IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received April 13, 2021, accepted April 16, 2021, date of publication April 20, 2021, date of current version May 4, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3074541

Multi-Cascade Perceptual Human Posture Recognition Enhancement Network

MENGLONG WU, DEXUAN DU, YUNDONG LI, (Member, IEEE), WENLE BAI, AND WENKAI LIU

School of Information Science and Technology, North China University of Technology, Beijing 100144, China

Corresponding author: Menglong Wu (wumenglong@126.com)

This work was supported by the National Natural Science Foundation of China under Grant 62071006.

ABSTRACT The current researches trend to adopt a low-resolution hot spot map to restore the original high-resolution representation to save computing cost, resulting in unsatisfactory detection performance, especially in human body recognition with a highly complex background. Aiming at this problem, we proposed a model of parallel connection of multiple sub-networks with different resolution levels on a high-resolution main network. It can maintain the network structure of a high-resolution hot spot map in the whole operation process. By using this structure in the human key point vector field network, the accuracy of human posture recognition has been improved with high-speed operation. To validate the proposed model's effectiveness, two common benchmark data sets of COCO key point data set and MPII human posture data set are used for evaluation. Experimental results show that our network achieves the accuracy of 72.3% AP and 92.2% AP in the two data sets, respectively, which is 3%-4% higher than those of the existing mainstream researches. In our test, only the accuracy of backbone's SimpleBaseline with ResNet-152 is close to ours, yet our network realized a much lower computing cost.

INDEX TERMS Artificial intelligence, convolution-net, DeepResolution-Net, pose recognition.

I. INTRODUCTION

Human posture estimation is one of the important applications of deep convolution neural network in behavior perception [1]–[3]. It has been widely utilized in pedestrian movement trend detection, health recovery training and other related fields, working by capturing image signals of human behaviors, tracking and judging different behavior postures [4]–[6].

At present, most of the mainstream human posture estimation networks used the character detector mechanism, which directly used the top-down single-person attitude estimation technology, such as the 3D-Mask R-CNN detection model proposed by He *et al.* [7], Johnson [8], and Huang and Zhong [9]. The Online Pose Tracking framework proposed by Guanghan Ning *et al.*, detected human candidate objects in the first frame and used a single-person posture classifier to track the position and posture of each candidate object. When the candidate object is lost, the framework will associate the current frame's detection data with the graph convolution network.

The associate editor coordinating the review of this manuscript and approving it for publication was Bo Pu.

However, these mentioned networks are limited by their network structure, and the computing overhead will increase linearly with the increase of the number of people in the same image [10]–[12]. Besides, the human detector may not accurately identify the position of the human body when it is occluded, resulting in missing some of the detected objects. Further, such a mechanism is more likely to cause the loss of detected objects in the blurred part since most of these networks are composed of a series of main networks from high-resolution to low-resolution [13], [14].

Aiming at this issue, we designed the MEPDN (Multi-Enhance-Pose-Detection-Net) network, which combines the high-speed bottom-up human key point vector field detection network and parallel multi-level high-resolution network (DeepResolution-Net). The basis of the MEPDN network is designed as a parallel multi-level concatenation neural network. The image is acting as the input of the two-branch CNN, and the high-resolution hot spot map is also utilized in each branch. The two CNN branches jointly predict the location of the key points and the affinity vector field of the key points. By constantly optimizing the operation results of the previous stage, the best results are finally obtained.

The MEPDN network changes the hot spot map of symmetrical structure in the vector field detection network of human key points by multi-stage convolution to a parallel multi-stage high-resolution hot spot map with step-by-step progression, which is referred to as DeepResolution-Net in this paper.

The DeepResolution-Net operation mechanism working by recalculating the subnetworks with symmetrical resolution, changing it to make it paralleled by sub-networks with multiple levels of resolution, and then continuously adding sub-networks with different resolutions in parallel of each layer. In that way, it is possible to keep part of the convolution sub-network operating at the highest resolution all the time. Therefore, this kind of network model is different from the traditional symmetrical high-to-low resolution network model such as Hourglass [15], it can provide more powerful performance in recognition and computing speed.

On the other hand, the introduction of high-resolution network can enable the MEPDN to identify the vector relationship between key points more accurately. We used the high-resolution feature hotspot map of parallel multi-scale fusion in each layer of the cascade network, and allow the parallel networks in the MEPDN to exchange information with each other to achieve multi-scale fusion and feature extraction. The final estimated key point is the output of the high-resolution backbone network. This parallel network can connect high-resolution and low-resolution networks instead of serial connections as in the previous method. Therefore, the method can maintain high-resolution rather than restore the resolution through a low-to-high process to make the predicted Heatmap more accurate in space.

Our contributions can be mainly concluded in the following three points:

(1) Proposed an enhanced human posture estimation network, based on multi-level concatenation convolution neural network and step-by-step progressive multi-level resolution network.

(2) Applied step-by-step multi-resolution network to the calculation of human key point affinity field for the first time, which improved the energy efficiency of human key point affinity field calculation under high complex background to some extent.

(3) Balanced the computational cost and recognition accuracy, which provided certain advantages in the mainstream bottom-up human posture estimation network.

In the second chapter of this paper, we compare algorithm logic between the designed the MEPDN network and the two representative mainstream attitude recognition networks, and briefly describe the implementation methods of different attitude detection networks.

For the further method demonstration of the network, we put it in the third chapter, in which we will logically demonstrate the methods used in the implementation of the MEPDN network and explain the two parts of the network architecture of the MEPDN network from the aspect of mathematical logic.

In the fourth chapter, we will compare the mainstream network performance indicators, compare the network performance in various scenarios in two mainstream test sets, and achieve gratifying results in some key indicators of attitude recognition network performance.

II. RELATED WORK

At present, many studies have proposed their solutions to the problem of human posture recognition. For example, the DeepPose [16] network from Alexander Toshev *et al.* defines the human posture estimation problem as a key point regression problem. By using the DNN network to capture the correlation information of each human key point, the regression calculation is carried out for each key point. In this way, the explicit design feature extractor and local detector can be avoided. There is no need for DNN network to establish the topology for the key points, which makes the whole method easier to implement.

In the AlphaPose [17] network proposed by Hao-Shu Fang *et al.*, SSD-512 is used for human detection and Stacked Hourglass for attitude estimation. The network is mainly composed of three parts:

(1) Symmetric Spatial Transformer Network (SSTN) symmetric space transformation network: used to extract single-person regions from inaccurate Bounding Box.

(2) Non-maximum suppression of NMS parameterized attitude: it is used to solve the redundancy in the calculation.

(3) Pose-Guided Proposals Generator (PGPG): It used to enhance the training data of AlphaPose network.

The above two representative networks also show two different recognition models currently used in the field of human posture estimation:

(i) Top-down recognition model [15], [18], [19]. Those kinds of models selected the location of the detection object given by MASK R-CNN and other network models firstly. Then estimated the attitude of each instance object respectively.

For example, the 3D-Mask R-CNN model proposed by Georgia Gkioxari *et al.* The basic network of the 3D-Mask R-CNN model is a standard ResNet network model, which is extended to 3D. The network model generated a 3D feature blob for Tube Proposal Network (TPN) to generate Proposal tubes. In the model, RPN generated a 2D attitude suggestion box, TPN generated the 3D attitude suggestion tube. These Tubes are used to extract regional features from 3D Feature-blob, with a spatio-temporal RoI-Align mechanism.

These Proposals are fed into a classifier, and another key point detector is used to predict the heat map of the key points. However, such a design suffered a big drawback in calculation consumption. When the number of identified objects in the picture is large, the detector is required to operate repeat calculation in one image, which makes the operation time increased linearly with the increase of the number of identified objects in the image.

(ii) Bottom-up recognition model [20], [21]. Those kinds of models predicted the hot spot map of the key points in the

human body firstly, and returned the spatial position of the key points. Then, selected position with the highest Gaussian value as the key point of the network model prediction and expressed the spatial correlation between various parts of the body as a graphic model of a tree structure.

For example, the bottom-up network structure proposed by Pishchlin *et al.* [22]–[24], jointed marks candidate key points and associated them with each object. However, it may take several hours for the network to finish the operation since it is an NP-HARD problem to deal with the integer linear programming problem on a fully connected graph. In other networks, since the main network adopted the method of high-to-low resolution, the key point prediction error may occur and the correlation between variables may not be captured by the tree structure model when the identified object in the image is too small.

The HigherHRNet [25] network model proposed by Bowen Cheng *et al.* is a bottom-up human posture estimation network based on HRNet and SimpleBaseline [26]. The network adopts the method of multi-scale fusion. Through heatmap aggregation, in the inference stage, it solves the training problem caused by the diversity of detected objects in a bottom-up network to a certain extent.

The network uses high-resolution feature pyramid learning scale perception representation, which enables to achieve multi-resolution supervision in network training and multi-resolution aggregation function in logical reasoning. More accurate detection results can be obtained in small human target detection. However, this method still has some limitations in the output feature map. Because the output feature map is related to the data set used by the network, repeated experiments need to be carried out according to the selected data set to determine the size of the output feature map with the best performance.

Through the clustering problem of graph theory nodes, the model effectively used non-maximum suppression to express the optimization problem as an Integer Linear Program (ILP) problem. It can be effectively solved in network computing. However, since the model utilized both fast R-CNN (for human body detection) and ILP (for human posture estimation), the computational complexity of the DeepCut network model is very large.

Zhang *et al.* [45] tried to improve the network performance by re-decoding the hot spot map. They decoded the hot spot map predicted by the network into the coordinates of the key points of the human body in the original image space, and proposed a more principled distributed perceptual decoding method. However, this kind of key point coordinate representation network based on distributed perception will make the network vary with the size of the input image, which will have a great impact on the amount of computation.

The solutions of traditional human posture estimation networks are mostly focused on probabilistic graphic models and picture structure models [27]–[29]. But the utilization of deep convolution neural networks can provide more efficient results for human body key point estimation. However, most networks adopted a similar classification sub-network structure with reduced resolution [2], [30], [31], which is composed of the main body that produces the same resolution as its input and the regression used to estimate the hot spot map.

For example, Hourglass and its derivative networks used a symmetrical high-to-low resolution network. It combined with a structure body with the same resolution as the input elements of the network and a regression device used to estimate the hot spot atlas. Such networks utilized the hot spot atlas to estimate the location of the key points in the human body and then restore it to high resolution in the output phase, which may result in the loss of some information during the operation phase.

In addition to improving the resolution of the hot spot map in the classification sub-network in different ways, Yanrui Bin *et al.* [49] proposed Semantic Data Augmentation (SDA). By pasting segmented body parts with various semantic granularity, it can enhance the effect of human body detection in the image when the target in the detected image is seriously occluded. Compared with the human body key point affinity field method adopted by the MEPDN in this paper, the method of logically deducing the position coordinates of other key points through the known key points, the SDA network needs to preprocess the image. If the occlusion range is large or the resolution of the input image is high, the pasting segmented body parts with various semantic granularity processing may take longer. It will also lead to a linear increase in network computing overhead.

Different from others, Ke *et al.* [46] is focused on improving the hourglass model of deep conv-deconv with four key improvements: (i) multi-scale supervised learning, terminal multi-scale regression network; (ii) key point masking training method; (iii) structure-aware loss; (iv) key point masking training scheme. These improvements make the network play a role in complex multi-person recognition scenes such as scale change, occlusion and so on. However, this kind of multi-scale structure-aware network may have higher requirements for the data used in the network training stage. It makes the network training more tedious, which will indirectly lead to the increase of network computing.

The MEPDN network we designed in this paper is a bottom-up target detection network. In the network, the idea of partial affinity field (PAF) is used to calculate the correlation between key points, and the correlation score is given through a set of PAF. At the same time, the 2D vector field is used to encode the position and direction of the limbs in the image domain. Such a network can learn the relationship between the human body key point position and the human body affinity field simultaneously in two branches CNN. As to further enhance the recognition accuracy of the network in the responsible background, we used the main network in

the two parallel branches to maintain high resolution in the whole operation.

III. METHODOLOGY

In this chapter, we will introduce the method of the MEPDN in three parts. Part A will be an overview of the MEPDN network, which will involve the overall design architecture of the MEPDN network. It will systematically show the underlying design concept of the MEPDN network for the first time, including the structure of the multi-level concatenated convolution neural network and the step-by-step multi-resolution hot spot atlas. It will introduce the multi-level concatenated convolution neural network, which is the core of the MEPDN network. In part B, we will start from the human body key point affinity field used by the MEPDN network to predict human key points, gradually expand and carry on the mathematical logic deduction. We believe this will help to show the theoretical basis of the superior performance of MEPDN in human body key point detection. Part C will show the step-by-step multi-resolution main network, which we call DeepResolution-Net. It is an essential way for MEPDN to control computing overhead while maintaining high enough recognition accuracy. It works with a multi-level concatenated convolution neural network and finally constructs the overall structure of the MEPDN network.

A. INTRODUCTION OF THE MEPDN NETWORK

The MEPDN network adopted the parallel multi-level convolution neural network as the underlying design architecture. At the same time, the human body key point detection and human posture estimation of parallel operation in the network are enhanced by high-resolution main network parallel multi-level different resolution sub-networks.

The proposed pedestrian posture detection network, focused on improving the ability of vehicles to detect pedestrians and reducing the operation time of the network in the case of high complexity and multi-objectives. It is based on a multi-stage parallel convolution neural network and step-by-step progressive multi-resolution hot spot map. This network structure can improve the recognition accuracy of the network.

The bottom-up mechanism is adopted in the multi-cascade parallel convolution neural network in recognition, which first tracks and locates the human body's key points, and then connects the posture structure of the human body at anchor. In this way, the operation time of the network will not increase linearly with the increase of detected targets in the image. Besides, by integrating the step-by-step multi-resolution hot spot map, recognizing human key points in the complex background is not accurate enough can be alleviated.

Unlike the previous attitude recognition network, which is from high resolution to low resolution, and then restored to a high-resolution symmetrical structure in the output phase, the stepped structure enables the network to maintain high-resolution output throughout the operation process. Meanwhile, in the high-resolution main network, different resolution subnetworks are continuously introduced into the exchange unit to realize the exchange of information across the network. As a result, the network retains more image details in operation and does not increase the operation time significantly.

When designing the MEPDN network, we strive to set most of the parameters to learnable parameters. It will significantly help the network to have better adaptability in different data sets. However, it is inevitable that some network parameters are non-learnable parameters. Through many experiments, we finally determine the non-learnable parameter values that can bring the best performance in COCO data sets and MPII data sets. We will make an additional explanation when we calculate the non-learnable parameters in the mathematical formula.

In this part, we focused on the mathematical logic of the human key point affinity field and multi-stage high-resolution network. We combined these two techniques for the first time to realize multi-person attitude recognition in high-precision and complex background, to achieve better results in the follow-up experimental phase.

The MEPDN network is inspired by the human affinity field and the confidence map network of key points [32], [33]. The network preset the pixel positions of $\mathbf{p}$ human body key points, as $\mathbf{Y_p} \in \boldsymbol{Z} \in \mathbf{R}^2$, where $\mathbf{Z}$ is a set of coordinates of all human body key points in the input image. The network will be continually trained so that it can eventually predict the location of all key points in the human body.

Fig.1 gives a two-branches multi-level CNN architecture, as follows.

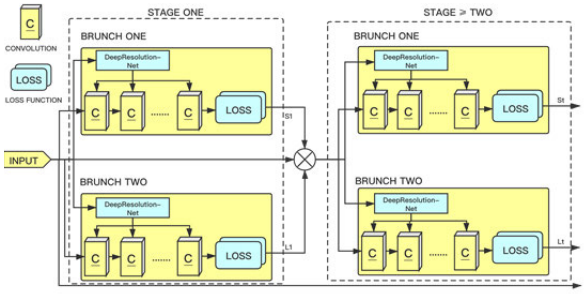


FIGURE 1. Schematic diagram of the MEPDN network structure.

Fig.1 is only a schematic diagram of the subordinate structure of the DeepResolution-Net to the original multi-stage concatenation convolution structure. In actual operation, DeepResolution-Net encapsulated each multi-cascade sub-network into a whole, which is formed by continuously parallel step-by-step progressive multi-level resolution sub-network. Each stage of the network is composed of subnetworks of the multi-class sequence predictor $\mathbf{g_x}(\cdot)$. Training will enable the classifier to predict the location of each key point in different levels of the network model.

The introduction of the DeepResolution-Net structure can replace the sequential multi-resolution sub-network structure of the original multi-stage concatenation structure. Therefore,

the resolution in each stage of the sub-network will no longer be a high-to-bottom process, but a step-by-step parallel connection of multiple sub-networks with different levels of resolution in the continuous operation.

When the loss function reaches the expected threshold, the convolution operation of the picture in the MEPDN network based on a multi-level concatenation structure will be terminated. Therefore, the final number of the convolution stages is not fixed, depending on the stage at which the loss function finally reaches the threshold.

The multi-level connection structure of the DeepResolution-Net is started with a high-resolution sub-network model, and multiple resolution subnetworks will be added step by step in operation. As shown in **Fig. 2:**

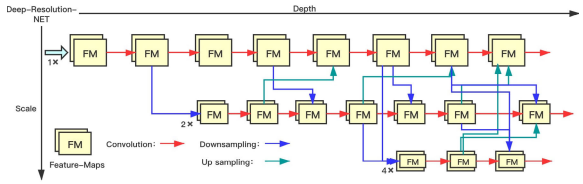


FIGURE 2. The schematic diagram of DeepResolution-Net network model.

The feature-Maps in DeepResolution-Net network are still multi-stage volume integrators. Each Feature-Maps in the graph is the characteristic graph of sub-network with different resolutions. These sub-network structures contain multi-level concatenated convolution neural network structures.

The downsampling multiple here is the non-learnable parameter. After many tests, we finally determine that the hot spot map of the original resolution is used in the first layer network. The hot spot map in the second layer network is obtained by twice downsampling of the first layer network. In the third layer network, the hot spot map is obtained by quadruple sampling using the first layer network. This three-layer network structure can not only maintain the high-speed computing ability of the network, but also greatly improve the image detection ability of the network.

B. HUMAN KEY POINT DETECTION AND PAF NETWORK

In each stage, the classifier $\mathbf{g_x}(\cdot)$ will predict the confidence value for each key point of the human body based on the features extracted from $\mathbf{Z}$ in the image, and the information from the classifier of the previous stage from the neighborhood around each $\mathbf{Y_x}$ in the current stage. The confidence values generated by the classifier in the first stage are as follows:

$$g_1 = (x_z) \rightarrow \{b_1^p(Y_z = z)\}_{p\in\{0...P\}} \tag{1}$$

$\boldsymbol{x_z}$ in Equation (1) represents the image features extracted from the position $\mathbf{Z}$ of the input image. Equation (1) indicates that the feature information extracted by the classifier in each stage of the MEPDN network and the data obtained by the classifier in the previous stage are input to the next stage of the network. In this way, a multi-level cascaded network structure is formed.

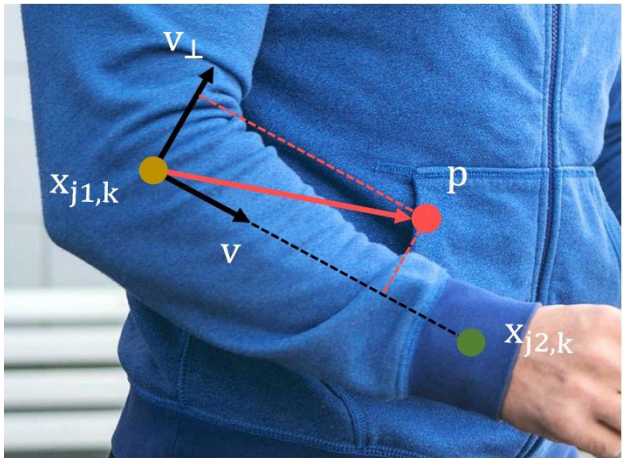


FIGURE 3. Schematic diagram of PAF vector field.

$\mathbf{b_1^p}(\mathbf{Y_p} = \mathbf{z})$ is the prediction score of the classifier $\mathbf{g_t}(\cdot)$ for the image position $\mathbf{z}$ to allocate the $\mathbf{p}$ part in the first stage. In the network, the first branch is used to generate the predictive confidence map $\mathbf{S^T}$ in each stage. The second branch is used to generate the affinity field PAFS $\mathbf{L^t}$. After the operation of each stage, the confidence map $\mathbf{S^T}$ and the affinity field PAFS $\mathbf{L^t}$ will be connected in the two branches and put into the next stage.

The image is first analyzed by initialized and fine-tuned the first 10 layers of the VGG-19, in which the feature map $\mathbf{F}$ will be generated. Then, input feature map $\mathbf{F}$ to the top branch of the network to generate a set of confidence maps $\mathbf{S^1} = \boldsymbol{\rho^1}(\mathbf{F})$. Input bottom branch to generate affinity field $\mathbf{L^1} = \boldsymbol{\varphi^1}(\mathbf{F})$, where $\boldsymbol{\rho^1}$ and $\boldsymbol{\varphi^1}$ are the deduced CNN network that deduced by the key points and affinity field of the human body in the first stage of the network. In subsequent phases, $\boldsymbol{\rho^1}$ and $\boldsymbol{\varphi^1}$, will be continuously generated from the previous phase. They will be used together with feature map F for joint computation and producing more accurate predictions in the process.

$$S^t = \rho^t(F, S^{t-1}, L^{t-1}), \quad \forall t \geq 2 \tag{2}$$

$$L^t = \varphi^t(F, S^{t-1}, L^{t-1}), \quad \forall t \geq 2 \tag{3}$$

Among them, $\boldsymbol{\rho^1}$ and $\boldsymbol{\varphi^1}$ will be used in the CNN network of stage $\boldsymbol{t}$ for inference calculation. We represent all the evaluated belief values of the p part at each image location $z = (u, v)^T$ in the image as $\boldsymbol{b_t^p} \in \boldsymbol{R^{w\times h}}$, where $\boldsymbol{w}$ and $\boldsymbol{h}$ are the width and height of the image respectively, as shown in (4).

$$b_t^p[u, v] = b_t^p(Y_p = z) \tag{4}$$

The confidence graph set of all the key points of the human body in the image is represented as $\boldsymbol{b_t} = \boldsymbol{R^{w\times h\times(P+1)}}$.

To predict the position of human key points and PAF confidence map more accurately in each iteration, we apply a loss function in two parallel branches. The loss function is applied after multi-level convolution. So the original convolution structure can be spatially weighted to solve the problem that some data sets cannot completely mark all characters.

The loss function is described as follows:

$$f_s^t = \sum_{j=1}^{J}\sum_{p} W(p)\cdot \left\| S_j^t(p) - S_j^*(p) \right\|_2^2 \quad (5)$$

$$f_L^t = \sum_{c=1}^{C}\sum_{p} W(p)\cdot \left\| L_c^t(p) - L_c^*(p) \right\|_2^2 \quad (6)$$

$$\left\| S_j^t(p) - S_j^*(p) \right\|_2^2 = \left(\sqrt[2]{\sum_j \left| S_j^t(p) - S_j^*(p) \right|^2} \right)^2 \quad (7)$$

S_j^* is the basic fact partial confidence graph, and L_c^* is the basic fact partial affinity vector field. **W** is a binary mask with $\mathbf{W(p)} = 0$ when the annotation is missing at an image location **p**. In the network, the gradients will be adding periodically to solve the problem of the disappearance of gradient. Each confidence map is a 2D representation of a specific human key point at each pixel position. Hence, if there is a human key point that should be detected in the image, the confidence of this pixel will reach a peak. When it is applied in a multi-person target scene, the relative confidence peaks are supposed to appear in each key point of each human target.

In the network, the corresponding confidence graph $\mathbf{S_{j,k}^*}$ is generated firstly for each human target **k**. Let $\boldsymbol{x_{j,k}} \in \mathbf{R^2}$ be the basic real position of the body part $\boldsymbol{j}$ of human $\boldsymbol{k}$ in the image, $\boldsymbol{x_z}$ contains all the $\boldsymbol{x_{j,k}}$. The value at the location of $\boldsymbol{p} \in \boldsymbol{R^2}$ in $\boldsymbol{S_{j,k}^*}$ is defined as:

$$S_{j,k}^*(p) = \exp\left(-\frac{\left\| p - x_{j,k} \right\|_2^2}{\sigma^2}\right) \quad (8)$$

The basic fact confidence map to be predicted by the network is a collection of confidence maps through operators, as shown in (9).

$$S_j^*(p) = \max_k S_{j,k}^*(p) \quad (9)$$

By measuring the confidence of the relationship between the key points of the human body, the MEPDN can judge which key points belong to the same person. Undoubtedly, the simplest method is to mark the midpoint of the connection of key points as the marking attributes of different human bodies. However, it may lead to deviation when body is overlapped, and resulting in misjudgment of human body structure.

To solve this problem, the PAF method is adopted in this paper. PAF is a 2D vector field of each human limb. For the pixel region of a limb belonging to the same individual, all pixels in the region will point to another key point from one key point in the region.

Let $\boldsymbol{x_{j_1,k}}$ and $\boldsymbol{x_{j_2,k}}$ be the basic real positions of body parts $\boldsymbol{j_1 j_2}$ in the body $\boldsymbol{c}$ of character $\boldsymbol{k}$. If point $\boldsymbol{p}$ is on the limb, then the value at $\boldsymbol{L_{c,k}^*(p)}$ refers to the unit vector from $\boldsymbol{j_1}$ to $\boldsymbol{j_2}$. The basic real partial affinity vector field $\boldsymbol{L_{c,k}^*}$ at image point $\boldsymbol{p}$ is defined in (10).

$$L_{c,k}^*(p) = \begin{cases} v & \text{(if p on limb c,k)} \\ 0 & \text{(otherwise)} \end{cases} \quad (10)$$

Here, $v = (x_{j_2,k} - x_{j_1,k}) \big/ \left\| x_{j_2,k} - x_{j_1,k} \right\|_2$ is the unit vector in the corresponding limb direction of the target. The point set on the limb is the point $\boldsymbol{p}$ in $0 \le v \cdot (p - x_{j_1,k}) \le l_{c,k}$ and $\left| v_\perp \cdot (p - x_{j_1,k}) \right| \le \sigma_1$, defined as the point within the segment distance threshold. Besides, the limb width $\boldsymbol{\sigma_1}$ is the distance in pixels, the limb length is $l_{c,k} = \left\| x_{j_2,k} - x_{j_1,k} \right\|_2$, and the $\mathbf{v}_\perp$ is a vector perpendicular to **v**.

The affinity force field of all people of the average image in partial groundtruth is defined as follows, where $\boldsymbol{n_c(p)}$ is the number of non-zero vectors of all k individuals at point $\boldsymbol{p}$.

$$L_c^*(p) = \frac{1}{n_c(p)} \sum_k L_{c,k}^*(p) \quad (11)$$

The correlation between the candidate component detections is measured by calculating the line integral on the corresponding PAF along the line segment connecting the position of the candidate component, as described in (12) and (13).

$$E = \int_{u=0}^{u=1} L_c(p(u)) \cdot \frac{d_{j_2} - d_{j_1}}{\left\| d_{j_2} - d_{j_1} \right\|_2} du \quad (12)$$

$$p(u) = (1-u) d_{j_1} + u d_{j_2} \quad (13)$$

For the two candidate component locations $\boldsymbol{d_{j_1}}$ and $\boldsymbol{d_{j_2}}$, the predictive component affinity field $\boldsymbol{L_c}$ is sampled along the line segment to measure their associated confidence. After executed **NMS** (Non-Maximum Suppression) operation to the predicted confidence map, a group of discrete candidate body parts can be obtained. For each part, there are multiple candidates since multiple people are in the image. A large possible limb combination can be defined by those multiple candidates. Through the above integral formula, the score of each candidate limb can be calculated.

Firstly, the discrete candidate limbs defined in (14) are obtained according to the predictive confidence map.

$$D_J = \{ d_j^m : \text{for } j \in \{1 \ldots J\}, m \in \{1 \ldots N_j\} \} \quad (14)$$

In (14), $\boldsymbol{N_j}$ is the number of candidates for part $\boldsymbol{j}$, $\mathbf{d_j^m} \in \mathbf{R^2}$ is the position of the $\boldsymbol{m}$ detection candidate points of part $\boldsymbol{j}$.

The matching goal is to connect the candidate parts and the other candidate parts of the same person. First, the variable $\boldsymbol{z_{j_1 j_2}^{mn}} \in \{0, 1\}$ is defined to indicate whether there is a connection between the two candidate parts $\boldsymbol{d_{j_1}^m}$ and $\boldsymbol{d_{j_2}^n}$. The connection set of all candidate parts is described as:

$$Z = \{ z_{j_1 j_2}^{mn} : \text{for } j_1, j_2 \in \{1, J\}, \; m \in \{1 \ldots N_{j_1}\}, \quad n \in \{1 \ldots N_{j_2}\} \} \quad (15)$$

$$\max_{Zc} E_c = \max_{Zc} \sum_{m \in D_{j_1}} \sum_{n \in D_{j_2}} E_{mn} \cdot z_{j_1 j_2}^{mn} \quad (16)$$

Then the two body parts $\boldsymbol{j_1}$ and $\boldsymbol{j_2}$, corresponding to **limb** $\boldsymbol{c}$, are considered separately, aiming to find the graph matching method with the highest total affinity value. The total affinity value is defined as follows.

C. DeepResolution-Net

We further improved the multi-level convolution network and introduced a step-by-step progressive multi-resolution

network structure with Deephigh-Net structure, as described in **Fig. 4**.

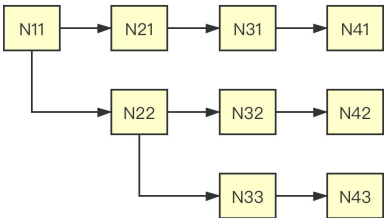


FIGURE 4. Schematic diagram of DeepResolution-Net network.

The resolution of the parallel subnetwork in the latter stage is superimposed by the resolution of the previous stage and a new network with a lower resolution. Therefore, a unit that can exchange data information in a multi-level sub-network is required. Here, we utilized a switching unit mechanism.

Fig.4 shows a schematic diagram of the structure of the multi-resolution hot spot map of the MEPDN network. We start with a main network: $N_{11} \rightarrow N_{12} \rightarrow N_{13} \rightarrow N_{14}$ (this represents the four stages of a high-resolution main network).

In each stage, we access parallel subnetworks $N_{22} \rightarrow N_{32} \rightarrow N_{42}$ with smaller resolutions that are different from those of the main network, and then on the basis of the subnetworks, we access a layer of the lowest resolution subnetworks $N_{33} \rightarrow N_{43}$ in parallel to form a three-layer parallel network structure. These three-layer structures transmit the hot spot map information at the same time, which makes the classifier get more hot spot map information and increases the accuracy of network prediction.

The number of network layers here is a non-learnable parameter. After many tests, it is determined that the three-layer network structure can maximize the balance between the computational overhead and recognition accuracy of the MEPDN network.

The switching unit mechanism is used in cross-parallel subnets, which gives each subnet the right to repeatedly accept information from other parallel subnets. Here, we give a simple example to illustrate how the switching unit works in the network. We divide the third-level network into three switching blocks, each of which is composed of three parallel convolutional network units. A switching element on these parallel units is existed, as shown in **Fig. 5**.

$\mathbf{C_{sr}^{b}}$ represents the convolution unit of the b-th switch block at the r-th resolution in the s-level network, and this convolution unit is represented as $\boldsymbol{\epsilon}_{\mathbf{s}}^{\mathbf{b}}$. As can be seen from the figure, each switching unit has $\mathbf{s}$ mapping response inputs, denoted as $\{\mathbf{X_1}, \mathbf{X_2}, \ldots, \mathbf{X_s}\}$, while the outputs maintain $\boldsymbol{s}$ mapping responses and the same resolution. The mapping relationship between the input and output of the switching unit is described in (17).

$$\mathrm{Y_k} = \sum_{\mathrm{i}=1}^{\mathrm{s}} \mathrm{a}(\mathrm{X_i}, \mathrm{k}) \tag{17}$$

At the same time, the switching units across parallel subnetworks have additional output mappings, as described in (18).

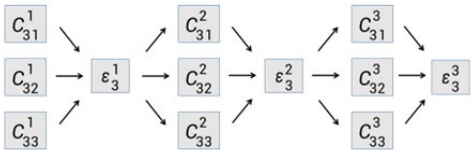


FIGURE 5. Schematic diagram of DeepResolution-Net network across neuron switching units.

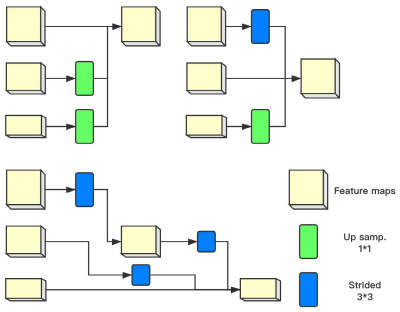


FIGURE 6. Structure diagram of cross-neuron exchange unit.

$$\mathrm{Y_{s+1}} = \mathrm{a}(\mathrm{Y_s}, \mathrm{s}+1) \tag{18}$$

The function $\mathbf{a}(\mathbf{X_i}, \mathbf{k})$ represents the set of $\mathbf{X_i}$ sampled up and down from resolution $\mathbf{i}$ to $\mathbf{k}$.

Here we simply return the heat map from the high-resolution representation of the output of the last switching unit. And use the mean square error function as the loss function to compare the predicted hot spot map with the real value. By applying the 2D Gaussian function and centering on the groudtruth position of each key point, a base truth thermal mpa with a standard deviation of 1 pixel is generated. The schematic diagram of the cross-neuron exchange unit is shown in **Fig. 6**.

As shown in **Fig.6**, for example, in the 3×3 convolution of a step switching unit, 2 times step size and 2 times downsampling is utilized. For the 3×3 convolution of two consecutive step switching units, 2 times step size and 4 times downsampling is utilized. For up-sampling, simple nearest-neighbor sampling is utilized, and then 1×1 convolution is adopted to align data channels.

IV. EXPERIMENT

In this part, experimental results and analyses of proposed work are given, tested on two mainstream datasets, namely COCO 2017 dataset and MPII dataset.

A. KEY POINT DETECTION IN COCO DATA SET

COCO data set is large and rich object detection, segmentation and subtitle data set. This data set aims at scene understanding and is mainly intercepted from complex daily scenes. The target in the image is calibrated by accurate segmentation. The images include 91 types of targets, 328000 images and 2500000 labels. By far, it has the largest dataset of semantic segmentation, containing 80 categories

TABLE 1. Coco validation set: Performance comparison of THE MEPDN network with CPN, Hourglass and other networks.

Method	Pretrain	Backbone	Input	#Params	GFLOPs
CPN[31]	Y	ResNet-50	256*192	27.0M	6.21
CPN+OHKM [31]	Y	ResNet-50	256*192	27.3M	6.21
HigherHRNet [49]	Y	HRNet-W48	256*192	63.8M	-
Hourglass[15]	N	Hourglass	256*192	25.1M	14.3
SimpleBaseline[1][26]	Y	ResNet-50	256*192	34.1M	8.92
SimpleBaseline[2][26]	Y	ResNet-101	256*192	53.3M	12.3
SimpleBseline[3][26]	Y	ResNet-152	256*192	68.3M	15.2
DSRL[47]	Y	HRNet-32	256*192	-	7.12
Ours	**Y**	**MEPDN**	**256*192**	**21.1M**	**26.3**

TABLE 2. Coco validation set: Performance comparison of the MEPDN network with CPN, Hourglass and other networks [AP].

Method	AP	$AP^{[50]}$	$AP^{[75]}$	$AP^{[M]}$	$AP^{[L]}$
CPN[31]	68.2	-	-	-	-
CPN+OHKM[31]	68.1	-	-	-	-
Hourglass[15]	66.9	-	-	-	-
HigherHRNet[25]	70.5	88.2	75.1	66.6	72.3
SimpleBaseline[1][26]	70.4	88.6	78.3	77.2	76.3
SimpleBaseline[2][26]	71.2	89.3	79.3	78.1	77.1
SimpleBseline[3][26]	72.0	89.3	79.8	78.9	77.8
MDN_3[48]	62.9	85.1	69.4	58.8	71.4
Dark[45]	68.4	88.6	77.4	66.0	74.0
Ours	**72.3**	**86.1**	**79.4**	**78.8**	**78.1**

and more than 330000 images, of which 200000 are tagged, and the number of individuals in the entire dataset is more than 1.5 million.

On the basis of COCO2017 Training data set, we supplemented the human body key point identification data in a special environment. It combined these training data with COCO2017 as the training data set of our network. In the performance evaluation of the network structure, we used val2017 data set and test-dev2017 data set. The val2017 data set contained about 5000 images and the test-dev2017 data set contained 20k images.

The Network performance standard OKS is defined in (19):

$$\mathrm{OKS} = \frac{\sum_I \exp(-d_i^2/2s^2k_i^2\sigma(v_i > 0))}{\sum_i \sigma(v_i > 0)} \tag{19}$$

OKS (Object Key point Similarity) is the similarity of key points. In the task of evaluating the key points of the human body, the quality of the key points predicted by the network is not only calculated by the simple Euclidean distance, but also by adding a certain scale to calculate the similarity between the two points.

In (19), $\boldsymbol{d_i}$ is the Euclidean distance between the detected key points and the corresponding ground reality, $\boldsymbol{v_i}$ is the real visibility sign of the ground, $\boldsymbol{s}$ is the object proportion, and $\boldsymbol{k_i}$ is the constant of each key point that controls the attenuation.

In Table 1 and Table 2, we compared the data of the MEPDN network with that of other mainstream human posture recognition networks.

In Table 1, AP refers to the average precision, that is, the average precision of each class in multi-class prediction.

TABLE 3. Coco test-dev set: Performance comparison of the MEPDN network with CPN, Hourglass and other networks.

Bottom-up:keypoint detection and grouping					
Method	AP	$AP^{[50]}$	$AP^{[75]}$	$AP^{[M]}$	$AP^{[L]}$
OPENPOSE[21]	61.8	84.9	67.5	68.2	66.5
Associative Embedding[35]	65.5	86.8	72.3	72.6	70.2
PersonLab[36]	68.7	89.0	75.4	75.5	75.4
MultiPoseNet[37]	69.6	89.3	76.6	76.3	73.5
Top-down:human detection and single-person keypoint detection					
Mask-RCNN[7]	63.1	87.3	68.7	71.4	-
CPN[31]	72.1	91.4	80.0	77.2	78.5
Integral Pose Regression[38]	67.8	88.2	74.5	74.0	-
SimpleBaseline[26]	73.7	91.9	81.1	80.0	79.0
Ours	**73.9**	**91.3**	**80.6**	**79.1**	**78.2**

In Cascade R-CNN, the equivalent values of AP [42] and AP [75] indicate that the IoU threshold of detector is greater than 0.5,0.75 (AP at OKS = 0.50,AP at OKS = 0.75), etc. $AP^{[M]}$: AP for medium objects: $32^2 <$ area $< 96^2$; $AP^{[L]}$: AP for large objects: area $> 96^2$.

During the experiments, we trained the MEPDN network from scratch, and the input size of the picture is 256×192, which is the same as that of other mainstream neural networks. Finally, the experiments get a score of 72.3% AP, which is nearly 4% higher than that of CPN. Compared with the Hourglass network, the MEPDN network has a lead of more than 5% in AP. Even comparing with the best-performed network, the SimpleBaseline network, which adopted ResNet-152 as a backbone, the proposed network also maintained a similar level of AP score.

The comparison of our network and the mainstream human posture estimation networks on the COCO test set is shown in Table 3. We can see that due to the utilization of DeepHigh-Net structure, the MEPDN network has an AP score, which is 10% higher than the OpenPose with convolutional cascade network, and outperforms other bottom-up networks included in the data. Corresponding to the top-down network, only the SimpleBaseline of a backbone with ResNet-152 can reach the same AP score comparing to ours, which demonstrated the proposed network is better than other bottom-up attitude estimation networks.

B. KEY POINT DETECTION IN MPII DATA SET

MPII is a dataset used to evaluate human posture estimates and related benchmarks, included about 25000 images and more than 40,000 people with annotated joints, which collecting images using established classifications of human activity. The dataset covered 410 human behaviors and each image provides an active tag. Each image in the dataset is from a YouTube video, and it also provides an associated unannotated framework. In addition, the annotations of the test set include body part occlusion, 3D torso and head orientation.

In the experiment, we used the standard measure, PCK^h (header to normalize the correct key probability). PCK is the percentage of the difference between the key points of the human body detected by the attitude estimation network

TABLE 4. Testing on MPII dataset.

Method	Hea	Sho	Elb	Wri	Hip	Kne	Ank	Total
Insafutdinov et al.[20]	96.8	95.2	89.3	84.4	88.4	83.4	78.0	88.5
Wei et al.[18]	97.8	95.0	88.7	84.0	88.4	82.8	79.4	88.5
Bulat et al.[30]	97.9	95.1	89.9	85.3	89.4	85.7	81.7	89.7
Newell et al.[15]	98.2	96.3	91.2	87.1	87.4	87.4	83.6	90.9
Sun et al.[39]	98.1	96.2	91.2	87.2	89.8	87.4	84.1	91.0
Tang et al.[40]	97.4	92.1	92.1	87.7	90.2	87.7	84.3	91.2
Luvizon et al.[41]	98.1	92.0	92.2	87.5	90.6	88.0	82.7	91.2
Chu et al.[32]	98.5	96.3	92.0	88.1	90.6	88.0	84.9	91.5
Chou et al.[42]	98.2	96.8	91.9	88.0	91.3	89.1	86.0	91.9
Yang et al.[43]	98.5	96.7	92.5	88.7	91.1	88.6	86.0	92.0
Ke et al.[46]	98.5	96.8	92.7	88.4	90.6	89.3	86.3	92.1
Tang et al. [2]	98.4	96.9	92.6	88.7	91.8	89.4	86.2	92.3
Dark[45]	97.2	95.9	91.2	86.7	89.7	86.7	84.0	96.0
SDA[49]	97.2	96.3	91.2	86.9	90.0	87.2	83.7	90.8
Ours	**97.6**	**95.3**	**96.1**	**87.1**	**90.2**	**89.0**	**88.7**	**91.7**

TABLE 5. #Params and GFLOPs of some top-performed methods reported in Table 4.

Method	#Params	GFLOPs	PCKh@0.5
Insafutdinov et al.[20]	42.6M	41.2	88.5
Newell et.al.[15]	25.1M	19.1	90.9
Yang et al.[43]	28.1M	21.3	92.0
Dark[45]	63.6M	32.9	90.6
Tang et al. [2]	15.5M	15.6	92.3
Ours	**21.1M**	**26.3**	**91.7**

and the normalized groundtruth of a small part of the size of the human torso. For MPII evaluation, this 100% is usually normalized using a small portion of the head size of the detected object, expressed as PCK^h.

That is, the Euclidean distance between the predicted key points and the real key points is within the range of alpha $\times$ l pixels, where alpha $= 0.5$, and l is the size of Head. Specifically, the diagonal length of the boundary box of $l = 0.6 \times$ Head is calculated in detail.

In Table 4, we can see that the MEPDN network has achieved an average improvement of nearly 5% compared with other networks at the key points with a large range of movement (such as Kne and Ank). This is mainly due to the fact that the MEPDN network uses the human key point affinity field to predict the relationship between the key points and learn their implicit spatial relationships. At the same time, the MEPDN network also achieves an average score of about 95 in the prediction of key points of the human upper limb, which is also at the top level in the comparison network.

In Table 5, we make an extended comparison of networks with higher overall scores than the MEPDN. We can see that these networks have a lead of less than 1% over the MEPDN in total scores, but an average of 36% behind in # Params.

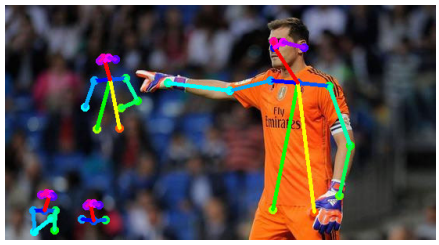
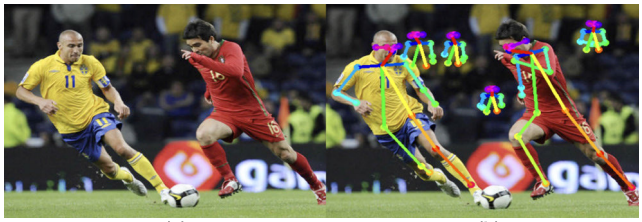


FIGURE 7. Experimental results of bust image.



(a) (b)

FIGURE 8. Experimental results of images containing two portraits:(a) Original image (b) Processed image.

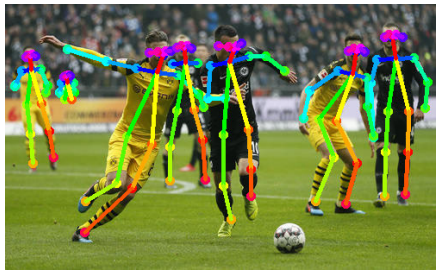


FIGURE 9. Experimental results containing multiple portraits.

In GFLOPs, the MEPDN have a lead of more than 29% in average. This is enough to show that the MEPDN is still extremely ahead of these top networks.

C. ACTUAL OPERATION RESULT

In this part, we will focus on the running results of the MEPDN network under some representative scenarios, which will more intuitively show the performance of the MEPDN network outside of the data.

In Fig.7, we choose a photo containing a single bust. In this photo, we can clearly see that all the key points of the portrait in the close range are accurately marked and correctly connected. At the same time, some relatively clear portraits in the blurred background are also captured by the MEPDN network. Because DeepResolution-Net makes the image do not lose too many detail pixels in multi-stage processing. So that these fuzzy background portraits can also be correctly identified to a certain extent.

In Fig.8, we use a close-up image made up of multiple people. We can see that the two portraits in the close range are accurately marked, and some relatively clear fuzzy portraits in the background are also marked.

In Fig.9, we increase the number of portraits in close-up images. When the MEPDN network is used to process the image, due to the use of a multi-level convolution neural network, this bottom-up human posture network will not increase the computing time with the increase of the number

of portraits in the image, so we do not have more computational overhead in different image processing.

V. CONCLUSION

In this paper, we proposed a human posture recognition network that combined a multi-stage concatenated convolution neural network and step-by-step progressive multi-resolution main network. This network finds a balance between computational cost and recognition accuracy. Finally it becomes the current the MEPDN network. The recognition accuracy of this network model is higher than that of most mainstream networks in high complex backgrounds. And it still maintains a low computing overhead, which makes it easy for the network to run on devices with ordinary computing power. Of course, the network still has a lot of promotion space to improve.

At present, the MEPDN network is a forward propagation non-end-to-end network, the output of each stage convolution will be the input information of the next stage. The network identification results of key points will be continuously optimized in the multi-stage optimization, and the prediction results of each stage will be extracted and optimized in the next stage. In theory, the use of end-to-end network structure may optimize the prediction results of the network in each stage. But in terms of overall performance, the network already has the ability to gradually optimize the test results in multiple stages. The introduction of end-to-end architecture may not improve the overall performance of the network, and this back propagation may lead to an increase in network computing overhead. There is no doubt that this is a very interesting idea, and we will explore how to introduce end-to-end architecture into the MEPDN networks in the future. And we will be deeply interested in how to control computing overhead in this case.

REFERENCES

[1] X. Li, H. Wang, L. Yi, L. J. Guibas, A. L. Abbott, and S. Song, "Category-level articulated object pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 3703–3712, doi: 10.1109/CVPR42600.2020.00376.

[2] W. Tang, P. Yu, and Y. Wu, "Deeply learned compositional models for human pose estimation," in *Proc. ECCV*, Sep. 2018, pp. 190–206.

[3] Y. Liu, K. Chen, C. Liu, Z. Qin, Z. Luo, and J. Wang, "Structured knowledge distillation for semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 2599–2608, doi: 10.1109/CVPR.2019.00271.

[4] C. Liu *et al.*, "Auto-DeepLab: Hierarchical neural architecture search for semantic image segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 82–92, doi: 10.1109/CVPR.2019.00017.

[5] H. Eun, J. Moon, J. Park, C. Jung, and C. Kim, "Learning to discriminate information for online action detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 806–815, doi: 10.1109/CVPR42600.2020.00089.

[6] C.-Y. Wu, R. Girshick, K. He, C. Feichtenhofer, and P. Krähenbühl, "A multigrid method for efficiently training video models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 150–159, doi: 10.1109/CVPR42600.2020.00023.

[7] K. He, G. Gkioxari, P. Dollar, and R. B. Girshick, "Mask R-CNN," in *Proc. ICCV*, 2017, pp. 2980–2988.

[8] W. J. Johnson, "Adapting Mask-RCNN for automatic nucleus segmentation," May 2018, *arXiv:1805.00500*. [Online]. Available: https://arxiv.org/abs/1805.00500

[9] Z. Huang, Z. Zhong, L. Sun, and Q. Huo, "Mask R-CNN with pyramid attention network for scene text detection," Nov. 2018, *arXiv:1811.09058*. [Online]. Available: https://arxiv.org/abs/1811.09058

[10] M. Xu, C. Zhao, D. S. Rojas, A. Thabet, and B. Ghanem, "G-TAD: Sub-graph localization for temporal action detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 10153–10162, doi: 10.1109/CVPR42600.2020.01017.

[11] Y. Li, Z. Wang, L. Wang, and G. Wu, "Actions as moving points," in *Proc. ECCV*. Cham, Switzerland: Springer, Oct. 2020, pp. 68–84.

[12] J. Tang, J. Xia, X. Mu, B. Pang, and C. Lu, "Asynchronous interaction aggregation for action detection," in *Proc. ECCV*. Cham, Switzerland: Springer, Nov. 2020, pp. 71–87.

[13] J. Wu, Z. Kuang, L. Wang, W. Zhang, and G. Wu, "Context-aware RCNN: A baseline for action detection in videos," in *Proc. ECCV*. Cham, Switzerland: Springer, Nov. 2020, pp. 440–456.

[14] Y. Cai, Z. Wang, Z. Luo, B. Yin, A. Du, H. Wang, X. Zhang, X. Zhou, E. Zhou, and J. Sun, "Learning delicate local representations for multi-person pose estimation," in *Proc. ECCV*. Cham, Switzerland: Springer, Dec. 2020, pp. 455–472.

[15] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *Proc. ECCV*, 2016, pp. 483–499.

[16] A. Toshev and C. Szegedy, "DeepPose: Human pose estimation via deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 1653–1660, doi: 10.1109/CVPR.2014.214.

[17] H. Fang, S. Xie, Y. Tai, and C. Lu, "RMPE: Regional multi-person pose estimation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2017, pp. 2353–2362, doi: 10.1109/ICCV.2017.256.

[18] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional pose machines," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4724–4732.

[19] V. Belagiannis and A. Zisserman, "Recurrent human pose estimation," in *Proc. 12th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, 2017, pp. 468–475, doi: 10.1109/FG.2017.64.

[20] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, and B. Schiele, "DeeperCut: A deeper, stronger, and faster multi-person pose estimation model," in *Proc. ECCV*, 2016, pp. 34–50.

[21] Z. Cao, S.-E. Wei, Y. Sheikh, and T. Simon, "Realtime multi-person 2D pose estimation using part affinity fields," Apr. 2017, *arXiv:1611.08050*. [Online]. Available: https://arxiv.org/abs/1611.08050

[22] L. Pishchulin, E. Insafutdinov, S. Tang, B. Andres, M. Andriluka, P. V. Gehler, and B. Schiele, "DeepCut: Joint subset partition and labeling for multi person pose estimation," Apr. 2016, *arXiv:1511.06645*. [Online]. Available: https://arxiv.org/abs/1511.06645

[23] L. Pishchulin, A. Jain, M. Andriluka, T. Thormählen, and B. Schiele, "Articulated people detection and pose estimation: Reshaping the future," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 3178–3185, doi: 10.1109/CVPR.2012.6248052.

[24] L. Pishchulin, M. Andriluka, P. Gehler, and B. Schiele, "Strong appearance and expressive spatial models for human pose estimation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 3487–3494, doi: 10.1109/ICCV.2013.433.

[25] 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/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 5385–5394, doi: 10.1109/CVPR42600.2020.00543.

[26] B. Xiao, H. Wu, and Y. Wei, "Simple baselines for human pose estimation and tracking," in *Proc. ECCV*, 2018, pp. 472–487.

[27] Y. Yang and D. Ramanan, "Articulated pose estimation with flexible mixtures-of-parts," in *Proc. CVPR*, Jun. 2011, pp. 1385–1392.

[28] L. Pishchulin, M. Andriluka, P. Gehler, and B. Schiele, "Poselet conditioned pictorial structures," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 588–595.

[29] X. Chu, W. Ouyang, H. Li, and X. Wang, "Structured feature learning for pose estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4715–4723.

[30] A. Bulat and G. Tzimiropoulos, "Human pose estimation via convolutional part heatmap regression," in *Proc. ECCV* (Lecture Notes in Computer Science), vol. 9911. Cham, Switzerland: Springer, 2016, pp. 717–732.

[31] Y. Chen, Z. Wang, Y. Peng, Z. Zhang, G. Yu, and J. Sun, "Cascaded pyramid network for multi-person pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7103–7112, doi: 10.1109/CVPR.2018.00742.

[32] X. Chu, W. Yang, W. Ouyang, C. Ma, A. L. Yuille, and X. Wang, "Multi-context attention for human pose estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5669–5678.

[33] G. Ning and H. Huang, "LightTrack: A generic framework for online top-down human pose tracking," May 2019, *arXiv:1905.02822*. [Online]. Available: https://arxiv.org/abs/1905.02822
[34] C.-Z. Guan, "Realtime multi-person 2D pose estimation using ShuffleNet," in *Proc. 14th Int. Conf. Comput. Sci. Educ. (ICCSE)*, Aug. 2019, pp. 1302–1310.
[35] A. Newell, Z. Huang, and J. Deng, "Associative embedding: End-to-end learning for joint detection and grouping," in *Proc. NIPS*, 2017, pp. 2274–2284.
[36] 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. ECCV*. Cham, Switzerland: Springer, Sep. 2018, pp. 282–299.
[37] M. Kocabas, S. Karagoz, and E. Akbas, "MultiPoseNet: Fast multi-person pose estimation using pose residual network," in *Proc. ECCV*, in Lecture Notes in Computer Science, vol. 11215. Cham, Switzerland: Springer, Oct. 2018, pp. 437–453.
[38] X. Sun, B. Xiao, F. Wei, S. Liang, and Y. Wei, "Integral human pose regression," in *Proc. ECCV*, 2018, pp. 536–553.
[39] K. Sun, C. Lan, J. Xing, W. Zeng, D. Liu, and J. Wang, "Human pose estimation using global and local normalization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5600–5608.
[40] Z. Tang, X. Peng, S. Geng, L. Wu, S. Zhang, and D. N. Metaxas, "Quantized densely connected U-nets for efficient landmark localization," in *Proc. ECCV*, 2018, pp. 348–364.
[41] D. C. Luvizon, H. Tabia, and D. Picard, "Human pose regression by combining indirect part detection and contextual information," *Comput. Graph.*, vol. 85, pp. 15–22, 2019.
[42] C. Chou, J. Chien, and H. Chen, "Self-adversarial training for human pose estimation," in *Proc. Asia–Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA ASC)*, Nov. 2018, pp. 17–30, doi: 10.23919/APSIPA.2018.8659538.
[43] W. Yang, S. Li, W. Ouyang, H. Li, and X. Wang, "Learning feature pyramids for human pose estimation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1290–1299.
[44] L. Ke, M. Chang, H. Qi, and S. Lyu, "Multi-scale structure-aware network for human pose estimation," in *Proc. ECCV*. Cham, Switzerland: Springer, Oct. 2018, pp. 731–746.
[45] F. Zhang, X. Zhu, H. Dai, M. Ye, and C. Zhu, "Distribution-aware coordinate representation for human pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 7091–7100, doi: 10.1109/CVPR42600.2020.00712.
[46] L. Ke, H. Qi, M. Chang, and S. Lyu, "Multi-scale supervised network for human pose estimation," in *Proc. 25th IEEE Int. Conf. Image Process. (ICIP)*, 2018, pp. 564–568, doi: 10.1109/ICIP.2018.8451114.
[47] L. Wang, D. Li, Y. Zhu, L. Tian, and Y. Shan, "Dual super-resolution learning for semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 3773–3782, doi: 10.1109/CVPR42600.2020.00383.
[48] A. Varamesh and T. Tuytelaars, "Mixture dense regression for object detection and human pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 13083–13092, doi: 10.1109/CVPR42600.2020.01310.
[49] Y. Bin, X. Cao, X. Chen, Y. Ge, Y. Tai, C. Wang, J. Li, F. Huang, C. Gao, and N. Sang, "adversarial semantic data augmentation for human pose estimation," in *Proc. ECCV*. Cham, Switzerland: Springer, Nov. 2020, pp. 606–622.

MENGLONG WU received the Ph.D. degree in communications and information systems from the Beijing University of Posts and Telecommunications, China, in 2015. He is currently an Associate Professor with the School of Information Science and Technology, North China University of Technology, Beijing. His research interests include wireless communication, signal processing, machine learning, and neural networks.

DEXUAN DU received B.S. degree in electronic and information engineering from the North China University of Technology, in 2018, where he is currently pursuing the degree in electronics and communication engineering.

YUNDONG LI (Member, IEEE) received the Ph.D. degree in communication and information system from Beihang University, Beijing, China, in 2005. He is currently an Associate Professor with the School of Information Science and Technology, North China University of Technology, Beijing. His current research interests include machine learning, neural networks, computer vision, and building damage assessment.

WENLE BAI was born in Shanxi, China, in 1967. He received the Ph.D. degree in communication engineering from the Beijing University of Posts and Telecommunication, China, in 2006. He is currently working as a Professor with the North China University of Technology. His research interests include wireless communication, statistical signal processing, and multi-user communication. He has published about 20 articles in the related areas.

WENKAI LIU received the Ph.D. degree from the Institute of Semiconductors, Chinese Academy of Sciences, in 2002. He is currently a Professor with the North China University of Technology. His research interests include optical signal processing, machine learning, and neural networks.