CE-HigherHRNet: Enhancing Channel Information for Small Persons Bottom-Up Human Pose Estimation

M. Y. Li, J. Zhao

Abstract—**Accurately detecting the keypoints of small persons in an image using bottom-up multi-person pose estimation algorithms is exceptionally difficult owing to scale variation challenges. HigherHRNet initially solved the challenge of multi-player scale change pose estimation. However, because it uses repeated cross-scale fusion, owing to inherent defects in channel reduction, semantic information is lost. Furthermore, the aliasing effects produced by the miscellaneous feature maps formed after cross-scale fusion have a significant impact on the detection accuracy of small persons. In this paper, we propose a novel bottom-up human pose estimation algorithm based on HigherHRNet, called Channel-Enhanced HigherHRNet (CE-HigherHRNet). CE-HigherHRNet comprises three main components: a multi-scale sub_pixel skip fusion module, a lightweight attention mechanism (with channel attention enhanced and spatial attention modules), and a high-resolution feature pyramid with an added Dupsampling module. The lightweight attention mechanism optimizes the feature map after each fusion. Deconvolution is replaced with Dupsampling, which strengthens the network's scale awareness and makes it more sensitive to robust scale changes. The average precision (AP) of CE-HigherHRNet on the COCO test-dev dataset was 71.9% (an improvement of 1.4% compared with HigherHRNet). Furthermore, the average detection accuracy of small persons was 68.1% AP (an improvement of 1.5% AP). These results verify that the proposed CE-HigherHRNet is more robust in processing scale changes and has a stronger ability to handle crowded environments. Thus, it is more accurate in positioning small persons in images and human bodies in crowded environments.**

Index Terms—**Channel enhancement, Deep learning, HigherHRNet, Pose estimation.**

I. Introduction

In the area of computer vision, human pose estimation [1] has been identified as a research and application hotspot. In human pose estimation, the objective is to estimate and detect human joint points in images or videos and output relevant information on human limbs, such as the position of each joint point, their corresponding category information, the connection relationship between the body parts, and even the outline of the human body. These data are then used by machines to automatically determine the position of persons in a scene, and to understand the behavior of the human body.

Currently, human pose estimation is predominantly used in the following areas: behavior recognition, human–computer interaction, intelligent security, motion capture, and training robots. Many factors affect the detection of human joints, such as occlusion, clothing, the person's environment, the angle of the camera, and the distance between the camera and the person. Traditional methods use multi-angle depth cameras and radars to estimate and track the posture of the human body. However, with the rapid development of deep learning and convolutional neural networks (CNNs) in recent years, it has become feasible to reliably infer the keypoints of the human body from images without the need for additional professional acquisition equipment. Consequently, pose estimation for the human body utilizing CNNs is an active research area and significant progress has been made [2],[3],[4].

HigherHRNet [5] uses HRNet as its backbone, and makes repeated multi-scale fusions by repeating the paralleling interaction of subnetworks. In the multi-scale fusion, a 1*1 convolutional layer is needed to reduce the number of channels in high-level feature maps, which causes channel information loss, and consequent semantic information loss. Because cross-scale feature maps have semantic differences, straight fusing after linear interpolation results in aliasing effects when performing cross-scale fusion [6],[7], which, in turn, complicates the positioning and recognition tasks [8].

In this study, we focused on solving the problems of channel reduction in high-level feature maps and aliasing effects when performing cross-scale fusion, and the question of how to generate higher-resolution prediction heatmaps to restore the keypoints of small persons. To this end, we propose an efficient HigherHRNet variant called Channel-Enhanced HigherHRNet (CE-HigherHRNet).

Predicting the keypoints of small persons presents two major issues. The first issue is scale variations, which necessitates improving the keypoint prediction accuracy for small persons without reducing that of persons with large statures. The second issue is the generation of a high-resolution, high-quality heatmap to precisely locate the small

Manuscript received August 31, 2021; revised February 05, 2022. This work was supported by the Natural Science Foundation of Liaoning Province under Grant 20180551048.

M. Y. Li is a Master's Student of School of Computer Science and Software Engineering, University of Science and Technology Liaoning, Anshan 114051, China (e-mail: 381593836@qq.com).

J. Zhao is a Professor of School of Computer Science and Software Engineering, University of Science and Technology, Liaoning, Anshan 114051, China (corresponding author, phone: 86-13998086167; e-mail: zhaoji@ustl.edu.cn).

person's keypoints. The bottom-up approach [9], [10] focuses on how to group the keypoints of the human body and only uses feature maps with a single resolution to forecast the heatmap of keypoints. The models based thereon do not have the ability to sense scale.

By increasing the resolution of the input image, PersonLab [10] produces heatmaps with a higher resolution. The keypoint detection accuracy for small persons continues to improve as the resolution increases. However, when the input image resolution is overly high, the keypoint detection accuracy of a large-scale human body falls significantly, whereas that of a small person does not improve further. Moreover, the above operations are seriously restricted by hardware conditions, and exponential increases in the required video memory may occur. HigherHRNet can generate higher-resolution feature maps by deconvolution. However, owing to the inherent problems of deconvolution [11], [12], HigherHRNet [5] cannot generate a character map with the same dimensions as the input image. Rather, it requires choosing the policy of introducing multi-resolution heatmap aggregation in the inference process.

We assessed our proposed method on the COCO dataset; the details are presented in Section IV. Specifically, the improved HigherHRNet achieved an average precision (AP) of 71.9% on the COCO test-dev dataset without any post-processing. Compared with HigherHRNet, it improves the AP by 1.4%. In addition, we observed that we can obtain gains for both small persons and large persons. For small persons, our method improves the AP by 1.5% compared to HigherHRNet, and the detection accuracy of large-scale human bodies also increases by 1.2% AP. These results validate our theory. Meanwhile, we achieved 66.6% AP on the CrowdPose dataset. This result shows that the bottom-up approach is more advantageous in crowded environments.

The contributions of this study are primarily in three areas:

- We propose a new channel-enhanced backbone that solves the problem of channel information loss caused by multi-scale fusion; it incurs a small computational burden only.
- We propose an improved lightweight attention mechanism based on the convolutional block attention module (CBAM) [13], apply the channel attention mechanism to the new feature maps obtained after each fusion, and optimize all the feature maps produced to eliminate the influence of aliasing during multi-scale fusion.
- We introduce an upsampling method that incurs less computation than deconvolution to generate a predictive heatmap that is consistent with the input image resolution. The method recovers keypoints of small persons lost in the low-resolution heatmaps to strengthen the scale perception ability of the network.

II. Related Work

A. Human Pose Estimation

Human pose estimation methods can be classified into two main approaches: top-down methods [4], [14], [15], [16], [17], [18] and bottom-up methods [9], [10]. The first step of the top-down approach is to gather the full human body instance frame by detecting a human target [19], and then extract the human joints based on this frame. The first step of the bottom-up approach [20] is to directly extract the full image of the human body joints and distribute the full image keypoints to the corresponding human instances in the Heuristic Post-Processing method. At present, deep CNNs provide the mainstream keypoints detection solutions [21], which are mainly divided into two types: (1) direct return to the locations of the keypoints [22] and (2) estimating the keypoint heatmaps [23] and then selecting the position with the highest heat value in the heatmap as the keypoints. Human pose estimation is the prework of behavior recognition [24], and it is typically used together with object detection methods [25] to detect human behavior [26]. It is widely used in the field of gesture recognition [27], [28].

Since 2012, AlexNet, a deep learning method has been applied to image classification, image detection, and image segmentation problems. In 2014 [29], for the first time, a CNN was successfully used to solve the problem of single-person pose estimation. Because of the background time required, the network structure was relatively simple, and this facilitated the use of some traditional skeleton ideas. The proliferation of deep learning prior to 2016 also launched human pose estimation into its prime time. Starting from convolutional pose machines (CPM) [30], CNN can model feature representation as well as spatial position information for keypoints. The position of the corresponding keypoints can be determined by locating the maximum response position through the channels on the predicted heatmap. The entire network of CPM has multiple stages, and each stage is designed with a small network to extract features. Then, a supervision signal is added at the end of each stage.

In July of 2016, Stacked Hourglass Networks [4] first put forward a network structure using multi-scale features to identify posture. The previous pose estimation network structure made predictions solely by using the last layer of the convolution feature, which can cause loss of information. In 2016, the very important dataset COCO appeared. OpenPose [9], proposed by the Carnegie Mellon University (CMU) team, took first place in the COCO competition that year. It first determined the position of each joint in the images by utilizing CPM as its component, and then used part affinity field (PAF) to construct the keypoints of the human body.

In 2017, Chen proposed the cascaded pyramid network (CPN) [17], which uses a network to detect coarse results (GlobalNet), and then refines (RefineNet) these results. In 2018, Li proposed multi-stage pose estimation network （MSPN）[31], in which a skip connection was added to two adjacent stages for better transmission of information. In 2018, Xiao proposed SimpleBaseline [14], which is a simple but very effective baseline network that uses deconvolution to expand the resolution of the feature map, and replaces the previous commonly used upsampling method with linear interpolation. It has made scholars seriously consider how to obtain a high-resolution prediction heatmap.

Papandreou et al. at PersonLab [10] used expanded ResNet

[32] and direct learning to categorize the keypoints of each pair of 2D grouping keys bias field. In 2019, Wang Jingdong et al. proposed HRNet [15], [16], which emphasized the importance of spatial resolution. The network is divided into numerous stages, but the most detailed spatial information is always conserved because of parallel connections. PifPaf [33] positioned keypoints locations through part intensity field (PIF), and then used PAF to associate the human body keypoints, thereby forming a complete human posture. In 2020, Wang et al. proposed HigherHRNet [5], a scale-aware high-resolution network with multi-resolution supervision for training and multi-resolution heatmap aggregation for inference. It overcomes the issue of scale change in bottom-up multi-person pose estimation, and more precisely detects the keypoints of small persons in the image.

B. Network Architecture

Fig.1. depicts the structure of the original HigherHRNet [5]. With HRNet as the backbone, it commences with a high-resolution subnetwork, progressing the subnets from high to low resolution, and then connecting the multi-resolution subnetworks in parallel. Throughout the process, information is exchanged several times between parallel multi-resolution subnets, resulting in repetitive multi-scale fusion. Then, to generate a high-resolution heatmap, it creates a high-resolution feature pyramid, which starts at 1/4 resolution. Such a new feature map has the highest resolution in the backbone, and it is generated by a deconvolution that is as large as the backbone output feature map. The targets of various resolutions are then assigned to the corresponding feature pyramid level using the multi-resolution supervision strategy. At the same time, the multi-resolution heatmap aggregation strategy is introduced in the inference process to generate a high-resolution heatmap with scale awareness.

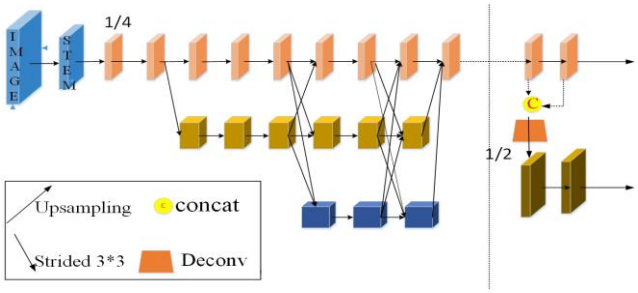


Fig. 1. The original HigherHRNet [5] network structure.

C. High-quality High-resolution Feature Maps

The main method used conventionally to generate a high-resolution feature map is the encoder–decoder method [3], [17]. The encoder captures contextual information and the decoder restores the high-resolution representation through a series of linear interpolation upsampling operations, and skip connection of encoders with the same resolution. Deconvolution can produce higher-quality feature maps for predicting heatmaps, as shown by SimpleBaseline [14]. HRNet [15], [16] has a plurality of branches composed of different resolutions, with a resolution branch for capturing contextual information in terms of enhanced features, and for promoting contextual information classification and positioning performance. High-resolution branches are used to capture spatial information and, through repeated multi-scale fusion between branches, high- resolution feature maps with rich semantics can be used to predict heatmaps.

III. Proposed Method

In this section, we present our proposed CE-HigherHRNet, which reduces channel information loss, optimizes the feature maps generated after fusion, and generates higher-quality high-resolution feature maps. The structure is primarily made up of three elements: multi-scale sub_pixel skip fusion, lightweight attention mechanism (which includes a channel attention enhanced module and a spatial attention module), and a high-resolution feature pyramid with added Dupsampling module. The details are presented below.

A. Network Architecture

Fig. 2. shows the overall network structure of CE-HigherHRNet. It is composed of an improved backbone and improved high-resolution feature pyramid. According to the settings of HRNet, the overall backbone is divided into four stages of high-resolution parallel subnets, which we denote as $\{C_1, C_2, C_3, C_4\}$, and their resolution is the input of the image resolution {1/4/, 1/8/, 1/16/, 1/32}. We do not use the general method of feature fusion in the multi-scale fusion.

To decrease the number of channels in the feature map, a 1*1conv is applied first, followed by upsampling for fusion. Because the core purpose of our change is to make the most of all the rich semantic information of the C_{44} and C_{33}, the above method will cause serious aliasing effects owing to the loss of C_{44} and C_{33} channel information during the repeated multi-scale fusion. Therefore, there is a need to improve the backbone. $\{N_1, N_2, N_3\}$ form a high-resolution feature pyramid, in which N_1 is the backbone's final feature map, and N_2, N_3 are the higher-resolution feature maps generated by Dupsampling [12]. We use three-scale feature maps to obtain the final prediction heatmap, and also use the attention mechanism to make the model express the characteristics of the keypoints of small persons. Then, we average the heatmaps of all the scales to make the final prediction.

B. Multi-scale Sub_Pixel Skip Fusion (MSSF)

The specific method of multi-scale fusion of the fourth stage C_4 in HigherHRNet is shown in Fig.3. (a). First, $\{F_{44}, F_{43}, F_{42}\}$ is generated by reducing the channel dimensions of the feature map at the $\{C_{44}, C_{43}, C_{42}\}$ layer with low-resolution by 1*1 conv, so that the number of channels of $\{F_{44}, F_{43}, F_{42}\}$ and C_{41} are consistent. The feature maps $\{F_{44}, F_{43}, F_{42}\}$ are upsampled so that all of the feature maps have the same resolutions. Then, fusion is performed to add the corresponding elements. The 1*1 convolutional layer is utilized to decrease the high-level feature map's channel count. This causes loss of channel information and also loss of semantic information. When performing cross-scale fusion, because cross-scale feature maps possess semantic

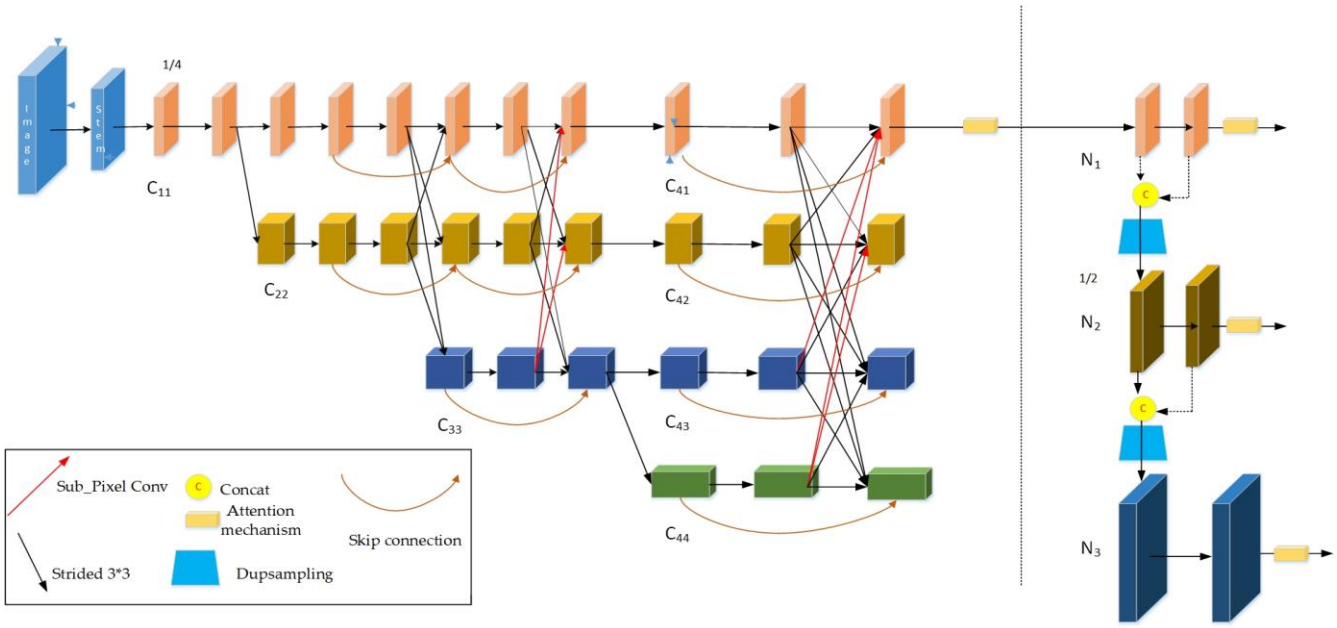


Fig. 2. A schematic of the overall network structure of CE-HigherHRNet. The overall network is composed of an improved backbone and an improved high-resolution feature pyramid.

differences, direct fusion after linear interpolation results in aliasing effects.

We observed that the rich channel information of C_{44} is not fully utilized. Based on this discovery, we developed rich channel information with low-resolution feature maps to enhance the final capability of the model. Specifically, we designed a novel fusion method aroused by sub_pixel conv [11], which is an upsampling approach, also known as Pixel Shuffle. Fig. 3. (b) shows its operation; it converts feature maps with the shape of $H_1*W_1*C_1 \bullet r^2$ into a feature map with the shape of $rH_2*rW_2*C_2$. The formula can be defined mathematically as follows:

$$PS(F) = F[x/r],[y/r],\ C \bullet r \bullet mod(y,r) + C \bullet \mathrm{mod}(x,r) + c \quad (1)$$

Where F is the input feature map, r is the upscaling factor, and $PS(F)_{x,y,c}$ represents the output feature pixel point on the coordinates (x, y, c).

We introduce MSSF to directly upsample low-resolution feature maps without channel reduction. When upsampling by using sub_pixel conv, the generated high-resolution feature maps may not be reliable. Therefore, we introduce skip connections, and fuse the output from the previous layer with the feature map again to generate high-quality feature maps, as shown in Fig. 3. (c). In this manner, the feature aggregation makes the spatial position information in the high-resolution feature maps obtained by the fusion more accurate, and the generated low-resolution feature maps have complete semantic information and stronger characterization capabilities.

Mutual information exchange can ensure that the network considers both spatial location information and feature abstract information, and can also enhance the ability of network feature information dissemination while reducing the difficulty of training. The MSSF only uses rich information channels of feature maps {C_{44} , C_{33} } with low resolution, rather than replacing all the upsampling operations with sub_pixel conv, because the fusion of

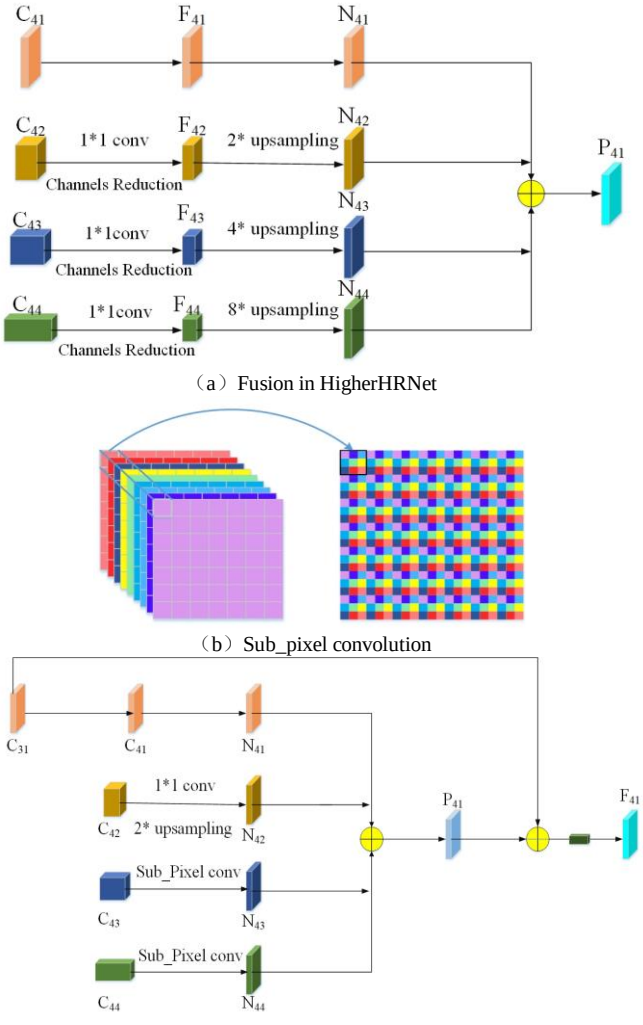


Fig. 3. (a) Shows the fourth stage C_4 fusion method in HigherHRNet. (b) Shows the working principle of Sub_Pixel, (c) Introduces our Multi-scale.

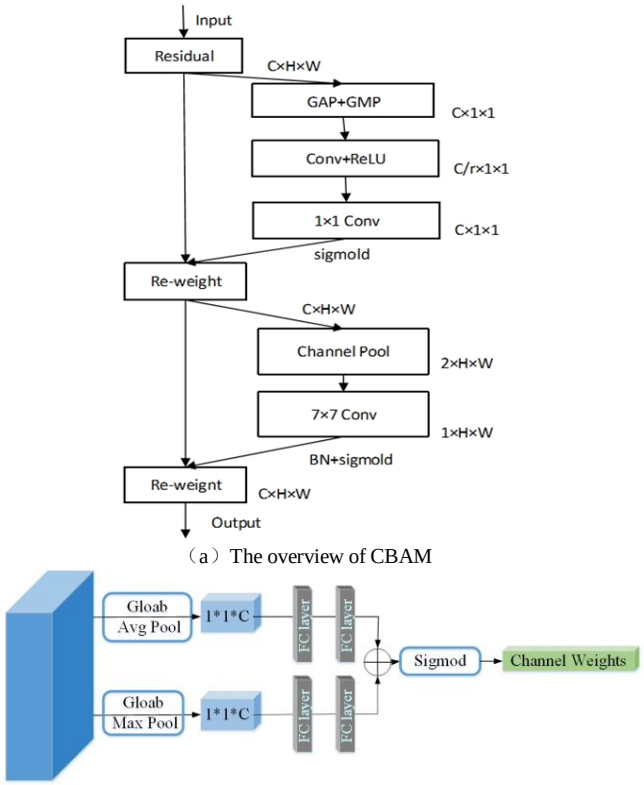
adjacent layers can only generate slight aliasing effects, which can be solved by adding a residual block [6]. Therefore, we only need to process the feature map of multi-scale cross fusion and incorporate it into F_{41}. This is depicted as follows:

$$F_i = \begin{cases} \varphi(C_i) + ps\left(\bar{\varphi}(C_{i+1})\right) + ps\left(\bar{\varphi}(C_{i+2})\right), i = 2 \\ \varphi(C_i) + \varphi(C_i) + ps\left(\bar{\varphi}(C_{i+2})\right), \quad i = 1 \end{cases} \tag{2}$$

Where i represents the index of the layer in the entire backbone, and $\bar{\varphi}$ represents the channel conversion. r=2 means that the sub_pixel conversion factor is two, which doubles the generated spatial resolution. We propose a novel lightweight channel with the attention enhanced module to optimize the final generated feature maps to eliminate the negative consequences of aliasing effects.

C. Lightweight Convolutional Block Attention Module

Cross-scale feature maps have differences in semantic information due to the loss of channel information, and the mixed features can cause aliasing effects. Therefore, we designed a lightweight channel attention mechanism based on the channel attention mechanism of CBAM, and applied it to the final feature maps generated by MSSF, because the third-stage high-resolution subnet and the fourth-stage high-resolution subnet in the backbone require three-scale and four-scale feature fusion, respectively. At the same time, skip connection is also made and the aliasing effect becomes more obvious. The simplest way to eliminate the impact of the aliasing effect is to introduce the channel attention mechanism after the feature fusion. Therefore, in this study, an improved channel attention mechanism based on the CBAM [13] was developed so that the final generated feature map has no aliasing effect. The overall structure is shown in Fig. 4. (a). This study also introduces a focus mechanism mixed by the channel attention and spatial attention mechanisms into a high-resolution pyramid. The passage has a mechanism for focusing the mixing mechanism. Thus, the final feature maps are more sensitive to the position and identification of keypoints that are not easily tested in small persons.



（a）The overview of CBAM

（b）The Channel Attention Enhanced Module (CAEM).

Fig.4. (a) Overall network structure of the CBAM, (b) Illustration of Channel Attention Enhanced Module (CAEM).

1) Channel Attention Enhanced Module

We also propose a lightweight channel attention mechanism, called the channel attention enhanced module (CAEM), based on CBAM. Fig. 4. (b) depicts the structure. First, global average pooling and global max pooling of the input feature map are utilized to aggregate the feature map's spatial information to create two different spatial context descriptors, F_{avg}^{c} and F_{max}^{c}, which respectively portray average pooling features and max pooling features. Each pixel on the feature map receives feedback from the average pooling. When calculating the gradient of the back propagation, the maximum pooling only has gradient feedback for the maximum response in the feature maps. Then these two features are respectively sent to two parallel fully-connected layers. Finally, the fully-connected layer's output feature vector is combined through the corresponding element-wise operation, and the sigmoid function is used for the activation operation. The combined vector is mapped to the interval (0–1) to obtain the final channel attention feature map. The method can be expressed as follows:

$$CA(F) = \sigma\left(FC_1\left(AvgPool(F)\right) + FC_2\left(MaxPool(F)\right)\right) \tag{3}$$

2) Spatial Attention Module

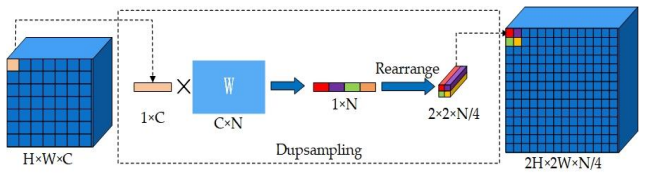
In this study, the spatial attention mechanism of CBAM [13] is used. CBAM utilizes the spatial relationship of features to obtain a spatial attention map. In contrast to the channel attention map, the spatial attention map concentrates on the position information of the keypoints of the human body [13]. It is a supplement to the channel attention map. This study utilizes the spatial attention mechanism to focus on the human body's keypoints that are more difficult to detect for small persons in an image. To calculate the spatial attention, first the feature map output by the channel attention mechanism is utilized as the input feature map of the module to make a channel-based global average pooling and global maximum pooling to aggregate the channel information of one feature map, and then generate two 2D pictures: $F_{avg}^{s} \in R_1 \times H \times W$ and $F_{max}^{s} \in R_1 \times H \times W$. Based on the channel, the resulting two 2D pictures undergo a concat operation. Then, after 7*7 conv, the dimensionality is reduced to one channel, and spatial attention feature maps are generated through batch normalization (BN) and Sigmoid functions. The spatial attention is calculated as follows:

$$\mathrm{SA}(F) = \sigma\left(f^{7\times7}\left(\left[AvgPool(F); MaxPool(F)\right]\right)\right) \tag{4}$$

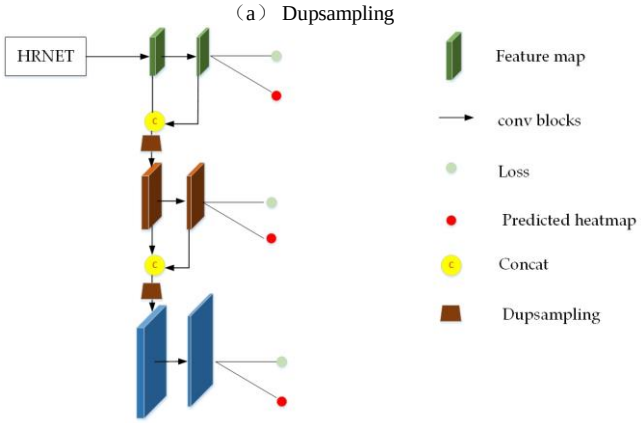
Where σ represents the Sigmoid function, $f^{7\times7}$ represents a convolution operation, and the convolution kernel's size is 7×7.

D. High-Resolution Feature Pyramid.

This section introduces Dupsampling [12] based on the problem of deconvolution. The workflow is shown in Fig.5.



（a） Dupsampling



（b） The overall structure of the high-resolution feature pyramid

Fig.5. (a) The working process of Dupsampling is introduced. (b) Overall structure of the high-resolution feature pyramid.

(a). A high-resolution feature pyramid is designed to predict heatmaps. When deconvolution is upsampling, it is necessary to pad the image (i.e., fill in zeros). In fact, the white zero-filled part of the image is invalid information, and it does not facilitate optimization of the gradient. For the keypoints of a large body in an image, the upsampling operation of filling in zeros makes the feature maps generate a checkerboard pattern, and impact the generated feature maps. Thus, the detection accuracy of the large body declines, and the deconvolution is highly computation-intensive. In considering the above problems, we replaced the deconvolution with Dupsampling to create a higher-resolution feature map for predicting heatmaps. Fig. 5. (b) depicts the Dupsampling network structure. There are feature maps with three resolutions in this article. They are the 128 × 128 feature maps output by the backbone, and feature maps with resolutions of 256 × 256 and 512 × 512 generated by Dupsampling. Because the computational cost of the deconvolution is very high, if the high-resolution feature pyramid in the HigherHRNet generates feature maps with three resolutions, the computational cost will increase significantly. The addition of a second deconvolution module will produce a large precision drop of 0.8% AP for large persons. However, for the Dupsampling, the idea of depth separable convolution is a borrowed idea. First, all the channels of a pixel are reshaped, then rearranged by *W* learned through the network, and finally enlarged by the upscale factor r, so that the amount of calculation is very small compared to that of the deconvolution. Even if the second Dupsampling module is used, the overall number of parameters will be slightly less than that of a deconvolution. Moreover, because the Dupsampling does not have a zero-filling operation, it will not cause a checkerboard effect in the newly generated feature map. When using heatmap aggregation strategies, the experimental results in Section IV show that after generating a feature map with the same input image size by using two Dupsampling modules, there is no significant drop in the detection accuracy of the big human body, rather it improves slightly. The precision of small targets ire also improved. Hence, it demonstrates that CE-HigherHRNet is a pose estimator with recognizable scale.

IV. Experiments and Results

A. Dataset and Evaluation Metrics

The COCO dataset [20] contains more than 200,000 images and 250,000 persons with seventeen keypoints labels. It is compartmentalized into train/val/test-dev sets, which contain 57,000, 5,000, and 20,000 images, respectively. This article presents our ideas on the CrowdPose [35] dataset. The bottom-up method is well known to exhibit better performance for crowded people. The CrowdPose [35] dataset incorporates 20,000 images. The training, validation, and testing are divided in the ratio of 5:1:4. In addition, crowded scenes have more datasets than COCO datasets. Keypoints detection needs to simultaneously detect human targets and locate the coordinates of human keypoints, which is a task in which detection and keypoints estimation are performed at the same time. Microsoft designed a novel evaluation metric, the object keypoint similarity (OKS), which is expressed as follows:

$$OKS = \frac{\sum_i exp\left(-d_i^2 / 2s^2k_i^2\right)\delta\left(v_i > 0\right)}{\sum_i \delta\left(v_i > 0\right)} \tag{5}$$

Where d_i is the Euclidean distances between the corresponding ground truth of each human keypoint and the detected keypoints, v_i is the ground truth's visibility in the flag (predicted by the detector; v_i is not used), *s* is the factor of the object, k_i represents the normalized factor of the *i*-th human body keypoints. The performance metrics follow the standard COCO-style mean AP (mAP) metrics under different intersection over union (IoU) thresholds, ranging from 0.5 to 0.95 with intervals of 0.05 [5].

B. Training Details

We followed the same training procedure as the original HigherHRNet [5], and used the same dataset for the data enhancement processing. Images with input size 640 × 640 were randomly rotated ([-30◦, 30◦]), scaled ([0.75, 1.5]), translated ([-40, 40], and flipped. In the experiments, all the models actualized using the PyTorch framework and the Adam optimizer [34] was used for the training. The batch size was 16, number of training epochs 300, and initial learning rate 1e-3. In the 210th and 260th training epochs, the learning rate decreased to 1e-4 and 1e-5, respectively.

C. Results on COCO2017

Table I summarizes the bottom-up method results on the COCO 2017 test-dev dataset, and introduces two test methods, the single-scale test and multi-scale test. From Table I, we designed two image preprocessing methods consistent with HigherHRNet to compare them with the other bottom-up human pose estimation methods [5]. One uses an

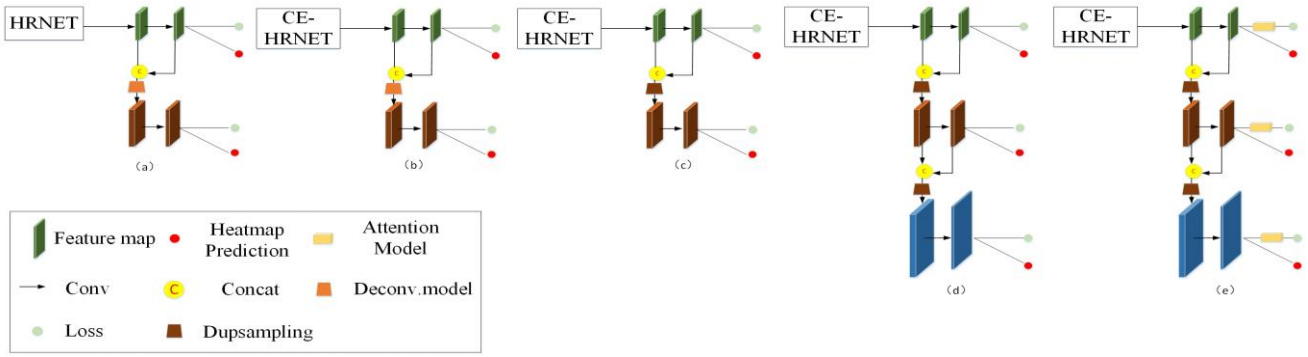


Fig. 6. (a) The original HigherHRNet's high-resolution feature pyramid network. (b) Replace HRNET with the improved CE-HRNET. (c) Use dupsampling module to replace the deconvolutional module to generate feature maps with input image resolution 1/2. (d) On the basis of (c), generate a feature map that is consistent with the resolution of the input image. (e) Add an improved CBAM attention mechanism at the end to form the final network structure.

TABLE I
COMPARISONS OF BOTTOM-UP METHODS.

Method	Backbone	Input size	#Params	GFLOPs	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}
w/ single-scale test									
OpenPose	VGG-19	368	-	-	61.8	84.9	67.5	57.1	68.2
Hourglass	Hourglass	512	277.8M	206.9	56.6	81.8	61.8	49.8	67.0
PersonLab	ResNet-152	1401	68.7M	405.5	66.5	88.0	72.6	62.4	72.3
PifPaf	ResNet-152	-	-	-	66.7	-	-	62.4	72.9
Bottom-up HRNet	HRNet-W32	512	28.5M	38.9	64.1	86.3	70.4	57.4	73.9
HigherHRNet	HRNet-W32	512	28.6M	47.9	66.4	87.5	72.8	61.2	74.2
HigherHRNet	HRNet-W48	640	63.8M	68.4	69.8	88.2	75.1	64.4	74.2
Ours	CE-HRNet-W32	512	31.9M,	52.8	67.3	87.5	73.8	62.5	75.3
Ours	CE-HRNet-W48	640	67.3M	160.7	71.0	88.4	76.9	66.3	75.9
w/ multi-scale test									
Hourglass	Hourglass	512	277.8M	206.9	65.5	86.8	72.3	60.6	72.6
PersonLab	ResNet-152	1401	68.7M	405.5	68.7	89.0	75.4	64.1	75.5
SPM	Hourglass	384	-	-	66.9	85.5	72.9	62.6	73.1
HigherHRNet	HRNet-W48	640	63.8M	154.3	70.5	89.3	77.2	66.6	75.8
Ours	CE-HRNet-W48	640	67.3M	160.7	71.9	89.2	78.8	68.1	77.0

input image with a resolution of 512 × 512 and 32 channels in the first layer of the backbone, the other uses an input image with a resolution of 640 × 640 and 48 channels in the first layer of the backbone.

It can be seen that in the single-scale test, although our method used the attention mechanism many times, and more residual blocks than HigherHRNet were utilized to refine the feature maps' features, the parameters of our method (+5.49%) and computational complexity (+4.14%) increased slightly, but our results compared to those of HigherHRNet improved by 0.9% AP. In the multi-scale test, our results, compared to the HigherHRNet, improved by 1.4% AP. The detection accuracy for small persons improved significantly, reaching 68.1% AP, whereas the detection accuracy for large persons also increased by 1.2% AP. This proves that our method is effective in human pose estimation, especially for small humans. The detailed results and analysis of the ablation experiments are given in the next section.

Table II summarizes the top-down method results of the COCO 2017 test-dev dataset. The performance of CE-HigherHRNet is similar to some top-down methods. We used the same testing strategy that is used in bottom-up methods. The results obtained are very close to the results of the top-down method, which has been known to have a very slow inference speed. However, owing to the existence of the bounding boxes of the target detection algorithm, and the less crowded and occluded environment in the COCO dataset, its detection accuracy is much better than that of bottom-up methods.

TABLE II
COMPARISONS OF TOP-DOWN METHODS.

Method	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}
Mask-RCNN	63.1	87.3	68.7	57.8	71.4
G-RMI	68.5	87.1	75.5	65.8	73.3
SimpleBaseline	73.7	91.9	81.8	70.3	80.0
CPN	72.1	91.4	80.0	68.7	77.2
AlphaPose	72.3	89.2	79.1	68.0	78.6
CFN	72.6	86.1	69.7	78.3	64.1
CE-HigherHRNet-W48	71.9	89.2	78.8	68.1	77.0

D. Ablation Experiments

To better comprehend the increased gain of our proposed method, we conducted ablation experiments on the COCO 2017 test-dev dataset for each individual component. All of

the comparative experiments used pretreatment approaches to train the input image with a resolution of 640 × 640 and 48 channels in the backbone's first layer. Fig.6 illustrates all the compositions of our experiment, and the results are shown in Table III, Table IV, and Table V.

Effect of Multi-scale Sub_Pixel Skip Fusion. We conducted ablation experiments for the MSSF. The results are shown in Table III and Table IV, which include sub_pixel conv, skip connection, and three components of the CAEM. When performing a separate sub_pixel conv (Fig. 6. (b)), the final precision decreased by 1.3% AP compared to that used in HigherHRNet as the baseline Fig.6. (a)). The precision of a small person decreased by 1.5% AP, and that of a large human body decreased by 0.7% AP. Although the problem of information loss of the channel has been solved, the aliasing effect is very serious because it causes a decrease in the detection accuracy. The skip connection and CAEM need to be added in feature fusion to eliminate the aliasing effects. The results prove that our proposed MSSF can solve the problem of aliasing effects.

TABLE III
ABLATION STUDY OF MULTI-SCALE SUB_PIXEL SKIP FUSION.

Method	w/ sub_pixel conv	w/ skip connection	w/ CAEM	AP	AP^M	AP^L
HigherHRNet				69.8	64.4	74.2
Ours	√			68.5	62.9	73.5
Ours	√	√		69.3	64.0	74.2
Ours	√	√	√	70.3	65.1	74.9

TABLE IV
ABLATION STUDY OF CE-HIGHERHRNET'S COMPONENTS.

Method	w/ MSSF	w/ LCBAM	w/ DUpsampling	AP	AP^M	AP^L
HigherHRNet				69.8	64.4	74.2
Ours	√			70.3	65.1	74.9
Ours	√	√		70.8	65.3	75.4
Ours	√	√	√	71.0	66.3	75.9

Effect of high-resolution feature pyramid. The high-resolution feature pyramid comprises two components: lightweight attention module and Dupsampling module. After the Dupsampling, we added four basic blocks [32] to the two generated resolution feature maps to refine the features of the feature map. We conducted two sets of ablation experiments.

TABLE V
ABLATION STUDY OF DIFFERENT HEATMAP SIZE.

Method	Heatmap resolution	AP	AP^M	AP^L
HigherHRNet	256	69.8	65.4	76.4
Ours	256	69.9	65.6	76.4
Ours	512	70.1	66.0	76.7

The results in Table V show that to verify the effectiveness of higher-resolution predictive heatmaps, we had to set CE-HigherHRNet to the same setting as HigherHRNet, without adding the lightweight convolutional block attention module (LCBAM) at the end (Fig. 6.(c)).

It can be seen that when the two generated prediction heatmaps of 160*160 and 320*320 aggregate with multi-resolution, the precision of small persons was slightly improved (0.2%) whereas that of large persons did not change. However, when we generated the input image of the same size feature map (Fig. 6. (d)), the precision of small persons increased by 0.8% AP, whereas that of large persons improved by 0.7% AP. The results show that 1) prediction with higher resolution is beneficial to detect the keypoints of small persons, and the bottom-up methods require the ability to sense scale. 2) Dupsampling can generate better feature maps than deconvolution, and the calculation is 20% of that of the deconvolution. In further experiments, it can be seen that adding LCBAM (Fig. 6. (e)) behind the predicted heatmaps can make the precision of a small person improve by 1% AP, and that of a large body improve by 0.5% AP. These results prove that our method indeed has scale awareness.

Effect of Input image resolution. An experimental analysis was conducted to determine whether the input image resolution has an impact on the performance of the model. The input image resolution of CE-HigherHRNet was set to 512 × 512, 640 × 640, and 768 × 768, and the number of channels in the first layer of the network was set to 48. All the methods used single-scale test.

TABLE VI
ABLATION STUDY OF CE-HIGHERHRNET WITH DIFFERENT TRAINING IMAGE SIZE.

Train Size	AP	AP^M	AP^L
512	69.2	63.0	75.6
640	71.0	66.3	75.9
768	70.8	66.9	73.6

When performing the test, all the image resolutions were cropped to be consistent with the input image resolution for testing. From Table VI, it can be seen that when the input image resolution was increased to 640 × 640, the AP was significantly increased by 1.9%, of which only 0.3% AP was for the large human body, and more increase resulted for the small persons, with an increase of 3.3% AP. At the same time, it was found from experiments that if the input image resolution is further increased to 768 × 768, the mAP will not continue to increase; the detection accuracy of the small persons will increase slightly, but that of the large human body will drop significantly by 2.3% AP. This also verifies that, as mentioned for HigherHRNet, setting the input image to 640 × 640 works best. Hence, in this study, the training input image resolution was set to 640 × 640.

TABLE VII
COMPARISON OF THE INFERENCE TIME.

Method	Input size	Inference time
HRNet	640×640	689ms
HigherHRNet	640×640	154ms
Ours	640×640	119ms

Table VII shows the results of inference speed comparison. It compares the inference speed of three network models: HRNet, HigherHRNet, and CE-HigherHRNet. For fair comparison, we set 32 channels in the first layer of the

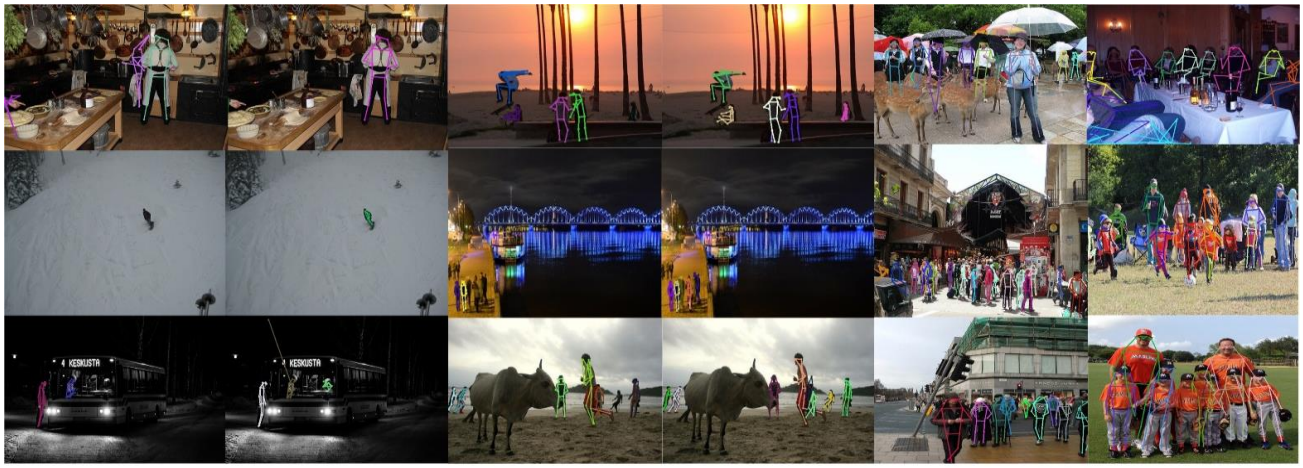


Fig. 7.（a) (c) are the test results of HigherHRNet, (b) (d) are the test results of CE-HigherHRNet, (e) (f) are the test results of crowded people.

backbone, and the input image resolution both were 640 by 640. It can be seen that because HRNet is a two-stage method, the accuracy is slightly higher than that of the proposed model, but the inference speed is very slow. Compared with HigherHRNet, the inference speed of CE-HigherHRNet is improved by 35 ms, and it has a slight improvement in speed and accuracy.

E. Results on CrowdPose

The precision of the top-down methods has always been better than the bottom-up methods. However, for the problem of crowded people, the bottom-up methods are the better choice because the top-down methods need to select the human body instance frame to estimate the human body pose, which is very difficult for crowds. As can be seen in Table VIII, the top-down method [17], [18] that performs favorably on the COCO dataset does not perform well on the CrowdPose dataset.

To verify the robustness of CE-HigherHRNet in a crowded scene, we used the CE-HigherHRNet-W48 on CrowdPose for training and testing, and report the performance on the test set. The evaluation indices all followed the COCO, and were trained and tested by setting the input image resolution to 640 × 640.

TABLE VIII
COMPARISON OF TOP-DOWN AND BOTTOM-UP METHODS ON THE CROWDPOSE TEST DATASET.

Method	AP	AP^{50}	AP^{75}	AP^{E}	AP^{M}	AP^{H}
Top-down methods						
Mask-RCNN	57.2	83.5	60.3	69.4	57.9	45.8
SimpleBaseline	60.8	81.4	65.7	71.4	61.2	51.2
AlphaPose	61.0	81.3	66.0	71.2	61.4	51.1
Top-down with refinement						
SPPE	66.0	84.2	71.5	75.5	66.3	57.4
Bottom-up methods						
HigherHRNet-W48	65.9	86.4	70.6	73.3	66.5	57.9
CE-HigherHRNet-W48	66.6	86.9	72.0	76.1	67.5	58.7

AP^{E} is easy, representing Crowd Index range (0–0.1), AP^{M} is medium, representing Crowd Index range (0.1–0.8), AP^{H} is hard, representing Crowd Index range (0.8–1).

The results are shown in Table VIII. Our CE-HigherHRNet is much better than the top-down pose estimation method. AlphaPose is the best top-down method. Our proposed method is also slightly improved compared to the refined SPPE. However, it can be seen that our main enhancement is in the moderately crowded AP^{M}. We achieved 1% AP improvement for AP^{H} with a congestion factor exceeding 0.8% AP, thus proving that performance under extreme overcrowding situations is still a very difficult problem.

F. Results Comparison

We also compare the qualitative results between HigherHRNet and CE-HigherHRNet in Fig.7. (a). Our experiment was mainly aimed at the keypoints detection of small persons. As shown in the first row of Fig.7.(a), because the aliasing effect makes the model's positioning of the keypoints of the human body inaccurate and this kind of problem can be solved by our method, the third row of Fig.7.(b) shows that our method can detect drivers in a dark environment. It can be seen from Fig.7.(b) and Fig.7.(d) that our method can predict smaller people regardless of the environment they are in. In crowded, dimly lit, and other environments, the proposed model can estimate the human body pose very accurately. However, it can be seen that in the first row of Fig.7.(e), the human foot is occluded and the model misidentified the deer's hoof as a human foot. Therefore, CE-HigherHRNet still has room for improvement in severely occluded environments. In summary, CE-HigherHRNet has stronger discrimination ability, better performance, and is more precise in positioning the keypoints of the human body. All the images were selected from the COCO2017 val dataset. We compared the detection performance with a threshold of 0.5.

V. CONCLUSION

In this paper, we proposed CE-HigherHRNet, a new bottom-up method for human pose estimation based on HigherHRNet. Channel reduction causes loss of channel information in low-resolution feature maps, and the fused

feature maps engender aliasing effects. We proposed a new channel-enhanced backbone to solve this problem. Specifically, we use sub_pixel conv to directly perform channel reduction and upsampling, which solves the problem of channel loss. We employ an enhanced attention mechanism based on CBAM to optimize the fused feature map every time, and skip connect the feature map generated after fusing this layer and the former layer, to solve the problem of aliasing effects. For the high-resolution feature pyramid, we introduce Dupsampling instead of deconvolution to generate higher-resolution prediction heatmaps to strengthen the model's detection of small persons, and also to strengthen the network's scale perception ability. In this study, multiple sets of experiments were conducted for objective performance analysis of the proposed method. The results indicate that CE-HigherHRNet is a significant step forward in addressing the pose estimation challenges.

REFERENCES

[1] W. Ouyang, X. Chu, and X. Wang, "Multi-source deep learning for human pose estimation," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, 20–23 June 2014, pp. 2329–2336.

[2] A. Bulat and G. Tzimiropoulos, "Human pose estimation via convolutional part heatmap regression," in *Proc. Haptics: Science, Technology, Applications*, London, UK, 4–7 July 2016, vol. 9911, pp. 717–732.

[3] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *Proc. Europ. Conf. Comput. Vis. (ECCV)*, Amsterdam, The Netherlands, 11–14 October 2016, pp. 483–499.

[4] J. J. Tompson, A. Jain, Y. LeCun, and C. Bregler, "Joint training of a convolutional network and a graphical model for human pose estimation," in *Proc. Adv. Neural Inf. Process. Syst*. MIT Press: Cambridge, MA, USA, 2014, vol. 27, pp. 1799–1807.

[5] B. Cheng, B. Xiao, J. Wang, H. Shi, T. S. Huang, and L. Zhang, "HigherHRNet: Scale-aware representation learning for bottom-up human pose estimation," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, USA, 13–19 June 2020, pp. 5386–5395.

[6] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 2117–2125.

[7] C. Guo, B. Fan, Q. Zhang, S. Xiang, and C. Pan, "Augfpn: Improving multi-scale feature learning for object detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, USA, 13–19 June 2020, pp. 12595–12604.

[8] J. Cao, Q. Chen, J. Guo, and R. Shi, "Attention-guided context feature pyramid network for object detection," arXiv preprint arXiv:2005.11475,2020

[9] Z. Cao, T. Simon, S. E. Wei, and Y. Sheikh, "Realtime multi-person 2D pose estimation using part affinity fields," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Hawaii, USA, 21–26 July 2017, pp. 7291–7299.

[10] G. Papandreou, T. Zhu, L. C. Chen, S. Gidaris, J. Tompson, and K. Murphy, "Personlab: Person pose estimation and instance segmentation with a bottom-up, part-based, geometric embedding model," in *Proc. Europ. Conf. Comput. Vis. (ECCV),*, Munich, Germany, 8–14 September 2018, pp. 269–286.

[11] W. Shi, J. Caballero, F. Huszár, J. Totz, A. P. Aitken, R. Bishop, and Z. Wang, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, USA, June 26–July 1 2016, pp. 1874–1883.

[12] Z. Tian, T. He, C. Shen, and Y. Yan, "Decoders matter for semantic segmentation: Data-dependent decoding enables flexible feature aggregation," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, USA, 16–19 June 2019, pp. 3126–3135.

[13] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Europ. Conf. Comput. Vis. (ECCV)*, Munich, Germany, 8–14 September 2018, pp. 3–19.

[14] B. Xiao, H. Wu, and Y. Wei, "Simple baselines for human pose estimation and tracking," in *Proc. Europ. Conf. Comput. Vis. (ECCV)*, Munich, Germany, 8–14 September 2018, pp. 466–481.

[15] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, USA, 16–19 June 2019, pp. 5693–5703.

[16] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, and B. Xiao, "Deep high-resolution representation learning for visual recognition," in *Proc. IEEE Trans. Pattern Anal. Mach. Intell*, 2020, pp. 1-1.

[17] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Intl. Conf. Comput. Vis.* (ICCV), Venice, Italy, 22–29 October 2017, pp. 2980–2988.

[18] H.S. Fang, S. Xie, Y. W. Tai, and C. Lu, "RMPE: Regional multi-person pose estimation," in *Proc. IEEE Intl. Conf. Comput. Vis.* (ICCV), Venice, Italy, 22–29 October 2017, pp. 2334–2343.

[19] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Sys*t, 2015, vol. 28, pp. 91–99.

[20] U. Iqbal and J. Gall, "Multi-person pose estimation with local joint-to-person associations," in *Proc. Europ. Conf. Comput. Vis. (ECCV)*, Amsterdam, Netherlands, 8–16 October 2016, pp. 627–642.

[21] G. Gkioxari, A. Toshev, and N. Jaitly, "Chained predictions using convolutional neural networks," P in *Proc. Europ. Conf. Comput. Vis. (ECCV)*, Amsterdam, Netherlands, 8–16 October 2016, pp. 728–743.

[22] A. Toshev and C. Szegedy, "Deeppose: Human pose estimation via deep neural networks," *in Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Columbus, USA, 20–23 June 2014, pp. 1653–1660.

[23] X. Chu, W. Ouyang, H. Li, and X. Wang, "Structured feature learning for pose estimation," *in Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, USA, June 26–July 1 2016, pp. 4715–4723.

[24] Y. Tan, W. Yan, S. Huang, D. Du, and L. Xia, "A motion deviation image-based phase feature for recognition of thermal infrared human activities," *Engineering Letters*, vol. 28, no.1, pp. 48–55, 2020

[25] C. Sharma, S. Singh, G. Poornalatha, and K. B. Ajitha Shenoy, "Performance analysis of object detection algorithms on YouTube video object dataset," *Engineering Letters*, vol. 29, no.2, pp. 813–817, 2021

[26] A. W. R. Emanuel, P. Mudjihartono, and J. A. M. Nugraha, "Snapshot-based human action recognition using OpenPose and deep learning," *IAENG Intl. J. Comput. Sci*, vol. 48, no.4, pp. 862–867, 2021

[27] J. Li, M. Yang, Y. Liu, Y. Wang, Q. Zheng, and D. Wang, "Dynamic hand gesture recognition using multi-direction 3D convolutional neural networks," *Engineering Letters*, vol. 27, no.3, pp. 490–500, 2019

[28] Q. Zheng, X. Tian, S. Liu, M. Yang, H. Wang, and J. Yang, "Static hand gesture recognition based on gaussian mixture model and partial differential equation," *IAENG Intl. J. Comput. Sci*, vol. 45, no.4, pp. 569–583, 2018

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

[30] S. E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional pose machines," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, USA, June 26–July 1 2016, pp. 4724–4732.

[31] W. Li, Z. Wang, B. Yin, Q. Peng, Y. Du, T. Xiao, and J. Sun, "Rethinking on multi-stage networks for human pose estimation," arXiv: 1901.00148,2019.

[32] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, USA, June 26–July 1 2016, pp. 770–778.

[33] S. Kreiss, L. Bertoni, and A. Alahi, "PifPaf: Composite fields for human pose estimation," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*) , Long Beach, USA, 16–19 June 2019, pp. 11977–11986.

[34] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," arXiv: 1412.6980,2014. 5.

[35] J. Li, C. Wang, H. Zhu, Y. Mao, H.S. Fang, and C. Lu, "CrowdPose: Efficient crowded scenes pose estimation and a new benchmark," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, USA, 16–19 June 2019, pp. 10863–10872.