Research Article

Lanfei Zhao and Zhihua Chen*

CRNet: Context feature and refined network for multi-person pose estimation

https://doi.org/10.1515/jisys-2022-0060
received December 31, 2021; accepted May 01, 2022

Abstract: Multi-person pose estimation is a challenging problem. Bottom-up methods have been greatly studied because the prediction speed of top-down methods is related to the number of people in the input image, making these methods difficult to apply in real-time environments. To solve the problems of scale sensitivity and quantization error in bottom-up methods, it is necessary to have a model that can predict multi-scale keypoints and refine quantization error. To achieve this, we propose context feature and refined network for multi-person pose estimation (CRNet), which can effectively solve the problems of scale sensitivity and quantization error in bottom-up methods. We use a multi-scale feature pyramid and context feature to achieve scale invariance of the network. We extract global and local features and then fuse them by attentional feature fusion (AFF) to obtain context feature that adapt to multi-scale keypoints. In addition, we propose an efficient refined network to solve the problem of quantization error and use multi-resolution supervised learning to further improve the prediction accuracy of CRNet. Comprehensive experiments are conducted on two benchmarks: COCO and MPII datasets. The average precision of CRNet reached 72.1 and 80.2%, respectively, surpassing most state-of-the-art methods.

Keywords: convolutional neural network, multi-scale feature, attentional feature fusion, refined network, multi-resolution supervision

1 Introduction

Multi-person pose estimation is a challenging problem in the field of computer vision. It aims to locate all keypoints of human joints (such as elbows, wrists, and knees) or body parts from an input image and combine them into independent person instance. It is widely applied in the fields of human–computer interaction, human behavior recognition, security monitoring, etc. [1]. The application of the convolutional neural network (CNN) has greatly improved the prediction accuracy of multi-person pose estimation [2–5]. Multi-person pose estimation methods based on CNN can be divided into top-down and bottom-up methods.

Because the prediction speed of the top-down methods is affected by the number of people in the natural scene and the prediction accuracy depends heavily on the person detector, it is difficult to apply to the real-time scene. Currently, the bottom-up methods are receiving more and more attention. The bottom-up methods simultaneously predict all the keypoints of human joints in the input image. However, the multiple scale of keypoints in an image is one of the main reasons to limit bottom-up methods. Therefore, the extraction of context feature containing multi-scale information is being widely studied. In addition, the

* **Corresponding author: Zhihua Chen,** The Higher Educational Key Laboratory for Measuring & Control Technology and Instrumentations of Heilongjiang Province, Harbin University of Science and Technology, Harbin 150080, China, e-mail: chancwah@163.com
Lanfei Zhao: The Higher Educational Key Laboratory for Measuring & Control Technology and Instrumentations of Heilongjiang Province, Harbin University of Science and Technology, Harbin 150080, China, e-mail: zlf82563144@hrbust.edu.cn

state-of-the-art methods also face the problem of quantization error. Although there are some methods [2,6] to solve the quantization error, these methods are difficult to adapt to the keypoints of all scales.

In this work, we propose a bottom-up multi-person pose estimation method based on context feature and refined network (CRNet), aiming at enhancing scale invariance and optimizing quantization error of the network. First, a context feature extraction method is proposed, which can extract global and local features, and the attention mechanism is used for multi-scale feature fusion to enhance the scale invariance of the network. Second, a quantization error refined network is proposed, which has the ability of automatically repairing quantization error and multi-resolution supervision is used to facilitate the proposed CRNet learning. Finally, the effectiveness of the proposed methods is verified on two mainstream multi-person pose estimation benchmark datasets: COCO dataset [7] and MPII dataset [8].

Our contributions are summarized as follows:

(1) We propose the context feature to overcome the difficulty of detecting multi-scale keypoints and improve the prediction accuracy of the model.
(2) We propose a refined network to cope with the inherent quantization error problem of the bottom-up methods to further improve the accuracy of multi-person pose estimation.
(3) Compared with the state-of-the-art methods, our CRNet model achieves competitive results on two mainstream benchmarks.

The remainder of this article is structured as follows:

Section 2 introduces the related work of multi-person pose estimation and the network architecture of HRNet. Section 3 discusses the proposed method, where Sections 3.1 and 3.2, respectively, introduce the structure of context feature and refined network, and Section 3.3 introduces our CRNet model. The experimental setup and result analysis are described in detail in Section 4. Section 5 summarizes the proposed methods and the effectiveness of the network.

2 Related work

2.1 Top-down methods

The top-down methods combine the single-person pose estimation with the object detection algorithm, which uses the person detector to detect all human instances from the image and then estimates the pose for each person. G-RMI [9] used the full convolutional residual network to estimate the pose of each detected person body based on the faster R-CNN. CPN [10] proposes a two-stage cascaded pyramid network, which introduces online difficult keypoints mining technology to predict the difficult keypoints. The Simple Baseline network [11] uses some deconvolution layers to increase the output feature maps resolution of the residual network. The main idea of the high-resolution network (HRNet) [12] is to maintain high resolution by connecting multi-resolution subnets in parallel, which allows HRNet achieves state-of-the-art results in multiple benchmark datasets.

2.2 Bottom-up methods

Different from the top-down methods, the bottom-up methods predict all the human keypoints in the image and then assign the keypoints to an individual person by grouping algorithm. The DeepCut [13] model is the first bottom-up multi-person pose estimation method, which uses the integer linear program algorithm to solve the association problem of keypoints. OpenPose [14] uses the two-branch network to predict the heatmaps and part affinity fields of keypoints, respectively. Associative Embedding [15] is an end-to-end multi-person pose estimation model based on the hourglass network, which can efficiently predict the

heatmaps and grouping information of keypoints simultaneously. The main idea of HigherHRNet [16] is to use HRNet as a feature extraction network. The deconvolution layers similar to Simple Baseline is used to output the heatmap of keypoints with higher resolution and combined with the grouping method of associative embedding. This model can effectively predict the small-scale keypoints.

2.3 HRNet

HRNet [12] uses a stem structure that consists of two strided convolutional layers to quickly downsample the input image by four times and takes the result of downsampling as the input of the main body. As shown in Figure 1, the main body of HRNet is divided into four stages. A high-resolution subnet is taken as the first stage, and a new stage is formed by gradually adding high-resolution subnets to low-resolution subnets, and the multi-resolution subnets are connected in parallel. Therefore, the parallel subnet of the latter stage is composed of the multi-resolution subnet of the previous stage and a low-resolution subnet. To improve the robustness of the network, multi-scale feature fusion units are included in each stage of HRNet, so that subnets with different resolutions can interact with each other.

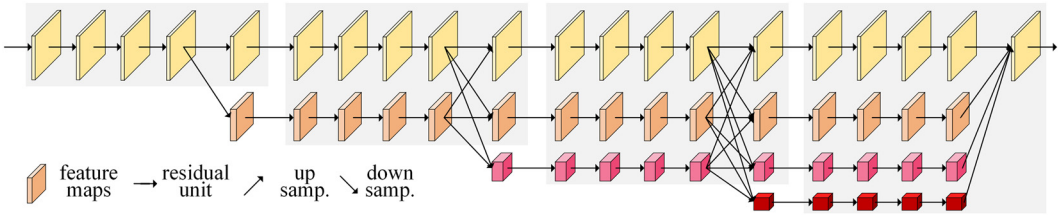

Figure 1: The structure of HRNet.

Although HRNet is a powerful backbone network that can estimate conventional human pose well, the prediction accuracy of multi-scale keypoints needs to be improved. The coarse downsampling of stem structure leads to the loss of small-scale keypoints information, which makes it difficult to estimate the pose of small-scale person. HRNet cannot obtain global semantic information, which leads to incorrect estimation of large-scale human pose. For this reason, we propose the context feature to enhance the scale invariance of HRNet. In addition, a refined network is proposed to solve the inherent quantization error problem of the bottom-up methods.

3 Our method

3.1 Context feature

The purpose of the context feature is to enable the network to adapt to multi-scale human keypoints and enhance the scale invariance. As shown in Figure 2, the context feature extraction module (CFEM) is mainly composed of two parts: (a) multi-scale feature extraction and (b) attentional feature fusion, in which the multi-scale feature extraction module is responsible for extracting multi-scale features, and the attention feature fusion module uses the attention mechanism to fuse multi-scale features.

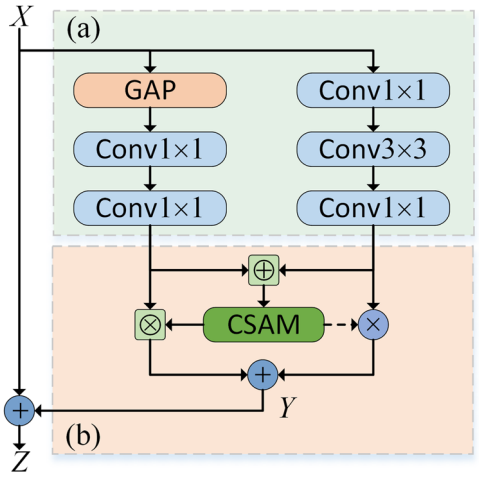

Figure 2: Context feature extraction module. (a) Multi-scale feature extraction, (b) attentional feature fusion.

3.1.1 Multi-scale feature extraction

The receptive field is very important for multi-person pose estimation [3,17]. To solve the problem that CNN cannot effectively extract global semantic information, we propose a multi-scale feature extraction module that can extract global and local context information simultaneously. As shown in Figure 2(a), there are two branches in the multi-scale feature extraction module: the left branch extracts global information, and the right branch extracts local information. For global information, we use global average pooling (GAP) to compress the spatial information of input feature $X \in \mathbf{R}^{H\times W\times C}$ with C channels and feature maps of size $H \times W$, and then use two point-wise convolution layers to learn the global information. In addition, batch normalization and ReLU nonlinear activation unit follow each convolution to obtain global feature $F_g \in \mathbf{R}^{1\times 1\times C}$ containing global context information. Formally, the process can be summarized as follows:

$$F_g = f_g(g(X); W_g), \tag{1}$$

where $f_g(\cdot)$ denotes global convolution and W_g is the corresponding parameter set. Given an input feature X, $X_c(i, j)$ represents the information of coordinate position (i, j) in the c channel, and the GAP $g(X)$ is calculated by

$$g(X) = \frac{1}{W \times H}\sum_{i=1}^{W}\sum_{j=1}^{H} X_c(i, j). \tag{2}$$

For local information, we use a similar method to the bottleneck residual to better extract local feature containing more detailed information. Given an input feature $X \in \mathbf{R}^{H\times W\times C}$, the local feature $F_l \in \mathbf{R}^{H\times W\times C}$ can be computed as follows:

$$F_l = f_l(X; W_l), \tag{3}$$

where $f_l(\cdot)$ denotes local convolution and W_l is the corresponding parameter set.

3.1.2 Attentional feature fusion

The feature fusion based on the attention mechanism can effectively highlight the information-rich features [18]. Since attentional feature fusion (AFF) [19] only uses channel attention, we add a spatial attention mechanism on the basis of AFF to construct an attentional feature fusion module to fuse multi-scale features. As shown in Figure 2(b), the attentional feature fusion module takes the global feature F_g and

local feature F_l from the multi-scale feature extraction module as input and then merges them into the channel-spatial attention module (CSAM) to calculate the attention weight. The attention weight is multiplied by the input feature to obtain the attention feature, and the attention feature is fused to obtain the context feature $Y \in \mathbf{R}^{H\times W\times C}$. The resulting feature with powerful expressiveness contains global and local context information that can effectively predict multi-scale human keypoints. Let $M(\cdot)$ represent the CSAM, and the process can be expressed as follows:

$$Y = M(F_g \oplus F_l) \otimes F_g + (1 - M(F_g \oplus F_l)) \times F_l, \tag{4}$$

where $\oplus$ and $\otimes$ represent the addition and multiplication of the broadcast mechanism. In Figure 2(b), the dashed line represents $1 - M(F_g \oplus F_l)$. It is worth noting that the output of $M(F_g \oplus F_l)$ is a real number between 0 and 1, so is the output of $1 - M(F_g \oplus F_l)$, which allows attentional feature fusion module to weight F_g and F_l with attention.

As shown in Figure 3, both channel and spatial attention mechanisms are used in CSAM. Given an intermediate feature $X' \in \mathbf{R}^{H\times W\times C}$ as input, the output feature $X'' \in \mathbf{R}^{H\times W\times C}$ refined by CSAM can be obtained as follows:

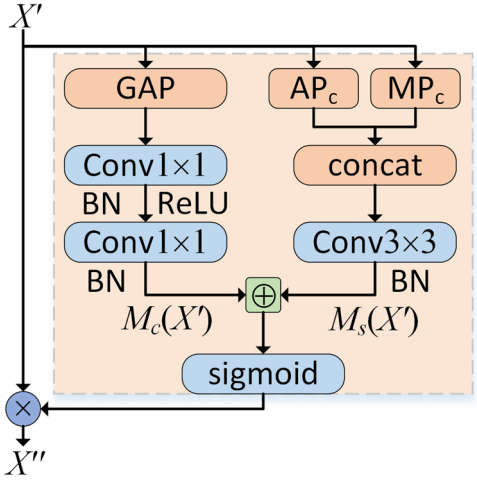

Figure 3: Channel-spatial attention module.

$$X'' = s(M_c(X') \oplus M_s(X')) \times X', \tag{5}$$

where $M_c(X')$ and $M_s(X')$ represent channel attention and spatial attention, respectively. $s(\cdot)$ denotes the sigmoid activation function.

Channel attention is generated based on the relationship between feature channels. First, GAP is used to spatially compress the input features, and then the dependency relationship between channels is learned through two point-wise convolutional layers. The process can be expressed as follows:

$$M_c(X') = f_c(g(X'); W_c), \tag{6}$$

where $f_c(\cdot)$ denotes channel convolution and W_c is the corresponding parameter set.

Spatial attention is generated based on the spatial dependence between features. Maximum pooling MP_c and average pooling AP_c along the channel direction are performed and then concatenating the output feature. A 3×3 convolutional layer is used to learn the spatial relationship of feature to obtain spatial attention $M_s(X')$. The process can be expressed as follows:

$$M_s(X') = f_s([\mathrm{AP}_c(X'); \mathrm{MP}_c(X')]; W_s), \tag{7}$$

where $f_s(\cdot)$ denotes spatial convolution and W_s is the corresponding parameter set.

3.2 Refined network

Aiming at the problem of quantification error, the previous methods add an offset vector to the maximum activation position of the predicted heatmaps. A standard offset is used in the stacked hourglass network [2], which is equal to 1/4 of the unit vector of the maximal activation to the second maximal activation direction. However, this method only relies on empirical formulas without the support of theoretical derivation. DARK [6] performs Taylor series expansion at the maximum position of the heatmaps to obtain the location of predicted keypoints. When the ground true position of keypoints is far from the predicted maximum position of the heatmap, it will not meet the Taylor series expansion condition. Therefore, this method is difficult to deal with some keypoints with a large range of motion, such as wrists, ankles, etc. In this work, we utilize the powerful nonlinear expression ability of CNN to propose a refined network to solve the quantization error problem. The goal of the refined network is to find a nonlinear function $f_{\text{refine}}(\cdot)$ that maps the keypoint positions with error to precise positions. Given the heatmap $H^h \in \boldsymbol{R}^{\frac{H}{4}\times\frac{W}{4}\times N}$ that contains information about N keypoints from the backbone network, the refined heatmap $H^r \in \boldsymbol{R}^{H\times W\times N}$ by the refined network can be computed as follows:

$$H^r = f_{\text{refine}}(H^h; W_r), \tag{8}$$

where W_r is the parameter set of the refined network.

To balance efficiency and accuracy, as shown in Figure 4, the refined network uses two deconvolution layers to upsample the heatmaps output by the backbone network to the input image size. Four bottleneck residual blocks are used for refinement learning, and skip connection across blocks are used to enhance the refined feature. Finally, a point-wise convolution layer is used for channel number matching to obtain a calibrated heatmap.

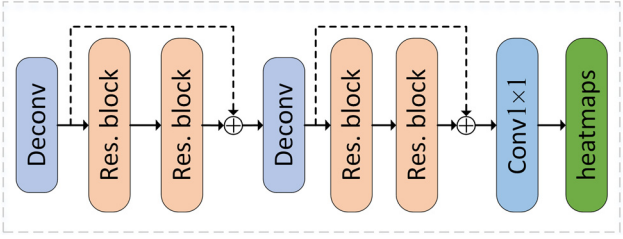

Figure 4: The refined network.

3.3 CRNet

Our CRNet takes HRNet [12] as the backbone and improves it with the proposed context feature and refined network. The overall architecture of CRNet is shown in Figure 5.

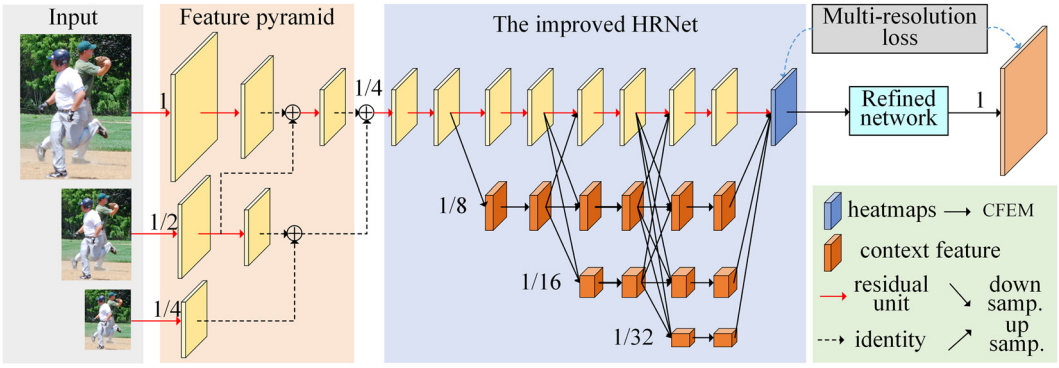

Figure 5: The overall architecture of CRNet.

(1) Due to the coarse downsampling operation of stem in HRNet, the important detailed position information is lost and the importance of low-level features is ignored. To solve this problem, we use a feature pyramid module instead of the stem structure for downsampling. The module accepts input in three scales, which can better extract low-level and small-scale feature, and alleviates the problem that small-scale keypoints cannot be detected due to coarse downsampling.
(2) The proposed CFEM is used to replace the residual blocks of 1/8, 1/16, and 1/32 resolution subnets in HRNet and retain the residual blocks of 1/4. The improved HRNet can extract global and local features and better fuse multi-scale features via the attention mechanism to enhance the scale invariance of the network.
(3) A refined network is added to the end of the improved HRNet, so that the network can calibrate the quantization error and output more accurate keypoints heatmaps. In addition, multi-resolution loss function is used to supervise CRNet learning. In this work, the mean square error loss function is used to calculate the loss of the network. L^h and L^r are the loss functions of two different scales corresponding to the improved HRNet and refined network, respectively. The total loss of CRNet can be computed as follows:

$$\text{Loss} = \alpha L^h + (1 - \alpha)L^r, \tag{9}$$

where α denotes balance factor, and the prediction accuracy is highest when it is set to 0.9 through ablation experiment. L^h and L^r are, respectively, calculated by

$$L^h = \frac{1}{N}\sum_{i=1}^{N}||\widehat{H_i^h} - H_i^h||_2^2, \tag{10}$$

$$L^r = \frac{1}{N}\sum_{i=1}^{N}||\widehat{H_i^r} - H_i^r||_2^2, \tag{11}$$

where $\widehat{H_i^h} \in \mathbf{R}^{\frac{H}{4}\times\frac{W}{4}}$ and $\widehat{H_i^r} \in \mathbf{R}^{H\times W}$ are, respectively, the label heatmap of the ith keypoint with two different scales. Let x denote the 2D coordinates of the heatmap and $\widetilde{x_i}$ denote the ground true coordinates of the ith keypoint, $\widehat{H_i}$ can be generated by

$$\widehat{H_i}(x) = \frac{1}{2\pi\sigma^2}\exp\left(-\frac{||x - \widetilde{x_i}||_2^2}{2\sigma^2}\right), \tag{12}$$

where σ is the standard deviation, which is set to 1 in this work.

4 Experiments

Experiments are performed on two mainstream multi-person pose estimation benchmarks (COCO dataset [7] and MPII dataset [8]) to verify the effectiveness of the proposed methods in enhancing network scale invariance and optimizing quantization error.

4.1 Experiment setup

4.1.1 Datasets

We evaluate the proposed CRNet model on two widely adopted 2D benchmark datasets: COCO dataset [7] and MPII dataset [8]. The train/val/test sets of COCO keypoint detection dataset contain 57, 5, and 20 k images, respectively. The standard evaluation metric of COCO is Object Keypoint Similarity, which can be calculated by

$$\mathrm{OKS} = \frac{\sum_{i=1}^{17} \exp\left(\frac{-d_i^2}{2s^2k_i^2}\right)\delta(v_i > 0)}{\sum_{i=1}^{17}\delta(v_i > 0)}, \tag{13}$$

where i is the index of all keypoints, d_i represents the Euclidian distance between the predicted value of keypoint and the ground true value, v_i denotes whether the keypoint is visible, s represents the scale of the object, k_i represents a constant controlling attenuation corresponding to each keypoint, and $\delta(\cdot)$ is the step function. We report the standard average precision: AP^{50} (AP at OKS = 0.50), AP^{75}, mAP (the mean of AP scores at 10 OKS positions, 0.50, 0.55, ..., 0.90, 0.95); AP^M for medium objects and AP^L for large objects.

The MPII Human Pose dataset contains 5,602 images containing multi-person, of which 3,844 were used to train the CRNet model and the remaining 1,758 images were used for model testing. The dataset also provides more than 28,000 annotated single-person pose samples, and each person is annotated with 16 keypoints. The MPII dataset uses the mean average precision (mAP) of keypoint detection to evaluate model accuracy.

4.1.2 Data augmentation

We use traditional data augmentation strategies for multi-person pose estimation. For the COCO dataset, the methods we used to enhance the training samples included random rotation of −30 to 30 degrees, random scaling of 0.75 to 1.5 times, random cropping the image to the size of 512 × 512 or 640 × 640, and random flipping. For 512 × 512 input size, the CRNet will generate ground true heatmaps with resolutions of 128 × 128 and 512 × 512, respectively, while for 640 × 640 input size, the CRNet will generate ground true heatmaps with resolutions of 160 × 160 and 640 × 640, respectively. For the MPII dataset, we enhance the training samples that included random cropping of the image to the size of 384 × 384, random rotation of −40 to 40 degrees, random scaling of 0.7–1.3 times, and random flipping.

4.1.3 Training

For the COCO dataset, Adam optimizer [20] is used to update parameters. The base learning rate is 1×10^{-3} and dropped to 1×10^{-4} and 1×10^{-5} on the 200th epochs and 260th epochs, respectively. We trained the proposed CRNet model for a total of 300 epochs. For the MPII dataset, we randomly selected 350 samples from the multi-person training set as a validation set during training and used the remaining and single-person samples for model training. By using RMSprop [21] as the training optimizer and setting the base learning rate to 0.003, the total epoch of training was 250 and the learning rate was reduced by two times at 150, 170, 200, and 230 epochs.

4.2 Results on COCO dataset

We evaluated CRNet on the COCO test-dev set. In Table 1, we can see the comparison results of the proposed CRNet model and the state-of-the-art bottom-up methods. The HRNet implemented by the bottom-up approach is still strong with mAP reaching 64.1%, which is close to Associative Embedding network, but its parameters and FLOPs are only 10 and 19%, respectively, of Associative Embedding network. In our CRNet model, when HRNet-W32 is used as the backbone, mAP is 69.2% better than most networks and slightly lower than HigherHRNet with HRNet-W48. However, our CRNet is more advantageous in terms of parameters and FLOPs. When HRNet-W48 was used as the backbone, CRNet achieved the best accuracy and mAP reached 72.1%. Compared with HigherHRNet, the accuracy for medium scale and large scale was improved by 1.9 and 1.1%, respectively. Indicating that the proposed CRNet can effectively predict multi-scale human pose.

Table 1: Comparisons with state-of-the-arts on the COCO test-dev set

Method	Backbone	Input size	Params	GFLOPs	mAP	AP^{50}	AP^{75}	AP^{M}	AP^{L}
OpenPose [14]	—	—	—	—	61.8	84.9	67.5	57.1	68.2
AE [15]	Hourglass	512	277.8M	206.9	65.5	86.8	72.3	60.6	72.6
PensonLab [22]	ResNet-152	1401	68.7M	405.5	68.7	89.0	75.4	64.1	75.5
PifPaf [23]	—	—	—	—	66.7	—	—	62.4	72.9
HRNet*	HRNet-W32	512	28.5M	38.9	64.1	86.3	70.4	57.4	73.9
SPM [24]	Hourglass	—	—	—	66.9	88.5	72.9	62.6	73.1
HigherHRNet [16]	HRNet-W48	640	63.8M	154.3	70.5	89.3	77.2	66.6	75.8
CRNet(Ours)	HRNet-W32	512	30.2M	54.7	69.2	87.6	73.6	64.8	74.5
	HRNet-W48	640	68.5M	197.4	**72.1**	**89.8**	**78.3**	**68.5**	**76.9**

* represents a bottom-up implementation.

4.3 Results on MPII dataset

Table 2 presents the evaluation result of the proposed CRNet on MPII test-dev set. Table 2 presents that the previous methods have achieved high accuracy when dealing with relatively fixed or large-sized keypoints or parts such as the head and shoulders. However, the prediction accuracy of relatively flexible keypoints or parts such as wrists and ankles is still low. The reason is that these keypoints or parts are easy to be occluded and the spatial location is changeable. It is necessary for the network to learn multi-scale context information and refine the quantization error, but the previous methods are difficult to achieve at the same time. The proposed CRNet combines multi-scale context feature and refined network, which can predict multi-scale keypoints and refine the quantization error inherent in the bottom-up methods. Our CRNet model achieves 80.2% mAP on MPII multi-person test-dev set, surpassing the previous state-of-the-art models. As shown in Figure 6, the prediction accuracy of CRNet is better than previous methods in most keypoints, among which elbow, wrist, knee, and ankle are improved by 1.7, 1.9, 1.5, and 2.1% respectively. It demonstrates that the proposed CRNet can better deal with difficult key points or parts.

Table 2: Comparison with state-of-the-arts on the MPII test-dev set

Method	Head	Sho.	Elb.	Wri.	Hip.	Knee	Ank.	mAP
Insafutdinov et al. [25]	88.8	85.2	75.9	64.9	74.2	68.8	60.5	74.3
Duan et al. [26]	88.4	86.3	70.4	63.4	73.6	72.5	66.7	74.6
Cao et al. [14]	91.2	87.6	77.7	66.8	75.4	68.9	61.7	75.6
Newell et al. [15]	92.1	89.3	78.9	69.8	76.2	71.6	64.7	77.5
Fieraru et al. [27]	91.8	**89.5**	80.4	69.6	**77.3**	71.7	65.5	78.0
Nie et al. [24]	89.7	87.4	80.4	72.4	76.7	74.9	68.3	78.5
CRNet(Ours)	**92.3**	88.9	**82.1**	**74.3**	76.9	**76.4**	**70.4**	**80.2**

Bold values represent the best results for the corresponding metric.

4.4 Ablation experiments

4.4.1 Effectiveness of each component in CRNet

To analyze the effectiveness of each individual component in the proposed CRNet, we perform a series of ablation experiments on the COCO validation set using HRNet-W32 as the backbone network. Figure 7 describes in detail the five network structures adopted in the ablation experiment. Figure 7(a) shows the original HRNet, Figure 7(b) represents the addition of feature pyramid structure, Figure 7(c) shows the improved HRNet using context feature, Figure 7(d) adds a refined network, and Figure 7(e) uses multi-resolution supervision. The experimental results are presented in Table 3.

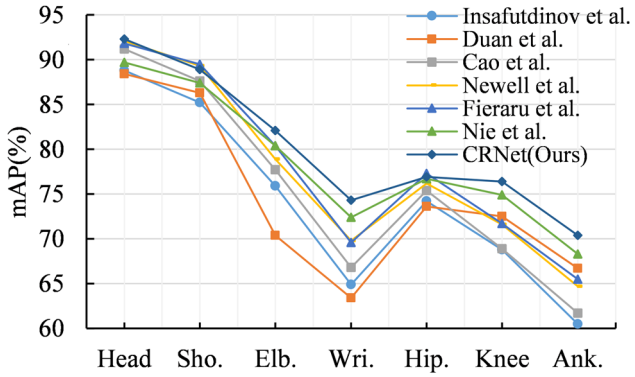

Figure 6: Comparison of prediction accuracy of all keypoints.

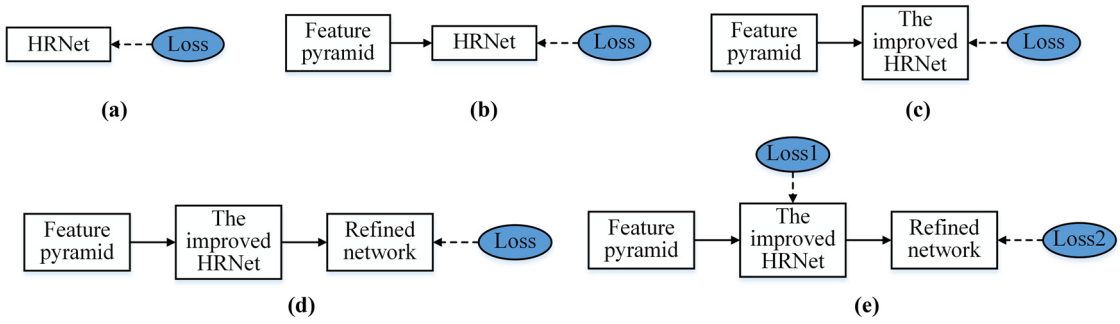

Figure 7: The network structure of the ablation experiment.

4.4.1.1 Feature pyramid structure

We use a feature pyramid to replace stem structures in HRNet for downsampling. It can be seen from Table 3 that mAP is improved by 0.6% after the feature pyramid is used. This is because stem structure ignores the important low-level features, while feature pyramid structure retains them. In addition, the feature pyramid can alleviate the problem of small-scale keypoints information loss caused by coarse downsampling.

4.4.1.2 Context feature

To solve the scale-sensitive problem of HRNet, we propose a context feature module and use it to replace the residual block in HRNet. Table 3 shows that mAP reaches 67.5% after using context features. Both AP^M and AP^L are greatly improved, indicating that compared with HRNet, our method could better predict multi-scale keypoints.

4.4.1.3 Refined network

After adding the refined network, the mAP decreased by 0.8%, which was caused by the lack of multi-resolution supervision. The network still regarded the refined network as a part of the backbone network. However, the HRNet based on context features is powerful enough, and it is prone to overfitting if the refined network is added.

Table 3: Results of ablation experiments on the COCO validation set

Network	Feature pyramid structure	Context feature	Refined network	Multi-resolution supervision	mAP	AP^M	AP^L
HRNet					63.9	56.7	72.4
CRNet(Ours)	√				64.5	57.2	72.5
	√	√			67.5	62.9	74.1
	√	√	√		66.7	62.6	73.7
	√	√	√	√	**68.8**	**65.1**	**74.6**

Bold values represent the best results for the corresponding metric.

4.4.1.4 Multi-resolution supervision

After using the refined network and multi-resolution supervision, the mAP reached 68.8%. The reason is that the multi-resolution supervision can effectively distinguish the improved HRNet from the refined network. The former is responsible for extracting important features and outputting heatmaps, while the latter is responsible for refining the heatmaps. Table 3 shows that AP^M has a significant improvement, demonstrating that the quantization error of keypoints at small and medium scales is more serious than that at large scales.

4.4.2 The impact of balance factor

To verify the influence of the balance factor in equation (9), the value of α is gradually increased from 0 to 1 with a stride of 0.1. HRNet-W32 is used as the backbone for training, and the test results on the COCO validation set are shown in Figure 8. mAP is highest when α has a value of 0.9, so the value of α is set to 0.9 in this work.

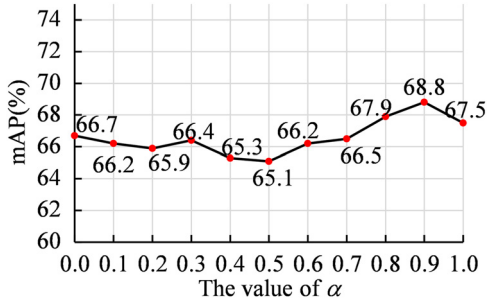

Figure 8: The impact of the balance factor on the network.

4.4.3 The impact of input size

We propose a context feature to solve the problem of scale sensitivity. In the context feature, we consider the features of multiple scales and use the attention mechanism to achieve multi-scale fusion. To verify the effectiveness of context feature for enhancing network scale invariance, we evaluated the impact of input image size on HRNet and the improved HRNet based on context feature on the COCO validation set. In the experiments, HRNet-W32 is used as the backbone network, and three different resolutions of 256×256, 384×384, and 512×512 are input, respectively. Two important conclusions can be drawn from the observations in Table 4: (a) As the input size decreases, the prediction accuracy of HRNet and the improved

Table 4: The effect of input size on the COCO validation set

Network	Input size	mAP	AP^{50}	AP^{75}	AP^{M}	AP^{L}
HRNet	256 × 256	55.9	79.4	61.7	53.2	64.4
HRNet + Context feature		**61.2**	**83.7**	**67.6**	**57.5**	**68.5**
HRNet	384 × 384	61.7	84.8	67.9	55.5	69.3
HRNet + Context feature		**65.8**	**86.4**	**71.3**	**60.3**	**73.2**
HRNet	512 × 512	63.9	85.3	70.1	56.7	72.4
HRNet + Context feature		**67.5**	**88.6**	**72.9**	**62.9**	**74.1**

Bold values represent the best results for the corresponding metric.

HRNet decreases to varying degrees, indicating that the problem of scale sensitivity does exist. (b) However, compared with HRNet, our method has less precision loss, especially when the resolution is reduced to a lower level. It is proved that context feature can indeed enhance scale invariance and have better robustness to different scales.

4.4.4 Effectiveness of the refined network

In this work, we propose a refined network to address quantization error. To verify the effectiveness of the refined network, we compare it with several offset methods. These include no offset (resulting directly in the heatmaps maximum activation), standard offset [2], and DARK [6]. In the experiment, HRNet-W32 is used as the backbone network, and the input resolution is 512 × 512. The experimental results on the COCO validation set are presented in Table 5, from which we can draw the following two important conclusions: (a) The standard offset method brings 1.4% mAP improvement, indicating that the original HRNet does have quantization errors, and the offset method can suppress this error to a certain extent. (b) Compared with DARK, our refined network improves 0.5% mAP, which proves that the refined network can suppress the quantization error better than the offset methods, thereby further effectively improving the prediction accuracy of the model.

Table 5: Effectiveness of the refined network on the COCO validation set

Network	mAP	AP^{50}	AP^{75}	AP^{M}	AP^{L}
HRNet + no offset	63.9	85.3	70.1	56.7	72.4
HRNet + standard offset	65.3	85.7	70.8	59.3	73.0
HRNet + DARK	66.4	86.2	71.6	61.4	73.8
HRNet + refined network	**66.9**	**87.4**	**72.5**	**62.3**	**73.9**

Bold values represent the best results for the corresponding metric.

4.5 Visualization of inference results

The visualization of the inference results of the proposed CRNet for conventional pose on COCO and MPII datasets is shown in Figure 9. Figure 9 shows that the proposed CRNet has excellent results for conventional human pose estimation. The visualization of multi-scale human pose inference results is shown in Figure 10, where Figure 10(a) is the output result of HRNet and Figure 10(b) is the output result of the proposed CRNet. In Figure 10(a), we can see that HRNet will have incorrect or unestimable problems when estimating the keypoint of large scale or small scale. The main reason is that the stem module causes the loss of small-scale keypoints and HRNet cannot obtain the global context information for the large-scale human body. The feature pyramid structure and context feature proposed in this work can effectively solve the scale

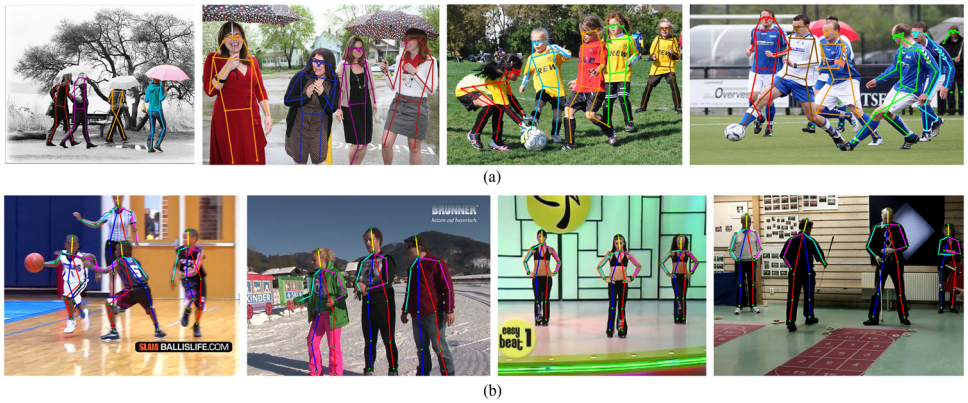

Figure 9: Visualization of conventional poses. (a) Visualization of inference results of COCO data set. (b) Visualization of inference results of MPII data set.

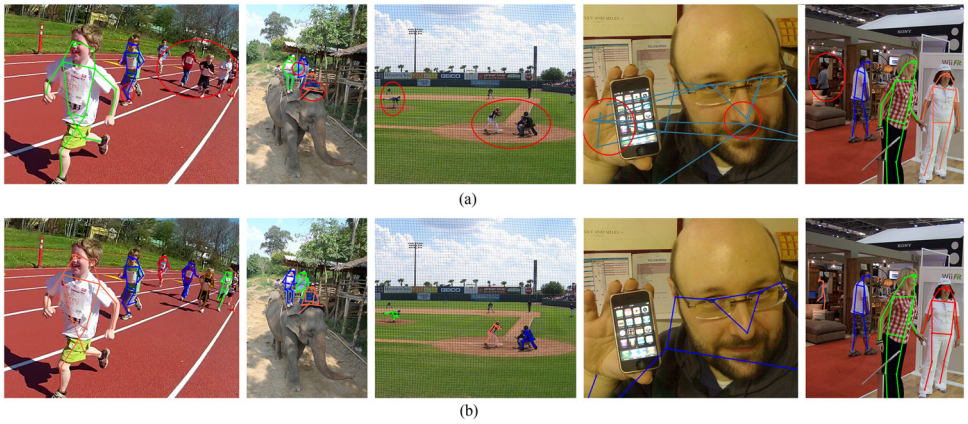

Figure 10: Visualization of multi-scale poses. (a) Visualization of multi-scale pose inference results in HRNet. (b) Visualization of multiscale pose inference results in our CRNet.

sensitivity problem in HRNet. Furthermore, the prediction accuracy of multi-person pose estimation is further improved by the refined network proposed in this work.

5 Conclusion

Aiming at the problem of scale sensitivity and quantization error in bottom-up multi-person pose estimation tasks, we proposed a context feature and refined network for multi-person pose estimation based on HRNet(CRNet). We use multi-scale feature pyramid and context feature to solve multi-scale variation challenges. We extract multi-scale features and fuse them with the attentional feature fusion method to obtain context feature, which can effectively enhance the scale invariance of the network. In addition, we propose a simple but efficient refined network to solve the quantization error problem and CRNet is trained by multi-resolution supervision. The average precision of CRNet on COCO and MPII multi-person test-dev sets was 72.1 and 80.2%, respectively, which outperforms most bottom-up state-of-the-art methods.

Conflict of interest: The authors state no conflict of interest.

References

[1] Chen Y, Tian Y, He M. Monocular human pose estimation: A survey of deep learning-based methods. Computer Vis Image Underst. 2020;192:1–20.

[2] Newell A, Yang K, Deng J. Stacked hourglass networks for human pose estimation. Proceedings of the European Conference on Computer Vision. Amsterdam, Netherlands, Berlin: Springer; 2016, October 8–16. p. 483–99.

[3] Chu X, Yang W, Ouyang W, Ma C, Yuille AL, Wang X. Multi-context attention for human pose estimation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, Piscataway, USA: IEEE; 2017, July 21–26. p. 1831–40.

[4] Nie X, Feng J, Xing J, Xiao S, Yan S. Hierarchical contextual refinement networks for human pose estimation. IEEE Trans Image Process. 2018;28(2):924–36.

[5] Wang Z, Liu G, Tian G. A parameter efficient human pose estimation method based on densely connected convolutional module. IEEE Access. 2018;6:58056–63.

[6] Zhang F, Zhu X, Dai H, Ye M, Zhu C. Distribution-aware coordinate representation for human pose estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, Piscataway, USA: IEEE; 2020, June 16–20. p. 7093–102.

[7] Lin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D, et al. Microsoft coco: Common objects in context. Proceedings of the European Conference on Computer Vision. Zurich, Switzerland, Berlin: Springer; 2014, September 5–12. p. 740–55.

[8] Andriluka M, Pishchulin L, Gehler P, Schiele B. 2d human pose estimation: New benchmark and state of the art analysis. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Columbus, Piscataway, USA: IEEE; 2014, June 23–28. p. 3686–93.

[9] Papandreou G, Zhu T, Kanazawa N, Toshev A, Tompson J, Bregler C, et al. Towards accurate multi-person pose estimation in the wild. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, Piscataway, USA: IEEE; 2017, July 21–26. p. 4903–11.

[10] Chen Y, Wang Z, Peng Y, Zhang Z, Yu G, Sun J. Cascaded pyramid network for multi-person pose estimation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, Piscataway, USA: IEEE; 2018, June 19–23. p. 7103–12.

[11] Xiao B, Wu H, Wei Y. Simple baselines for human pose estimation and tracking. Proceedings of the European Conference on Computer Vision. Munich, Berlin, Germany: Springer; 2018, September 8–14. p. 466–81.

[12] Sun K, Xiao B, Liu D, Wang J. Deep high-resolution representation learning for human pose estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, Piscataway, USA: IEEE; 2019, June 15–21. p. 5693–5703.

[13] Pishchulin L, Insafutdinov E, Tang S, Andres B, Andriluka M, Gehler PV, et al. Deepcut: Joint subset partition and labeling for multi person pose estimation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, Piscataway, USA: IEEE; 2016, June 26–July 1. p. 4929–37.

[14] Cao Z, Simon T, Wei SE, Sheikh Y. Realtime multi-person 2d pose estimation using part affinity fields. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, Piscataway, USA: IEEE; 2017, July 21–26. p. 7291–9.

[15] Newell A, Huang Z, Deng J. Associative embedding: End-to-end learning for joint detection and grouping. Adv Neural Inf Process Syst. 2017;30:2277–87.

[16] Cheng B, Xiao B, Wang J, Shi H, Huang TS, Zhang L. Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, Piscataway, USA: IEEE; 2020, June 16–20. p. 5386–95.

[17] Li J. Research on bottom-up approaches for multi-person pose estimation. PhD thesis. Hefei: University of Science and Technology of China; 2021.

[18] Su K, Yu D, Xu Z, Geng X, Wang C. Multi-person pose estimation with enhanced channel-wise and spatial information. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, Piscataway, USA: IEEE; 2019, June 15–21. p. 5674–82.

[19] Dai Y, Gieseke F, Oehmcke S, Wu Y, Barnard K. Attentional feature fusion. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. (virtual), Piscataway: IEEE; 2021, January 5–9. p. 3560–9.

[20] Kingma DP, Ba J. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980; 2014.

[21] Tieleman T, Hinton G. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. COURSERA: Neural Netw Mach Learn. 2012;4(2):26–31.

[22] Papandreou G, Zhu T, Chen LC, Gidaris S, Tompson J, Murphy K. Personlab: Person pose estimation and instance segmentation with a bottom-up, part-based, geometric embedding model. Proceedings of the European Conference on Computer Vision. Munich, Berlin, Germany: Springer; 2018, September 8–14. p. 269–86.

[23] Kreiss S, Bertoni L, Alahi A. Pifpaf: Composite fields for human pose estimation. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, Piscataway, USA: IEEE; 2019, June 15–21. p. 11977–86.

[24] Nie X, Feng J, Zhang J, Yan S. Single-stage multi-person pose machines. Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul, Korea, Piscataway: IEEE; 2019, October 27–November 2. p. 6951–60.
[25] Insafutdinov E, Andriluka M, Pishchulin L, Tang S, Levinkov E, Andres B, et al. Arttrack: Articulated multi-person tracking in the wild. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, Piscataway, USA: IEEE; 2017, July 21–26. p. 6457–65.
[26] Duan P, Wang T, Cui M, Sang H, Sun Q. Multi-person pose estimation based on a deep convolutional neural network. J Vis Commun Image Representation. 2019;62:245–52.
[27] Fieraru M, Khoreva A, Pishchulin L, Schiele B. Learning to refine human pose estimation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition workshops. Salt Lake City, Piscataway, USA: IEEE; 2018, June 19–23. p. 205–14.