Signal, Image and Video Processing (2023) 17:643–650
https://doi.org/10.1007/s11760-022-02271-7

ORIGINAL PAPER

Human pose estimation based on feature enhancement and multi-scale feature fusion

Dandan Cao[1] · Weibin Liu[1] · Weiwei Xing[2] · Xiang Wei[2]

Received: 12 September 2021 / Revised: 4 March 2022 / Accepted: 16 May 2022 / Published online: 18 June 2022

Abstract

The human pose estimation has been greatly improved with the development of deep neural network. However, there are some challenges in this task, such as the occlusions in images and various scales of the human body. In this study, we propose a novel convolutional neural network architecture based on dual attention mechanism and multi-scale feature fusion to generate keypoints prediction and estimate the location of human body parts in images. Firstly, the feature enhancement module(FEM) performs local feature enhancement process for each feature map of the network using the double-attention mechanism, where channel attention is used to filter out the channels that need more attention and spatial attention is used to enhance the local features of each feature map at the spatial level. Secondly, we design a multi-scale feature fusion(MSFF) module by using the cascade of atrous convolution to aggregate contextual information and enhance the expressiveness of features. The multi-scale contextual information is increased by expanding the perceptual field, which helps to detect adjacent keypoints. Finally, we introduce an improved upsampling module that jointly uses upsampling2D and transposed convolution to better regress the obtained feature maps to higher resolution and output heatmaps. Extensive experiments on MPII and COCO human pose estimation benchmarks demonstrate the effectiveness of our network.

Keywords Human pose estimation · Convolutional neural network · Feature enhancement · Multi-scale feature fusion · Upsampling module

1 Introduction

Human pose estimation is one of the most fundamental and challenging research topics in computer vision, aiming at locating keypoints of the human body and accurately recognizing the human pose in given images. The research of human pose estimation is of great importance, and the technology of human pose estimation has a wide range of applications such as video surveillance [1], virtual reality, intelligent, and human-computer interaction. Therefore, the study of human pose estimation has also attracted much attention from computer vision researchers [2]. However, due to occlusions, flexible body poses, and the variety of human body appearance scales, the human pose estimation is challenging and difficult.

For the study of human pose estimation, the current research methods are divided into two categories: model-based human pose estimation and deep learning-based human pose estimation. The model-based methods rely on hand-crafted features such as Histogram of Gradient (HOG)-based features, Motion Boundary Histograms (MBH) features, and Scale-Invariant Feature Transform (SIFT) features. These methods have poor generalization ability and suffer from disadvantages such as large training data and poor robustness. The deep learning-based methods mainly use Convolutional Neural Networks(CNN) to extract features. Pfister et al. [3] were the first to propose the use of CNN to obtain features of images for behavior analysis. With the development of deep learning technology, deep learning-based human pose estimation has become more and more mature and has become the main method to solve the problem of human posture estimation. Despite the great success of these methods, there are still drawbacks. Most of the current work [4–6] uses continuous downsampling to obtain a

✉ Weibin Liu
wbliu@bjtu.edu.cn

[1] Institute of Information Science, Beijing Jiaotong University, Beijing 100044, China

[2] School of Software Engineering, Beijing Jiaotong University, Beijing 100044, China

larger perceptual field to fuse multi-scale information, and the information of some small keypoints is lost and cannot be reconstructed to accurately localize these small keypoints. We use different sampling rates of cavity convolution for sampling to increase the perceptual field to obtain multi-scale information and enhance the ability of the model to obtain multi-scale information.

In this paper, we propose a novel network framework for human pose estimation with feature enhancement module (FEM) and multi-scale feature fusion (MSFF) to obtain more important details and fuse multi-scale contextual information. Firstly, we use the residual neural network [7] (ResNet) as the backbone network to extract image features. Then, the feature enhancement module jointly uses channel attention mechanism and spatial attention mechanism to enhance the local feature map from the backbone network. Then, the features of different scales are fused to obtain contextual information of the human body at different scales to improve the network accuracy. We introduce the atrous convolution in the network to increase the perceptual field of feature extraction to obtain a larger range of feature information, so as to capture richer multi-scale detail features. Finally, the upsampling module has been redesigned to better restore the resolution of the image.

The main contributions of this work can be summarized as follows:

(1) We combine the channel attention mechanism with the spatial attention mechanism in the feature enhancement module in the feature extraction to enhance the extracted detail features of human keypoints.
(2) The multi-scale feature fusion module is designed to enrich the scale information of the obtained features by using the improvement atrous spatial pyramid pooling module. And we use the atrous convolution in the backbone to expand the receptive field of the feature map.
(3) The upsampling module is redesigned by jointly using upsampling2D and transposed convolution to better recover the obtained feature maps to high resolution and to predict the human body keypoints heatmaps.

The rest of this paper is organized as follows. Section 2 of the paper presents the work related to human pose estimation, including single-person pose estimation methods and multi-person pose estimation methods. Section 3 describes our method. Section 4 shows the experimental procedure and results. The conclusion is presented in Sect. 5.

2 Related work

Single-person pose estimation methods. Single-person pose estimation is to locate the joint positions of a single person in the input image. According to the different formulas of human pose estimation tasks, single-person pose estimation methods can be divided into regression-based methods and heatmap detection-based methods. The regression-based method generates joint coordinates directly by mapping from images to human joint coordinates in an end-to-end manner. Deeppose proposed by Alexander Toshev et al. [8] firstly uses deep convolutional networks (DNN) to obtain the global features of the keypoints of the human body, then maps the coordinates of the keypoints from the input image, and finally uses the cascaded refinement regressor to refine the positions. Since it is very difficult to predict joint coordinates directly from the input image, Sun et al. [9] proposed a structure perception regression method based on bone representation. The bone-based representations based on body structure information lead to more accurate results than the methods only using joint positions.

The heatmap detection-based methods aim to predict body parts or joints through heatmaps where each two-dimensional Gaussian distribution centered on a joint position represents a joint position. YannLeCun et al. [10] proposed the framework of the combined depth network and tree structure model, which is the first human pose estimation method using heatmap detection. Newell et al. [4] proposed a stacked hourglass architecture based on stacked residual modules. Tang and Wu [11] proposed a part-based branching network (PBN) based on the grouping of human keypoints and designed a multi-stage structure for refinement.

Both methods have their advantages and disadvantages. The heatmap detection-based methods contain rich pixel information supervision and have better robustness compared to the regression-based methods. However, due to the pooling operation, the resolution of heatmap representation is low, which greatly affects the accuracy of keypoint coordinate estimation. Therefore, in this paper, we cleverly adopt the upsampling2D and transposed convolution to regress the heatmap to obtain more accurate keypoint coordinates.

Multi-person pose estimation methods. Since the number of human bodies in the input image is unknown, multi-person pose estimation needs to handle the task of detecting the human body and locating keypoints, which can be divided into top-down and bottom-up methods. The top-down methods first detect the people in the image using the person detector and then perform single-person pose estimation for each person detected. In order to recover the low-resolution feature maps to high resolution and output heatmaps, Xiao et al. [12] proposed to simply add several deconvolution layers after the last convolutional layer of the feature extraction network. Cascaded pyramid network (CPN) proposed by Chen et al. [6] includes GlobalNet for locating simple keypoints and RefinedNet for locating hard keypoints, and employs online hard keypoints mining strategy. Sun et al. [5] proposed HR-Net with multi-scale feature fusion that connects

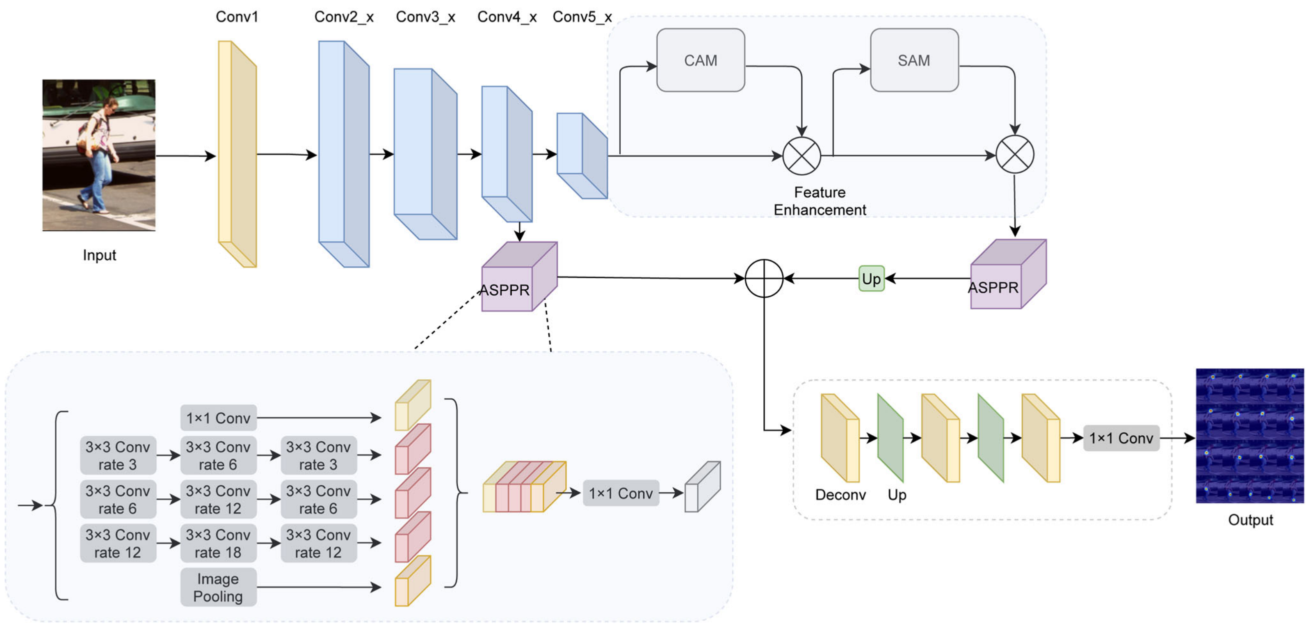

Fig. 1 The overall framework of proposed method

multiple multi-resolution sub-networks in parallel to maintain the high-resolution representation of the features during the process.

The bottom-up methods directly predict all joint positions of all people in the image and then connect them into independent human skeletons. Deepcut proposed by Pishchulin et al. [13] first detects all candidate body parts using Fast R-CNN [14], then annotates all candidate parts into corresponding part classes, and then assembles them into a complete skeleton by clustering. Cao et al. [15] proposed to encode the predicted partial affinity (PAF) domains to combine all estimated body candidate joints into a human body. Kreiss et al. [16] designed the PifPaf network containing Partial Intensity Fields (PIF) and Partial Association Fields (PAF), the former for improving the accuracy of the heatmap at high resolution, and the letter for connecting the keypoints.

As the performance of current object detectors is getting better, the top-down methods are easier to implement and have better performance. In this paper, we adopt the top-down approach by using the Faster-RCNN [14] detector to detect the human body in the image and deliver the cropped image to the pose estimation network. The attention mechanism is added to the pose estimation network for local feature enhancement, and the multi-scale feature fusion module is designed to fuse information from features of different resolutions to obtain richer feature representation.

3 Proposed method

3.1 Backbone network

The overall framework is shown in Fig. 1. We adopt ResNet to extract features from input images as the backbone network. ResNet replaces the traditional deep neural network with convolutional and pooling layers. It introduces skip connections to adapt the input from the previous layer to the next layer without modifying the input. It is the most common backbone network for image feature extraction, and it has also been used for pose estimation. Our model removes the fully connected (FC) layer from ResNet and retains the conv1, conv2_x, conv3_x, conv4_x, and conv5_x.

3.2 Feature enhancement module

In the feature enhancement module, we jointly use the channel attention mechanism and the spatial attention mechanism. For the process of extracting features for human pose estimation, more attention should be paid to the useful information in the input images. We integrate channel attention mechanisms to selectively inhibit or enhance channel information and spatial attention mechanisms to focus on the useful position information in the feature maps.

Based on the literature [17], we introduce the feature enhancement module into the human pose estimation task. The structure is shown in Fig. 2. In this study, we add Channel Attention Module (CAM) and Spatial Attention Module (SAM) over the conv5_x. It takes the original features F output from the last layer of the ResNet as the input of the

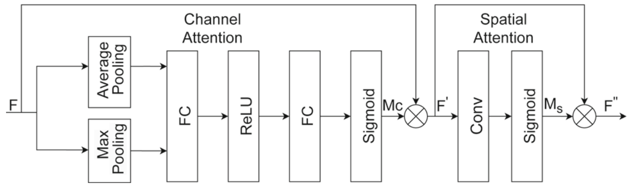

Fig. 2 Diagram of feature enhancement module

channel attention module and squeezes the feature map in the spatial dimension by using global average pooling and global maximum pooling. Then, feature maps F^c_{avg} and $F^c_{\max}$ are sent to the shared network of multi-layer perceptron (MLP) consisting of a hidden layer to reduce computational overhead. The channel attention map $M_c(F)$ is output by the sigmoid function. In short, the process is shown in Equation (1).

$$\begin{aligned} M_c(F) &= \sigma(MLP(AvgPool(F)) + MLP(MaxPool(F))) \\ &= \sigma(W_1(W_0(F^c_{\text{avg}})) + W_1(W_0(F^c_{\max}))) \end{aligned} \tag{1}$$

where F^c_{avg} and $F^c_{\max}$ represent the output of the global average pooling and global maximum pooling, σ denotes the sigmoid function, $W_0 \in R^{C/r \times C}$, and $W_1 \in R^{C \times C/r}$ are the weights of the MLP.

The spatial attention module mainly acquires the channel module features as the input and then performs the average pooling and the maximum pooling to generate two-dimensional feature maps:F^s_{avg} and $F^c_{\max}$. The obtained values are used to indicate which region features in the image are worthy of attention. Then, the two-dimensional feature maps are connected to perform the dimensionality reduction by 7×7 convolution. Then, the sigmoid activation is used to obtain the weight coefficient $M_s(F)$ needed to enhance the spatial information as follows:

$$\begin{aligned} M_s(F) &= \sigma(f^{7\times7}([AvgPool(F); MaxPool(F)])) \\ &= \sigma(f^{7\times7}([F^s_{\text{avg}}; F^s_{max}])) \end{aligned} \tag{2}$$

where F is the feature map enhanced by the channel attention module. F^s_{avg} and $F^s_{\max}$ are two descriptions obtained by performing global maximum pooling and average pooling, and σ represents the sigmoid function.

The feature enhancement module can learn key information of channel dimension and spatial dimension respectively, and adaptively reassign the weights of features to enhance the channel information and spatial information in the feature map. By adding the feature enhancement module over the last layer of ResNet, the network can obtain more effective features during the learning process, which helps to improve the localization accuracy of joints in pose estimation.

3.3 Multi-scale feature fusion

Due to the various sizes of human bodies in the image, it is difficult to achieve accurate localization of some smaller size human keypoints by using fixed-size features. In order to improve the processing capability of the network for small size human keypoints, we design the multi-scale feature fusion module. This module automatically extracts the contextual information of the feature map to identify keypoints that are not obvious and occluded by expanding the receptive field by using the atrous convolution with different expansion rates respectively. Atrous Convolutions introduce the dilation rate, which defines the spacing of the values as the convolution kernel processes the data to increase the feature perceptual field and maintain the resolution of the images. The atrous spatial pyramid pooling (ASPP) [22], which is commonly used in semantic segmentation, mainly performs pyramid-like convolution operations on the incoming convolutional feature maps by using atrous convolution [23] with different expansion rates. It increases the feature perceptual field and captures multi-scale context information simultaneously without decreasing the resolution.

In order to better capture the contextual multi-scale information of the input features and improve the detection accuracy of human keypoints, this study designs an improved atrous spatial pyramid pooling module ASPPR, as shown in Fig. 1. By gradually increasing the expansion rate, richer semantic information suitable for large-scale human keypoint localization is obtained. Subsequently, the expansion rate is reduced to capture local information suitable for small-scale human keypoint localization to enhance the extraction of detailed features. The improved atrous spatial pyramid pooling module contains five independent branches, where one convolutional branch consists of a 1×1 convolutional layer, three atrous convolutional branches consist of three atrous convolutional layers with different expansion rates, with convolutional kernel size of 3×3 and expansion rates of $(3, 6, 3)$, $(6, 12, 6)$, and $(12, 18, 12)$; one image pooling branch consists of an average pooling layer, a 1×1 convolutional layer, and an upsampling layer. By gradually increasing the expansion rate and then reducing the expansion rate, we can obtain richer semantic information and local information suitable for small-scale human keypoints locating. In the experiment, we found that setting the parameters in this way can obtain better performance without increasing too much computation. Finally, the feature maps output from the five branches are fused with features.

We design the multi-scale feature fusion network as shown in Fig. 1. The multi-scale feature fusion module first inputs the features $C5'$ extracted by the feature enhancement module to the ASPPR module and performs the upsampling operation to ensure the same feature size as the output of the previous layer, and then inputs the feature map $C4$ extracted

Table 1 The Comparison of PCKh @0.5 on the MPII validation set. The results of the proposed method are bold and placed in the last two rows of the table

Model	Backbone	Head	Sho	Elb	Wri	Hip	Kne	Ank	Mean
Hourglass [4]	–	96.0	96.3	90.3	85.4	88.8	85.0	81.9	89.2
Simplebaseline [12]	ResNet-50	96.4	95.3	90.0	83.2	88.4	84.0	79.6	88.5
Simplebaseline [12]	ResNet-152	97.0	95.9	90.3	85.0	89.2	85.3	81.3	89.6
LFP [18]	–	96.8	96.0	90.4	86.0	89.5	85.2	82.3	89.6
DLCM [19]	–	95.6	95.9	90.7	86.5	89.9	86.6	82.5	89.8
AL [20]	–	96.5	96.0	90.5	86.0	89.2	86.8	83.7	89.9
PGCN [21]	ResNet-50	–	–	–	83.6	–	–	80.8	88.9
Ours	**ResNet-50**	**96.5**	**95.4**	**90.0**	**83.5**	**88.6**	**84.3**	**80.3**	**88.9**
Ours	**ResNet-152**	**97.1**	**95.9**	**90.4**	**85.3**	**89.4**	**85.7**	**81.5**	**89.9**

by conv4_x in the backbone network to the ASPPR module to obtain the multi-scale context information, and performs the matrix summation operation according to the channel dimension and the result after upsampling of the feature map $C5'$ to obtain the fused feature map. The deep feature and shallow feature are combined to get the target feature which contains both detailed information and semantic information, which greatly improves the accuracy of joints location.

3.4 Improvement upsampling module

In the improvement upsampling module, we implement deconvolution layers to regress obtained feature maps to a higher resolution. At the end of the network, several layers of deconvolution are used to generate heatmaps for keypoints. In order to achieve better restoration effect, upsampling2d is used to enlarge the image to the required size, and then, the deconvolutional layer is performed to form a new upsampling module. The module uses upsampling2d to expand the image size by two times, and then, the deconvolutional layer with batch normalization and ReLU [24] activation is used. Through convolution operation, the resolution of the rough image can be restored and the robustness of the network can be improved.

4 Experiment

In this section, we show our experimental results. We evaluated our model on MPII [25] and COCO [26] datasets.

4.1 Dataset and evaluation metrics

MPII: The MPII dataset [25] is the standard benchmark for 2D human pose estimations. The images are collected from online videos covering a wide range of activities and annotated by humans for 16 joints. It contains 25,000 training images. The evaluation metric is the Percentage of Correct Keypoints (PCK).

COCO: The COCO dataset [26] presents imagery data with various human poses, different body scales, and occlusion patterns. It contains more than 200,000 images and 250,000 human instances labeled with 17 keypoints. The COCO evaluation defines the object keypoint similarity (OKS) and uses the mean average precision (AP) over 10 OKS thresholds as the main competition metric. The OKS is calculated from the distance between predicted points and ground-truth points normalized by the scale of the person.

4.2 Implementation details

A Mean Squared Error(MSE) is used to compare the predicted heatmaps with ground-truth heatmaps H_k generated by applying a two-dimensional normalized Gaussian center at the ground-truth position of the kth joint. The ground-truth human box is made as height:width = 4:3, then cropped out of the image and adjusted to a fixed resolution of 256×192 or 384×288. Data augmentation includes rotation(± 40 degrees), scale ($\pm 30\%$) and flip. Our pose estimation network model is initialized by pre-training on the publicly released ImageNet [27] classification task. In the training for pose estimation, our basic learning rate is set to 1e$-$3, which drops to 10 times at 90 epochs and 120 epochs, for a total of 140 training epochs. We use the Adam [28] optimizer in the training process. Our proposed model is implemented in Pytorch. All experiments are run on a server with an Nvidia GEFORCE GTX 2080Ti GPU.

4.3 Results

Results on MPII. We compare the performance of our method and some previous methods on the MPII dataset. The PCKh@0.5 is used to evaluate the performance. For simplicity, we directly use the PCKh scores in [5]. Table 1 shows the final results on MPII validation set. As shown in the table, our method obtains the PCKh score of 89.9% on the MPII validation set, which is comparable to previous methods.

Table 2 Comparison with the 8-stage Hourglass, CPN, SimpleBaseline, ECSN and PGCN on the COCO val2017 dataset. The results of the proposed method are bold and placed in the last four rows of the table

Method	Backbone	Input Size	AP
8-stage Hourglass [4]	–	256 × 192	66.9
8-stage Hourglass [4]	–	256 × 256	67.1
CPN [6]	ResNet-50	256 × 192	68.6
CPN [6]	ResNet-50	384 × 288	70.6
SimpleBaseline [12]	ResNet-50	256 × 192	70.4
SimpleBaseline [12]	ResNet-50	384 × 288	72.2
SimpleBaseline [12]	ResNet-101	256 × 192	71.4
SimpleBaseline [12]	ResNet-101	384 × 288	73.6
ECSN [29]	ResNet-50	256 × 192	72.1
ECSN [29]	ResNet-50	384 × 288	73.8
PGCN [21]	ResNet-50	256 × 192	71.1
PGCN [21]	ResNet-50	384 × 288	72.9
Ours	**ResNet-50**	**256 × 192**	**72.3**
Ours	**ResNet-50**	**384 × 288**	**73.9**
Ours	**ResNet-101**	**256 × 192**	**72.5**
Ours	**ResNet-101**	**384 × 288**	**74.5**

Results on COCO. In the COCO evaluation, we followed the SimpleBaseline[12] and used the human detector whose AP is 56.4 to provide human bounding boxes. Table 2 shows our result with other competitive methods on the COCO validation dataset. The human detection AP reported in 8-stage Hourglass [4], CPN [6], ECSN [29] and PGCN [21] is 55.3. Compared with the baseline [12], our method achieves better performance with the same backbone network. Compared with the typical methods such as Hourglass [4] and CPN [6], our method has a significant improvement. Our method also has a certain improvement in performance compared with more advanced methods such as ECSN [29] and PGCN [21]. While OHKM is used in ECSN [29], our method is comparable to ECSN [29], which means our method is simpler. Therefore, we can conclude that our method has comparable results.

Table 3 lists the results of our method and other methods on the COCO test-dev dataset. As observed, our method outperforms CPN by 1.2 AP for the input size of 384 × 288. SimpleBaseline, ECSN and our method use the same backbone of ResNet-50. For the input size of 256 × 192, our method is compared with ECSN. And our method outperforms SimpleBaseline by 1.6 AP for the input size of 256×192. This demonstrates the effectiveness of our method. The final effect of the proposed network on the keypoints of human body in COCO dataset images is shown in Fig. 3.

Table 3 Comparison results on the COCO test-dev dataset. The results of the proposed method are bold and placed in the last two rows of the table

Method	Backbone	Input size	AP
Mask-RCNN [30]	ResNet-50-FPN	–	63.1
G-RMI [31]	ResNet-101	353 × 257	64.9
CPN [6]	ResNet-Inception	384 × 288	72.1
SimpleBaseline [12]	ResNet-50	256 × 192	70.0
SimpleBaseline [12]	ResNet-50	384 × 288	71.5
ECSN [29]	ResNet-50	256 × 192	71.4
ECSN [29]	ResNet-50	384 × 288	73.2
Ours	**ResNet-50**	**256 × 192**	**71.6**
Ours	**ResNet-50**	**384 × 288**	**73.3**

4.4 Ablation Study

To examine the influence of feature enhancement, multi-scale feature fusion module and redesigned upsampling module, we perform the ablation study on the COCO validation dataset. We adopt the ResNet-50 as the backbone network with the 256 × 192 input size. Table 4 shows the experimental results when we gradually apply different components, i.e., feature enhancement module, multi-scale feature fusion, and redesigned upsampling module one by one. As can be observed, feature enhancement can lead to 0.6 AP improvement than the baseline model. This suggests that the joint attention mechanism helps the network to better focus on important information and enhance the extracted features. When multi-scale feature fusion is applied, the aggregation for information can further increase AP by 0.9, which indicates that increasing the receptive field to obtain multi-scale contextual information and performing information fusion can lead to more accurate human keypoints localization results. Finally, the proposed improvement upsampling module is shown able to expand image size for better recovery of high resolution with the improvement of 0.4 AP. In order to show the independent effect of the three modules, we verify the performance effect of different combinations. It can be observed that the feature enhancement module and multi-scale feature fusion module play an important role in improving the performance.

5 Conclusion

This paper proposes a novel convolutional neural network architecture to locate the positions of the human body keypoints in images. By using the proposed network, more importable feature information can be obtained by the use of the feature enhancement module, and the fusion of multi-scale feature can refine the prediction for the positions of

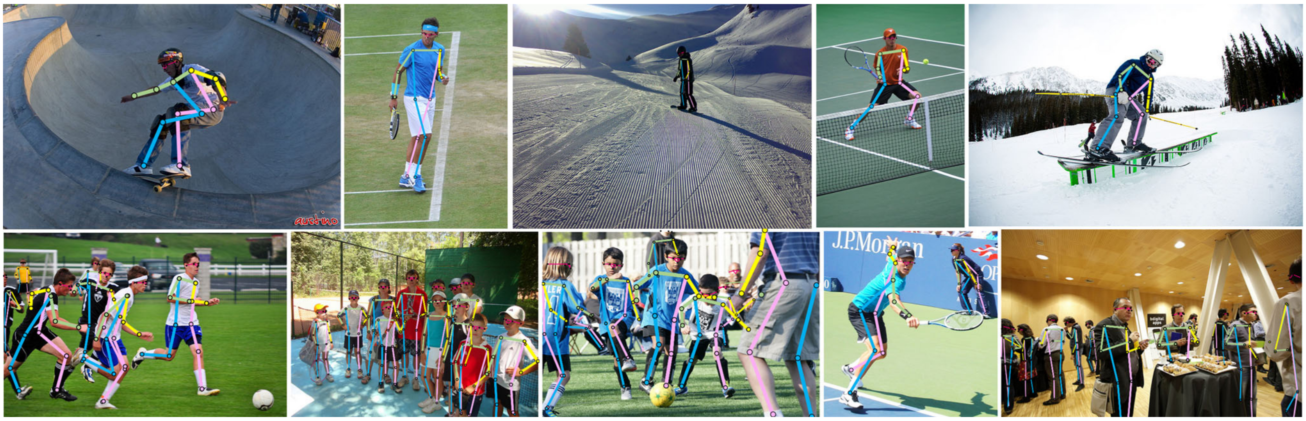

Fig. 3 Qualitative results of some example images in the COCO datasets

Table 4 Ablation analysis of proposed method

FEM	MSFF	Improvement upsampling	AP
×	×	×	70.4
√	×	×	71.0
√	√	×	71.9
×	√	√	71.3
√	×	√	71.2
√	√	√	72.3

The experiment is evaluated on COCO validation set

the keypoints. In redesigned upsampling module, upsampling2D and transposed convolution are jointly used to better regress the obtained feature maps to higher resolution and output heatmaps. We compare experiments with different methods on two widely used datasets, i.e. MPII and COCO. Based on the experimental results, we found that fine feature extraction and utilization is an important task in human pose estimation. The final results also indicate that our method has better performance. In future work, we will focus more on reducing the number of parameters in the network to achieve real-time human pose estimation while improving network performance.

Acknowledgements This research is partially supported by the Beijing Natural Science Foundation (No. 4212025), National Natural Science Foundation of China (Nos. 61876018, 61906014, 61976017).

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Ethical approval This article does not contain any studies with human participants and/or animals performed by any of the authors.

Informed consent There is no informed consent for this study.

References

1. Miki, D., Abe, S., Chen, S., Demachi, K.: Robust human pose estimation from distorted wide-angle images through iterative search of transformation parameters. Signal Image Video Process. **14**(4), 693–700 (2020)
2. Chen, Y., Tian, Y., He, M.: Monocular human pose estimation: a survey of deep learning-based methods. Comput. Vis. Image Underst. **192**, 102897 (2020)
3. Pfister, T., Simonyan, K., Charles, J., Zisserman, A.: Deep convolutional neural networks for efficient pose estimation in gesture videos. In: Asian Conference on Computer Vision, pp. 538–552. Springer (2014)
4. Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. In: European conference on computer vision, pp. 483–499. Springer (2016)
5. 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, pp. 5693–5703 (2019)
6. Chen, Y., Wang, Z., Peng, Y., Zhang, Z., Yu, G., Sun, J.: Cascaded pyramid network for multi-person pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7103–7112 (2018)
7. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
8. Toshev, A., Szegedy, C.: Deeppose: human pose estimation via deep neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1653–1660 (2014)
9. Sun, X., Shang, J., Liang, S., Wei, Y.: Compositional human pose regression. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2602–2611 (2017)
10. Tompson, J., Jain, A., LeCun, Y., Bregler, C.: Joint training of a convolutional network and a graphical model for human pose estimation. arXiv preprint arXiv:1406.2984 (2014)
11. Tang, W., Wu, Y.: Does learning specific features for related parts help human pose estimation? In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1107–1116 (2019)
12. Xiao, B., Wu, H., Wei, Y.: Simple baselines for human pose estimation and tracking. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 466–481 (2018)
13. Pishchulin, L., Insafutdinov, E., Tang, S., Andres, B., Andriluka, M., Gehler, P.V., Schiele, B.: Deepcut: Joint subset partition and labeling for multi person pose estimation. In: Proceedings of the

IEEE Conference on Computer Vision and Pattern Recognition, pp. 4929–4937 (2016)

14. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: towards real-time object detection with region proposal networks. Adv. Neural. Inf. Process. Syst. **28**, 91–99 (2015)
15. Cao, Z., Simon, T., Wei, S.E., Sheikh, Y.: Realtime multi-person 2d pose estimation using part affinity fields. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7291–7299 (2017)
16. Kreiss, S., Bertoni, L., Alahi, A.: Pifpaf: Composite fields for human pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11977–11986 (2019)
17. Woo, S., Park, J., Lee, J.Y., Kweon, I.S.: Cbam: convolutional block attention module. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 3–19 (2018)
18. Yang, W., Li, S., Ouyang, W., Li, H., Wang, X.: Learning feature pyramids for human pose estimation. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1281–1290 (2017)
19. Tang, W., Yu, P., Wu, Y.: Deeply learned compositional models for human pose estimation. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 190–206 (2018)
20. Ryou, S., Jeong, S.G., Perona, P.: Anchor loss: Modulating loss scale based on prediction difficulty. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 5992–6001 (2019)
21. Bin, Y., Chen, Z.M., Wei, X.S., Chen, X., Gao, C., Sang, N.: Structure-aware human pose estimation with graph convolutional networks. Pattern Recogn. **106**, 107410 (2020
22. Chen, L.C., Papandreou, G., Schroff, F., Adam, H.: Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587 (2017)
23. Yu, F., Koltun, V.: Multi-scale context aggregation by dilated convolutions. arXiv preprint arXiv:1511.07122 (2015)
24. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Adv. Neural Inf. Proc. Syst. **25** (2012)
25. Andriluka, M., Pishchulin, L., Gehler, P., Schiele, B.: 2d human pose estimation: new benchmark and state of the art analysis. In: Proceedings of the IEEE Conference on computer Vision and Pattern Recognition, pp. 3686–3693 (2014)
26. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European Conference on Computer Vision, pp. 740–755. Springer (2014)
27. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. Int. J. Comput. Vis. **115**(3), 211–252 (2015)
28. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
29. Su, K., Yu, D., Xu, Z., Geng, X., Wang, C.: Multi-person pose estimation with enhanced channel-wise and spatial information. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5674–5682 (2019)
30. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2961–2969 (2017)
31. Papandreou, G., Zhu, T., Kanazawa, N., Toshev, A., Tompson, J., Bregler, C., Murphy, K.: Towards accurate multi-person pose estimation in the wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4903–4911 (2017)