SAMKR: Bottom-up Keypoint Regression Pose Estimation Method Based On Subspace Attention Module

1st Linwei Chen
National and Local Joint Engineering Lab of Computer Aided Design, School of Software Engineering
Dalian University
Dalian, China
chenlinwei@s.dlu.edu.cn

2nd Dongsheng Zhou*
National and Local Joint Engineering Lab of Computer Aided Design, School of Software Engineering
Dalian University
Dalian, China
Corresponding:zhouds@dlu.edu.cn

3rd Rui Liu
National and Local Joint Engineering Lab of Computer Aided Design, School of Software Engineering
Dalian University
Dalian, China
liurui@dlu.edu.cn

4th Qiang Zhang
National and Local Joint Engineering Lab of Computer Aided Design, School of Software Engineering
Dalian University
Dalian, China
zhangq@dlu.edu.cn

Abstract—As a hot research issue in computer vision, 2D human pose estimation plays an important role in human-computer interaction, intelligent monitoring, 3D human pose estimation and so on. Aiming at the problem of human scale inconsistency in the situation of multi-person, a keypoint regression method based on subspace attention module (SAMKR) is proposed in this paper for the 2D human pose estimation. Firstly, each keypoint is divided into independent regression branches, and then the feature mapping in each keypoint regression branch is evenly divided into a specified number of feature mapping subspaces, and different attention mappings are derived for each feature mapping subspace. By learning different attention maps in each feature subspace, multi-scale feature representation can be effectively improved. The experimental results show that SAMKR achieved 74.4 AP score on the CrowdPose test set, which may lead to an improvement of + 7.1AP, and reached 70.4 AP score on the COCO test-dev data set, which was 0.4AP higher than the baseline.

Keywords—2D human pose estimation, Bottom-up paradigm, keypoint regression, subspace attention

I. Introduction

Human pose estimation is an important direction in the field of computer vision, which is widely used in motion recognition, human-computer interaction, animation, intelligent monitoring and other fields. Nowadays, human pose estimation includes many research branches, including 2D human pose estimation, 3D human pose estimation, video human pose estimation and multi-view human pose estimation. Among them, single-image 2D human pose estimation is the basis of two-segment 3D human pose estimation, video human pose estimation and multi-view human pose estimation. The improvement of 2D attitude estimation performance will also bring a lot of impetus to the development of these branches.

In recent years, 2D human pose estimation has gradually developed from single-person estimation to multi-person estimation. However, due to occlusion, complex background, limb crisscross, non-uniform body scale and other reasons, multi-person human pose estimation is still a challenging problem. According to the different estimation paradigms, multi-person attitude estimation can be divided into two ways: top-down and bottom-up. The top-down method first detects all the people in the image and normalizes each person to a similar size, and then detects the key nodes for each person. The bottom-up approach first detects all the keypoints of all the human body, and then groups them into human instances. Compared with the top-down method, the bottom-up method is faster and this is beneficial for real-time pose estimation. Especially for the scenes of crowd, it is difficult to segment each person correctly because of the high overlap of the detection objects, which leads to the inability of top-down method. Therefore, the bottom-up estimation method is more suitable in this case. However, due to the influence of perspective effect, the human body in different position has different scale in an image, which makes it difficult to estimate the human pose accurately.

How to capture the multi-scale feature information of human in the image becomes the key to improving the accuracy of human pose estimation. Recently, Rajat Saini et al. [1] proposed the ULSAM method that divided a feature map into several sub-feature spaces, and added an independent attention map to each sub-feature space. It has a better effect on dealing with fine-grained image classification. Inspired by ULSAM, combined with DEKR[2] keypoint regression scheme, this paper proposes a method to extract multi-scale feature information of keypoints based on subspace Attention Module ,which is named as Subspace Attention Module keypoint regression (SAMKR). The main idea of SAMKR is to divide multiple feature subspaces on each independent keypoint regress branch, and add independent lightweight attention blocks to each feature subspace to learn the individual attention mapping of each keypoint feature space. And the proposed model structure can learn multi-scale features more effectively.

Specifically, SAMKR uses HRNet [3,4] as the backbone to aggregate multi-resolution feature channels. Firstly, SAMKR divides independent regression branches for each keypoint. On the regression branch of each keypoint, the feature map of the branch is adaptively convoluted to activate pixels in the keypoint region. Then the activated feature map is divided into several feature subspaces. In each feature subspace, the deep separable convolution operation is first carried out, and then the attention map with only one filter is used to extract features to reduce the number of parameters and the amount of calculation. Each subspace will derive different attention maps, and finally SAMKR will aggregate the feature maps of all feature subspaces. In this way, different attention maps are learned for each keypoint regression branch to achieve feature extraction in multi-scale and the complementary advantages are formed with adaptive convolution activation spatial pixels. In addition, this paper follows the method of using the keypoint heatmap to match the regression pose and improve the detection accuracy of dense keypoints. The keypoint heatmap and the central heatmap will be output in the network. Experiments show that SAMKR method can effectively improve the accuracy of keypoint regression, and it will be more effective for the scenes of intensive crowds- and multi-scale crowds.

2022 International Joint Conference on Neural Networks (IJCNN) | 978-1-7281-8671-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/IJCNN55064.2022.9891995

978-1-7281-8671-9/22/$31.00 ©2022 IEEE

The contribution of this paper can be summarized as follows:

- In this paper, a bottom-up human pose estimation method SAMKR is proposed. This method can make each independent keypoint regression branch learn the multi-scale attention feature map better, so as to improve the representation ability of multi-scale features of the network, and the advantage is more obvious in crowded scenes.
- In this paper, the influence of the internal structure of the subspace attention module on the performance of SAMKR network detection is studied in detail. Through a lot of derivation and control experiments, the relationship between the parameter configuration of the subspace attention module and the performance of SAMKR is demonstrated, and the optimal configuration scheme of SAMKR is summarized.
- The proposed method is verified on COCO dataset and CrowdPose dataset. The experimental results show that the SAMKR method obtains better detection results on the CrowdPose dataset with a score of 74.4 AP. In addition, the performance tested on the COCO dataset also reached the SOTA level, and the score reached 70.4AP in the COCO test-dev dataset.

II. Related Work

A. Bottom-up multi-person pose estimation

As the progress of detection technology, the paradigm of top-down human pose estimation [3,5-13] has achieved satisfactory results in the conventional situation, but it requires additional calculations in the situation of crowded crowd. In contrast, bottom-up human pose estimation[2,14-21], although slightly less accurate than top-down methods, is more efficient and more conducive for real-time detection and intensive crowd- detection.

In recent works of bottom-up human pose estimation, many studies have focused on the methods of optimizing grouping [8, 14, 16, 22, 23, 24]. Openpose[14] can quickly correspond human keypoints to individuals through Part Affinity Fields. PifPaf [22] realizes the location and connection of human nodes by predicting the Part Intensity Field information of each location. Associative embedding [23] integrates detection and grouping tasks into an end-to-end network for the first time. Personlab [24] integrates human keypoint detection and instance segmentation into a network, and corresponded all keypoints to their respective instances through greedy decoding algorithm.

This paper aims at the problem of multi-scale in bottom-up human pose estimation to enhance the multi-scale feature extraction ability of the network. So in the process of keypoint regression, it could maintain good accuracy in each scale .

B. Heatmap Regression

In recent years, with the application of convolutional neural network (CNN), human pose estimation has made remarkable progress. Human pose estimation is usually regarded as a regression task and the early CNN method trend to predict the keypoints of single human pose estimation directly. However, it is a challenging task as the inherent fuzziness of keypoints that annotated manually. The subsequent heatmap representation has solved the above problems well. The method based on heatmap [3, 5, 13, 15, 17, 25, 26] aims to construct a real heatmap by placing a two-dimensional Gaussian kernel on the marker points of the real keypoint coordinates. The pixel values in the heatmap are usually regarded as the probabilities of corresponding pixels as keypoints. This is not only easy to implement, but also well adapted to the ambiguity of manual annotation. Nowadays, the method based on heatmap regression has been widely used in human pose estimation.

HRNet [3, 4] maintains high-resolution feature maps throughout the network instead of recovering from low-resolution ones, making keypoint heatmaps more accurate. HigherHRNet [15] uses HRNet as the base network to predict heatmaps by adding a deconvolution module to generate higher resolution feature maps. Zhengxiong Luo et al. [17] focused on the related problems of heatmap regression in dealing with various human scales and label ambiguity, and proposed a scale-adaptive and weight-adaptive heatmap review scheme.

C. Keypoint Regression

Keypoint regression aims to achieve keypoint detection by minimizing the direct loss between the predicted keypoint and the ground truth. Some recent studies [2, 18, 20, 21, 27, 28] have focused on the dense keypoint regression. CenterNet [21] belongs to the anchor-free series of target detection, which directly detects the central point of the target and realizes the balance between speed and accuracy. Point-set anchors [20] acquires features on a set of points that may be closer to the regression target to obtain more information . DEKR [2] learns the representation of each keypoint from its corresponding keypoint region. PRTR [27] is a method of modeling the Federated Space and appearance of keypoints by constructing a cascade transformer.

However, unlike heatmap regression, keypoint regression is inherently more difficult to optimize due to the high accuracy required for gesture recognition. In order to improve the result of regression , this paper adopts the method of matching the position of the keypoint with the last keypoint detected from the heatmap . In addition, this method can also obtain additional performance gains with the help of some of the latest improved schemes of heatmap regression.

D. Attention mechanisms

Inspired by the fact that human visual system can find key areas naturally and efficiently, attention mechanism is introduced into the computer vision system. At present, attention mechanism has been involved in various tasks such as classification, detection and semantic segmentation.

SENet [29] learns the feature weight according to loss, in which the effective feature maps are assigned with heavy weight and the invalid feature maps are assigned with light weight. CBAM [30] applies a simple and effective convolution attention module in the feedforward convolution

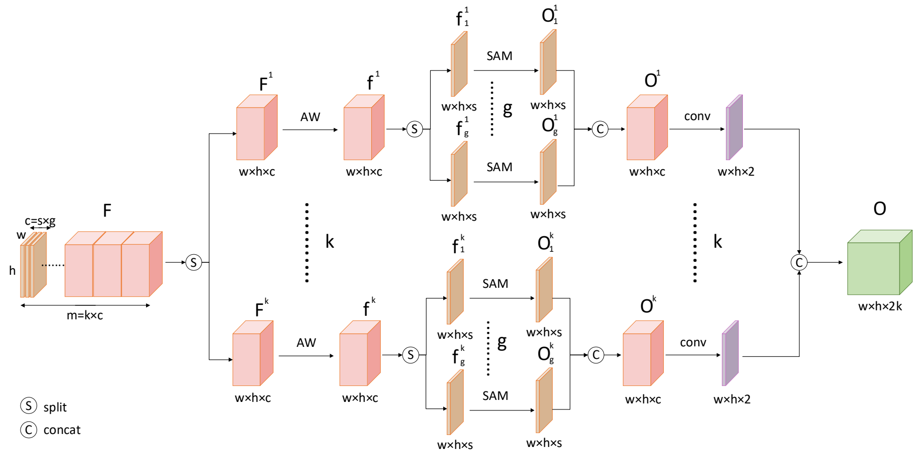

Fig.1. The complete structure of regression keypoint migration module.

neural networks. GENet [31] focuses on how to extract more effective context relevance from feature maps, so as to regulate the information between them. ULSAM [1] can learn individual attention maps of each feature subspace, in which the cross-channel information and multi-scale multi-frequency features can be learned efficiently. Xiao Chu et al. [32] introduced the attention mechanisms into pose estimation, and the attention mechanism is realized by conditional random field.

The method in this paper aims to capture the multi-scale feature information in the image by multiple spatial attention modules. Each attention map is only responsible for feature capture at its own scale. Finally, the final keypoint regression results can be improved by the aggregated information of multi-scale .

III. Approach

2D multi-person pose estimation is to detect the keypoint of human body in the image, and then deduce the human body poses. In this paper, a bottom-up human pose estimation method SAMKR is proposed based on the thought of subspace attention module. In this section, we will introduce the method in detail.

A. Framework of SAMKR

HRNet. In this paper, HRNet [3,4] is used as the backbone. Through HRNet, the image starts from the high resolution of the first stage, and each stage generates a new parallel branch with the current minimum resolution of 1/2. After the fourth stage, the output four parallel channels with different resolutions are combined to form the feature map F^x. Next, an offset feature map O for keypoint regression is generated and a heatmap is used to evaluate the pose for dense keypoint regression.

Offset. The experiments in recent work [2] show that the effect of grouped regression in which some keypoints are grouped into a single branch, is worse than that of the single keypoint regression, and the independent regression scheme of the keypoints is proved to be helpful in improving the regression quality of the keypoints. SAMKR follows the independent regression scheme of keypoints, and the overall flow chart of is shown in Fig.1. Where w, h resolution represents the width and height of the feature map, m represents the number of initial feature map channels of the module, k represents the number of keypoints, g represents the number of feature subspaces divided in the regression branch of each keypoint, s represents the number of feature channels in each feature subspace, AW represents the adaptive convolution module, SAM represents the attention module in the separation feature subspace.

First, feature map F with a specified number of channels (k× C) is generated from the spliced feature map F^x through a 1×1 convolution operation. Then, the feature map F is divided into k keypoint regression branches $[F^1, F^2, \cdots, F^k]$ according to the number of keypoints k, as shown in (1). And each keypoint is matched with an independent regression branch .

$$[F^1, F^2, \cdots, F^k] = \mathrm{Chunk}(F, \mathrm{k}) \tag{1}$$

This paper follows the method of adaptive convolution[2] to activate the pixels in the region. In each keypoint regression branch, the adaptive convolution operation is first performed, and the pixels in the keypoint region can be activated by adaptive convolution, and the activated feature map f^i is output, as shown in (2).

$$f^i = AW(F^i) \tag{2}$$

ULSAM[1] proposed by Rajat Saini et al. is a attention module for compact neural networks By learning the individual attention mapping of each feature subspace, it is sufficient for multi-scale feature learning and can learn cross-channel information effectively. Inspired by ULSAM, this paper uses SAM (Subspace Attention Module) module to enhance the multi-scale feature extraction of the model, and optimizes the problem of human scale inconsistency in bottom-up pose estimation. As shown in (3), the activated feature map f^i is divided into g feature subspaces $[f_1^i, f_2^i, \cdots, f_g^i]$. Here, the number of the keypoint branch is denoted by i, and the number of the feature subspace divided by the regression branch is denoted by j. In (4), new feature map O_j^i is obtained after feature map f_j^i learns different attention graphs in SAM module.

$$[f_1^i, f_2^i, \cdots, f_g^i] = \mathrm{Chunk}(f^i, \mathrm{g}) \tag{3}$$

$$O_j^i = SAM(f_j^i) \tag{4}$$

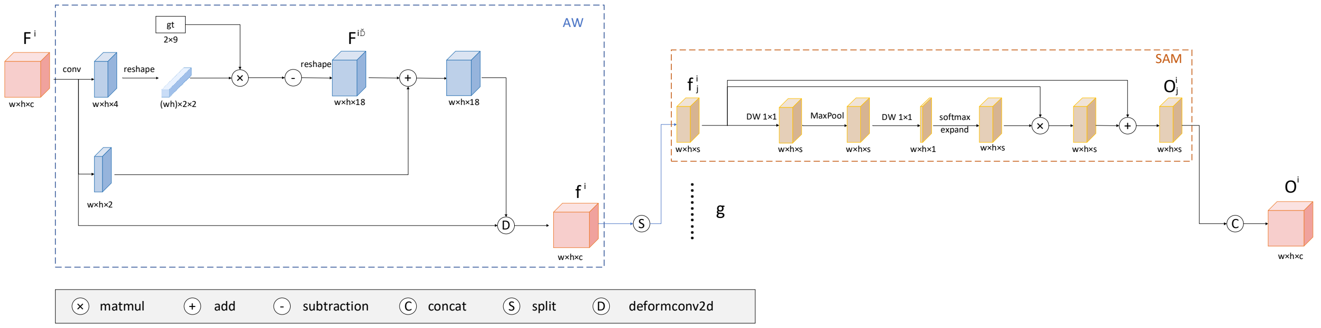

Fig .2. Internal structure diagram of regression keypoints in SAMKR network.

In (5), concat$(O_1^i, \cdots O_g^i)$ indicates that all the new feature maps derived from all the feature subspaces on this branch are spliced together. In (6), concat $(O^1, \cdots O^k)$ indicates that the offset feature map output in the regression branches of all keypoints are spliced together. $Conv^{1\times1}$ means that 1×1 convolution kernel is used for convolution operation, and the number of channels in the feature map of the final feature is generated as offset feature graph.and the final output feature map O is obtained.

$$O^i = Concat\left(O_1^i, O_2^i, \cdots, O_g^i\right) \quad (5)$$

$$O = \mathrm{Concat}(Conv^{1\times1}(O^1, O^2, \cdots, O^k)) \quad (6)$$

Heatmap. Following the method of [2], a separate branch is used to predict the corresponding heatmap and center heatmap for each keypoint. The central heatmap indicates the confidence of each pixel belongs to the center of the human body, and the heatmap is used to evaluate and rank the regression candidate position. When HRNet-w32 is used as the backbone network, the heatmap loss function is set according to the method of adaptive heatmap regression [17]. When using HRNet-w48 as the backbone network, the heatmap loss function is set according to the reference[2].

B. *Details of SAMKR*

Fig.2 shows the internal details of the keypoint offset regression in SAMKR. The operation of adaptive convolution module is shown in (7) and (8), in which ⊗ represents the element-by-element multiplication of matrices, ⊕ represents the addition of matrices, and ⊖ represents the subtraction of matrices. Gt is a 2 × 9 matrix set in the network, representing the offset of the regular 3 × 3 region relative to the central pixel. The offset of the activated pixel relative to the central position is calculated, and then the activated feature map f^i is obtained through the deformable convolution.

$$F^{i'} = \mathrm{Conv}(F^i)\otimes\mathrm{Gt} \ominus \mathrm{Gt} \quad (7)$$

$$f^i = \mathrm{DeformConv}(F^{i'}\oplus\mathrm{Conv}(F^i), F^i) \quad (8)$$

Divide f^i into g feature subspaces on average, and the feature map in each feature subspace is denoted as f_j^i. The specific operation is shown in (9), where DW^1 represents the deep convolution with 1 × 1 convolution kernel, maxpool represents the maximum pooling with 3 × 3 kernel size and 1 filling, and PW^1 is the point-by-point convolution with only one filter. In the feature subspace, the initial feature map f_j^i first performs deep separable convolution operation, and SAMKR performs maximum pooling operation after deep convolution to collect spatial information of keypoints. Since deep convolution is an independent operation of each channel, the multi-channel feature extraction of single pixel is realized by point-by-point convolution of a single filter to realize the weighted combination of multi-channel features. Finally, the jump connection with original f_j^i is implemented to form a new feature map O_j^i.

$$O_j^i = \mathrm{softmax}(PW^1(\mathrm{maxpool}(DW^1(f_j^i))))\otimes f_j^i \oplus f_j^i \quad (9)$$

C. *Analysis of parameter setting*

According to the introduction of the previous section, the total number of channels in the input feature map is denoted as m, the number of channels in the keypoint regression branch is denoted as c, the number of feature subspaces divided on each regression branch is denoted as g, the number of channels in each feature subspace is denoted as s, and the number of keypoints is denoted as k. Different parameter settings correspond to different network structures. This paper takes the coco dataset as an example, and sets the number of channels of the initial feature map F to 272. Some typical cases are analyzed as follow:

- [c=m, 1<g<c] indicates that the network is a single branch structure. Different from multi-branch structure, the decoupling characteristics of each keypoint must be learned implicitly, and the representation of one keypoint cannot be clearly decoupled from other keypoints. 1 < g < c indicates that the strategy of multiple subspaces is adopted. Although it is conducive to improving the multi-scale representation, it cannot replace the decoupling learning between keypoints. The network framework is shown in Fig.3 (a).

- [c=m/k, g=1] indicates that the network adopts an independent keypoint regression strategy with a single feature space and a unique attention map. A single attention map cannot capture the complex relationship and multi-scale features in the whole feature graph. Therefore, the final performance of the network is not satisfactory The network framework is shown in Fig.3 (b).

- [c=m/k, g=c] indicates that the network adopts an independent keypoint regression strategy in which a separate feature subspace is distributed and a separate attention map is generated for each channel.Because there is only one channel, the attention map cannot learn the feature information across the channels, and the whole process of generating the attention map is simplified to a simple nonlinear transformation. The network framework is shown in Fig.3 (c).

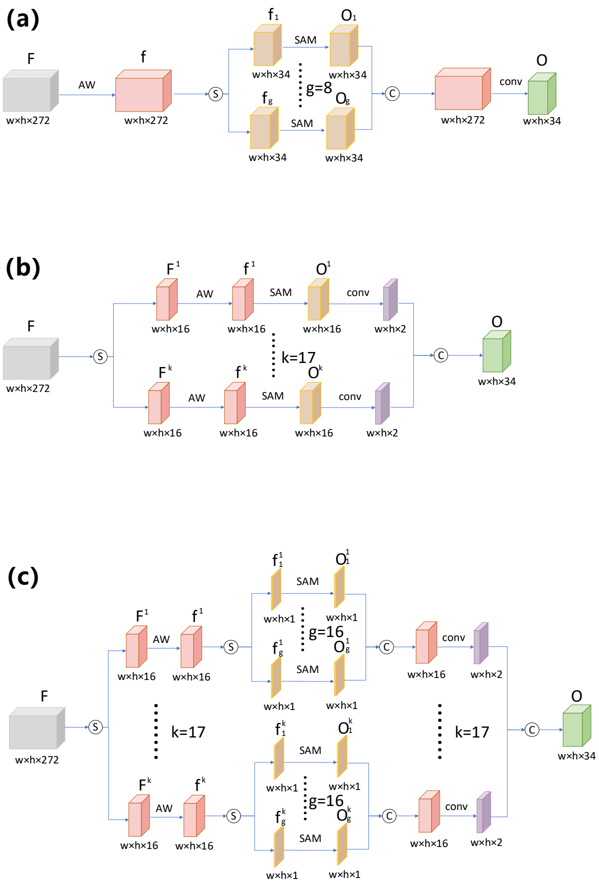

Fig. 3. (a), (b) and (c) respectively show the internal structure of the module under Settings[c=m, 1<g<c], [c=m/k, g=1] and [c=m/k, g=c].

- [c=m/k, 1<g<c] indicates that the network adopts the strategy of independent keypoints regression strategy , and when dividing multiple feature subspaces, each subspace contains more than one feature channel. In this way, there are multiple feature channels in each feature subspace, and each attention map can learn cross-channel information. After aggregating multiple attention graphs, the multi-scale representation of keypoints is enhanced. The structure of SAMKR network is this kind of configuration, and the network framework is shown in Fig. 1.

IV. Experiments

A. setting

Dataset. This paper evaluates the performance of keypoint detection task on COCO dataset. The COCO dataset contains a total of 200,000 images and 250,000 human pose instances marked with 17 keypoints. In this paper, the centralized training model of train2017 includes 57,000 pictures and 150,000 human pose examples. val2017 contains 5,000 images, the set of test-dev2017 contains 20,000 images. The performance of the model trained in this paper will be evaluated on val2017 and test- dev2017 datasets.

Evaluation metric. This paper used Object Keypoint Similarity (OKS) for COCO pose estimation. We report average precision and average recall scores with different thresholds and different object sizes: AP, AP^{50}, AP^{75}, AP^{M}, AP^{L}, AR, AR^{M} and AR^{L}.

Training. According to the setting of [23,33], this paper uses a series of data expansion methods, including random rotation ([$-30^{o}, 30^{o}$]), random scaling ([0.75,1.5]), random translation ([-40,40]), and random horizontal flipping. This paper cuts the images to 512 × 512 resolution for HRNet-W32 and 640 × 640 resolution for HRNet-W48. The Adam[34] optimizer is used to improve training results. The basic learning rate is set to $1e^{-3}$ and is dropped to $1e^{-4}$ and $1e^{-5}$ at the 90th and 120th iterations, respectively. The final training will end at 140 iterations. When using HRNet-W32 as the backbone, the loss function is set by adaptive heatmap regression[17].–When using HRNet-W48 as the backbone , the loss function is set in the same way as the baseline method DEKR [2].

B. Analysis

In this paper, comparative experiments are carried out on several classical structures under different parameter settings of SAMKR, and experiments are also carried out on the optimization of structure settings in SAMKR.

Comparative experiment. Several parameter settings corresponding to the different network structures that have been discussed in section III are designed to verify the performance of SAMKR. HRNet-W32 is used as the main network in the experiment, and the experimental results are shown in table 1. In the table, each line corresponds to a different network structure. From the table , we can find, the second and third network structure achieved lower AP scores (64 and 64.2 respectively). This is mainly because they adopt the separate feature space/subspace and separate attention map, which makes them unable to learn the cross-channel features. Compared with the first structure, the fourth structure has an improvement of 1.6 AP (67.3 and 68.9 respectively). This proves that the strategy of independent keypoints regression is effective.

The optimal structure of SAMKR. Next, we will explore the influence of parameter settings on the performance of SAMKR , and summarizes the parameter settings with better performance.

A control experiment was conducted in the coco2107 validation test set (as shown in Table 2). Taking HRNet-W32 as an example, the output channel is revised to 272 (= 17 × 16) after a 1 × 1 convolution. In this way, each regression branch contains 16 channels, which is convenient for grouping and will not increase too much width. Since the parameters of s, g, c need to meet the relationship of c=g×s. We first fix the number of regression branch channels of keypoints c=16 three combinations of g and s are denoted as 2_SAMKR_16, 4_SAMKR_16 and 8_SAMKR_16 respectively. It can be clearly seen from the experimental results that when g is set to 8, the performance of SAMKR network is significantly better than the others. Then we fix the number of feature subspaces g=8 , three combinations of c and s are denoted as 8_SAMKR_8, 8_SAMKR_16 and 8_SAMKR_24. It can be clearly seen from the experimental results that the performance of 8_SAMKR_16 is significantly better than the others. It is easy to notice that, the performance of combinations 8_SAMKR_8 and 8_SAMKR_24 are even worse than 2_SAMKR_16 and 4_SAMKR_16. This shows that only considering the

Table 1. conducts experiments on coco validation according to some typical structure designs in SAMKR

m	c	g	s	AP
272	272	8	34	67.3
272	16	1	16	64
272	16	16	1	64.2
272	16	8	2	**68.9**

Table 2. designs the control experiment according to the parameter configuration of SAMKR.

Models	g	c	s	AP	APM	APL
2_SAMKR_16	2	16	8	64.4	56.8	75.6
4_SAMKR_16	4	16	4	64.1	56.4	75.6
8_SAMKR_16	8	16	2	**68.9**	**62.8**	**78.2**
8_SAMKR_8	8	8	1	62.9	56.5	72.8
8_SAMKR_24	8	24	3	63.5	55.1	76.1
4_SAMKR_8	4	8	2	63.9	57.1	74.2
12-SAMKR_24	8	24	2	65.7	58.6	76.5

parameter g can not make the network to be optimal. At last, we fix the number of channels in each feature subspaces = 2, three combinations of g and c are denoted as 4_SAMKR_8, 8_SAMKR_16 and 12_SAMKR_24. From the experimental results, it is still the combination of 8_SAMKR_16 that achieves the optimal result. This is mainly because the setting of 4_SAMKR_8 has the problem of too few branch channels, so much information is lost in the operation of $\mathrm{Conv}^{1\times1}$, and only 63.9AP is reached. Although the setting of 12_SAMKR_24 has increased the number of subspaces and the total number of channels of branches has also increased significantly, it has not improved the performance either, and the AP has only reached 65.7AP.

Through a large number of comparative experiments, this paper obtains the performance of the model trained under each parameter setting, and summarizes that the setting of 8_SAMKR_16 achieves the best detection performance in the COCO data set. According to the comparative experiment, it is clear that the performance of SAMKR is not determined by one parameter. It is determined by the total channel number c in the keypoint branch, the channel number s in the feature subspace divided in the branch, and the input size of the image.

C. *COCO Keypoint Detection*

Results on COCO validation2017. As shown in Table 3, this paper compares the performance of the proposed method on test set COCO validation2017 with the latest and most representative bottom-up method.

Compared with other methods, the performance of SAMKR is further improved. Using HRNet-W32 as the backbone, the SAMKR achieved 68.9AP score under the input size of 512 × 512. Compared with Centernet-HG whose model size is much larger than HRNet-W32, the improvement of SAMKR is 4.9AP. In addition, it is 0.9AP higher than the baseline method DEKR and achieves the same performance as the state-of-the-art bottom-up heatmap regression network SWAHR. When using the HRNet-W48 as the backbone, the SAMKR achieved 71.6 AP score under the input size of 640 × 640, which is 0.6 AP higher than the baseline method and 0.8 AP higher than SWAHR.

Results on COCO test-dev2017. Comparison results on the Coco test-dev2017 dataset are shown in Table 4. SAMKR achieves a 67.9AP score based on the HRNet-W32. .It provides a 0.6AP improvement over baseline DEKR, and is comparable to the SWAHR. When taking the HRNet-W48 as the backbone , the AP score of SAMKR reaches 70.4AP. Under the same input size, it exceeds 2AP compared to HrHRNet, 0.4AP compared to DEKR, and 0.2AP compared to SWAHR.

The comparison results on COCO data set prove that the method SAMKR proposed in this paper shows excellent performance on this data set. It reaches the most advanced level of bottom-up 2D human pose estimation, and improves the multi-scale feature extraction ability of the network.

D. *CrowdPose Keypoint Detection*

In order to further prove the effectiveness of SAMKR method, we carried out experiments on CrowdPose database also. The CrowdPose dataset has more crowd aggregation phenomena and is more challenging for human pose estimation. The multi-scale problem of human body is also more serious. The CrowdPose training set has 10,000 images, the validation set has 2,000 images, and the test set consists of 20,000 images.

We uses the same standard average accuracy based on OKS as the evaluation criterion. The CrowdPose dataset is divided into three congestion levels : easy, medium and hard, corresponding to AP^E , AP^M and AP^H , respectively. This paper reports the following indicators : AP, AP^{50} , AP^{75} , AP^E , AP^M, AP^H . CrowdPose training and testing methods follow the training and testing methods used in COCO.

The results of SAMKR and other SOTA methods on the CrowdPose test set are shown in Table 5. When the SAMKR network takes HRNet-W32 as the backbone, it reaches 73AP score which is 7.3AP higher than that of the baseline method DEKR. When the SAMKR network takes HRNet-W48 as the backbone, it reaches 74.4 AP score. Compared with the baseline method DEKR, it has an improvement of 7.1 AP . Compared with the most advanced bottom-up heatmap regression method SWAHR, it has an improvement of 2.8 AP.

From the experimental results on the CrowdPose dataset, the SAMKR method not only exceed the baseline method based on keypoint regression, but also exceed the method based on heatmap regression, and the improvement is obvious. This shows that the SAMKR method is more suitable for the challenging scenes of crowded crowd, and it is helpful to deal with scale-mixed scenes through efficient multi-scale feature learning.

E. *Visualization results and application*

We also experimented with computing time, testing it on the same device and under the same test set, all experiments were done on a RTX3090. The experimental results are shown in Table 6. The calculation time of both our method and DEKR method is much smaller than that of SWAHR method, indicating that the method based on key point regression has a great advantage in the calculation speed. Combined with the accuracy and computation time of our method, it shows that our method has achieved a good balance between performance and computation speed. Fig. 4 shows some visualization results of SAMKR on Crowdpose dataset. As shown in Fig. 4, the visualization results show that the pose of the human body is correctly estimated in most cases. The network can better cope with occlusion, human multi-scale, limb entanglement and other challenges. Fig. 5 shows the application results of this method in the

Table 3. The performance of SAMKR network tested on the coco validation2017 dataset

Method	**Input Size**	**#Params**	**GFLOPs**	**AP**	AP^{50}	AP^{75}	AP^{M}	AP^{L}	**AR**	AR^{M}	AR^{L}
Heatmap based single-scale testing											
PifPaf [22]	-	-	-	67.4	-	-	-	-	-	-	-
HGG [16]	512	-	-	60.4	83.0	66.2	-	-	64.8	-	-
PersonLab [24]	601	-	-	54.1	76.4	57.7	40.6	73.3	57.7	43.5	77.4
PersonLab [24]	1401	68.7	405.5	66.5	86.2	71.9	62.3	73.2	70.7	65.6	77.9
HrHRNet-W32 [15]	512	28.5	47.9	67.1	86.2	73	-	-	-	61.5	76.1
HrHRNet-W48 [15]	640	63.8	154.3	69.9	87.2	76.1	-	-	-	65.4	76.4
SWAHR(HrHRNet-W32) [17]	512	28.6	48	**68.9**	**87.8**	74.9	**63**	77.4	-	-	-
SWAHR(HrHRNet-W48) [17]	640	63.8	154.6	70.8	**88.5**	76.8	66.3	77.4	-	-	-
Regression based single-scale testing											
CenterNet-DLA [21]	512	-	-	58.9	-	-	-	-	-	-	-
CenterNet-HG [21]	512	227.8	206.9	64.0	-	-	-	-	-	-	-
DEKR((HRNet-W32) [2]	512	29.6	45.4	68	86.7	74.5	62.1	77.7	73	66.2	82.7
DEKR((HRNet-W48) [2]	640	65.7	141.5	71	88.3	77.4	66.7	78.5	76	70.6	84
Ours(HRNet-W32)	512	29.6	44.9	**68.9**	87.6	74.9	62.8	**78.2**	**73.4**	**66.6**	**83.1**
Ours(HRNet-W48)	640	65.7	140.9	**71.6**	**88.5**	**78.2**	**67.2**	**78.9**	**76.5**	**71.2**	**84.3**

Table 4. The verification results of SAMKR in COCO test-dev2017 set.

Method	**Input size**	**AP**	AP^{50}	AP^{75}	AP^{M}	AP^{L}	**AR**	AR^{M}	AR^{L}
Heatmap based single-scale testing									
Openpose [14]	-	61.8	84.9	67.5	57.1	68.2	66.5	-	-
Hourglass [8]	512	56.6	81.8	61.8	49.8	67	-	-	-
PifPaf [22]	-	66.7	-	-	62.4	72.9	-	-	-
PersonLab [24]	1401	66.5	88	72.6	62.4	72.3	71	66.1	77.7
HrHRNet-W32 [15]	512	66.4	87.5	72.8	61.2	74.2	-	-	-
HrHRNet-W48 [15]	640	68.4	88.2	75.1	64.4	74.2	-	-	-
SWAHR(HrHRNet-W32) [17]	512	67.9	**88.9**	74.5	**62.4**	75.5	-	-	-
SWAHR(HrHRNet-W48) [17]	640	70.2	**89.9**	76.9	65.2	77	-	-	-
Regression based single-scale testing									
CenterNet-DLA [21]	512	57.9	84.7	63.1	52.5	67.4	-	-	-
CenterNet-HG [21]	512	63	86.8	69.6	58.9	70.4	-	-	-
SPM [18]	-	66.9	88.5	72.9	62.6	73.1			
DEKR((HRNet-W32) [2]	512	67.3	87.9	74.1	61.5	76.1	72.4	65.4	81.9
DEKR((HRNet-W48) [2]	640	70	89.4	77.3	65.7	76.9	75.4	69.7	83.2
Ours(HRNet-W32)	512	**67.9**	88.4	**74.7**	62	**76.5**	**72.8**	**65.9**	**82.2**
Ours(HRNet-W48)	640	**70.4**	89.5	**77.4**	**66.3**	**77.2**	**75.7**	**70.2**	**83.3**

Table 5. The verification results of this method in CrowdPose test set.

Method	**AP**	AP^{50}	AP^{75}	AP^{E}	AP^{M}	AP^{H}
OpenPose [14]	-	-	-	62.7	48.7	32.3
HrHRNet-W48 [15]	65.9	86.4	70.6	73.3	66.5	57.9
SWAHR(HrHRNet-W48) [17]	71.6	88.5	77.6	78.9	72.4	63
DEKR(HRNet-W32) [2]	65.7	85.7	70.4	73	66.4	57.5
DEKR(HRNet-W48) [2]	67.3	86.4	72.2	74.6	68.1	58.7
Ours(HRNet-W32)	**73**	**89.6**	**78.7**	**80.7**	**74**	**62.8**
Ours(HRNet-W48)	**74.4**	**89.8**	**80.4**	**81.5**	**75.8**	**64**

Table 6. the comparison between the SAMKR method and the most advanced method in computing time.

Models	**Backbone**	**Cost time(ms)**
SWAHR	HRNet-W48	816.3
DEKR	HRNet-W48	137.8
SAMKR	HRNet-W48	164.4

actual pose extraction task. In Fig. 5, the detection of motion video by SAMKR still has a good effect without frame correlation. It means that the SAMKR network model is available in the actual scene.

V. Conclusion

In this paper, a bottom-up human pose estimation method SAMKR is proposed based on separated spatial attention module keypoint regression. The main idea of SAMKR is to use multiple attention modules to extract multi-scale information from each single regression keypoint branch.

Experimental results on COCO and CrowdPose datasets show that compared with the state of the art methods, SAMKR could achieve better results. Especially for the crowded crowd, the improvement is obvious. The internal design of the separate spatial attention module can be modified freely. Not only the parameter configuration but also the internal function modules can be adjusted freely. So, how to optimize the network to achieve better results is our future research direction.

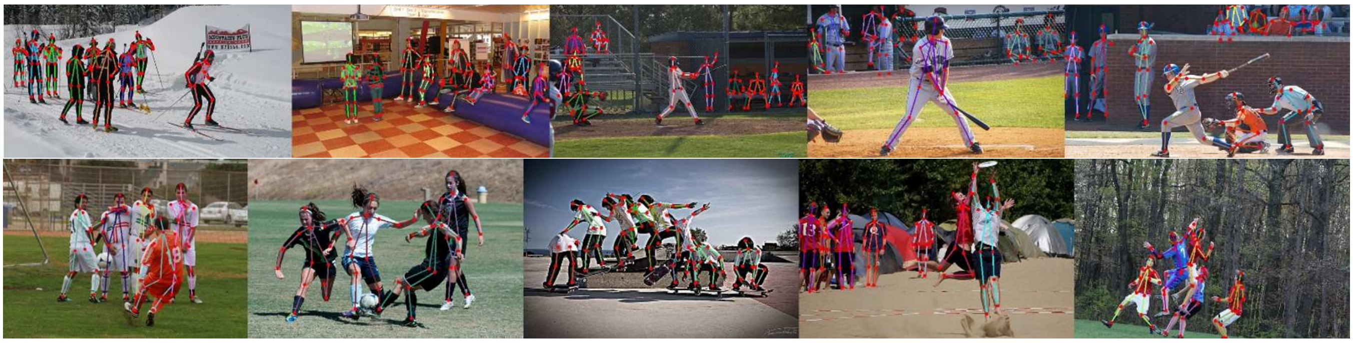

Fig. 4. shows the visualization results of SAMKR network on CrowdPose dataset. It can be seen that in complex scenes, SAMKR can better cope with multi-scale differences, occlusion, dense, limb entanglement and other challenges.

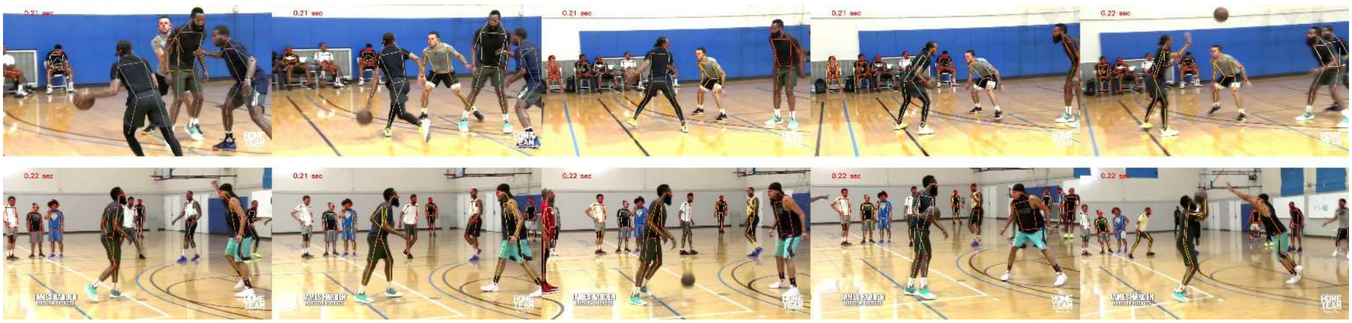

Fig. 5. shows the visualization results of SAMKR network on a basketball motion video without applying any before-and-after frame connections.

ACKNOWLEDGEMENT

This work was supported by the Key Project of NSFC (Grant No. U1908214), Special Project of Central Government Guiding Local Science and Technology Development(Grant No. 2021JH6/10500140), the Program for Innovative Research Team in University of Liaoning Province (Grant No. LT2020015), the Support Plan for Key Field Innovation Team of Dalian(Grant No. 2021RT06); the Support Plan for Leading Innovation Team of Dalian University (Grant No. XLJ202010); the Science and Technology Innovation Fund of Dalian (Grant No. 2020JJ25CY001).

REFERENCES

[1] R. SAINI, N. K. JHA, B. DAS, S. MITTAL, C. K. MOHAN. Ulsam: Ultra-lightweight subspace attention module for compact convolutional neural networks[C]. Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2020:1627-1636.

[2] Z. GENG, K. SUN, B. XIAO, Z. ZHANG, J. WANG. Bottom-Up Human Pose Estimation Via Disentangled Keypoint Regression[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021:14676-14686.

[3] K. SUN, B. XIAO, D. LIU, J. WANG. Deep high-resolution representation learning for human pose estimation[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019:5693-5703.

[4] J. WANG, K. SUN, T. CHENG, B. JIANG, C. DENG, Y. ZHAO, et al. Deep high-resolution representation learning for visual recognition[J]. IEEE transactions on pattern analysis and machine intelligence, 2020.

[5] Y. CAI, Z. WANG, Z. LUO, B. YIN, A. DU, H. WANG, et al. Learning delicate local representations for multi-person pose estimation[C]. European Conference on Computer Vision, 2020:455-472.

[6] H.-S. FANG, S. XIE, Y.-W. TAI, C. LU. Rmpe: Regional multi-person pose estimation[C]. Proceedings of the IEEE international conference on computer vision, 2017:2334-2343.

[7] K. HE, G. GKIOXARI, P. DOLLáR, R. GIRSHICK. Mask r-cnn[C]. Proceedings of the IEEE international conference on computer vision, 2017:2961-2969.

[8] A. NEWELL, K. YANG, J. DENG. Stacked hourglass networks for human pose estimation[C]. European conference on computer vision, 2016:483-499.

[9] G. PAPANDREOU, T. ZHU, N. KANAZAWA, A. TOSHEV, J. TOMPSON, C. BREGLER, et al. Towards accurate multi-person pose estimation in the wild[C]. Proceedings of the IEEE conference on computer vision and pattern recognition, 2017:4903-4911.

[10] K. SU, D. YU, Z. XU, X. GENG, C. WANG. Multi-person pose estimation with enhanced channel-wise and spatial information[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019:5674-5682.

[11] J. WANG, X. LONG, Y. GAO, E. DING, S. WEN. Graph-pcnn: Two stage human pose estimation with graph pose refinement[C]. European Conference on Computer Vision, 2020:492-508.

[12] S.-E. WEI, V. RAMAKRISHNA, T. KANADE, Y. SHEIKH. Convolutional pose machines[C]. Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, 2016:4724-4732.

[13] B. XIAO, H. WU, Y. WEI. Simple baselines for human pose estimation and tracking[C]. Proceedings of the European conference on computer vision (ECCV), 2018:466-481.

[14] Z. CAO, T. SIMON, S.-E. WEI, Y. SHEIKH. Realtime multi-person 2d pose estimation using part affinity fields[C]. Proceedings of the IEEE conference on computer vision and pattern recognition, 2017:7291-7299.

[15] B. CHENG, B. XIAO, J. WANG, H. SHI, T. S. HUANG, L. ZHANG. Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020:5386-5395.

[16] S. JIN, W. LIU, E. XIE, W. WANG, C. QIAN, W. OUYANG, et al. Differentiable hierarchical graph grouping for multi-person pose estimation[C]. European Conference on Computer Vision, 2020:718-734.

[17] Z. LUO, Z. WANG, Y. HUANG, L. WANG, T. TAN, E. ZHOU. Rethinking the Heatmap Regression for Bottom-up Human Pose Estimation[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021:13264-13273.

[18] X. NIE, J. FENG, J. ZHANG, S. YAN. Single-stage multi-person pose machines[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019:6951-6960.

[19] L. PISHCHULIN, E. INSAFUTDINOV, S. TANG, B. ANDRES, M. ANDRILUKA, P. V. GEHLER, et al. Deepcut: Joint subset partition and labeling for multi person pose estimation[C]. Proceedings of the

IEEE conference on computer vision and pattern recognition, 2016:4929-4937.

[20] F. WEI, X. SUN, H. LI, J. WANG, S. LIN. Point-set anchors for object detection, instance segmentation and pose estimation[C]. European Conference on Computer Vision, 2020:527-544.

[21] X. ZHOU, D. WANG, P. KRäHENBüHL. Objects as points[J]. arXiv preprint arXiv:1904.07850, 2019.

[22] S. KREISS, L. BERTONI, A. ALAHI. Pifpaf: Composite fields for human pose estimation[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019:11977-11986.

[23] A. NEWELL, Z. HUANG, J. DENG. Associative embedding: End-to-end learning for joint detection and grouping[C]. In NeurIPS, pages 2274–2284, 2017.1,3,6,7,8, 2016.

[24] G. PAPANDREOU, T. ZHU, L.-C. CHEN, S. GIDARIS, J. TOMPSON, K. MURPHY. Personlab: Person pose estimation and instance segmentation with a bottom-up, part-based, geometric embedding model[C]. Proceedings of the European Conference on Computer Vision (ECCV), 2018:269-286.

[25] Y. CHEN, Z. WANG, Y. PENG, Z. ZHANG, G. YU, J. SUN. Cascaded pyramid network for multi-person pose estimation[C]. Proceedings of the IEEE conference on computer vision and pattern recognition, 2018:7103-7112.

[26] F. ZHANG, X. ZHU, H. DAI, M. YE, C. ZHU. Distribution-aware coordinate representation for human pose estimation[C]. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020:7093-7102.

[27] K. LI, S. WANG, X. ZHANG, Y. XU, W. XU, Z. TU. Pose Recognition with Cascade Transformers[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021:1944-1953.

[28] X. NIE, J. FENG, J. XING, S. YAN. Pose partition networks for multi-person pose estimation[C]. Proceedings of the european conference on computer vision (eccv), 2018:684-699.

[29] J. HU, L. SHEN, G. SUN. Squeeze-and-excitation networks[C]. Proceedings of the IEEE conference on computer vision and pattern recognition, 2018:7132-7141.

[30] S. WOO, J. PARK, J.-Y. LEE, I. S. KWEON. Cbam: Convolutional block attention module[C]. Proceedings of the European conference on computer vision (ECCV), 2018:3-19.

[31] J. HU, L. SHEN, S. ALBANIE, G. SUN, A. VEDALDI. Gather-excite: Exploiting feature context in convolutional neural networks[J]. arXiv preprint arXiv:1810.12348, 2018.

[32] X. CHU, W. YANG, W. OUYANG, C. MA, A. L. YUILLE, X. WANG. Multi-context attention for human pose estimation[C]. Proceedings of the IEEE conference on computer vision and pattern recognition, 2017:1831-1840.

[33] B. X. BOWEN CHENG, J. WANG, H. SHI, T. S. HUANG, L. ZHANG. Bottom-up Higher-Resolution Networks for Multi-Person Pose Estimation[J].

[34] D. P. KINGMA, J. BA. Adam: A method for stochastic optimization[C]. In Y oshua Bengio and Yann LeCun,editors,ICLR, 2015.6,8, 2014.