	**17.2**	**41.9**	**73.8**	**66.3**	**73.7**	**67.0**	**57.6**

Table 5. Comparison with other methods on multi-person pose estimation benchmarks.

Methods	Dir.	Dis.	Eat	Gre.	Phon.	Pose	Pur.	Sit	SitD.	Smo.	Phot.	Wait	Walk	WalkD.	WalkP.	Avg
Baselines																
Regression	46.5	53.5	47.9	46.9	53.3	46.4	48.9	64.3	64.6	53.1	54.2	47.5	40.9	53.4	45.2	51.7
RLE	47.1	51.3	50.4	48.0	53.0	45.8	47.4	65.0	69.0	53.7	52.0	46.7	40.9	51.2	44.7	51.6
Integral Pose	45.3	51.5	47.5	45.5	53.2	42.6	47.1	63.9	63.8	53.0	53.8	46.1	40.0	51.3	43.2	50.6
SAR-3D	46.5	51.1	46.9	45.3	49.7	45.1	46.9	58.7	63.2	50.0	51.1	46.1	40.5	51.3	44.1	49.4
SAR-decouple	42.2	49.5	46.8	44.8	52.2	42.5	43.8	60.7	65.1	50.7	52.6	44.7	37.8	50.0	41.6	48.9
SoTA methods																
Sun [22]	52.8	54.8	54.2	54.3	61.8	53.1	53.6	71.7	86.7	61.5	67.2	53.4	47.1	61.6	63.4	59.1
PoseNet [15]	50.5	55.7	50.1	51.7	53.9	46.8	50.0	61.9	68.0	52.5	55.9	49.9	41.8	56.1	46.9	53.3
Sun [23]	47.5	**47.7**	49.5	50.2	51.4	43.8	46.4	58.9	65.7	49.4	55.8	47.8	38.9	49.0	43.8	49.6
RLE [10]	43.3	51.0	**44.5**	44.5	51.7	43.1	46.0	59.2	63.7	49.6	52.5	44.1	37.5	50.5	41.2	48.6
Ours (SAR)	**40.9**	47.8	44.8	**43.4**	**50.3**	**41.0**	**42.6**	**58.2**	**61.2**	**48.9**	**50.0**	**42.9**	**36.2**	**48.3**	**40.2**	**47.1**

Table 6. Comparison with other methods on Human3.6M benchmark.

4.4. Monocular 3D Human Pose Estimation

In this section we show that SAR can also generalize to perform 3D keypoint localization. We conduct experiments on Human3.6M [6], a large-scale indoor benchmark for 3D human pose estimation. For evaluation, MPJPE is reported to measure the error of the predicted keypoints and groundtruth in 3D space. Following [10], we use (S1, S5, S6, S7, S8) for training and (S9, S11) for evaluation.

SAR for 3D keypoint. We propose two variants of SAR to locate 3D keypoint. The first is SAR-3D that generates 3D feature map $\mathbf{F}_{3D} \in \mathbb{R}^{c\times d\times h\times w}$ from $\mathbf{F}$ to provide 3D spatial location prior to locate target, which is similar to Integral Pose [23] that also generates 3D heatmap. Benefited by the continuous output of SAR, a small sized $\mathbf{F}_{3D}$ can already produce promising results, *e.g.*, $d = h = w = 16$. For the second variant, we decouple the depth dimension from spatial dimension and estimate it with a separate localization branch, which is denoted as SAR-decouple. For x, y coordinates, we directly adopt 2D estimation pipeline. For z dimension, we convert the feature map $\mathbf{F}$ to a 1D feature map $\mathbf{F}_z \in \mathbb{R}^{c\times d}$ to locate z-dim of keypoint, $d = 64$.

Baselines. We first compare SAR with several 3D keypoint localization methods, including regression, RLE [10] and Integral Pose [29]. All experiments are conducted under the same setting. As shown in Table 6, both two variants of SAR achieve superior 3D keypoint localization accuracy, which is lower than Integral Pose by 1.2 and 1.7 MPJPE. SAR-decouple achieves better performance than SAR-3D. We think that it is because SAR-decouple with large spatial size can generate more accurate localization results, which is consistent with experiments in Sec. 4.1.

Comparison with other methods. In Table 6 we also compare SAR with recent methods under the same setting, *e.g.*, with flip test. We can observe that SAR reduces the MPJPE from 48.6 to 47.1, which is better than most image-based 3D human pose estimation methods.

5. Conclusion

This work proposes a novel and effective regression method by integrating spatial location prior to relieve the difficulty of direct regression. We introduce spatial-aware regressor and selector with a unified objective to achieve spatial-aware regression. Comprehensive experiments on four different keypoint localization tasks and seven benchmarks demonstrate the promising effectiveness and generalization capability of the proposed method.

Acknowledgement This work is supported in part by Natural Science Foundation of China under Grant No. U20B2052, 61936011, in part by the Okawa Foundation Research Award.

References

[1] Joao Carreira, Pulkit Agrawal, Katerina Fragkiadaki, and Jitendra Malik. Human pose estimation with iterative error feedback. In *CVPR*, 2016. 2

[2] Ziteng Gao, Limin Wang, and Gangshan Wu. Mutual supervision for dense object detection. In *ICCV*, 2021. 2

[3] Zigang Geng, Ke Sun, Bin Xiao, Zhaoxiang Zhang, and Jingdong Wang. Bottom-up human pose estimation via disentangled keypoint regression. In *CVPR*, 2021. 7, 8

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 2, 6

[5] Junjie Huang, Zheng Zhu, Feng Guo, and Guan Huang. The devil is in the details: Delving into unbiased data processing for human pose estimation. In *CVPR*, 2020. 1

[6] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE TPAMI*, 2013. 8

[7] Sheng Jin, Lumin Xu, Jin Xu, Can Wang, Wentao Liu, Chen Qian, Wanli Ouyang, and Ping Luo. Whole-body human pose estimation in the wild. In *ECCV*, 2020. 7

[8] Sven Kreiss, Lorenzo Bertoni, and Alexandre Alahi. Pifpaf: Composite fields for human pose estimation. In *CVPR*, 2019. 5

[9] Jiefeng Li, Can Wang, Hao Zhu, Yihuan Mao, Hao-Shu Fang, and Cewu Lu. Crowdpose: Efficient crowded scenes pose estimation and a new benchmark. In *CVPR*, 2019. 7

[10] Jiefeng Li, Siyuan Bian, Ailing Zeng, Can Wang, Bo Pang, Wentao Liu, and Cewu Lu. Human pose regression with residual log-likelihood estimation. In *ICCV*, 2021. 1, 2, 7, 8

[11] Jiefeng Li, Tong Chen, Ruiqi Shi, Yujing Lou, Yong-Lu Li, and Cewu Lu. Localization with sampling-argmax. *NeurIPS*, 2021. 2

[12] Yanjie Li, Shoukui Zhang, Zhicheng Wang, Sen Yang, Wankou Yang, Shu-Tao Xia, and Erjin Zhou. Tokenpose: Learning keypoint tokens for human pose estimation. In *CVPR*, 2021. 2

[13] Yanjie Li, Sen Yang, Peidong Liu, Shoukui Zhang, Yunxiao Wang, Zhicheng Wang, Wankou Yang, and Shu-Tao Xia. Simcc: A simple coordinate classification perspective for human pose estimation. In *ECCV*, 2022. 6, 7

[14] 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. 2, 5, 7

[15] Gyeongsik Moon, Ju Yong Chang, and Kyoung Mu Lee. Camera distance-aware top-down approach for 3d multi-person pose estimation from a single rgb image. In *ICCV*, 2019. 8

[16] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *ECCV*, 2016. 2

[17] Alejandro Newell, Zhiao Huang, and Jia Deng. Associative embedding: End-to-end learning for joint detection and grouping. In *NeurIPS*, 2017. 7, 8

[18] 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. 1, 2, 5

[19] Christos Sagonas, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *ICCVW*, 2013. 1

[20] Dahu Shi, Xing Wei, Liangqi Li, Ye Ren, and Wenming Tan. End-to-end multi-person pose estimation with transformers. In *CVPR*, 2022. 2

[21] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *CVPR*, 2019. 1, 2, 7

[22] Xiao Sun, Jiaxiang Shang, Shuang Liang, and Yichen Wei. Compositional human pose regression. In *ICCV*, 2017. 8

[23] Xiao Sun, Bin Xiao, Fangyin Wei, Shuang Liang, and Yichen Wei. Integral human pose regression. In *ECCV*, 2018. 1, 2, 7, 8

[24] Jonathan J Tompson, Arjun Jain, Yann LeCun, and Christoph Bregler. Joint training of a convolutional network and a graphical model for human pose estimation. *NeurIPS*, 2014. 2

[25] Alexander Toshev and Christian Szegedy. Deeppose: Human pose estimation via deep neural networks. In *CVPR*, 2014. 1, 2, 5, 7

[26] Dongkai Wang and Shiliang Zhang. 3d human mesh recovery with sequentially global rotation estimation. In *CVPR*, 2023. 1

[27] Dongkai Wang and Shiliang Zhang. Contextual instance decoupling for instance-level human analysis. *IEEE TPAMI*, 2023.

[28] Dongkai Wang, Shiliang Zhang, Yaowei Wang, Yonghong Tian, Tiejun Huang, and Wen Gao. Humvis: Human-centric visual analysis system. In *ACM MM*, 2023. 1

[29] Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. In *ECCV*, 2018. 1, 2, 5, 6, 7, 8

[30] Lumin Xu, Sheng Jin, Wentao Liu, Chen Qian, Wanli Ouyang, Ping Luo, and Xiaogang Wang. Zoomnas: searching for whole-body human pose estimation in the wild. *IEEE TPAMI*, 2022. 7

[31] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. Vitpose: Simple vision transformer baselines for human pose estimation. *arXiv preprint arXiv:2204.12484*, 2022. 1, 2

[32] Shiyu Xuan, Qingpei Guo, Ming Yang, and Shiliang Zhang. Pink: Unveiling the power of referential comprehension for multi-modal llms. *arXiv preprint arXiv:2310.00582*, 2023. 2

[33] Feng Zhang, Xiatian Zhu, Hanbin Dai, Mao Ye, and Ce Zhu. Distribution-aware coordinate representation for human pose estimation. In *CVPR*, 2020. 2, 5

[34] Song-Hai Zhang, Ruilong Li, Xin Dong, Paul Rosin, Zixi Cai, Xi Han, Dingcheng Yang, Haozhi Huang, and Shi-Min Hu. Pose2seg: Detection free human instance segmentation. In *CVPR*, 2019. 7

[35] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. *arXiv preprint arXiv:1904.07850*, 2019. 7, 8