The Visual Computer (2023) 39:3429–3440
https://doi.org/10.1007/s00371-023-02953-4

ORIGINAL ARTICLE

A lightweight pose estimation network with multi-scale receptive field

Shuo Li[1,2] · Ju Dai[2] · Zhangmeng Chen[1,2] · Junjun Pan[1,2]

Accepted: 9 June 2023 / Published online: 25 June 2023

Abstract
Existing lightweight networks perform inferior to large-scale models in human pose estimation because of shallow model depths and limited receptive fields. Current approaches utilize large convolution kernels or attention mechanisms to encourage long-range receptive field learning at the expense of model redundancy. In this paper, we propose a novel Multi-scale Field Lightweight High-resolution Network (MFite-HRNet) for human pose estimation. Specifically, our model mainly consists of two lightweight blocks, a Multi-scale Receptive Field Block (MRB) and a Large Receptive Field Block (LRB), to learn informative multi-scale and long-range spatial context information. The MRB utilizes group depthwise dilation convolutions with varied dilation rates to extract multi-scale spatial relationships from different feature maps. The LRB leverages large depthwise convolution kernels to model large-range spatial knowledge at the low-level features. We apply MFite-HRNet to single-person and multi-person pose estimation tasks. Experiments on COCO, MPII, and CrowdPose datasets demonstrate that our network outperforms current state-of-the-art lightweight networks in either single-person or multi-person pose estimation tasks. The source code will be publicly available at https://github.com/lskdje/MFite-HRNet.git.

Keywords Human pose estimation · High-resolution network · Lightweight model · Multi-scale receptive field

1 Introduction

Existing high-resolution networks with large model capacities have achieved outstanding performance in 2D Human Pose Estimation (HPE) [1–3] thanks to the maintenance of high-resolution feature maps and multi-scale feature information fusion. However, high model complexity makes those networks computationally prohibitive, hindering training and challenging inference. This work aims to create a fantastic lightweight network for HPE with relatively adequate performance.

Early endeavors to facilitate HPE mainly focus on employing multi-scale or high-resolution features. The hourglass network [1] leverages the stacked hourglass modules with residual blocks to advance pose estimation. The Cascaded Pyramid Network (CPN) [2] utilizes the GlobalNet to incorporate multi-scale feature learning to estimate simple keypoints and the RefineNet to combine all pyramid features to predict hard keypoints. The high-resolution network (HRNet) [3] argues the benefit of keeping high-resolution branches and integrating different resolution features at different levels, effectively enhancing information extraction abilities. However, the above methods retain considerable parameters and computation complexities, demanding high prerequisites for hardware devices and affecting training and inference speeds.

To speed up model training and reasoning, some researchers have begun to shift their attention to lightweight network design. Compressing DNN models into compact ones suitable for edge devices, while maintaining comparable performance to the original, is the essence of lightweight networks [4]. The MobileNet [5] is one of the pioneering work proposed for embedded devices. The core idea is to decompose convolution, i.e., depthwise convolution and pointwise convolution, which can effectively lessen the network parameters. ShuffleNet [6] is another milestone of lightweight networks. It founds that 1×1 convolution is a substantial

✉ Ju Dai
daij@pcl.ac.cn

✉ Junjun Pan
pan_junjun@buaa.edu.cn

Shuo Li
leeshuo@buaa.edu.cn

Zhangmeng Chen
zhmchen@buaa.edu.cn

[1] Beihang University, Haidian, Beijing 100191, China

[2] Peng Cheng Laboratory, Nanshan, Shenzhen 518000, China

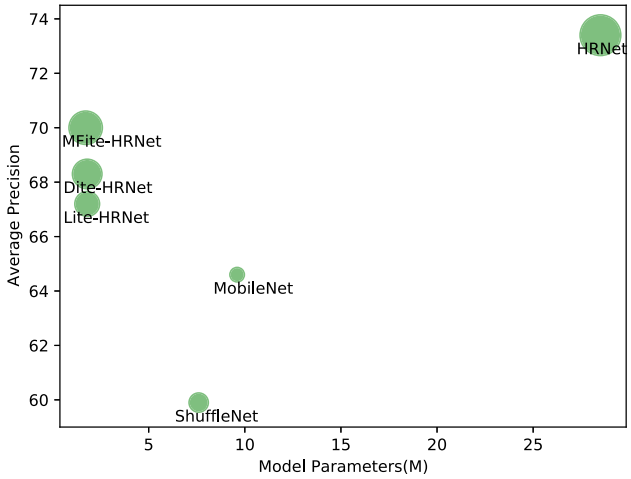


Fig. 1 Comparisons with state-of-the-art lightweight pose estimators regarding model parameters, performance, and receptive fields. The green circles represent receptive fields' the larger radius of a circle, the larger its receptive field

computational bottleneck in MobileNet [5] and other models. Therefore, pointwise convolution with grouping and channel shuffle is presented to reduce model complexity and improve expressiveness. Although ShuffleNet and MobileNet have been applied in HPE, their performance is unappealing with relatively larger model parameters and computations compared to the latest lightweight models designed for HPE.

In light of these limitations, lightweight models based on the promising HRNet [3] become prevailing. The Small HRNet[1] is formulated by simply reducing the depths and widths of HRNet [3], but the performance is unsatisfactory. Lite-HRNet [7] incorporates the Shuffle-ResBlock [6, 8] into a high-resolution network, replacing 1×1 convolution with a conditional channel weighting block to decrease massive computations. However, the slighter network width and depth severely restrain the whole receptive field. Inspired by Lite-HRNet [7], Dite-HRNet [9] advances the model's capability to extract long-range spatial information with larger convolution kernels. Multi-kernel size Dynamic Split Convolution (DSC) is leveraged to dynamically capture multi-scale context information and optimize the trade-off between capacities and performance. Nonetheless, through numerous empirical experiments, we found a more promising manner to allocate appropriate convolution kernels for different model depths. As illustrated in Fig. 1, the lightweight networks, i.e., MobileNet, Lite-HRNet, Dite-HRNet, and MFite-HRNet, possess fewer model parameters and smaller receptive fields than HRNet. Further, our MFite-HRNet, with smaller parameters and larger receptive fields, achieves the best performance compared with other lightweight models. Therefore, expanding the overall perceptive field and customizing different receptive fields are expected to wells balance model's capabilities and complexity.

In this paper, we propose a novel Multi-scale Field Lightweight High-resolution Network (MFite-HRNet) for human pose estimation. Specifically, our model mainly consists of two lightweight blocks, i.e., the Multi-Scale Receptive Field Block (MRB) and the Large-Scale Receptive Field Block (LRB), to enrich feature expressiveness. The MRB utilizes Group Depthwise Dilation Convolutions (GDDC) with the same kernel and varied dilation rates to model multi-scale spatial relationships. Since GDDC can leverage the same parameters to attain larger perceptive fields, MRB enables learning informative human pose patterns and benefits from a better balance between performance and complexity. The LRB manipulates the features of all input channels and models the longer-range spatial dependencies with a 7×7 depthwise convolution kernel. Through MRB and LRB, our network achieves state-of-the-art performance, including single-person pose estimation on COCO [10] and MPII [11] datasets, as well as multi-person pose estimation on COCO and CrowdPose [12] datasets, compared with other lightweight models when the size of model parameters is consistent.

We summarize the main contributions as follows:

- We propose a novel lightweight model, i.e., MFite-HRNet, for human pose estimation. The network is formulated based on the HRNet, mainly consisting of the MRB and LRB blocks to expand receptive fields and reduce model parameters.
- The MRB concentrates on learning multi-scale context feature via group depthwise dilation convolutions with different dilation rates. The LRB is designed to model long-range spatial feature with large depthwise convolution kernel. The two lightweight blocks are leveraged as the foundational components of MFite-HRNet.
- Through adequate balancing between model performance and parameters, experiments on COCO, MPII, and CrowdPose datasets demonstrate that our MFite-HRNet outperforms state-of-the-art lightweight human pose estimation models.

2 Related work

Human pose estimation The HPE task aims to estimate the spatial positions of keypoints for each person in a scene and construct a hinged skeleton representation. Since human numbers in an image are unspecified, existing methods for HPE are approximately classified into single-person-based [3, 13–15] and multi-person- based [2, 16–20] mod-

[1] Small HRNet is available at https://github.com/HRNet/HRNet-Semantic-Segmentation. It simply reduces the depths and widths of the original HRNet.

els. Multi-person pose estimation approaches can be further categorized into top-down [1, 2, 16, 21] and bottom-up [17–20, 22, 23] patterns. The core concepts behind single-person and top-down multi-person pose estimation are essentially the same. That is, persons in natural scenes are first located; then, single-person pose estimation models can be used to predict each person's keypoint positions. On the contrary, bottom-up multi-person methods directly estimate all keypoint heatmaps for all persons simultaneously, following grouping strategies, e.g., associative embedding [24] and part affinity fields [22], leveraged to articulate the same person's keypoints to form a complete skeleton. In this work, we conduct experiments for both single-person and bottom-up multi-person pose estimation tasks with the proposed MFite-HRNet.

Lightweight single-person pose estimation Existing lightweight models for single-person pose estimation are mainly formulated based on two types of backbones. The first one is usually established on prevalent lightweight classification models such as ShuffleNet series [6, 8] and MobileNet and its variants [5, 25, 26], replacing the last fully connection (FC) layer with several upsampling layers and adding shortcut connections for the same scale features. The second type is mainly built based on the HRNet [3] by decreasing its depths and widths. For example, Lite-HRNet [7] combines HRNet with ShuffleNet and provides the approach of cross-resolution weight calculation. Dite-HRNet [9] presents the Dynamic Split Convolution(DSC) and Dynamic Kernel Aggregation(DKA) modules, which are built on the ShuffleNet and use multi-larger convolution kernels to extract long-range spatial correlations. In addition, DeepLab [27] utilizes dilation convolution with different dilation rates to improve the module's ability to extract multi-scale information. Different from the above lightweight pose estimation works and inspired by DeepLab, we design the lightweight MRB to leverage multi-scale depth-wise dilation convolutions, which can effectively facilitate feature extraction while maintaining an equivalent amount of parameters.

Lightweight multi-person pose estimation In general, bottom-up multi-person pose estimation networks perform much faster than top-down ones. Therefore, lightweight multi-person pose estimation endeavors are mainly developed using a bottom-up strategy. The relative lightweight models HigherHRNet-W24 and HigherHRNet-W16 are constructed by directly reducing the model depth and width of HigherHRNet [17], which comprises an HRNet [3] as the backbone and formulates the upsampled high-resolution output. EfficientHRNet [28] has seen attractive progress for multi-person pose prediction, which employs the EfficientNet with a high-resolution structure as the backbone. LitePose [29] utilizes the MobileNet as the backbone and leverages a UNet structure to deconvolute and upsample the high-level feature maps with enhanced results. However, none of the above methods attempt to improve the multi-scale feature extraction capability in the basic block to optimize the model performance. Therefore, we formulate the MRB block to leverage group depthwise dilation convolution with different dilation rates for multi-scale feature learning.

3 Methodology

3.1 Overview of MFite-HRNet

The proposed MFite-HRNet is a high-resolution lightweight pose estimator established based on HRNet [3]. The overall framework is illustrated in Fig. 2, which mainly consists of four stages (different columns) and four branches (different rows). To balance model performance and complexity, we design two novel lightweight blocks, *i.e.,* LRB and MRB, and incorporate them into the network. Precisely, LRB is embedded into stage 1 to extract large-range spatial information at low-level feature maps. MRB is utilized for stages 2, 3, and 4 to learn multi-scale context informative knowledge.

As shown in Fig. 2, given an input image $\mathbf{I} \in H \times W \times 3$, we first utilize a 3×3 Convolution (Conv) with stride 2 to downsample and transform it into a C-dimension feature space. Then, LRB is leveraged to extract large-scale spatial features with the feature maps adapted to $H/4 \times W/4 \times C$. The output of LRB is fed into stage 2 branch 1. Meanwhile, it is downsampled by a 3×3 DepthWise Convolution (DWConv) with stride 2 and expanded by 1x1 convolution to increase nonlinearity. The downsampled feature maps become $H/8 \times W/8 \times 2C$ and serve as the input of stage 2 branch 2. In stages 2, 3, 4, we leverage MRB as the fundamental block. Within each stage, a module is formulated by sequentially stacking the same MRB$-G(G \in 1, 2, 3, 4)$ twice and then fusing feature maps from different branches through fusion block. The fusion block in our model is similar to HRNet [3], replacing regular convolution with Depthwise Separable Convolution (DSConv). $\times M_i, i \in 2, 3, 4$ in Fig. 2 refers to the number of module repetitions for the corresponding stage and is dependent on the configuration of MFxk-HRNet-18 and MFxk-HRNet-30 in Table 1. After stage 4, the feature maps of four branches are upsampled to the resolution of branch 1 and then merged. We employ a 3×3 convolution to predict 2D poses. In the following, we elaborate on the design details of LRB and MRB blocks.

3.2 Multi receptive field block

Our MFite-HRNet is inspired by the HRNet [3]. Although HRNet demonstrates excellent performance in HPE tasks, the two stacked 3×3 convolutions of the basic block (shown in Fig. 3a) utilized in stages 2, 3, and 4 can be further optimized with fewer parameters. Therefore, Small HRNet

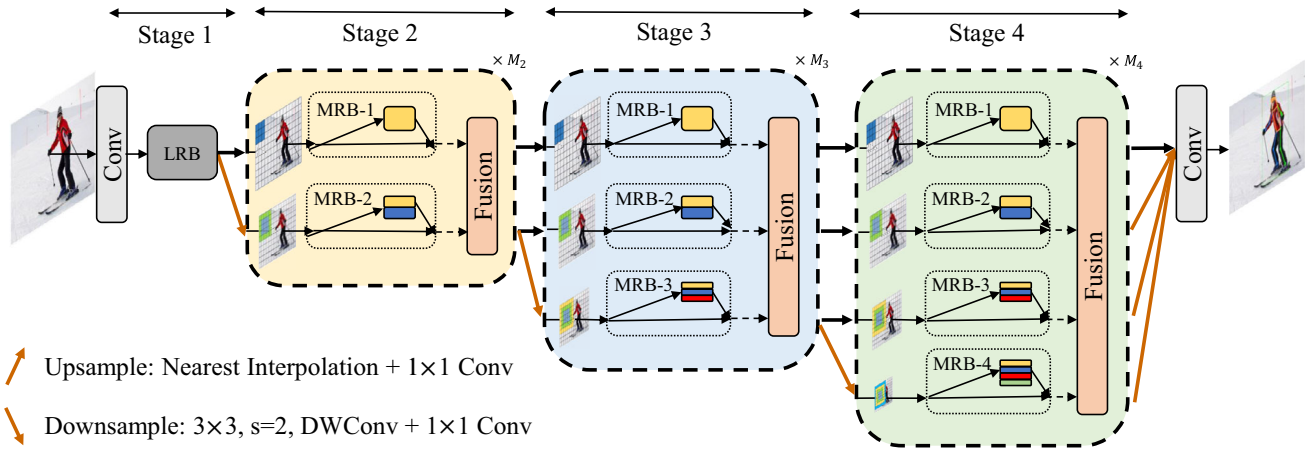


Fig. 2 Framework of the proposed MFite-HRNet. It is established based on the HRNet, LRB, and MRB which constitutes the basic blocks. The LRB aims to model long-range spatial information, and the MRB focuses on multi-scale spatial knowledge learning for different-resolution feature maps. $\times M_i, i \in 2, 3, 4$ refers to the number of repetitions of modules in different stages

Table 1 Structure of MFite-HRNet. Stage 1 has a convolution layer and a LRB Block. There exists a sequence of modules between stage 2 and stage 4, with each module consisting of a group of MRB blocks and a fusion block. k in MFxk-HRNet-N is the base kernel size in MRB, N denotes the number of layers, and $\times M_i$ signifies the module repetition numbers (referred to $\times M_i$ in Fig. 2.)

Stages	Operator	Feature resolutions	Output channels	Modules ($\times M_i$)	
				MFxk-HRNet-18	MFxk-HRNet-30
Image		256 × 192	3		
Stage 1	Conv	128 × 96	32	1	
	LRB	64 × 48	32		
Stage 2	MRB-1, MRB-2	64 × 48, 32 × 24	48, 96	2	3
	Fusion Block				
Stage 3	MRB-1, MRB-2, MRB-3	64 × 48, 32 × 24, 16 × 12	48, 96, 144	4	8
	Fusion Block				
Stage 4	MRB-1, MRB-2, MRB-3, MRB-4	64 × 48, 32 × 24, 16 × 12, 8 × 6	48, 96, 144, 192	2	3
	Fusion Block				

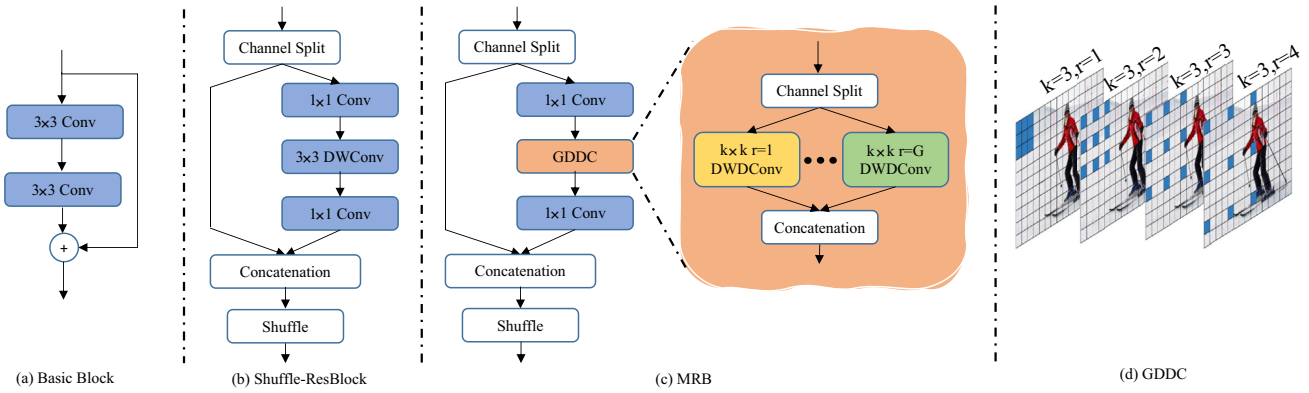


Fig. 3 Comparisons of different basic blocks designed for our MFite-HRNet in stage 2, 3, and 4. **a** Basic block proposed in ResNet [30]. **b** Shuffle-Resblock proposed in ShuffleNet [6]. **c** Our MRB block. **d** GDDC Visualization. k and r denote the kernel size and dilation rate

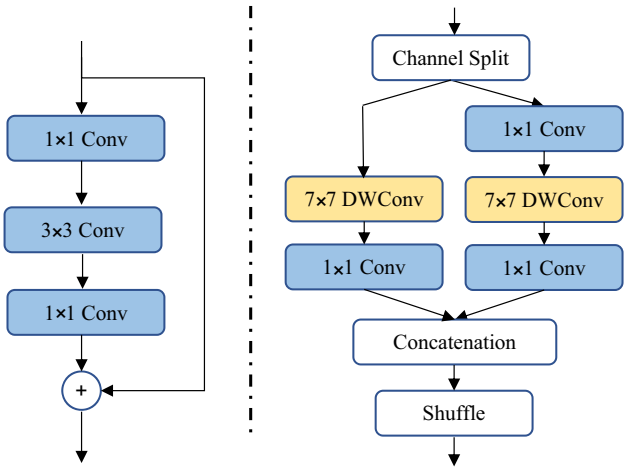
endeavors to reduce the network depths and channel widths of HRNet while keeping the basic block to achieve lightweight. However, the model performance deteriorates seriously after pruning. The results reflect that the basic block in HRNet may not satisfy lightweight requirements. Meanwhile, the Shuffle-ResBlock (shown in Fig. 3b) has been validated to be a convincing design for lightweight models in various vision tasks. It starts by splitting the input features into two segments with equal channels and then leveraging DWConv to process only one segment. Shuffle-ResBlock significantly decreases model parameters and lessens computational efforts. Nevertheless, the receptive field of Shuffle-ResBlock is somewhat restricted, confining model learning ability.

We follow the insights of the Shuffle-ResBlock and present the Group DepthWise Dilation Convolution (GDDC) to replace the DWConv. GDDC is the core component of our lightweight MRB and is formulated by multiple DepthWise Dilation Convolutions (DWDConv) with different dilation rates (shown in Fig. 3c). GDDC firstly equally splits the input into G groups along feature channels. Then, DWDConv is leveraged to process each group with the dilation rate as the group index and output channels are the same as the input. At last, we concatenate the convolution results of the G groups as the output of GDDC.

It is well known that enlarging convolution kernels can benefit long-range feature learning and facilitate model performance. However, directly increasing the kernel size inevitably raises model parameters. Consequently, we design the GDDC component utilizing the same convolution kernel and different dilation rates for informative spatial feature learning, and the DWDConv used by GDDC maintains the same parameters compared with DWConv. The number of G in GDDC depends on where MRB is located on the branch (as shown in Fig. 2). The base kernel size can be flexibly configured according to needs. Thus, with the GDDC, MRB can effectively fuse multi-scale spatial context details and obtain a more extensive range of information than a 3×3 DWConv in the Shuffle-ResBlock. For the proposed framework, we design MFx3-HRNet and MFx5-HRNet based on the base kernel size of 3 and 5 to implement experiments. The details of MFx3-HRNet and MFx5-HRNet are listed in Table 1.

3.3 Large receptive field block

In HRNet, the input image is first quadruple downsampling with two sequential convolutions (kernel size $k = 3$, stride $s = 2$) before feeding into stage 1 (shown in Fig. 4a), in which four bottlenecks are stacked. However, the 3×3 convolution in the four bottlenecks maintains certain parameters. Therefore, we refer to the insight of the Shuffle-ResBlock and equally divide the input into two segments. Feature transformation is only implemented on one segment, and we replace the 3×3 convolution in the bottleneck with a 3×3 DWConv



(a) Bottleneck (b) LRB

Fig. 4 Comparisons of different basic blocks designed for our MFite-HRNet in stage 1. **a** Bottleneck proposed in ResNet [30]. **b** Our LRB block

to further reduce model parameters. As a result, it is found that such a modification leads to performance degradation.

To address the above issue and inspired by the fact that larger feature resolution and convolution kernels can effectively improve model representation, we first only twofold downsample an input image using a 3×3 convolution with a stride of 2. Then, the downsampled features are split into two parts with the same channels. Since averaging pooling tends to cause spatial information loss, we leverage a 7×7 DWConv with the stride of 2 for double downsampling in both parts to obtain the quadruple downsampling feature maps. To this end, we come up with our LRB block.

As illustrated in Fig. 4b, we divide the input channels into two segments to reduce the overall parameters. Then, a nonlinear transformation is performed on one segment using 1×1 convolution, 7×7 DWConv with stride 2, and 1×1 convolution for feature extraction. The other segment does not directly downsample as the Shuffle-ResBlock. Instead, it is processed by 7×7 DWConv with stride 2 and 1×1 convolution to enrich feature diversity. For each convolution operation, it is followed by batch normalization and ReLU. Finally, concatenation and channel shuffle are leveraged to fuse features.

4 Experiments

4.1 Settings

Datasets We investigate the performance of MFite-HRNet on three public widely used human pose estimation datasets, *i.e.,* COCO [10], MPII [11] and CrowdPose [12]. The COCO dataset [10] contains over 200K images and 250K person

instances marked with 17 keypoints. The training set includes 57K images and 150K person instances, while the validation and testing set compromises 5K images and 20K images. We train our model on the COCO train 2017 dataset, evaluate the network on the val2017 set, and test dev 2017 simultaneously. The MPII benchmark [11] possesses approximately 25K images with full-body pose annotations from real-world activities. There are over 40K individual instances, divided into 12K for testing and others for training. The CrowdPose dataset [12] is made up of 20K images with approximately 80K pedestrians labeled with 14 keypoints. This dataset encompasses more crowded scenes than the COCO dataset, posing more challenges to pose estimation methods.

Evaluation metrics Human pose estimation aims to predict body keypoints as close to ground truth as possible. To evaluate the predictions, we utilize the Object Keypoint Similarity (OKS), which calculates the matching degree between the predicted and ground truth values and is normalized by the human scale. The range of OKS varies from 0 to 1, with larger values indicating more accurate predictions. Based on OKS, the average precision (AP) is obtained as the average precision over ten positions at OKS values of 0.50, 0.55,..., 0.90, and 0.95. We also leverage the AP^{50} (AP at OKS = 0.5), AP^{75}, AP^{M}, and AP^{L} and average recall (AR) scores as the criteria. The above metrics are evaluated for the COCO and CrowdPose datasets. For the MPII dataset, we resort to the head-normalized Probability of Correct Keypoints (PCKh) at 50, *i.e.,* PCK@50 to assess the performance. Meanwhile, model parameters (Params (M)) and computation capacity (GFLOPs) are reported to evaluate model complexity.

Parameter settings We follow the default training and evaluation settings of MMPose [31], with an optimizer of Adam and the learning rate of $2e^{-3}$. We conduct experiments on 8 GeForce RTX 3090 GPUs for the single-person pose estimation task with each GPU processing 50 images. Following conventional procedures, image resolutions of 256×192 and 384×288 are leveraged to evaluate model performance. In contrast, for the multi-person trials, we implement experiments on 8 NVIDIA TESLA T4 and set the batch size as 40. To make a fair comparison with LitePose [29], we apply image resolutions of 256×256 and 488×488. We also present the results of Lite-HRNet [7] and Dite-HRNet [9] on the COCO dataset [10] with default parameters on multi-person task. We optimize those models using the default parameters.

Implementation details When estimating single-person pose for the COCO dataset [10], following the protocols of Lite-HRNet [7] and Dite-HRNet [9], we utilize the ground truth bounding boxes containing persons for training. While in the testing phase, we leverage a two-stage top-down person detector [14] to detect human instances first. For MPII [11], we obey the standard strategy to use the provided person boxes for experiments. We use post-Gaussian filters to estimate heatmaps and average the expected heatmaps of the original and flipped images. A quarter offset is applied from the highest answer to the second-highest response to establish each keypoint position. In multi-person pose estimation, predictions are made directly on a whole input image without human detection in advance. We modify our MFite-HRNet based on HigherHRNet [17] for the multi-person prediction and make comparisons with Lite-HRNet [7] and Dite-HRNet [9] at 256×256 and 488×488 resolutions.

4.2 Ablation study

To describe the insights in designing lightweight MRB and validate the effectiveness of MRB and LRB for MFite-HRNet, we perform ablation study on the COCO dataset with image at the resolution of 256×192. Experimental results on the validation set are reported in Table 2. Baseline-1 is constructed based on the small HRNet by replacing its basic block in stages 2, 3, and 4 with the Shuffle-ResBlock [6]. Furthermore, we regard the Lite-HRNet-18 [7] as our new baseline model, *i.e.,* Baseline-2.

Designing insights of MRB In the proposed MFite-HRNet, MRB is the fundamental block. It is formulated based on multi-scale DWDConv with the same convolution kernel (k = 3) and group dilation rates (r = 1, 2, 3, 4). The maximum equivalent convolution kernel sizes of MRB in different stages and branches range in 3, 5, 7, and 9. In contrast, we implement experiments using single-scale DWDConv to obtain equivalent convolution kernel sizes. To this end, we design two comparison experiments with the same model parameters as MFite-HRNet leveraging DWDConv. We progressively increase the dilation rate r from 2 to 4 in stages 2, 3, and 4 for Exp-1, while increasing r from 1 to 4 in branches 1, 2, 3, 4 for Exp-2. Though Exp-1 and Exp-2 have the maximum equivalent convolution kernel sizes as MRB, the single-scale dilation operation leads to spatial information loss. Therefore, the MRB we designed uses group depth-wise dilation convolutions with varied scales to improve the model's long-distance and multi-scale spatial information extraction ability.

Effectiveness of MRB MRB is leveraged to learn multi-scale spatial context information at different level features with lightweight parameters. To verify its superiority, we conduct experiments by replacing the DWConv (k=3) block of Baseline-1 in stage 2, 3, and 4 with our MRB. From Exp-4 in Table 2, we can safely conclude that with the proposed MRB, the AP metric has been improved by 1.4 compared to Baseline-1. Furthermore, for the Baseline-2, when we directly substitute the basic block of Lite-HRNet-18 with our MRB, the AP and AR of Exp-6 can achieve 1.0 and 0.9 enhancements. We also observe that Baseline-1 outperforms Baseline-2 slightly. The reasons are that Baseline-2 replaces the 1x1 convolution of Shuffle-ResBlock in Baseline-1 with conditional channel weighting, which

Table 2 Ablation studies on the COCO val2017 set

Experiments	Model	Params (M)	GFLOPs	AP	AR
Baseline-1	DWConv ($k = 3$) Stages 2/3/4	1.1	0.4	65.5	71.7
Exp-1	DWDConv ($k = 3$) Stage 2 ($r = 2$) Stage 3 ($r = 3$) Stage 4 ($r = 4$)	1.1	0.4	63.6	70.1
Exp-2	DWDConv ($k = 3$) Branch 1 ($r = 1$) Branch 2 ($r = 2$) Branch 3 ($r = 3$) Branch 4 ($r = 4$)	1.1	0.4	63.2	69.9
Exp-3	LRB Stage 1/2/3/4	1.23	0.5	67.7	73.7
Exp-4	w MRB, w/o LRB	1.06	0.37	66.9	73.0
Exp-5	MFite-HRNet (w MRB, w LRB)	1.06	0.37	67.1	73.1
Baseline-2	Lite-HRNet-18 [7]	1.1	0.2	64.8	71.2
Exp-6	Lite-HRNet-18 + MRB	1.1	0.2	65.8	72.1

The GFLOPs are computed with the input size 256 × 192. k = kernel size. r = dilation rate. MRB is formulated based on DWDConv with $k = 3$ and varied r at different stages and branches

reduces feature dimensions and then lifts them. Such a modification relieves computational effort but sacrifices some information. Nevertheless, the performance improvements in Exp-6 over Baseline-2 verify the advantages of our MRB block. We attribute the gains for MRB effectively reinforcing the model's ability to extract multi-scale long-range context knowledge, so as to benefit human pose estimation.

Effectiveness of LRB LRB is designed for extracting large-range spatial information. To verify its effectiveness, we replace all DWConv of stage 2/3/4 in Baseline-1 with LRBs and name it Exp-3. It can be seen from Table 2 that 2.2 AP and 0.2 AR gains have been obtained for Exp-3 relative to the Baseline-1 experiment. Comparing Exp-3 and Exp-5, we can further observe that the model constructed with total LRB blocks also performs better than the MFite-HRNet. Meanwhile, the larger convolution kernel of LRB increases the model parameters in contrast with the Exp-5 model. Therefore, we use MRB in stages 2/3/4 and LRB in stage 1 to achieve the best model parameters and performance balance.

4.3 Comparisons with state-of-the-arts for SPE

We conduct single-person pose estimation (SPE) on the COCO val2017, COCO test-dev2017 and MPII val sets and make comparisons with state-of-the-arts (SOTA).

COCO Val2017 set The comparisons with current cutting-edge methods on the COCO val2017 set are presented in Table 3. Our trained-from-scratch MFite-HRNet with different backbones and input sizes attain impressive performance gain with much less complexity, compared with methods based prevalent lightweight models, *i.e.,* MobileNetV2 [25] and ShuffleNetV2 [8]. Meanwhile, with nearly the same parameters, the proposed MFx5-HRNet-18 and MFx5-HRNet-30 yield 3.1 and 2.8 AP profits compared to Lite-HRNet-18 and Lite-HRNet-30, as well as 2.0 and 1.7 AP gains in contrast with Dite-HRNet-18 and Dite-HRNet-30 [9]. As for the Small HRNet, our progress achieves more than 12.7 for the AP metric. Compared to large networks such as Hourglass[1] and CPN [2], our network gets equivalent AP scores while negligibly modeling complexity. We further display some pose estimation results of MFx3-HRNet-30 in Fig. 5. It can be observed that our model performs well for large viewpoint changes, partial keypoint occlusions and multi-person scenes. Quantitative and qualitative results demonstrate the superiority of our MFite-HRNet.

COCO Test-dev2017 set Table 4 reports the comparison results of our model and state-of-the-art methods on the COCO test-dev2017 set. Our MFx5-HRNet-30 with the image resolution of 256×192 achieves a 71.2 AP score, which is significantly better than the large networks 8-stage Hourglass [1] and CPN [2] with fewer GFLOPs and parameters. Compared to Dite-HRNet-18 [9], MFx5-HRNet-18 improves AP by 1.0 points and AR by 0.9 points with the equivalent model complexity. Although both Dite-HRNet [9] and MFite-HRNet have improved Shuffle-ResBlock to learn long-range features, MFite-HRNet's GDDC leverages DWDConv with smaller feature channels in low-resolution branches, but more channels in high-resolution branches than Dite-HRNet's DWConv. Hence, MFite-HRNet achieves better performance with similar parameters compared to Dite-HRNet. In spite of having some performance gaps compared with some large networks [3, 32], our networks have considerably lower GFLOPs and model parameters. Those results demonstrate the superiority of our model for single-person pose estimation on the COCO test-dev2017 set.

MPII Val set Table 5 shows the results for our network and the most advancing lightweight networks on MPII val set. Our MFx5-HRNet-18 attains 87.5 PCKh@0.5, a higher accuracy with fewer parameters and it outperforms MobileNetV2 [25], MobileNetV3[26], ShuffleNetV2 [8], Small HRNet, Lite-HRNet [7], and Dite-HRNet [9] by 2.1, 3.2, 4.7, 7.3, 1.4, and 0.5 points, respectively. The improvement gaps for MFx5-HRNet-30 become larger in contrast with other networks as the model size grows. The performance gains with lightweight model complexity validate the effectiveness of our MFite-HRNet on MPII val set.

Table 3 Comparisons with SOTA lightweight models on the COCO val2017 set for SPE

Method	Backbone	Pretrain	Input size	#Params (M)	GFLOPs	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}	AR
Large networks											
8-stage Hourglass [1]	Hourglass	N	256 × 192	25.1	14.3	66.9	–	–	–	–	–
CPN [2]	ResNet-50	Y	256 × 192	27.0	6.2	68.6	–	–	–	–	–
Simple Baseline [14]	ResNet-50	Y	256 × 192	34.0	8.9	70.4	88.6	78.3	67.1	77.2	76.3
HRNet [3]	HRNet-W32	N	256 × 192	28.5	7.1	73.4	89.5	80.7	70.2	80.1	78.9
UDP [32]	HRNet-W32	Y	256 × 192	28.7	7.1	75.2	92.4	82.9	72.0	80.8	80.4
Small networks											
MobileNetV2 [25]	MobileNetV2	N	256 × 192	9.6	1.4	64.6	87.4	72.3	61.1	71.2	70.7
ShuffleNetV2 [8]	ShuffleNetV2	N	256 × 192	7.6	1.2	59.9	85.4	66.3	56.6	66.2	66.4
Small HRNet	HRNet-W18	N	256 × 192	1.3	0.5	55.2	83.7	62.4	52.3	61.0	62.1
Lite-HRNet [7]	Lite-HRNet-18	N	256 × 192	1.1	0.2	64.8	86.7	73.0	62.1	70.5	71.2
	Lite-HRNet-30	N	256 × 192	1.8	0.3	67.2	88.0	75.0	64.3	73.1	73.3
Dite-HRNet [9]	Dite-HRNet-18	N	256 × 192	1.1	0.2	65.9	87.3	74.0	63.2	71.6	72.1
	Dite-HRNet-30	N	256 × 192	1.8	0.3	68.3	88.2	76.2	65.5	74.1	74.2
MFite-HRNet	MFx3-HRNet-18	N	256 × 192	1.06	0.37	67.1	87.4	74.8	64.2	72.8	73.1
	MFx3-HRNet-30	N	256 × 192	1.72	0.6	69.2	88.2	76.8	66.4	75.3	75.2
MFite-HRNet	MF-5-HRNet-18	N	256 × 192	1.1	0.4	67.9	87.6	75.6	64.9	73.7	73.7
	MFx5-HRNet-30	N	256 × 192	1.79	0.65	**70.0**	**88.5**	**77.8**	**66.9**	**75.8**	**76.1**
MobileNetV2 [25]	MobileNetV2	N	384 × 288	9.6	3.3	67.3	87.9	74.3	62.8	74.7	72.9
ShuffleNetV2 [8]	ShuffleNetV2	N	384 × 288	7.6	2.8	63.6	86.5	70.5	59.5	70.7	69.7
Small HRNet	HRNet-W18	N	384 × 288	1.3	1.2	56.0	83.8	63.0	52.4	62.6	62.6
Lite-HRNet [7]	Lite-HRNet-18	N	384 × 288	1.1	0.4	67.6	87.8	75.0	64.5	73.7	73.7
	Lite-HRNet-30	N	384 × 288	1.8	0.7	70.4	88.7	77.7	67.5	76.3	76.2
Dite-HRNet [9]	Dite-HRNet-18	N	384 × 288	1.1	0.4	69.0	88.0	76.0	65.5	75.5	75.0
	Dite-HRNet-30	N	384 × 288	1.8	0.7	71.5	88.9	78.2	68.2	77.7	77.2
MFite-HRNet	MFx3-HRNet-18	N	384 × 288	1.06	0.83	69.2	88.1	76.4	65.7	75.6	74.8
	MFx3-HRNet-30	N	384 × 288	1.72	1.31	70.8	88.6	77.9	67.3	77.3	76.3
MFite-HRNet	MFx5-HRNet-18	N	384 × 288	1.1	0.89	70.4	88.3	77.5	66.8	76.7	76.2
	MFx5-HRNet-30	N	384 × 288	1.79	1.43	**72.1**	**88.9**	**78.4**	**68.7**	**78.2**	**77.7**

Pretrain = pretrain the backbone on the ImageNet classification task. Bold indicates the best results

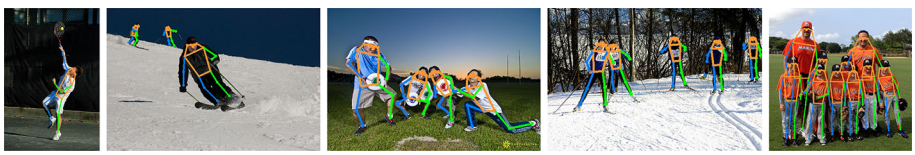


Fig. 5 Visualization of pose estimation results on COCO val2017 using MFx3-HRNet-18 model

4.4 Comparisons with state-of-the-arts for MPE

For the multi-person pose estimation (MPE), we implement experiments on the COCO val2017 set and CrowndPose test set. Tables 6 and 7 report contrast with the most advancing lightweight methods. To make fair comparisons, we train and test Lite-HRNet [7] and Dite-HRNet [9] using the same configurations as our MFite-HRNet.

COCO Val2017 set The results for MPE in COCO val2017 are presented in Table 6. Our approach achieves superior results at a resolution of 256 × 256, surpassing Lite-HRNet [7] by 4.4 and Dite-HRNet [9] by 7 regarding AP metric. When making contrast with Litepose [29], our model produces 1.1 AP profit. Notably, our model exhibits a remarkable performance improvement of 6.2 compared to EfficientHRNet [28]. Those results validate the advantages of MFite-HRNet for MPE task.

CrowdPose test set We evaluate our approach on the CrowdPose dataset with images at resolutions of 256 × 256 and 448 × 448 and report the results in Table 7. It can be observed that our approach outperforms Lite-HRNet [7] and Dite-HRNet [9] with large margins. Since our model

Table 4 Comparisons with SOTA models on the COCO test-dev2017 set for SPE

Method	Backbone	Pretrain	Input size	#Params (M)	GFLOPs	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}	AR
Large networks											
8-stage Hourglass [1]	Hourglass	N	256 × 192	25.1	14.3	66.9	–	–	–	–	–
CPN [2]	ResNet-50	Y	256 × 192	27.0	6.2	68.6	–	–	–	–	–
Simple Baseline [14]	ResNet-50	Y	256 × 192	34.0	8.9	70.0	90.9	77.9	66.8	75.8	75.6
HRNet [3]	HRNet-W32	N	256 × 192	28.5	7.1	73.4	89.5	80.7	70.2	80.1	78.9
UDP [32]	HRNet-W32	Y	256 × 192	28.7	7.1	75.2	92.4	82.9	72.0	80.8	80.4
Small networks											
MobileNetV2 [25]	MobileNetV2	N	256 × 192	9.6	1.4	64.6	87.4	72.3	61.1	71.2	70.7
ShuffleNetV2 [8]	ShuffleNetV2	N	256 × 192	7.6	1.2	59.9	85.4	66.3	56.6	66.2	66.4
Small HRNet	HRNet-W18	N	256 × 192	1.3	0.5	55.2	83.7	62.4	52.3	61.0	62.1
Lite-HRNet [7]	Lite-HRNet-18	N	256 × 192	1.1	0.2	63.7	88.6	71.1	61.1	68.6	69.7
	Lite-HRNet-30	N	256 × 192	1.8	0.3	66.7	89.9	74.9	63.9	71.9	72.7
Dite-HRNet [9]	Dite-HRNet-18	N	256 × 192	1.1	0.2	–	–	–	–	–	–
	Dite-HRNet-30	N	256 × 192	1.8	0.3	–	–	–	–	–	–
MFite-HRNet	MFx3-HRNet-18	N	256 × 192	1.06	0.37	66.4	89.5	74.4	63.8	71.4	72.3
	MFx3-HRNet-30	N	256 × 192	1.72	0.6	68.5	90.2	76.7	65.9	73.5	74.3
MFite-HRNet	MFx5-HRNet-18	N	256 × 192	1.1	0.4	67.4	89.7	75.5	64.5	72.6	73.1
	MFx5-HRNet-30	N	256 × 192	1.79	0.65	**69.4**	**90.6**	**77.6**	**66.7**	**74.4**	**75.1**
MobileNetV2 [25]	MobileNetV2	N	384 × 288	9.8	3.3	66.8	90.0	74.0	62.6	73.3	72.3
ShuffleNetV2 [8]	ShuffleNetV2	N	384 × 288	7.6	2.8	62.9	88.5	69.4	58.9	69.3	68.9
Small HRNet	HRNet-W18	N	384 × 288	1.3	1.2	55.2	85.8	61.4	51.7	61.2	61.5
Lite-HRNet [7]	Lite-HRNet-18	N	384 × 288	1.1	0.4	66.9	89.4	74.4	64.0	72.2	72.6
	Lite-HRNet-30	N	384 × 288	1.8	0.7	69.7	90.7	77.5	66.9	75.0	75.4
Dite-HRNet [9]	Dite-HRNet-18	N	384 × 288	1.1	0.4	68.4	89.9	75.8	65.2	73.8	74.4
	Dite-HRNet-30	N	384 × 288	1.8	0.7	70.6	90.8	78.2	67.4	76.1	76.4
MFite-HRNet	MFx3-HRNet-18	N	384 × 288	1.06	0.83	68.3	89.8	75.8	65.2	73.8	73.9
	MFx3-HRNet-30	N	384 × 288	1.72	1.31	70.1	90.5	77.7	66.9	75.8	75.6
MFite-HRNet	MFx5-HRNet-18	N	384 × 288	1.1	0.89	69.4	90.2	76.5	66.2	75.0	75.3
	MFx5-HRNet-30	N	384 × 288	1.79	1.43	**71.2**	**90.9**	**78.5**	**68.0**	**76.9**	**77.0**

Pretrain = pretrain the backbone on the ImageNet classification task. Bold indicates the best results

Table 5 Comparisons with SOTA lightweight models on the MPII val set for SPE

Method	#Params (M)	GFLOPs	PCKh
MobileNetV2 1× [25]	9.6	1.9	85.4
MobileNetV3 1× [26]	8.7	1.8	84.3
ShuffleNetV2 1× [8]	7.6	1.7	82.8
Small HRNet	1.3	0.7	80.2
Lite-HRNet-18 [7]	1.1	0.2	86.1
Lite-HRNet-30 [7]	1.8	0.4	87.0
Dite-HRNet-18 [9]	1.1	0.2	87.0
Dite-HRNet-30 [9]	1.8	0.4	87.6
MFx3-HRNet-18	1.06	0.4	87.3
MFx3-HRNet-30	1.72	0.6	88.0
MFx5-HRNet-18	1.1	0.4	87.5
MFx5-HRNet-30	1.79	0.6	**88.4**

Bold indicates the best results

does not utilize pre-trained models and elaborately tuning strategies such as Neural Architecture Search(NAS) [33], the performance is inferior to those of LitePose. Nonetheless, our approach performs better than other lightweight models under the same training configuration.

5 Conclusion

In this paper, we propose the novel lightweight MFite-HRNet for human pose estimation tasks. To address the problem of small receptive fields and lack of long-range spatial information modeling in existing lightweight networks, we formulate two novel lightweight basic blocks, i.e., LRB and MRB, which are, respectively, responsible for learning large-range spatial information at low-level feature maps and multi-scale spatial context at varied feature maps. With

Table 6 Comparisons with SOTA lightweight models on the COCO val2017 set for MPE

Method	Backbone	Pretrain	Input size	#Params (M)	GFLOPs	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}	AR
Large networks											
OpenPose [22]	–	–	–	–	–	61.8	84.9	67.5	57.1	68.2	–
Hourglass [1]	Hourglass	N	512 × 512	277.8	206.9	56.6	81.8	61.8	49.8	67.0	–
HigherHRNet [17]	HRNet-W32	N	512 × 512	28.6	47.9	67.7	87.0	73.8	61.9	76.3	72.3
Small networks											
EfficientHRNet [28]	$EfficientHRNet_{-4}$	Y	384 × 384	3.7	2.1	35.5	–	–	–	–	–
	$EfficientHRNet_{-2}$	Y	448 × 448	10.3	7.7	52.8	–	–	–	–	–
LitePose [29]	LitePose-XS	Y	256 × 256	1.7	1.2	40.6	–	–	–	–	–
Lite-HRNet [7]	Lite-HRNet-30	N	256 × 256	1.8	1.95	37.3	65.6	36.5	–	–	44.3
Dite-HRNet [9]	Dite-HRNet-30	N	256 × 256	1.8	1.95	34.7	63.2	33.3	–	–	41.6
MFite-HRNet	MFx3-HRNet-30	N	256 × 256	1.8	2.43	39.8	67.4	39.7	–	–	46.1
MFite-HRNet	MFx5-HRNet-30	N	256 × 256	1.8	2.43	**41.7**	**69.2**	**42.9**	–	–	**48.0**

Pretrain = pretrain the backbone on the ImageNet classification task. Bold indicates the best results

Table 7 Comparisons with SOTA lightweight models on the CrowdPose test set for MPE

Method	Backbone	Pretrain	Input size	#Params (M)	GFLOPs	AP	AP^{50}	AP^{75}	AR
Large networks									
OpenPose [22]	–	–	–	–	–	61.8	84.9	67.5	–
Hourglass [1]	Hourglass	N	512 × 512	277.8	206.9	56.6	81.8	61.8	–
HigherHRNet [17]	HRNet-W32	N	512 × 512	28.6	47.9	67.7	87.0	73.8	72.3
Small networks									
EfficientHRNet [28]	$EfficientHRNet_{-3}$	Y	416 × 416	5.3	4.3	46.1	79.3	48.3	–
	$EfficientHRNet_{-1}$	Y	480 × 480	13.0	14.2	56.3	81.3	59.0	–
LitePose [29]	LitePose-XS	Y	256 × 256	1.7	1.2	49.5	74.5	51.4	–
	LitePose-S	Y	448 × 448	2.7	5.0	**58.3**	**81.1**	**61.8**	–
Lite-HRNet [7]	Lite-HRNet-30	N	256 × 256	1.8	2.3	42.1	70.8	41.2	49.6
	Lite-HRNet-30	N	448 × 448	1.8	7.2	51.1	78.5	52.3	58.5
Dite-HRNet [9]	Dite-HRNet-30	N	256 × 256	1.8	2.3	41.8	70.6	40.7	49.5
	Dite-HRNet-30	N	448 × 448	1.8	7.2	51.1	78.6	52.0	58.6
MFite-HRNet	MFx3-HRNet-30	N	256 × 256	1.8	2.43	44.8	72.9	44.6	52.2
	MFx3-HRNet-30	N	448 × 448	1.8	16.2	55.0	79.9	56.2	62.3
MFite-HRNet	MFx5-HRNet-30	N	256 × 256	1.8	2.43	45.7	73.8	47.2	52.1
	MFx5-HRNet-30	N	448 × 448	1.8	16.2	56.1	81.1	58.4	62.8

Pretrain = pretrain the backbone on the ImageNet classification task. Bold indicates the best results

the proposed LRB and MRB blocks, our model performs superior compared with state-of-the-art lightweight pose estimation models on the COCO, MPII, and CrowdPose datasets.

Acknowledgements This research is supported by National Key R&D Program of China (No. 2022ZD0115902) and National Natural Science Foundation of China (Nos. 62102208, 62272017, U20A20195, 62172437).

Data Availability Data are available on reasonable request from the corresponding author.

Declarations

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

References

1. Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. In: ECCV, pp. 483–499 (2016)
2. Chen, Y., Wang, Z., Peng, Y., Zhang, Z., Yu, G., Sun, J.: Cascaded pyramid network for multi-person pose estimation. In: CVPR, pp. 7103–7112 (2018)

3. Sun, K., Xiao, B., Liu, D., Wang, J.: Deep high-resolution representation learning for human pose estimation. In: CVPR, pp. 5693–5703 (2019)
4. Wang, C.-H., Huang, K.-Y., Yao, Y., Chen, J.-C., Shuai, H.-H., Cheng, W.-H.: Lightweight deep learning: an overview. IEEE CONSUM ELECTR M, 1–12 (2022) doi:https://doi.org/10.1109/MCE.2022.3181759
5. Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H.: Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv PrePrint: arXiv:1704.04861 (2017)
6. Zhang, X., Zhou, X., Lin, M., Sun, J.: Shufflenet: An extremely efficient convolutional neural network for mobile devices. In: CVPR, pp. 6848–6856 (2018)
7. Yu, C., Xiao, B., Gao, C., Yuan, L., Zhang, L., Sang, N., Wang, J.: Lite-hrnet: A lightweight high-resolution network. In: CVPR, pp. 10440–10450 (2021)
8. Ma, N., Zhang, X., Zheng, H.-T., Sun, J.: Shufflenet v2: Practical guidelines for efficient cnn architecture design. In: ECCV, pp. 116–131 (2018)
9. Li, Q., Zhang, Z., Xiao, F., Zhang, F., Bhanu, B.: Dite-hrnet: Dynamic lightweight high-resolution network for human pose estimation. In: IJCAI, pp. 1095–1101 (2022)
10. Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: ECCV, pp. 740–755 (2014)
11. Andriluka, M., Pishchulin, L., Gehler, P., Schiele, B.: 2d human pose estimation: New benchmark and state of the art analysis. In: CVPR, pp. 3686–3693 (2014)
12. Li, J., Wang, C., Zhu, H., Mao, Y., Fang, H.-S., Lu, C.: Crowdpose: Efficient crowded scenes pose estimation and a new benchmark. In: CVPR, pp. 10863–10872 (2019)
13. Wei, S.-E., Ramakrishna, V., Kanade, T., Sheikh, Y.: Convolutional pose machines. In: CVPR, pp. 4724–4732 (2016)
14. Xiao, B., Wu, H., Wei, Y.: Simple baselines for human pose estimation and tracking. In: ECCV, pp. 466–481 (2018)
15. Sun, X., Xiao, B., Wei, F., Liang, S., Wei, Y.: Integral human pose regression. In: ECCV, pp. 529–545 (2018)
16. Fang, H.-S., Xie, S., Tai, Y.-W., Lu, C.: Rmpe: Regional multi-person pose estimation. In: ICCV, pp. 2334–2343 (2017)
17. Cheng, B., Xiao, B., Wang, J., Shi, H., Huang, T.S., Zhang, L.: Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation. In: CVPR, pp. 5386–5395 (2020)
18. Geng, Z., Sun, K., Xiao, B., Zhang, Z., Wang, J.: Bottom-up human pose estimation via disentangled keypoint regression. In: CVPR, pp. 14676–14686 (2021)
19. Jin, S., Liu, W., Xie, E., Wang, W., Qian, C., Ouyang, W., Luo, P.: Differentiable hierarchical graph grouping for multi-person pose estimation. In: ECCV, pp. 718–734 (2020)
20. Kreiss, S., Bertoni, L., Alahi, A.: Pifpaf: Composite fields for human pose estimation. In: CVPR, pp. 11977–11986 (2019)
21. Kendall, A., Grimes, M., Cipolla, R.: Posenet: A convolutional network for real-time 6-dof camera relocalization. In: ICCV, pp. 2938–2946 (2015)
22. Cao, Z., Simon, T., Wei, S.-E., Sheikh, Y.: Realtime multi-person 2d pose estimation using part affinity fields. In: CVPR, pp. 7291–7299 (2017)
23. Insafutdinov, E., Pishchulin, L., Andres, B., Andriluka, M., Schiele, B.: Deepercut: A deeper, stronger, and faster multi-person pose estimation model. In: ECCV, pp. 34–50 (2016)
24. Newell, A., Huang, Z., Deng, J.: Associative embedding: End-to-end learning for joint detection and grouping. In: NeurIPS, pp. 2277–2287 (2017)
25. Sandler, M., Howard, A.G., Zhu, M., Zhmoginov, A., Chen, L.: Mobilenetv2: Inverted residuals and linear bottlenecks. In: CVPR, pp. 4510–4520 (2018)
26. Howard, A., Pang, R., Adam, H., Le, Q.V., Sandler, M., Chen, B., Wang, W., Chen, L., Tan, M., Chu, G., Vasudevan, V., Zhu, Y.: Searching for mobilenetv3. In: ICCV, pp. 1314–1324 (2019)
27. Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. TPAMI **40**(4), 834–848 (2017)
28. Neff, C., Sheth, A., Furgurson, S., Tabkhi, H.: Efficienthrnet: Efficient scaling for lightweight high-resolution multi-person pose estimation. arXiv preprint arXiv:2012.14214 (2020)
29. Wang, Y., Li, M., Cai, H., Chen, W.-M., Han, S.: Lite pose: Efficient architecture design for 2d human pose estimation. In: CVPR, pp. 13126–13136 (2022)
30. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR, pp. 770–778 (2016)
31. Contributors, M.: OpenMMLab pose estimation toolbox and benchmark. https://github.com/open-mmlab/mmpose (2020)
32. Huang, J., Zhu, Z., Guo, F., Huang, G.: The devil is in the details: delving into unbiased data processing for human pose estimation. In: CVPR, pp. 5700–5709 (2020)
33. Cai, H., Chen, T., Zhang, W., Yu, Y., Wang, J.: Efficient architecture search by network transformation. In: AAAI, vol. 32 (2018)

Shuo Li received the B.S. degree in School of electronic and information engineering from Tiangong university, Tianjin, China, in 2016. He worked at the Research Institute under Beijing Institute of Technology from 2016 to 2020. He is currently a M.S candidate in School of Software, Beihang University, Beijing, China. His research interest is in deep learning, human pose estimation, computer vision.

Ju Dai is currently a Research Assistant in Peng Cheng Laboratory (PCL), Shenzhen, China. She received both B.S. and M.S. degree in Electronic Engineering, China University of Geosciences (CUG), Wuhan, China, in 2011 and 2014, respectively, and the Ph.D. degree in Signal Processing in Dalian University of Technology (DUT), Dalian, China, in 2020. She worked in PCL as Postdoctoral Research Fellow from 2020 to 2022. Her research inter ests include human pose estimation, motion prediction, computer animation, person re-identification and saliency detection.

Zhangmeng Chen received the B.S. degree in School of Computer Science and Technology, Beijing Institute of Technology, Beijing, in 2012, the M.S degree in School of Computer and Information Technology, Beijing Jiaotong University, Beijing, in 2016. He is currently a Ph.D candidate in School of Computer Science, Beihang University. His research interest is in deep learning, 3D human pose estimation, action recognition.

Junjun Pan is currently a professor in School of Computer Science, Beihang University. He received both B.S and M.S degree in School of Computer Science, Northwestern Polytechnical University, China. In 2006, he studied in National Centre for Computer Animation (NCCA), Bournemouth University, UK as Ph.D. candidate with full scholarship. In 2010, he received the Ph.D. degree and worked in NCCA as Postdoctoral Research Fellow. From 2012 to 2013, he worked as a Research Associate in Center for Modeling, Simulation and Imaging in Medicine, Rensselaer Polytechnic Institute, USA. In November 2013, he was appointed as Associate Professor in School of Computer Science, Beihang University, China. His research interests include virtual surgery and computer animation.