A Lightweight Top-Down Multi-Person Pose Estimation Method Based on Symmetric Transformation and Global Matching

YANGYANG LI\textsuperscript{1}, (Senior Member, IEEE), DANQING YANG\textsuperscript{1}, YANQIAO CHEN\textsuperscript{2}, CHENG PENG\textsuperscript{1}, ZHENXIANG SUN\textsuperscript{1}, AND LICHENG JIAO\textsuperscript{1}, (Fellow, IEEE)

\textsuperscript{1}Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education, School of Artificial Intelligence, Xidian University, Xi’an 710071, China

\textsuperscript{2}54th Research Institute of China Electronics Technology Group Corporation, Shijiazhuang 050081, China

Corresponding author: Yanqiao Chen (chenyanqiao2016@163.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 61772399 and Grant 62101517, and in part by the Key Research and Development Program in Shaanxi Province of China under Grant 2019ZDLGY09-05.

ABSTRACT

The top-down human pose estimation method usually faces the following problems: (i) The target detection result is not well applied in the pose estimation network. (ii) Difficulty of human detection in the crowded state. (iii) The complicated model leads to a long training time. Aiming at the issues above, a lightweight multi-person pose estimation method based on symmetric transformation and global matching is proposed. Symmetric transformation module adds spatial transformation network (STN) and spatial de-transformer network (SDTN) before and after the single-person pose estimation (SPPE) to extract high-quality single-person pose regions from inaccurate human candidate frames. Global matching method is used to transform the key point prediction problem into the optimal matching problem of the human body-key point graph, and solve the questions of misjudgment and error detection of pose estimation in the crowded state. Finally, depth wise separable convolution and inverted residual model are used to reduce the complexity of model, so as to improve the running speed while balancing the accuracy of the algorithm. Experiments show that the algorithm proposed in this paper not only enhance the overall performance of the multi-person posture estimation network in the crowded state, but also improves the running speed significantly, which further confirms the effectiveness and competitiveness of this algorithm.

INDEX TERMS

Human pose estimation, symmetric transformation, depthwise separable convolution, global matching method.

I. INTRODUCTION

Human pose estimation [1], [40], one of the basic research tasks of computer vision [2], has always attracted the attention of researchers, which is widely used in security monitoring, movie action special effects modeling, human behavior prediction tasks, human-computer interaction, virtual reality etc. Generally speaking, human pose estimation is defined as the positioning of human joints (also known as key points) in images or videos, and it is also defined as searching for a specific pose in the space of all joint poses.

According to different task test procedures, there are generally two research methods for multi-person pose estimation. One is the top-down research idea: firstly, each person is detected by the target detector to obtain the candidate frame of the human body, and then the candidate frame of the human body is sent to the single-person pose estimation network to obtain the result. Another research idea is the bottom-up approach: first the network is used to detect all the key points, and then judge which person each key point belongs to. At present, in the research of top-down multi-person human pose estimation [3]–[10], there are the following difficulties and challenges:

(1) The detection frame obtained by the target detection algorithm is deemed to be correct within the threshold range, but in fact it is interfered by other human body parts and even the target body structure is incomplete. When the detection frame is input to the subsequent human pose estimation, it adds difficulty to the work of the single-person pose estimation detector, resulting in false detection or detection.

The associate editor coordinating the review of this manuscript and approving it for publication was Mu-Yen Chen.
(2) It is difficult to detect human posture in a crowded state. In natural scenes, people may block each other and key points may overlap, which makes crowded crowds seem to be inevitable. In the follow-up key point prediction process, it is very easy to cause misjudgment and missed judgments.

(3) The model is complex and needs comparatively high training cost. The top-down processing method relies on the convolutional neural network to improve the processing results, and the depth of the convolutional neural network will significantly affect the performance. The deeper the network, the stronger its characterization ability, but this leads to more complex network models while pursuing performance.

In response to the above-mentioned difficulties and challenges, we have proposed corresponding solutions, which mainly include the following three aspects:

(1) Aiming at the problems of poor application effect of target detection results, complex model and long training time, a lightweight multi-person pose estimation network based on symmetric transformation is proposed. This method uses the human pose estimation network of spatial transformation [11], which adds a symmetric transformation module in the human detection, to extract high-quality single-person pose regions from inaccurate human candidate frames. This method improves the performance of the target detection network, so the single-person pose estimation network can obtain high-quality candidate frames. At the same time, taking advantage of the lightweight characteristics of the MobileNet [12], [13], using deep-separated convolution and inverted residual models to reduce the complexity of the network, so as to balance the accuracy while improving the speed of the algorithm.

(2) Aiming at the problem of difficult human body detection in crowded crowds, this paper proposes a global matching method by optimizing the loss function to solve the problems of misjudgment and missed detection of key points in the crowded state. The global matching method uses the Kuhn-Munkres(KM) matching algorithm [14] to transform the key point prediction problem into an optimal matching problem of the human body-key point graph.

(3) This paper has conducted a lot of experiments on the MPII [38] and MS COCO [34] dataset. The experimental results show that the algorithm proposed in this paper not only improves the quality of the human candidate frames and the overall performance of the multi-person post estimation network in the crowded state, but also improves the training speed while ensuring the performance, which further confirms the effectiveness and competitiveness of the algorithm in this paper.

II. RELATED WORK
A. CLASSIFICATION AND REGRESSION PROBLEMS

For classification and regression solving methods, the sparse probability regression scheme proposed by Raquel Urtasun and Trevor Darrell [15] is used to estimate the posture of the independently active human body. They developed an online method of GP sparsity, mainly by concentrating the local regression quantity at each test point, avoiding the inherent boundary problems of the offline method.

B. PROBABILISTIC GRAPHICAL MODEL

The traditional method of pose estimation task generally uses graph structure model, which mainly includes two models, one is tree model [16]–[18], the other is random forest model [19], [20]. The Multiple tree models proposed by Y. Wang and G. Mori [16] alleviate the limitations of a single tree structure model by combining the information provided across different tree models to carry out human pose estimation research. By locally capturing the set of constraints and trees that constitute the various parts of the model, a set of global constraints can be modeled, and easy-to-process algorithm learning and reasoning in these multi-tree models are provided.

C. DEEP LEARNING MODEL

With the development of deep learning, the effect of neural networks is getting better and better. So far, many basic deep learning frameworks have been improved and evolved based on these classic models to achieve specific needs, such as VGG [21], residual network ResNet [22] and InceptionNet [23] etc. After 2014, a large number of scholars began to introduce deep learning into human pose estimation research [1], so various models based on deep neural networks [24], [25] and convolutional neural networks [26], [27] began to appear. Its representative works include DeepPose [28], etc. In 2015, Flowing ConvNets [29] process the estimation problem as a target detection task and obtain the output result of heatmap to improve the robustness of the key point detection. This network mainly uses the spatial fusion model to extract the internal connections of the key points, but its shortcomings are also obvious: for example, the detection range is limited. The Convolutional Pose Machines(CPM) [30] proposed in 2016 has strong robustness. The main contribution of the network is to express the texture and spatial information of the human body’s posture through the mode of architecture. However, it costs a lot of training, real-time detection cannot be achieved. In the same year, Stacked Hourglass Networks [31] used convolutional neural networks and Residual Modules to get the top spot in the MPII analysis competition, which provided a basic model and research idea for later researchers.

With the continuous improvement of single-person pose estimation methods, multi-person pose estimation has begun to attract attention. The Deepcut [32] network proposed in 2016 uses a convolutional neural network to find all candidate joint points, and then cluster the nodes to determine which person each node belongs to, and mark each point to classify which part of the body belongs to. In 2017, He et al. proposed Mask R-CNN [5] based on Faster R-CNN [33], which added a substructure to the region of interest for prediction segmentation. Chen et al. [4] proposed a Cascade Pyramid Network (CPN) to refine the
process of pose estimation, which is the winner of COCO 2017 keypoint challenge. SimpleBaseline [9] consists of a deep backbone network and several deconvolutional layers. It can achieve pretty good performance on the COCO benchmark [34], although it is based on simple network architecture. Sun et al. [6] proposed a High-Resolution Network (HRNet) in 2019 achieved state-of-the-art performance, which could maintain high-resolution representations by connecting high-to-low resolution convolutions in parallel and repeatedly conducting multi-scale fusions. In next year, they present HigherHRNet [35] by using high-resolution feature pyramids, which can localize keypoints more precisely, especially for small person.

Most of these works [39], [47]–[52] focus on how to design the pose estimation network to improve the model accuracy, while ignoring the overall process in the practice of multi-person pose estimation task. The generation stage of human candidate frame has not been widely concerned. At the same time, this kind of method often needs expensive computing cost and is difficult to be applied in real time.

D. LIGHTWEIGHT POSE ESTIMATION
With the growing prosperity of multi-person pose estimation methods, networks are becoming more and more complex. Researchers are beginning to think about how to ensure the accuracy while reducing the complexity and improving the running speed of the model. Rafi et al. [36] proposed an efficient deep network architecture that can be trained efficiently on mid-range GPUs, but they didn’t conduct quantitative experiments on the model efficiency. Bulat et al. [37] binarized the network architecture for model compression and execution speedup to accommodate resource-limited platforms, which however suffers performance drop significantly. Simple baseline [9] proposed by Microsoft is a model of high precision and lightweight. Cheng et al. [10] proposed a Lightweight Pose Network (LPN) in 2020 following the architecture design principles of SimpleBaseline, which achieved higher running speed.

III. PROPOSED APPROACH
In this section, a lightweight multi-person pose estimation model based on symmetric transformation will be introduced, and the global matching method is used to transform the key point prediction problem into the optimal matching problem of the human body-key point graph, which further improves the overall performance of the model.

A. ARCHITECTURE OF THE PROPOSED MODEL
The architecture of the proposed model is shown in Fig. 1. Our framework is mainly composed of three networks: human detection network, symmetric transformation network and single-person pose estimation (SPPE) network. The whole process of the algorithm is divided into four steps: the first step is to use the Faster R-CNN [33] network of target detector to detect the human body, and use MobileNet V2 [13] as its backbone to obtain multiple candidate frames of the human body. The second step is to input the candidate frame of the human body into the symmetric transformation module. After affine transformation, a relatively clear and high-quality input suitable for a single-person pose estimator is generated. The third step is to input the processed human pose candidate frame into the SPPE, and the output result of SPPE is inversely transformed by using symmetrical transformation module to change the coordinates into the original coordinates, then output candidate key points to generate a human body-key point graph. The fourth step is to match the human body-key point graph with the global matching method to obtain the final multi-person pose estimation result.

1) SYMMETRIC TRANSFORMATION MODULE
This module uses the idea of spatial transformation network (STN) [11] to extract high-quality human candidate frames. STN as shown in Fig. 2(a), can be perfectly embedded in any standard neural network structure to provide spatial transformation capabilities. Unlike the pooling layer where the receiving domain is a fixed and localized layer, this network is an adaptive module mechanism. It adapts the transformation of the input data, can spontaneously transform the image or feature map in space and then perform the transformation on the entire feature map (non-local). This allows the network containing the space deformer to not only select the most relevant areas in the image, but also to convert these areas into standardized and expected poses, to simplify the subsequent recognition.

Symmetric transformation module uses STN to affine transform the output result of target detection to generate a higher quality human candidate frame. The formula of affine transformation is as follows:

\[
\begin{bmatrix}
X \\
Y
\end{bmatrix}
= T_\theta
\begin{bmatrix}
x \\
y
\end{bmatrix}
= \begin{bmatrix}
\theta_{11} & \theta_{12} & \theta_{13} \\
\theta_{21} & \theta_{22} & \theta_{23}
\end{bmatrix}
\begin{bmatrix}
x \\
y
\end{bmatrix}
\]

(1)

where \( T_\theta \) represents the affine transformation matrix, \( \theta_{11}, \theta_{12}, \theta_{21}, \theta_{22} \) are used to determine the rotation angle, \( \theta_{13}, \theta_{23} \) are used to determine the height and width of translation. \([X, Y]^T\) represents original coordinates, \([x, y]^T\) represents affine transformation coordinates. Similarly, the affine transformed coordinates also can be inverse transformed into original coordinates according to the affine transformation parameters, the inverse transformation formula is as follows:

\[
\begin{bmatrix}
X' \\
Y'
\end{bmatrix}
= T_\gamma
\begin{bmatrix}
x' \\
y'
\end{bmatrix}
= \begin{bmatrix}
\gamma_{11} & \gamma_{12} & \gamma_{13} \\
\gamma_{21} & \gamma_{22} & \gamma_{23}
\end{bmatrix}
\begin{bmatrix}
x' \\
y'
\end{bmatrix}
\]

(2)

where \( T_\gamma \) can be determined by \( T_\theta \), the formula is as follows:

\[
\begin{bmatrix}
\gamma_{11} & \gamma_{12} \\
\gamma_{21} & \gamma_{22}
\end{bmatrix} = \begin{bmatrix}
\theta_{11} & \theta_{12} \\
\theta_{21} & \theta_{22}
\end{bmatrix}^{-1}
\]

(3)

\[
\begin{bmatrix}
\gamma_{13} \\
\gamma_{23}
\end{bmatrix} = -1 \times \begin{bmatrix}
\gamma_{11} & \gamma_{12} \\
\gamma_{21} & \gamma_{22}
\end{bmatrix} \begin{bmatrix}
\theta_{13} \\
\theta_{23}
\end{bmatrix}
\]

(4)
That means, spatial de-transformer network (SDTN) as shown in Fig. 2(b), which is the inverse transformation network of STN, can be constructed to restore the human candidate frame. STN and SDTN form the symmetrical transform module together. This module does not require calibration of key points. It can perform rotation, translation, enlargement and reduction, and other geometric or spatial transformations on the input data according to different target tasks. At the same time, when the input images of the network have large spatial differences, this network can be embedded in the existing convolutional neural network to improve the accuracy of classification.

2) BACKBONE

In this paper, the backbone in Faster R-CNN is replaced by MobileNet V2 [12], which is improved on the basis of MobileNet V1 [13]. In addition to the depth separable convolution, MobileNet V2 adds an inverted residual block. Compared with the residual module, the depth convolution in the middle is wider, and except for the initial $1 \times 1$ convolution of ascending dimension, the number of $1 \times 1$ convolution channels for shortcut connections is less, and it is in an inverted state.

In this paper, the stacked hourglass network is used as the backbone of SPPE. It can capture the characteristics of images at multiple scales, which is very suitable for human pose estimation tasks. The basic unit in the stacked hourglass network is the residual module, and its structure is shown in Fig. 3(a). In order to reduce the amount of parameters of the model, this paper proposes a linear bottleneck in Fig. 3(b) to replace the standard $3 \times 3$ convolution with a $3 \times 3$ depthwise convolution in residual module. At the same time, because the ReLu activation function will filter out a lot of useful information in low dimensional features, we replace the reduced ReLu activation function with a linear activation function to reduce the loss of information.

B. GLOBAL MATCHING METHOD

In the state of dense crowds, when one person is close to another, there is a high probability that the network will only cut one person, and when the key points of the two people overlap, only one key point will be detected. As can be seen that, this approach is very unreasonable. According to the above problems, this paper designs the global matching method to transform the key point prediction problem into an optimal matching problem of the human body-key point graph based on KM algorithm, and the process is as follows in Algorithm 1:

1) LOSS FUNCTION

The traditional human body pose estimation network relies heavily on the results of human body detection. Its task goal is to locate and identify target joint candidate points according to a given body candidate frame, and the interference key points are generally directly suppressed to 0. But if the interference key point is mistaken as the target key point, and the real target point is treated as an interference point, which is suppressed to 0. It will be an irreversible error, and the missing key point will not be input to the subsequent processing steps.

The improved loss function proposed in this paper is to solve this problem. Using this loss function, a multi-peak heat map can be predicted, then the possible key points can be set as candidate key points, and the global candidate...
is a set \( C \) \( \in \mathbb{P} \), the standard deviation of the distribution is \( \sigma \), and the center is \( p^k \).

Then for the interference key points, we establish it as a set \( \Omega_i^k \), and the heat map of the interference key points is \( C_i^k \), which is composed of a mixed Gaussian distribution \( \sum_{p \in \Omega_i^k} G(p|\sigma) \). The final loss function is as follows:

\[
\text{Loss}_i = \frac{1}{K} \sum_{k=1}^{K} \text{MSE}[P_i^k, T_i^k + \mu C_i^k] \tag{5}
\]

where \( \mu \) is an attenuation coefficient of \([0,1] \). This model takes 0.5. The general heatmap loss function can be regarded as a special case of \( \mu=0 \).

2) HUMAN BODY-KEY POINT MATCHING GRAPH

Since highly overlapping human candidate frames tend to predict the same actual joints, in the actual network detection process, multiple key points will be considered as one key point, as shown in Fig. 4.

According to the loss function given in formula (5), a multi-peak Gaussian heatmap of a group of candidate key points can be obtained. First these candidate key points is grouped to obtain a set of candidate key points representing the same actual key point. Through high-quality key point prediction results, candidate key points representing the same key point are usually very close to each other. Therefore, these candidate key points can be grouped as follows, if the formula (6) is satisfied, they can be marked as the same point:

\[
\|P_i^k - p^k\|_2 \leq \text{min}(u_1^k, u_2^k) \delta^k \tag{6}
\]

where \( P_i^k \) and \( p^k \) represent two different key points, \( u_1^k, u_2^k \) is the response deviation of the two candidate key points on the Gaussian heatmap, and \( \delta^k \) is the parameter of the control deviation of the \( k \)-th key point. The reason for using \( \text{min}(u_1^k, u_2^k) \) instead of a constant threshold here is to ensure that: if and only if \( P_i^k \) and \( p^k \) fall into their respective control domains at the same time, they will be classified as a point, and a point can represent a group of points in the above-mentioned way.

Now by marking a set of points as a point, a set of candidate key points can be obtained:

\[
J = \{v_j^k: \text{for } k \in \{1, \ldots, K\}, j \in \{1, \ldots, N_k\}\} \tag{7}
\]

where \( k \) is the number of human joints, which is 17 in the MS COCO dataset. \( N_k \) is the number of candidate key points for a joint \( k \), and \( v_j^k \) is the \( j \)-th key point corresponding to joint \( k \).

Then a human body set is marked as:

\[
H = \{h_i: \forall i \in \{1, \ldots, M\}\} \tag{8}
\]

where \( h_i \) represents the \( i \)-th human body, and \( M \) is the total number of detected human candidate frames. The relationship between sets \( J \) and \( H \) is shown in Fig. 5.

After obtaining a human body set and a key point set, they can be connected together to form a human body-key point graph. If a key point \( v_j^k \) contains a candidate key point from the human body point \( h_i \), then an edge \( e_{ij}^k \) is established between them, and its weight is the response value of the candidate key point, denoted by \( w_{ij}^k \). In this case, an edge set \( \varepsilon = \{e_{ij}^k: \forall i, j, k\} \) was constructed. Finally, the human body-key point graph is expressed as:

\[
G = (H, J, \varepsilon) \tag{9}
\]
3) OBJECTIVE FUNCTION
After the above preparations are completed, the task of human pose estimation can be transformed into solving the optimal matching problem of the human body-key point graph, that is, the problem of maximizing the weight of all edges. The objective function is:

$$\max G_d^G = \sum_{i,j,k} w_{i,j}^k \cdot d_{i,j}^k$$

subject to

$$\sum_{j} d_{i,j}^k \leq 1, \quad \forall k \in \{1, \ldots, K\}, \forall i \in \{1, \ldots, M\}$$

$$\sum_{i} d_{i,j}^k \leq 1, \quad \forall k \in \{1, \ldots, K\}, \forall j \in \{1, \ldots, N_k\}$$

$$d_{i,j}^k \in \{0,1\} \quad \forall i, j, k$$

(10)

where $d_{i,j}^k$ indicates whether to keep the edge $e_{i,j}^k$ in the final graph, and the constraint conditions ensure that each human body can only match at most one $k$-th node. Finally, the KM algorithm is used to solve the optimal matching problem.

IV. EXPERIMENTS
A. DATASETS AND SETTINGS
This paper uses two datasets for evaluation, i.e., MPII and MS COCO.

1) MPII
MPII [38] is a dataset for human pose estimation tasks. The dataset contains about 25,000 images. In these images, there are more than 40,000 people with annotated human joint points. These images are collected systematically through daily human activities. In general, the dataset contains 410 human activities, and each image has a corresponding activity tag. The source of the image is a YouTube video. The entire dataset is divided into 25k training dataset samples and 3k validation dataset samples, which are divided into 16 types of labels(0-r ankle, 1-r knee, 2-r hip, 3-l hip, 4-l knee, 5-l ankle, 6-pelvis, 7-thorax, 8-upper neck, 9-head top, 10-r wrist, 11-r elbow, 12-r shoulder, 13-l shoulder, 14-l elbow, 15-l wrist). Each sample contains label information such as the identification ID, the coordinate information of the center point, and the real coordinate information of the human body joint points. The evaluation index for the MPII is PCK (Percentage of Correct Keypoints), and the specific calculation formula is as follows:

$$PCK^k_i = \frac{\sum_p \delta\left(\frac{d_{p}^i}{d_{p}^{def}} \leq T_k\right)}{\sum_p 1}$$

(11)

where $i$ represents the key point with ID $i$, $k$ represents the $k$-th threshold $T_k$, $p$ is the $p$-th person, $d_{p}^i$ represents the Euclidean distance between the predicted value of the key point with ID $i$ in the $p$-th person and the manually labeled value. $d_{p}^{def}$ is the scale factor of the $p$-th person, and different public datasets use different calculation methods for this factor. MPII uses the current person’s head diameter as the scale factor, that is, the Euclidean distance between the upper left point LT of the head Rect and the lower right point RB. $T_k$ is a manually set threshold, $T_k \in [0 : 0.01 : 0.1]$.

2) MS COCO
MS COCO is an image dataset created by Microsoft Corporation [34]. The images contain a total of 91 types...
of targets, 328,000 images and 2,500,000 annotation tags. The scene of this dataset is complex, and the instances in each image are accurately positioned and labeled. There are 80 categories and more than 330,000 images, 200,000 of which are labeled. The number of individuals in the entire dataset exceeds 1.5 million, and there are a total of 17 human key points (0-nose, 1-Leye, 2-Reye, 3-Lear, 4-Rear, 5-Lsho, 6-Rsho, 7-Lelb, 8-Relb, 9-Lwri, 10-Rwri, 11-Lhip, 12-Rhip, 13-Lkne, 14-Rkne, 15-Lank, 16-Rank), including training set (20,000 images and 25,000 human bodies with key points), validation set (15,000 human bodies and 1.7 million key points marked) and a test set (a challenge set of 20000 and a development set of 20000).

The evaluation criteria for the MS COCO mainly use average precision(AP) and the average recall rate(AR). Here a new variable is defined as a similarity measure, namely object keypoint similarity(OKS). The main idea of OKS is to calculate the weighted distance between the predicted positions of the human body joint points and the real positions of the joint points. The formula is as follows:

$$OKS_p = \frac{\sum_{i} \exp \left( -\frac{d_{pi}^2}{2\sigma_i^2} \right) \delta \left( v_{pi} = 1 \right)}{\sum_i \delta \left( v_{pi} = 1 \right)}$$

where \(p\) represents the \(p\)-th person, \(i\) represents the ID of the human joint point, \(\delta\) represents the normalization factor, which is obtained by calculating the standard deviation of the offset of the manually marked position, \(v_{pi}\) represents the state of the \(i\)-th human body joint points of the \(p\)-th person, which are divided into visible, occluded, and invisible, \(d_{pi}\) represents the Euclidean distance between the predicted position of the human body joint point and the real position, \(s_p^2\) represents the area of the human body detection frame.

Based on OKS, AP’s calculation method for key points of the human body is similar to the calculation method for target detection. The specific formula is as follows:

$$AP_S = \frac{\sum_p \delta \left( OKS_p > s \right)}{\sum_p 1}$$

TABLE 1. The software and hardware environment used in the experiment.

| Hardware | CPU | Intel(R) Core(TM) i7-7800xCPU @3.5GHz | 32GB |
|----------|-----|-------------------------------------|------|
|          | GPU | NVIDIA GeForce RTX 2080Ti           | 12GB |
| OS       | Linux Ubuntu 18.04.3 LTS            |      |

Software:

| Programming language | python 3.6.10 |
|----------------------|---------------|
| Deep learning framework | pytorch 1.1.0 |
| CUDA                 | cuda 10.0     |
Thus, the calculation formula of mean average precision (mAP) is as follows:

\[
mAP = \text{mean}[\text{AP}(0.50:0.05:0.95)]
\]  

(14)

B. IMPLEMENTATION DETAILS

In this experiment, the datasets used are MPII [38] and MS COCO [34], and the test indicators used in the experiment are PCK and mAP. The software and hardware environment used in the experiment are shown in Table 1.

The batch size used in this experiment is set to 32, the optimizer is RMSPropOptimizer, the decay and momentum are both set to 0.9, the learning rate is set to 4e-5, and the IOU is set to 0.5. The training process is terminated within 150 epochs. The loss function of the SPPE uses Mean Squared Error (MSE) to compare the predicted heatmap with the heatmap of the real image. A 2D Gaussian distribution is used around the center of the node with a standard deviation of 1, and the target detection parameter setting refers to Faster R-CNN.

C. PERFORMANCE EVALUATION

In order to verify the performance of the proposed method in human pose estimation, the proposed algorithm is compared with a number of excellent human pose estimation algorithms on the MPII and MS COCO. The following are the visual experimental results of the method in this paper, as shown in Fig. 6. It can be seen that in different scenarios, even when there are problems such as crowded people, part of the human body being occluded, or large changes in human body posture, the method in this paper can accurately identify each human body key point and form a key point graph for each person. What’s more, there are no false detections, missed detections, or connecting key points of different personnel together in the visualized results.

In the experiment on the MPII, the PCK evaluation index was used to calculate the total accuracy and correct percentage of each key point of the human body for Head, Shoulder, Elbow, Wrist, Hip, Knee, and Ankle. It can be seen from Table 2 that compared with other advanced methods, the method in this paper is superior to other methods in the prediction effect of most key points, and the accuracy rate has been significantly improved.

The experiment on the MS COCO uses the OSK evaluation index, where AP\text{50} represents the AP measurement value when the OSK threshold is 0.5, AP\text{75} represents the measurement value when the OSK threshold is 0.75; AP\text{M} represents the measured value of the target frame with a pixel area between 32\text{2} and 96\text{2}, and AP\text{L} represents the AP measured value of the target frame with a pixel area greater than 96\text{2}. It can be seen from Table 3 that compared to other existing state-of-the-art methods on the MS COCO, the method in this paper performs basically the same or even better than some methods, and compared to the best performed HRNet [6], there is some decline. What’s more, compared to the lightweight method LPN [10] proposed in 2020, its performance has been greatly improved.

In summary, it can be seen that the method proposed in this paper performs very well on MPII and MS COCO dataset. In order to reduce the complexity of the model and...
improve the running rate, our method sacrifices some accuracy requirements, but also achieves the average accuracy of some excellent methods. Experiments show that our method not only improves the running rate, but also achieves an acceptable result.

Meanwhile, in Table 4, the running rate of the algorithm in this paper is compared with other excellent multi-person pose estimation methods, where FPS represents Frames Per Second. The backbones used by SPPE in different methods are also given in Table 4. Because changing the backbone of the target detection algorithm will greatly reduce the accuracy during the test, the backbone of Faster R-CNN adopted in this paper is MobileNet V2, while the backbone in other top-down methods is still ResNet-50. The experimental results show that the running rate of the method in this paper far exceeds other methods, reaching a transmission rate of 20 frames per second. It can be seen that this method is superior to the current advanced multi-person pose estimation methods in both accuracy and operating speed.

### D. ABLATION STUDY

The method in this paper improves the human pose estimation network model from three perspectives, so several extra experiments are further performed to verify the effectiveness of each individual component of our proposed model. MPII and MS COCO are used in experiments of ablation study.

This paper adopts the ideas of symmetric transformation (ST), global matching method (GM) and the lightweight backbone network (LWNet) to improve the multi-person pose estimation network. In Table 5 and Table 6, the impact of each improved component on the performance of the algorithm is respectively discussed. Table 5 and Table 6 respectively show the results of ablation experiments on the MPII and MS COCO datasets. The experimental results show that:

1. When only the symmetric transformation module is used, the method in this paper is already superior to some other multi-person pose estimation methods. After using the symmetric transformation module on the MPII to improve the multi-person pose estimation network, the improved method is significantly better than the best method in the MPII in the comparison experiment. On the MS COCO dataset, compared to the best method HRNet, the accuracy rate is a little low.

2. After using the lightweight backbone, the running speed is greatly improved, but the accuracy has been reduced compared with using only symmetric transformation module. In the experiment of MPII, although the accuracy rate after weight reduction has decreased, it is still higher than other methods. But the accuracy on MS COCO is significantly lower than that of the best method. However, it can be seen that after symmetric transformation module reduces the influence of low-quality human candidate frames, our method can also achieve an acceptable result, which is significantly better than another lightweight method LPN.

3. The global matching method effectively compensates for the loss of precision caused by lightweight. In the experiments on the MPII and MS COCO, the improved method using symmetric transformation module and the global matching method has the best effect. Compared with only using symmetric transformation module module, the accuracy after adding the global matching method is better, which further verifies the effectiveness of this method. After adding lightweight, the accuracy has a little decreases, but it is still achieve acceptable results. This makes the algorithm in this paper far surpass other algorithms in terms of operating speed while ensuring accuracy.

### TABLE 4. Comparison of the running speed of different algorithms on the MPII.

| Method   | OpenPose [39] | G-RMI [3] | CPN [4] | SimpleBaseline [9] | HRNet-W32 [6] | HRNet-W48 [6] | LPN [10] | Ours |
|----------|---------------|-----------|---------|--------------------|---------------|---------------|----------|------|
| backbone | -             | ResNet-101| ResNet-Inception| ResNet-152        | 73.7          | 74.9          | 75.5     | 68.7 | 71.5 |
| mAP      | 61.8          | 68.5      | 71.1    | 73.7               | 74.9          | 75.5          | 68.7     | 71.5 |
| FPS      | 11            | 12        | 11      | 7                  | 6             | 4             | 18       | 20   |

### TABLE 5. Ablation study on MPII.

| Method | Head | Shoulder | Elbow | Wrist | Hip | Knee | Ankle | Total |
|--------|------|----------|-------|-------|-----|------|-------|-------|
| ST+LWNet | 87.5 | 85.1     | 78.2  | 69.3  | 75.2| 71.1 | 65.4  | 79.4  |
| ST+GM   | 93.2 | 90.0     | 86.0  | 79.6  | 81.4| 81.1 | 75.4  | 84.4  |
| ST      | 92.9 | 89.2     | 85.7  | 76.5  | 79.1| 79.2 | 69.9  | 83.1  |
| Ours    | 90.9 | 89.3     | 85.1  | 75.5  | 78.7| 79.4 | 72.1  | 82.3  |

### TABLE 6. Ablation study on MS COCO.

| Method   | mAP  | AP<sup>50</sup> | AP<sup>75</sup> | AP<sup>M</sup> | AP<sup>L</sup> |
|----------|------|-----------------|-----------------|----------------|----------------|
| ST+LWNet | 69.7 | 88.4           | 78.1            | 66.1           | 73.9           |
| ST+GM    | 73.2 | 90.6           | 81.1            | 69.8           | 79.6           |
| ST       | 72.3 | 90.2           | 80.0            | 68.7           | 78.6           |
| Ours     | 71.5 | 89.9           | 80.1            | 68.1           | 78.2           |
V. CONCLUSION

This paper mainly studies the top-down processing methods in multi-person pose estimation tasks, including how to overcome the poor application effect of target detection algorithms on single-person pose estimation networks, how to improve the accuracy of the network under crowded crowds, and how to improve the speed of the network under the premise of ensuring accuracy. In response to the above-mentioned difficulties and challenges, a lightweight multi-person pose estimation method based on symmetric transformation and global matching is proposed. The global matching method designs an optimal matching method of the human body-key point graph based on the KM algorithm. Experiments show that the algorithm proposed in this paper not only improves the quality of the candidate area frame and the overall performance of the multi-person posture estimation in the crowded state, but also improves the running speed while ensuring the network performance, which further confirms the effectiveness and competitiveness of the algorithm in this paper.

ACKNOWLEDGMENT

The authors would also like to express their sincere thanks to the editors and the anonymous reviewers for their valuable comments and contributions.

REFERENCES

[1] L. Sigal, “Human pose estimation,” in Computer Vision, K. Ikeuchi, Ed. Boston, MA, USA: Springer, 2014, doi: 10.1007/978-0-387-31439-6_584.
[2] R. Szeliski, Computer Vision: Algorithms and Applications. New York, NY, USA: Springer, 2011.
[3] G. Papandreou, T. Zhu, N. Kanazawa, A. Toshev, J. Tompson, C. Bregler, and K. Murphy, “Towards accurate multi-person pose estimation in the wild,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 3711–3719.
[4] 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.
[5] K. He, G. Gkioxari, P. Dollár, and R. Girshick, “Mask R-CNN,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2017, pp. 2980–2988.
[6] K. Sun, B. Xiao, D. Liu, and J. Wang, “Deep high-resolution representation learning for human pose estimation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 5686–5696.
[7] X. Sun, B. Xiao, F. Wei, S. Liang, and Y. Wei, “Integral human pose regression,” in Proc. Eur. Conf. Comput. Vis. (ECCV), Sep. 2018, pp. 536–553.
[8] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang, W. Liu, and B. Xiao, “Deep high-resolution representation learning for visual recognition,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 43, no. 10, pp. 3349–3364, Oct. 2021.
[9] B. Xiao, H. Wu, and Y. Wei, “Simple baselines for human pose estimation and tracking,” in Proc. Eur. Conf. Comput. Vis., Sep. 2018, pp. 472–487.
[10] Z. Zhang, J. Tang, and G. Wu, “Simple and lightweight human pose estimation,” 2019, arXiv:1911.10346.
[11] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, “Spatial transformer networks,” 2015, arXiv:1506.02051.
[12] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, “MobileNets: Efficient convolutional neural networks for mobile vision applications,” 2017, arXiv:1704.04861.
[13] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, “MobileNetV2: Inverted residuals and linear bottlenecks,” in Proc. CVPR, Jun. 2018, pp. 4510–4520.
[14] H. Zhu, M. C. Zhou, and R. Alkims, “Group role assignment via a Kuhn–Munkres algorithm-based solution,” IEEE Trans. Syst., Man, A. Syst. Human, vol. 42, no. 3, pp. 739–750, May 2012.
[15] R. Urtasun and T. Darrell, “Sparse probabilistic regression for activity-independent human pose inference,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2008, pp. 1–8.
[16] W. Yang and G. Mori, “Multiple tree models for occlusion and spatial constraints in human pose estimation,” in Proc. Eur. Conf. Comput. Vis. (ECCV), Oct. 2008, pp. 721–734.
[17] B. Sapp, A. Toshev, and B. Taskar, “Cascaded models for articulated pose estimation,” in Proc. Eur. Conf. Comput. Vis. (ECCV), Sep. 2010, pp. 406–420.
[18] F. Wang and Y. Li, “Beyond physical connections: Tree models in human pose estimation,” in Proc. CVPR, Jun. 2013, pp. 596–603.
[19] M. Donatone, J. Gall, C. Leistner, and L. Van Gool, “Human pose estimation using body parts dependent joint regressors,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2013, pp. 3041–3048.
[20] M. Sun, P. Kohli, and J. Shotton, “Conditional regression forests for human pose estimation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2012, pp. 3394–3401.
[21] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” 2014, arXiv:1409.1556.
[22] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2016, pp. 770–778.
[23] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, and A. Rabinovich, “Going deeper with convolutions,” in Proc. IEEE Conf. CVPR, Jun. 2015, pp. 1–9.
[24] X. Fan, K. Zheng, Y. Lin, and S. Wang, “Combining local appearance and holistic view: Dual-source deep neural networks for human pose estimation,” in Proc. CVPR, Jun. 2015, pp. 1347–1355.
[25] W. Ouyang, X. Chu, and X. Wang, “Multi-source deep learning for human pose estimation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2014, pp. 2337–2344.
[26] V. Belagiannis and A. Zisserman, “Recurrent human pose estimation,” in Proc. FG, May 2017, pp. 468–475.
[27] A. Jain, J. Tompson, M. Andriluka, G. W. Taylor, and C. Bregler, “Learning human pose estimation features with convolutional networks,” 2014, arXiv:1312.7302.
[28] A. Toshev and C. Szegedy, “DeepPose: Human pose estimation via deep neural networks,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2014, pp. 1653–1660.
[29] T. Pfister, J. Charles, and A. Zisserman, “Flowing convnets for human pose estimation in videos,” in Proc. IEEE Int. Conf. Comput. Vis., Dec. 2015, pp. 1913–1921.
[30] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, “Convolutional pose machines,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2016, pp. 4724–4732.
[31] A. Newell, K. Yang, and J. Deng, “Stacked hourglass networks for human pose estimation,” in Proc. Eur. Conf. Comput. Vis., Cham, Switzerland: Springer, Oct. 2016, pp. 483–499.
[32] L. Pishchulin, E. Insafutdinov, S. Tang, B. Andres, M. Andriluka, P. Gehler, and B. Schiele, “DeepCut: Joint subset partition and labeling for multi person pose estimation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 4929–4937.
[33] S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: Towards real-time object detection with region proposal networks,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 39, no. 6, pp. 1137–1149, Jun. 2017.
[34] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, “Microsoft coco: Common objects in context,” in Proc. Eur. Conf. Comput. Vis., Sep. 2014, pp. 740–755.
[35] B. Cheng, B. Xiao, J. Wang, H. Shi, T. S. Huang, and L. Zhang, “High-erHRNet: Scale-aware representation learning for bottom-up human pose estimation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 5385–5394.
[36] U. Rafi, B. Leibe, J. Gall, and I. Kostrikov, “An efficient convolutional network for human pose estimation,” in Proc. Brit. Mach. Vis. Conf., 2016, p. 2.
[37] A. Bulat and G. Tzimiropoulos, “Binarized convolutional landmark localizers for human pose estimation and face alignment with limited resources,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Oct. 2017, pp. 3726–3734.
[38] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele, “2D human pose estimation: New benchmark and state of the art analysis,” in Proc. IEEE CVPR, Jun. 2014, pp. 3686–3693.
[39] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, “Realtime multi-person 2D pose estimation using part affinity fields,” in Proc. IEEE Conf. CVPR, Jul. 2017, pp. 1302–1310.
C. Zheng, W. Wu, C. Chen, T. Yang, S. Zhu, J. Shen, N. Kehtarnavaz, and M. Shah, “Deep learning-based human pose estimation: A survey,” 2020, arXiv:2012.13392.

U. Iqbal, A. Milan, and J. Gall, “PoseTrack: Joint multi-person pose estimation and tracking,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 4654–4663.

E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, and B. Schiele, “DeeperCut: A deeper, stronger, and faster multi-person pose estimation model,” in Proc. Eur. Conf. Comput. Vis. (ECCV), Sep. 2016. pp. 34–50.

B. Andres, and B. Schiele, “ArtTrack: Articulated multi-person tracking in the wild,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017. pp. 1293–1301.

B. Sapp, A. Toshev, and B. Taskar, “CASCADED models for articulated pose estimation,” in Proc. Eur. Conf. Comput. Vis. (ECCV), Sep. 2010. pp. 406–420.

M. Fieraru, A. Khoreva, L. Pishchulin, and B. Schiele, “Learning to refine human pose estimation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW), Jun. 2018. pp. 205–214.

X. Nie, J. Feng, J. Zhang, and S. Yan, “Single-stage multi-person pose machines,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019. pp. 6950–6959.

F. Majeed, B. Wallace, M. Z. Uddin, and J. Torresen, “A robust human activity recognition approach using OpenPose, motion features, and deep recurrent neural network,” in Proc. Scand. Conf. Image Anal. (SCIA), 2019. pp. 299–310.

D. Freire-Obregón, M. Castrillón-Santana, P. Barra, C. Bisogni, and M. Nappi, “An attention recurrent model for human cooperation detection,” Comput. Vis. Image Understand., vols. 197–198, Aug. 2020. Art. no. 102991.

P. Barra, C. Bisogni, M. Nappi, D. Freire-Obregón, and M. Castrillón-Santana, “Gender classification on 2D human skeleton,” in Proc. 3rd Int. Conf. Bio-Eng. Smart Technol. (BioSMART), Apr. 2019. pp. 1–4.

A. Zeng, X. Sun, F. Huang, M. Liu, Q. Xu, and S. Lin, “SRNet: Improving generalization in 3D human pose estimation with a split-and-recombine approach,” in Proc. Eur. Conf. Comput. Vis. (ECCV), Nov. 2020. pp. 507–523.

Z. Chen, Y. Huang, H. Yu, B. Xue, and L. Wang, “Towards part-aware monocular 3D human pose estimation: An architecture search approach,” in Proc. Eur. Conf. Comput. Vis. (ECCV), Dec. 2020. pp. 715–732.

P. Weinzapfel, R. Brégier, H. Combaluzier, V. Leroy, and G. Rogez, “DOPE: Distillation of part experts for whole-body 3D pose estimation in the wild,” in Proc. Eur. Conf. Comput. Vis. (ECCV), Nov. 2020. pp. 380–397.

YANGQING YANG received the B.S. degree in mathematics and applied mathematics from Xidian University, Xi’an, China, in 2019, where she is currently pursuing the master’s degree with the Key Laboratory of Intelligent Perception and Image Understanding of the Ministry of Education of China, International Research Center of Intelligent Perception and Computation, School of Artificial Intelligence. Her current research interests include deep learning and person re-identification.

CHANGSHENG LI received the B.S. degree from Shanghai Jiaotong University, Shanghai, China, in 1982 and the M.S. and Ph.D. degrees in electronic engineering from Xi’an Jiaotong University, Xi’an, China, in 1984 and 1990, respectively. From 1990 to 1991, he was a Postdoctoral Fellow with the National Key Laboratory for Radar Signal Processing, Xidian University, Xi’an. Since 1992, he has been a Professor with the School of Electronic Engineering, Xidian University. He is currently the Director of the Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education of China, Xidian University. He is in charge of about 40 important scientific research projects, and published more than 20 monographs and 100 papers in international journals and conferences. His research interests include image processing, natural computation, machine learning, and intelligent information processing. Dr. Jiao is a member of the IEEE Xi’an Section Executive Committee, the Chairperson of awards and recognition committee, the Vice Board Chairperson of the Chinese Association of Artificial Intelligence, the Councilor of the Chinese Institute of Electronics, and the Committee Member of the Chinese Committee of Neural Networks. He is an Expert of Academic Degrees Committee of the State Council.

YANGQIAO CHEN was born in 1991. He received the Ph.D. degree in pattern recognition and intelligent system from Xidian University, Xi’an, China, in 2019. He is currently working as an Engineer with 54th Research Institute of China Electronics Technology Group Corporation, Shijiazhuang, China. His research interests include computer vision and telemeeter and telecontrol of UAV.

CHENG PENG received the B.E. degree in intelligent science and technology from Xidian University, Xi’an, China, in 2016, where he is currently pursuing the Ph.D. degree in pattern recognition and intelligent system with the Key Laboratory of Intelligent Perception and Image Understanding of the Ministry of Education of China, School of Artificial Intelligence. His research interests include deep learning, remote sensing image processing, and evolutionary computation.

ZHENXIANG SUN received the B.S. degree in intelligent science and technology from Xidian University, Xi’an, China, in 2018, and the master’s degree from the Key Laboratory of Intelligent Perception and Image Understanding of the Ministry of Education of China, International Research Center of Intelligent Perception and Computation, School of Artificial Intelligence, Xidian University, in 2021. His research interests include deep learning and human posture estimation.

LICHENG JIAO (Fellow, IEEE) received the B.S. degree from Shanghai Jiaotong University, Shanghai, China, in 1982, and the M.S. and Ph.D. degrees in electronic engineering from Xi’an Jiaotong University, Xi’an, China, in 1984 and 1990, respectively. From 1990 to 1991, he was a Postdoctoral Fellow with the National Key Laboratory for Radar Signal Processing, Xidian University, Xi’an. Since 1992, he has been a Professor with the School of Electronic Engineering, Xidian University. He is currently the Director of the Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education of China, Xidian University. He is in charge of about 40 important scientific research projects, and published more than 20 monographs and 100 papers in international journals and conferences. His research interests include image processing, natural computation, machine learning, and intelligent information processing. Dr. Jiao is a member of the IEEE Xi’an Section Executive Committee, the Chairperson of awards and recognition committee, the Vice Board Chairperson of the Chinese Association of Artificial Intelligence, the Councilor of the Chinese Institute of Electronics, and the Committee Member of the Chinese Committee of Neural Networks. He is an Expert of Academic Degrees Committee of the State Council.