Hindawi
Journal of Sensors
Volume 2021, Article ID 1484218, 10 pages
https://doi.org/10.1155/2021/1484218

Research Article

Multistage Polymerization Network for Multiperson Pose Estimation

Yu-Fei Bai,[1,2] **Hong-Bo Zhang**,[1,2] **Qing Lei**,[3] **and Ji-Xiang Du**[4]

[1]*School of Computer Science and Technology, Huaqiao University, Xiamen 361000, China*
[2]*Fujian Key Laboratory of Computer Vision and Machine Learning, Huaqiao University, Xiamen 361000, China*
[3]*Xiamen Key Laboratory of Computer Vision and Pattern Recognition, Huaqiao University, Xiamen 361000, China*
[4]*Fujian Key Laboratory of Big Data Intelligence and Security, Huaqiao University, Xiamen 361000, China*

Correspondence should be addressed to Hong-Bo Zhang; zhanghongbo@hqu.edu.cn

Received 7 October 2021; Revised 5 December 2021; Accepted 7 December 2021; Published 29 December 2021

Academic Editor: Cong-Bin Fan

Multiperson pose estimation is an important and complex problem in computer vision. It is regarded as the problem of human skeleton joint detection and solved by the joint heat map regression network in recent years. The key of achieving accurate pose estimation is to learn robust and discriminative feature maps. Although the current methods have made significant progress through interlayer fusion and intralevel fusion of feature maps, few works pay attention to the combination of the two methods. In this paper, we propose a multistage polymerization network (MPN) for multiperson pose estimation. The MPN continuously learns rich underlying spatial information by fusing features within the layers. The MPN also adds hierarchical connections between feature maps at the same resolution for interlayer fusion, so as to reuse low-level spatial information and refine high-level semantic information to obtain accurate keypoint representation. In addition, we observe a lack of connection between the output low-level information and the high-level information. To solve this problem, an effective shuffled attention mechanism (SAM) is proposed. The shuffle aims to promote the cross-channel information exchange between pyramid feature maps, while attention makes a trade-off between the low-level and high-level representations of the output features. As a result, the relationship between the space and the channel of the feature map is further enhanced. Evaluation of the proposed method is carried out on public datasets, and experimental results show that our method has better performance than current methods.

1. Introduction

Human pose estimation (HPE) can be understood as the position estimation of human skeletal joints, such as those in the head, left hand, and right foot. It is a fundamental yet challenging task in computer vision and has applications in many fields, such as human-computer interaction, action understanding, and autonomous driving. In recent years, great progress on HPE has been made with deep learning methods.

To obtain information that is beneficial for the locating and classification of skeleton joints, existing methods mainly perform interlevel or intralevel fusion of features. In interlevel fusion, the features of different layers of the neural network are fused, as shown in Figure 1(a). Conversely, intralevel fusion refers to the fusion of feature maps of different channels in the same layer, as shown in Figure 1(b). For example, Stacked hourglass [1] extracts feature of different levels for fusion and utilizes skip connections to effectively capture various spatial relationships of keypoints. The high-resolution network (HR-Net) [2] maintained the spatial information of high-resolution features through low-resolution features and enabled high-resolution subnets to continuously obtain semantic information provided by low-resolution features through dense connections. In the residual steps network (RSN) [3], the intralevel pyramid features were integrated to extract more detailed local spatial information to obtain delicate local representations and accurately locate keypoints.

Although some feature fusion methods have achieved improved performance, they only used one of the two fusion methods. The fusion of intralevel features can extract much

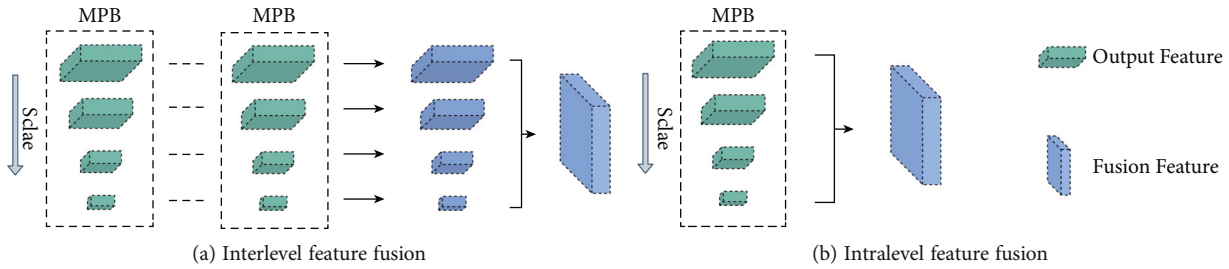

Figure 1: Illustration of intralevel feature fusion and interlevel feature fusion. (a) Interlevel feature fusion of same resolution. (b) Intralevel feature fusion of different scales.

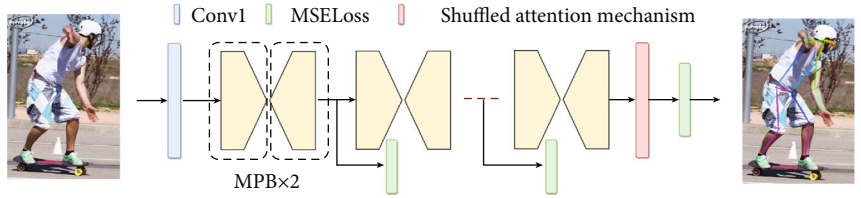

Figure 2: The framework of multistage polymerization network (MPN). Several multistage polymerization block (MPB) modules are cascaded. The shuffle attention mechanism (SAM) was used in the final stage.

more delicate local representations, thereby retaining more precise spatial information, which is critical to the localization of keypoints. However, mush unrecoverable information will be lost in the down and upsampling processes in intralevel fusion. Conversely, interlevel fusion can increase the capacity of the downsampling unit and thus reduce the loss of information. Therefore, it is effective to improve the accuracy of HPE by combining these two functional fusion methods. In existing multiperson pose estimation methods, there is little work employing intralevel fusion and interlevel fusion simultaneously. To improve the accuracy of HPE, this paper explores how to combine these two feature fusion methods.

To solve this problem, this paper proposes a novel multistage polymerization network (MPN). The framework of this MPN is shown in Figure 2. In the MPN, we use as the same intralevel fusion strategy as in the residual steps network (RSN). In the RSN, after channel cutting, the feature map was downsampled to different scales for intralevel fusion. On this basis, feature maps from different layers but of the same scale are fused by element-wise sum. To enhance intralevel fusion, feature connections are added between layers, and a cross-stage feature aggregation strategy is adopted to effectively propagate multiscale features from early stages to the current stage to further enrich the information contained in the current stage's features.

We notice that the network's output features usually directly enter into the attention mechanism for weighting, and thus, the network may ignore the cross-channel communication between high-level feature maps and low-level feature maps.

In order to solve this problem, this paper proposes a new attention module, the shuffle attention mechanism (SAM). The SAM uses a shuffle channel to enhance the cross-channel information exchange between the low-level and high-level information, thereby recalibrating the interdependence between the low-level and high-level feature maps. Experimental results also verify that the SAM can adaptively respond to important parts of the feature map.

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

(1) We propose a new MPN for HPE. The MPN enhances the image features by combining intralevel feature fusion with interlevel feature fusion, thereby improving the accuracy of HPE

(2) We propose a new attention mechanism SAM, which can strengthen the communication between different levels of feature maps and highlight the response of feature maps in spatial channels

The remainder of this paper is organized as follows. Section 2 introduces the related works, Section 3 describes the algorithms used to implement the proposed method, Section 4 presents and discusses the experimental results, and Section 5 concludes the paper.

2. Related Work

Previous research in human pose estimation was built based on the idea of part-based models, which use different configurations of parts to represent a person [4]. Current methods of human pose estimation can be divided into two categories: top-down approaches [1–3, 5–14] and bottom-up approaches [15–20]. Top-down approaches first obtain the position of the human body frame by a detector such as you only look once (YOLO) [21] or single shot multiBox detector (SSD) [22] and then detect the position of keypoints in the human region. In bottom-up approaches, all of the human keypoints in an image are detected directly, and then these parts are classified as human instances. We mainly

focus on feature fusion and strengthening feature connections in these methods and discuss the feature fusion issue from the aspect of efficient feature representation. Attention mechanism is also widely used in these methods. However, these methods directly input the feature map into the attention mechanism for weighting, without considering the communication between different semantic layers. We design the shuffle attention mechanism (SAM) module to strengthen the connection between different semantic layers through shuffle. Therefore, we also discuss the commonly used attention mechanisms in human pose estimation (HPE).

2.1. Human Pose Estimation Method. In recent years, heat map regression networks were applied to achieve multiperson pose estimation. The heat map of skeleton joint was first introduced in [23], which was designed to solve the problem of the coordinate prediction of the skeleton joint in traditional HPE methods. The space and context information of keypoints are lost in the coordinate prediction. But the heat map can solve this limitation well and become the most common form of skeleton representation. The key of heat map based methods is to design a network architecture to regress to regress heat map more effectively.

GRAPH-PCNN [24] proposed a two-stage framework based on the graph structure and the unrelated models. This method added a positioning subnet and a graph structure pose optimization module on the original heat map regression framework, in which the heat map was regressed by the network for rough positioning of the keypoints and providing a keypoint candidate set. The positioning subsystem was used to extract visual features of each kepoint in the candidate set and predict the final keypoint coordinates. Due to the resolution reduction of the heat map, there is a quantitative error in ground-truth heatmaps, which will lead to inaccurate model training and poor inference model performance. To solve these problems, Zhang et al. [25] proposed a new distribution sensing coordinate representation (DARK) for HPE. In DARK, Taylor expansion was applied to decode efficiently coordinates to generate unbiased heat map. Huang et al. [26] used the encoding-decoding process to generate keypoint heat map and regarded discrete pixel points as a metrics. However, this method had deviations in the data enhancement process. Therefore, a continuous measurement standard of unbiased data processing (UDP) is proposed in literature [27]. The continuous measurement standard was used as an image size measurement standard, which was defined as the distance between adjacent pixels in a particular space, thereby suppressing the positioning deviation caused by the approach to discrete measure. The case of occlusion will also affect the regress of the heat map; considering this, Qiu et al. [28] proposed an image guidance GCN network (IGP-GCN) which cascaded feature adaption. IGP-GCN network-integrated human structure and image context to optimize estimation results and learned the pose displacement by progressive manner. This made the IGP-GCN not only capture the posture structure information but also capture context image information simultaneously. In IGP-GCN, the occlusion joints can be inferred from the context information of the image and the pose structure clues.

2.2. Feature Fusion. Most previous work on multiperson pose estimation obtained rich feature information through interlayer connections or intralayer connections. The sequential architecture of convolutional pose machines (CPM) [14] used various connection strategies to implicitly capture spatial relations between key points and obtained a large receptive field through a larger estimator, thus, it can achieve a more refined spatial representation. The pyramid residual module (PRM) proposed by Cai et al. [3] enhances the invariance in scales of human components and shows great performance when using interlevel feature fusion. Newell et al. [17] proposed a U-shaped stacked hourglass network to obtain spatial connections between features of different resolutions through downsampling and skip connections. In addition, Chen et al. [5] used RefineNet combined with a cascaded pyramid network of interlayer features to maintain high-level and low-level information from multiscale feature maps. In high-resolution network (HR-Net) [2], four subnets were connected in parallel, and repeated cross-parallel convolution was used to perform multiscale fusion and enhance high-resolution representation. Meanwhile, the residual steps network (RSN) repeatedly enhanced the intralevel feature fusion to learn refined local representations. While these aforementioned methods have verified the effectiveness of interlayer feature fusion and intralayer feature fusion, exploring the combination of the two is rare in human pose estimation.

2.3. Attention Mechanism. The performance of attention mechanisms in computer vision is remarkable. Channel attention, spatial attention, and spatial attention combined with channel attention are the most used attention mechanisms at present.

2.3.1. Channel Attention. Squeeze-and-Excitation Network (SE-Net) [29] through the "Squeeze-and-Excitation" block can adaptively highlight the channel-wise feature maps by modeling the channel-wise statistics. The discriminative feature network (DFNet) [30] used global average pooling to introduce global context information and included a smooth network with global information and a channel attention model to improve intraclass consistency.

2.3.2. Spatial Attention. Kligvasser et al. [31] proposed a spatial activation function with depth-wise separable convolution. Zhao et al. [32] studied the spatial attention mechanism from the perspective of information flow. However, they only considered unilateral passage or space, while ignoring the combination of spatial attention and attention channels.

Spatial attention combined with channel attention: Spatial and Channel-wise Attention in Convolutional Networks (SCA-CNN) [33] proposed spatial and channel attention. Attention was not only in the channel coding but also in the spatial perspective to indicate what part of the feature map needed to be paid attention to.

Chu et al. [34] proposed a multiscale attention model multicontext attention (MCA) that improved the performance of pose estimation. Su et al. [12] proposed the Spatial and Channel-wise Attention Residual Bottleneck (SCARB)

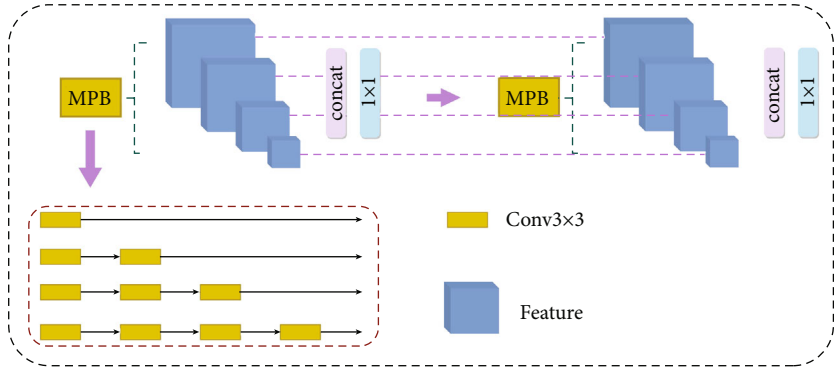

Figure 3: Framework of multistage polymerization block (MPB) module.

in multiperson pose estimation and studied the modeling order of space and channels. Meanwhile, Woo et al. [35] proposed a global average pool and largest pool channel attention module Convolutional Block Attention Module (CBAM). The dual attention network (DANet) [36] was proposed to adaptively integrate local features and global dependencies, the semantic dependency was modelled by a parallel channel dimension, and the space dimension has two kinds of attention module.

3. Method

The overall framework of the multistage polymerization network (MPN) is shown in Figure 2. It is a cascaded of several multistage polymerization block (MPB) modules. The shuffled attention mechanism (SAM) is used in the final stage. In this section, we will describe these modules in detail.

3.1. MPN: Multistage Polymerization Network. For the input image, the convolutional layer is applied to compute the feature maps. In this layer, there is a total of 104 convolution kernels. This layer is followed by the MPB network, which is designed to achieve intralayer fusion and interlayer fusion. The input feature maps of the MPB network are regularly sliced into four parts $F = \{f_1, f_2, f_3, f_4\}$ on the channel.

The MPB network is a cascade system of MPB modules. The sliced feature maps are fed into the first MPB module. Each MPB module consists of two operation blocks, which are designed according to the RSN [3] and are shown in Figure 3. In each block, the input feature maps are fed into the convolutional network. Four convolutional networks with different numbers of convolutional kernels are applied to generate features with different scales from the four input features, respectively. As shown in Figure 3, the number of the convolutional layers in these four convolutional networks is 4, 3, 2, and 1, respectively. All of these convolutional layers are built via a convolution operation.

Suppose that $x_1^i, x_2^i, x_3^i, x_4^i$ is the output of the first block in the i-th MPB module. For intralevel fusion, these output features are concatenated to generate block features X_i. These block features are upsampled by $[x]$ and fed into second operation block of the i-th MPB module. For the second block, the same operation as in the first block is applied, and its output is defined as $y_1^i, y_2^i, y_3^i, y_4^i$. Finally, the interlevel fusion between these blocks is applied to output the feature of each MPB M^i, as defined in the following equation.

$$M^i = \sum_{j=1}^{4} \left(x_j^i + y_j^i \right). \tag{1}$$

The MPB module refers to the idea of a supervision relay and performs loss calculation for each MPB module. First, we use the Gaussian kernel to spray all labels of key points onto the heat map Y, as defined in equation (2), where σ is the standard deviation of the object size-adaptation, (x, y) is the location of the heat map, and $(\tilde{x}, \tilde{y})$ is the true label coordinate. In this work, we build the heat map for each key point of the human skeleton independently. To obtain the output feature $\boldsymbol{M^i}$ of each MPB module, the keypoint prediction network, including upsample and two convolution operations, is applied to map the feature to the skeleton prediction heat map. Finally, the mean squared error (MSE) function is used to compute the prediction error of each MPB module, and the overall loss of the MPB network is defined as in equation (3).

$$Y = \exp\left(-\frac{(x-\tilde{x})^2 + (y-\tilde{y})^2}{2\sigma^2} \right), \tag{2}$$

$$L_{\text{MPB}} = \frac{1}{N \times K} \sum_{i=1}^{N} \sum_{j=1}^{K} \left(Y_j^i - \tilde{Y}_j \right)^2. \tag{3}$$

Here, N is the number of MPBs in the MPN, and K is the number of keypoints of the human skeleton. Y_j^i is the predicted heat map of the j-th keypoint by the i-th MPB module, and $\tilde{Y}_j$ is the ground truth heat map of the j-th keypoint.

The multistage polymerization block (MPB) module draws on the method of the residual steps networks (RSN) for intralevel fusion and uses cross-stage connections for interlevel fusion. The characteristic gradient gap formed by the tight connection structure is very narrow. In addition, channel information with different characteristics between different levels can complement and strengthen each other to obtain more precise spatial and semantic information.

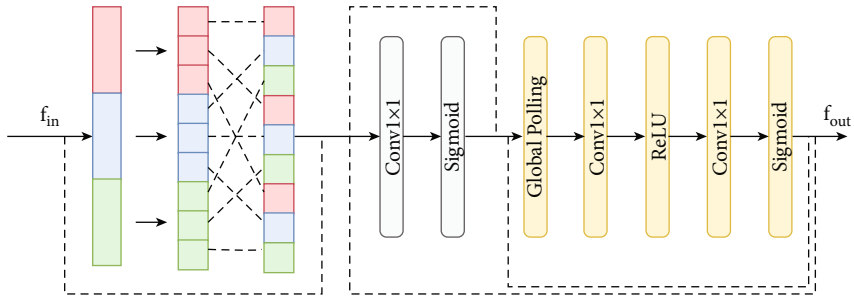

FIGURE 4: Architecture of shuffled attention mechanism (SAM). The path connection is represented by a dashed line, which is weighted in terms of channel and space.

3.2. SAM: Shuffled Attention Mechanism. The shuffled attention mechanism (SAM) is used in the last module of the multilevel network to shuffle and weight the output functions. As shown in Figure 4, the first module of the SAM is the channel shuffling of residual connections. After shuffling, a 1×1 convolutional operation and a Sigmoid activation function are applied to obtain the space attention α. The last part of the SAM is the channel attention, which consists of a global pooling, two 1×1 convolutional operations, a ReLU activation function, and a Sigmoid activation function to obtain the channel attention vector β.

3.3. Channel Shuffle Operation. To achieve the purpose of feature communication, we consider using a channel shuffle instead of dense pointwise convolution. As shown in Figure 4(a), the channel shuffle operation can be modelled as a process composed of "reshape-transpose-reshape" operations. Assuming that the input layer is divided into G groups, the input feature is reshaped into $G \times N$ dimensions, where N is the number of channels in each group. Then, the features are transposed into (N, G) dimensions to ensure that the input of the following group convolution operation comes from different groups. Finally, it is reshaped into dimensions (G, N) so that the information can flow between different groups. The shuffled feature is merged with the original by element-wise sum to form the output of the channel shuffle module.

Suppose the input of the SAM is f_{in}, this is also the output of the last MPB module. The channel shuffle can be formulated as in the following equation.

$$f_{\text{CS}}^{\text{out}} = \text{CS}(f_{\text{in}}) + f_{\text{in}}. \tag{4}$$

Here, $\text{CS}(\cdot)$ represents the channel shuffle operation, and $f_{\text{CS}}^{\text{out}}$ is the output of the channel shuffle module.

3.4. Attention Mechanism. Spatial attention: the feature map leads to undesirable results of keypoint locations due to the existence of areas in the spatial information that is not related to keypoints. The function of the spatial attention mechanism is to weight the feature map, reduce the interference of irrelevant areas, and adaptively highlight the areas related to the positioning task. The spatial-wise attention weight α is generated by a convolutional operation followed by a sigmoid function on the input. The spatial attention can be formulated as in the following equation.

$$\alpha = \text{Sigmoid}\left(\text{Conv}\left(W, f_{\text{out}}^{\text{CS}}\right)\right). \tag{5}$$

Here, $\text{Conv}(\cdot)$ denotes the convolution operation, and W is the learnable weight of the convolution operation. $\text{Sigmoid}(\cdot)$ is the Sigmoid activation function. Finally, the learned spatial attention weight α is rescaled, and the output is defined as in equation (6). $f_{\text{out}}^{\text{at}}$ is the output of the spatial attention mechanism.

$$f_{\text{out}}^{\text{at}} = f_{\text{out}}^{\text{CS}} \times (\alpha + 1). \tag{6}$$

3.4.1. Channel Attention. Each channel of the feature map is the feature activation of the corresponding convolutional layer. Since a convolution only operates in a local space, it is difficult to obtain enough information to extract the relationship between channels. Inspired by the Squeeze-and-Excitation Network (SENet) [29], which used excitation module to learn the weight of feature map of each convolutional layer, we regard channel attention as the process of adaptively selecting the convolutional layer.

In the squeeze step, the output feature of the spatial attention mechanism $f_{\text{out}}^{\text{at}}$ is used as the input of channel attention. We encode the entire spatial feature on a channel as a global feature and use global average pooling on $f_{\text{out}}^{\text{at}}$ to generate channel statistics $Z \in R^C$, as defined in the following equation.

$$Z_t = \frac{1}{H \times L} \sum_{i=1}^{H} \sum_{j=1}^{L} U_t(i, j). \tag{7}$$

Here, Z_t is the t-th element in Z, and U_t represents the output of the t-th convolution kernel in the channel attention network.

The squeeze operation obtains the global description characteristics, but we need another operation to capture the relationship between channels. It must be able to learn the nonlinear relationship between each channel. Moreover, the learned relationship is not mutually exclusive because multichannel features are allowed to instead of one-hot

form. Therefore, a Sigmoid gating mechanism is used for channel statistics Z, as defined in the following equation.

$$\beta = \text{Sigmoid}(\text{Conv}(W_2, \text{ReLU}(\text{Conv}(W_1, Z)))). \quad (8)$$

Here, $W_1 \in R^{C\times C}$ and $W_2 \in R^{C\times C}$ denote the learnable parameters in the two fully connected layers, and ReLU(·) denotes the ReLU activation function.

Finally, the channel attention weight β is learned by SAM. The output of the SAM can be generated by the following equation.

$$f_{\text{out}}^{\text{SAM}} = f_{\text{out}}^{\text{at}} \times (\beta + 1). \quad (9)$$

Like with the feature of the MPB, we turn the output feature of the SAM $f_{\text{out}}^{\text{SAM}}$ into an estimated keypoints Y_j^{SAM}. The loss of SAM module can be defined in the following equation.

$$L_{\text{SAM}} = \frac{1}{K}\sum_{j=1}^{K}\left(Y_j^{\text{SAM}} - \tilde{Y}_j\right)^2. \quad (10)$$

Here, Y_j^{SAM} is the heat map of the j-th keypoint predicted by the feature of the SAM. Finally, the overall loss function of the MPN is defined as in equation (11), which consists of the loss of MPB and SAM. In the training stage, the weights of the proposed method are obtained by minimizing the overall loss function.

$$\text{Loss} = L_{\text{MPB}} + L_{\text{SAM}}. \quad (11)$$

4. Experiments

4.1. Dataset and Experimental Settings

4.1.1. COCO Dataset. We evaluate our model on the challenging COCO dataset [37]. The COCO train2017 set, which includes 57 K images and 150 K person instances, is used to train the proposed model; the COCO minival dataset is used as the testing set. The input image is resized to 256×192.

4.1.2. MPII Dataset. The MPII human pose dataset is a state-of-the-art benchmark for the evaluation of human pose estimation. The dataset includes around 25 K images containing over 40 K people with annotated body joints. In this experiment, the data augmentation and the training strategy are set to be the same as in the COCO dataset, except that the input image size is 256×192.

4.1.3. Training Details. We implement the proposed MPN model in Pytorch, using 2 Nvidia GTX 2080Ti GPUs; the minimum batch size of each GPU is 8. The Adam optimizer is adopted, and the linear learning rate is gradually reduced from $5e-4$ to 0. The weight decays to $1e-5$. All images are rotated and scaled. The rotation range is set from -45 degrees to +45 degrees, and the zoom range is set from 0.7 to 1.35.

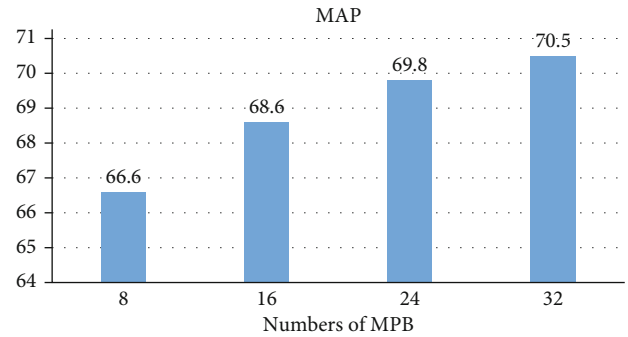

Figure 5: Ablation study on different numbers of MPBs (multistage polymerization block).

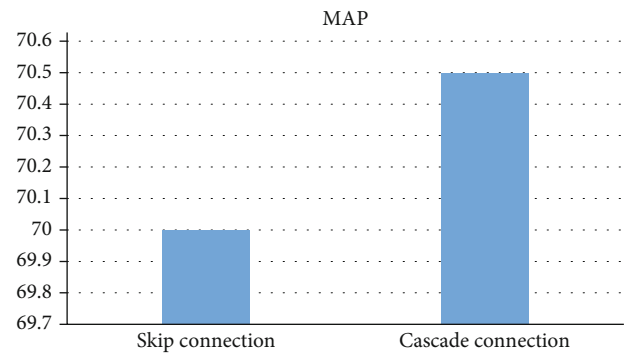

Figure 6: Multistage polymerization network (MPN) ablation experiment with cascade connection and skip connection.

Table 1: Ablation of the SAM and other attention mechanisms.

Method	mAP
MPN without attention mechanism	70.5
MPN + PRM	71.7
MPN + SCARB	71.9
MPN + SAM	72.3

Table 2: Assessing shuffle positions in the SAM.

Method	mAP
MPN without attention mechanism	70.5
MPN + SAM_A	72.2
MPN + SAM_B	72.3

4.1.4. Testing Details. We estimate the heat map using a Gaussian filter. We average the predicted heat maps of the original image with the results of the corresponding input image. A quarter offset in the direction from the highest response to the second-highest response is used to obtain the final keypoints. Similar to in the cascaded pyramid network (CPN) [5], the pose score is the product of the average score of the keypoints and the bounding box score.

Table 3: Results on the COCO test-dev dataset. "∗" denotes using ensembled models. AP50 and AP75 indicate that we set the threshold to 0.5 and 0.75, respectively. APM indicates that the size of the detected target in the image ranges from 322 to 922, and APL indicates that the target range is greater than 922.

Method	Backbone	mAP	AP50	AP75	APM	APL
CMU-pose [15]	—	61.8	84.9	67.5	57.1	68.2
Mask-RCNN [7]	ResNet-50-FPN	63.1	87.3	68.7	57.8	71.4
MSPN [38]	ResNet-50	71.5	90.2	79.2	68.2	77.6
Integral pose regression [39]	ResNet-101	67.8	88.2	74.8	63.9	74.0
G-RMI [11]	ResNet-101	68.5	87.1	75.5	65.8	73.3
RSN [3]	RSN-18	70.4	88.8	77.7	67.2	72.2
SimpleBaseline [13]	ResNet-50	71.3	89.9	78.9	68.3	77.4
Graph-PCNN [24]	HR32	71.5	89.0	79.0	68.4	77.6
SimpleBaseline + UDP [26]	ResNet-50	71.7	91.1	79.6	68.6	77.1
CSM [12]	ResNet-101	71.8	91.3	80.1	68.7	77.3
DARK [25]	HR48	71.9	89.1	79.6	69.2	78.0
CPN+ [5]	ResNet-inception	72.1	90.5	78.9	67.9	78.1
Ours	MPN-16	68.8	87.8	75.9	66.1	74.3
Ours	MPN-32	70.5	88.5	77.5	67.5	76.5
Ours	MPN − 32 + SCARB	71.9	89.2	78.8	69.1	77.8
Ours	MPN − 32 + SAM − B	72.3	89.2	79.4	69.4	78.3

Table 4: PCK_h@0.5 results on MPII test dataset.

Method	Hea	Sho	Elb	Wri	Hip	Kne	Ank	Mean
CPM [14]	97.8	95.0	88.7	84.0	88.4	82.8	79.4	88.5
Part heat map regression [40]	97.9	95.1	89.9	85.3	89.4	85.7	81.9	89.7
SimpleBaseline [13]	97.0	95.9	90.3	85.0	89.2	85.3	81.3	89.6
CFA [41]	96.1	95.7	91.3	86.4	89.2	87.5	83.6	90.0
MPN-32	96.3	94.6	87.1	80.7	87.3	82.2	77.8	87.2
MPN − 32 + SAM − B	96.4	96.0	90.7	85.4	89.8	86.6	82.1	90.1

4.1.5. COCO Evaluation Metric. The OKS-based mean average precision (mAP) is used as the evaluation indicator for the COCO dataset. According to the Euclidean distance d^2 between the detected key point and the corresponding ground truth, the OKS value is defined in the following equation.

$$\mathrm{OKS}_P = \frac{\sum_i \exp\left(-d_{pi}^2/\left(2S_P^2\sigma_i^2\right)\right)\delta\left(v_{\mathrm{p}i}=1\right)}{\sum_i \delta\left(v_{pi}=1\right)}. \tag{12}$$

Here, P represents the ID of a person in ground truth, pi represents the i-th keypoint of person P, $v_{pi}=1$ indicates that the i-th keypoint is visible, and S_P represents the square root of the area occupied by this person, which is calculated from the bounding box of person P. σ_i is the normalization factor of the i-th keypoint. And d_{pi}^2 represents the square of the Euclidean distance of the pi between the predicted value and the ground truth.

For a predicted person P, if the OKS value of this person OKS_P is higher than the threshold T, the prediction will be regarded as correct. The average precision is defined as in the following equation.

$$\mathrm{AP} = \frac{\sum_P \delta(\mathrm{OKS}_P > T)}{N}. \tag{13}$$

4.1.6. MPII Evaluation Metric. The percentage of correct key points (PCK) reports the percentage of keypoint detections falling within a normalized distance of the ground truth. PCK is defined in the following equation.

$$\mathrm{PCK}_i = \frac{\sum_P \delta\left(d_{pi}/d_p^{def} \le T_k\right)}{\sum_P 1}. \tag{14}$$

Here, PCK_i is the PCK value of the predicted results for the i-th keypoint, d_p^{def} represents the scale factor of the P-th person, and T_k is a threshold set to 0.5.

Figure 7: Prediction results on MPII (top row) and COCO (bottom row) datasets.

4.2. Ablation Study. In this section, we conduct an in-depth analysis of the structure of the proposed method. All of the ablation studies are performed on the COCO dataset.

4.2.1. The Numbers of MPB Modules. In this experiment, we explore the performance with different numbers of MPB modules. The comparison results are shown in Figure 5, where the number of MPB modules is set to 8, 16, 24, and 32. When the number of MPB modules reaches 32, the proposed method achieves the best performance, and the mAP is 70.5. With the continuous growth in the number of modules, the increase in the number of parameters will lead to an increase in the computational cost, and thus, we choose 32 as the ideal number of MPBs.

4.2.2. Cascade Connection and Skip Connection. To verify the effectiveness of the connection strategy in the MPB, we compare the cascade connection and skip connection. The comparison results are shown in Figure 6, and it is clear that the cascade connection produces better performance.

4.2.3. Ablation Study of the SAM. To verify the effectiveness of our SAM module, we compare it to existing modules: Spatial and Channel-wise Attention Residual Bottleneck (SCARB) and Pose Rene Machine (PRM). The input size is the default: 256×192. The results are shown in Table 1. It can be seen that our SAM results in a mAP improvement of 0.4 relative to SCARB, and a mAP improvement of 0.6 relative to PRM. We also analyze the impact of different shuffle positions on the performance of the SAM module. SAM-A puts the shuffle operation between the space and the channel, and SAM-B puts the shuffle operation in front of the space and the channel, as shown in Table 2. SAM-B results in the best mAP of 72.3, which is an improvement of 0.1 over SAM-A.

4.2.4. Comparison with the State-Of-the-Art Methods. To verify the effectiveness of our method, in this experiment, we compare the proposed model with the latest method on the COCO test-dev dataset. The comparison results are shown in Table 3. Without extra data for training, our single model can use MPN backbone network to reach a mAP 70.5, and by adding Spatial and Channel-wise Attention Residual Bottleneck (SCARB) to reach a mAP of 71.9, which is higher than CSM by mAP of 0.1. The results of adding SAM are higher than SCARB by a mAP of 0.6. These results show that our method is more effective.

We also validate the MPN on MPII test set. As shown in Table 4, adding the SAM module yields an improvement in mAP of 2.9, which further demonstrates the generalizability of our method.

Finally, Figure 7 shows the prediction results obtained by our MPN on the MPII and COCO datasets.

5. Conclusions

In this paper, we propose a top-down multistage polymerization network to handle multiperson pose estimation. The MPN learns exquisite key point representations through effective intralayer fusion and interlayer fusion. We also design a shuffled attention mechanism module. The shuffle aims to promote the cross-channel information exchange between pyramid feature maps while attention is carried out to make a trade-off between the low-level and high-level representations of the output features. Overall, we achieve a good result on two keypoint benchmarks.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

The authors would like to thank the anonymous reviewers for their valuable and insightful comments on an earlier version of this manuscript. This work was supported by the

Natural Science Foundation of China (nos. 61871196 and 62001176), National Key Research and Development Program of China (no. 2019YFC1604700), Natural Science Foundation of Fujian Province of China (nos. 2019J01082 and 2020J01085), and the Promotion Program for Young and Middle-aged Teacher in Science and Technology Research of Huaqiao University (no. ZQN-YX601).

References

[1] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *European Conference on Computer Vision*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds., pp. 483–499, Springer, 2016.

[2] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5693–5703, Long Beach, CA, 2019.

[3] Y. Cai, Z. Wang, Z. Luo et al., "Learning delicate local representations for multi-person pose estimation," in *European Conference on Computer Vision*, A. Vedaldi, H. Bischof, T. Brox, and J. M. Frahm, Eds., pp. 455–472, Springer, 2020.

[4] H.-B. Zhang, Q. Lei, B.-N. Zhong, J.-X. du, and J. L. Peng, "A survey on human pose estimation," *Intelligent Automation & Soft Computing*, vol. 22, no. 3, pp. 483–489, 2016.

[5] Y. Chen, Z. Wang, Y. Peng, Z. Zhang, G. Yu, and J. Sun, "Cascaded pyramid network for multi-person pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7103–7112, Salt Lake City, Utah, 2018.

[6] H. S. Fang, S. Xie, Y. W. Tai, and C. Lu, "Rmpe: regional multi-person pose estimation," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 2334–2343, Venice, Italy, 2017.

[7] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *In Proceedings of the IEEE International Conference on Computer Vision*, pp. 2961–2969, Venice, Italy, 2017.

[8] S. Huang, M. Gong, and D. Tao, "A coarse-fifine network for keypoint localization," in *In Proceedings of the IEEE International Conference on Computer Vision*, pp. 3028–3037, Venice, Italy, 2017.

[9] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, and B. Schiele, "DeeperCut: a deeper, stronger, and faster multi-person pose estimation model," in *European Conference on Computer Vision*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds., Springer, 2016.

[10] L. Pishchulin, E. Insafutdinov, S. Tang et al., "Deepcut: joint subset partition and labeling for multi person pose estimation," in *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4929–4937, Las Vegas, Nevada, 2016.

[11] G. Papandreou, T. Zhu, N. Kanazawa et al., "Towards accurate multi-person pose estimation in the wild," in *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4903–4911, Honolulu, Hawaii, 2017.

[12] K. Su, D. Yu, Z. Xu, X. Geng, and C. Wang, "Multi-person pose estimation with enhanced channel-wise and spatial information," in *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5674–5682, Long Beach, CA, 2019.

[13] B. Xiao, H. Wu, and Y. Wei, "Simple baselines for human pose estimation and tracking," in *Computer Vision – ECCV 2018. ECCV 2018*, vol. 11210 of Lecture Notes in Computer Science, , Springer.

[14] S. E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional pose machines," in *In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 4724–4732, Las Vegas, Nevada, 2016.

[15] Z. Cao, T. Simon, S. E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 7291–7299, Honolulu, Hawaii, 2017.

[16] M. Kocabas, S. Karagoz, and E. Akbas, "Multi PoseNet: fast multi-person pose estimation using pose residual network," in *Computer Vision–ECCV 2018. ECCV 2018*, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds., pp. 437–453, Springer, 2018.

[17] A. Newell, Z. Huang, and J. Deng, "Associative embedding: end-to-end learning for joint detection and grouping," http://arxiv.org/abs/1611.05424.

[18] F. Xia, P. Wang, X. Chen, and A. L. Yuille, "Joint multi-person pose estimation and semantic part segmentation," in *In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 6769–6778, Honolulu, Hawaii, 2017.

[19] X. Nie, J. Feng, J. Zhang, and S. Yan, "Single-stage multi-person pose machines," in *In Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6951–6960, Seoul, Korea, 2019.

[20] 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 *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 269–286, Munich, Germany, 2018.

[21] J. Redmon and A. Farhadi, "YOLO9000: better, faster, stronger," in *In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 7263–7271, Honolulu, Hawaii, 2017.

[22] W. Liu, D. Anguelov, D. Erhan et al., "SSD: single shot multi-Box detector," in *Computer Vision – ECCV 2016. ECCV 2016*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds., Springer, 2016.

[23] J. Tompson, A. Jain, Y. LeCun, and C. Bregler, "Joint training of a convolutional network and a graphical model for human pose estimation," in *In Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 1 (NIPS'14)*, pp. 1799–1807, MIT Press, Cambridge, MA, USA, 2014.

[24] J. Wang, X. Long, Y. Gao, E. Ding, and S. Wen, "Graph-PCNN: two stage human pose estimation with graph pose refinement," in *Computer Vision – ECCV 2020. ECCV 2020*, A. Vedaldi, H. Bischof, T. Brox, and J. M. Frahm, Eds., Springer, Cham, 2020.

[25] F. Zhang, X. Zhu, H. Dai, M. Ye, and C. Zhu, "Distribution-aware coordinate representation for human pose estimation," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7091–7100, 2020.

[26] J. Huang, Z. Zhu, F. Guo, and G. Huang, "The devil is in the details: delving into unbiased data processing for human pose estimation," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5699–5708, 2020.

[27] H. Dai, L. Zhou, and F. Zhang, "Joint COCO and Mapillary Workshop at ICCV 2019 keypoint detection challenge track technical report: distribution-aware coordinate representation for human pose estimation," arXiv, 2020: 2003.07232.

[28] L. Qiu, X. Zhang, Y. Li et al., "Peeking into occluded joints: a novel framework for crowd pose estimation," in *Computer Vision – ECCV 2020. ECCV 2020*, A. Vedaldi, H. Bischof, T. Brox, and J. M. Frahm, Eds., Springer, Cham, 2020.

[29] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 7132–7141, Salt Lake City, Utah, 2018.

[30] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, "Learning a discriminative feature network for semantic segmentation," in *In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 1857–1866, Salt Lake City, Utah, 2018.

[31] I. Kligvasser, T. R. Shaham, and T. Michaeli, "Learning a spatial activation function for efficient image restoration," in *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2433–2442, Salt Lake City, Utah, 2018.

[32] H. Zhao, Y. Zhang, S. Liu et al., "PSANet: point-wise spatial attention network for scene parsing," in *Computer Vision – ECCV 2018. ECCV 2018*, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds., Springer, Cham, 2018.

[33] L. Chen, H. Zhang, J. Xiao et al., "Sca-cnn: spatial and channel-wise attention in convolutional networks for image captioning," in *In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 5659–5667, Honolulu, Hawaii, 2017.

[34] X. Chu, W. Yang, W. Ouyang, C. Ma, A. L. Yuille, and X. Wang, "Multi-context attention for human pose estimation," in *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1831–1840, Honolulu, Hawaii, 2017.

[35] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, "CBAM: convolutional block attention module," in *Computer Vision – ECCV 2018. ECCV 2018*, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds., Springer, Cham, 2018.

[36] J. Fu, J. Liu, H. Tian et al., "Dual attention network for scene segmentation," in *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3146–3154, Long Beach, CA, 2019.

[37] T.-Y. Lin, M. Maire, S. Belongie et al., "Microsoft COCO: common objects in context," in *Computer Vision – ECCV 2014. ECCV 2014*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds., Springer, Cham, 2014.

[38] W. Li, Z. Wang, and B. Yin, "Rethinking on multi-stage networks for human pose estimation," 2019, arXiv: 1901.00148.

[39] X. Sun, B. Xiao, F. Wei, S. Liang, and Y. Wei, "Integral human pose regression," in *Computer Vision – ECCV 2018. ECCV 2018*, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds., Springer, Cham, 2018.

[40] A. Bulat and G. Tzimiropoulos, "Human pose estimation via convolutional part heatmap regression," in *Computer Vision – ECCV 2016. ECCV 2016*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds., Springer, Cham, 2016.

[41] Z. Su, M. Ye, and G. Zhang, "Cascade feature aggregation for human pose estimation," 2019, arXiv: 1902.07837.