IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.Doi Number

Lightweight Cross-fusion Network on Human Pose Estimation for Edge Device

First A. Xian Zhu[1,2], Second B. Xiaoqin Zeng[1], and Third C. Wei Ma[3,4]

[1]College of Computer and Information, Hohai University, Nanjing, Jiangsu 210098, China
[2]Department of Computer Science, Zijin College, Nanjing University of Science and Technology, Nanjing 210046, Jiangsu, China
[3]State Key Laboratory for Novel Software Technology, Department of Computer Science and Technology, Nanjing University, Nanjing, Jiangsu 210093, China
[4]Nanjing Institute of Tourism and Hospitality, Nanjing, Jiangsu 211100, China

Corresponding author: Second B. Xiaoqin Zeng (e-mail: zhuxian0920@yandex.com).

ABSTRACT The deployment of human pose estimation on edge devices are essential task in computer vision. Due to memory and storage space limitations, it is difficult for edge devices to maintain implementing Convolutional Neural Networks, which deployed large-scale terminal platforms with abundant computing resources. This paper proposed novel Lightweight Cross-fusion Network on Human Pose Estimation with information sharing. Using state-of-the-art efficient neural architecture, and Ghost Net, as the backbone, which are gradually applying a cross-information fusion network for key points extraction in the baseline and strengthen phases. As a result, the computational cost significantly reduces, while maintaining feature confidence more accurate and predicting key points heatmaps more precisely. Our network model can entirely execute on edge devices, and extensive self-comparison experiments have evaluated the architecture's effectiveness. The MS COCO 2017 dataset proved that the cross-fusion network is superior than other lightweight structures for pose estimation

Keywords: cross-fusion Network, human pose estimation, lightweight

I. INTRODUCTION

Real-time human pose estimation has received much attention in recent years due to its intelligent recognition feature, which offers critical application scenarios, including autonomous driving, intelligent security, human action recognition, etc. The robust Convolutional Neural Networks, offering simplicity and speed during learning and inference [1]. Such networks show to be the state-of-art approach in human pose estimation such as single-person pose estimation [2][3][4], multi-person pose estimation [5][6][7][8][9], video pose estimation, and tracking [10][11].

Improving accuracy and lightweight are the two main research goals of neural network design, On the one hand, the traditional Conventional Deep Neural Network obtains satisfactory accuracy by increasing the network layer to generate many parameters and floating-point operations. Therefore, such deep neural networks have significantly required computational resources, which exceed the power of many embedded developer kits, robots and edge devices. Network pruning [12][13], low-bit quantization[14][15], knowledge distillation[16] [17], and other methods apply to compact network structures, simultaneously, the lightweight network architecture of Shuffle Net[18][19]and Mobile Net[20][21][22]with utilizing the depth wise and pointwise convolution has achieved considerable success with fewer parameters and computation complexity.

On the other hand, the state-of-the-art lightweight network architecture reduced computation complexity and improved recognition accuracy by group convolutions. Learning multiple tasks has the advantages of reserving more intrinsic information. For instance, HR Net [23] gradually adopt high-to-low resolution from high-resolution subnetwork, and continuously fuse multi-resolution subnetworks by sharing the information through the whole process. Multi-task learning by exchanging information will help enforce a model with better generalizing ability than single-task learning [24].

Based on the above observation, we expressed a new architecture, namely Lightweight Cross-fusion Network, tailored explicitly for embedded development equipment and real-time target detection requirements. Our network comprises two set layer modules: backbone layer module and information cross-layer module [25]. The backbone layer module starts from the lightweight architecture, which reduces computation cost and parameters by adding the Ghost module [26]. Another layer module extracts the initial features obtained from the light network into two independent feature branches, eliminates interference information, and then shares sufficient information to

This work was supported in part by the 2018 Qing Lan Project of Jiangsu Province of China, Technology Zijin College (No. 2020ZRKX0401001, 2020ZRKX0401005), the University Science Research Project of Jiangsu Province (No. 17KJB520013,18KJB520023), the Project of Philosophy and Social Science Research in Colleges and Universities in Jiangsu Province (No.2020SJA0794)

enhance each component's feature extraction through cross-fusion.

Therefore, our lightweight cross-fusion network can maintain the real state of human pose estimation and can be applied to embedded applications, reducing parameters, and reducing immediate memory access. Consequently, for conventional network assessment from the COCO key points detection dataset [27], our work has two contributions: (I) We utilize the state-of-the-art efficient neural architecture, Ghost Net [26] as the backbone to generate more superior features by using fewer parameters. Thus, our approach can apply to embedded equipment for the Nvidia TX2 device. (ii) A new cross information network, performed repeatedly to boost the features representations in a multi-fuse setup, is proposed. Then, our approach effectively enhances the estimation accuracy with the help of improving the sharing information branch.

II. Related Work

A. Multi-person Pose Estimation

1) TOP-down

The matter of Multi-person pose estimation based top-bottom can first perform target detection, bounding box multiple people, and then fulfill single-person key points pose estimation on the marked target. As the detection targets increase, the computational cost will increase sharply. In Deep Cut [28], apply CNN to find all joint candidate points and then cluster to determine which person these combined points belong to and perform pose estimation. In [29][30] utilized the Resnet method in the detection stage and improved the recognition accuracy, Deeper Cut [29] reduced the candidate nodes based on Deep Cut [28] to increase the speed [30] clustered key points based on association and spatial information. Reference [31] adopt Faster RCNN for multi-person key points detection and image cropping, and then Resnet is used to predict and integrate dense heatmap and offset for each bounding box. RMPE [32] overcome the positioning error problem through a spatial mapping network and use the Stacked Hourglass method to recognize human posture. HR Net [23] maintains high-resolution expression as the backbone, performs high-to-low-resolution down-sampling in parallel at each stage, and finally performs multi-scale fusion and feature extraction. Based on the top-down approach, the recognition accuracy is high. However, the inference speed mainly depends on the number of the bounding box in the image.

2) BOTTOM-UP

Another method is bottom-up, which detects all joints in the detection target regardless of multiple people. It then identifies which person and which joint the key points belongs to according to the collective point relationship and connects them to get the human posture skeleton. [5] [30] [33] adopt CNN to predict the key points. In [5], realize fast joint point connection through graph theory. [30] Combine the top-down and bottom-up models, that is, the top-down method used to make a rough estimate of the human pose, and feed the bottom-up module for precise adjustment to obtain more accurate joint point positions. [33] improves the Open Pose [53] method, using dilated Mobile Net [50] for lightweight design and porting to edge devices. In [34], the association embedding algorithm introduces the joint points grouped by an end-to-end method. Since the bottom-up process does not require target detection, the recognition speed is much improved.

3) REALTIME PERSON POSE ESTIMATION

To predict the human pose in real-time, Reference [35] [36] adopt the adversarial learning framework and perform outdoor recognition. The 3D human pose structure [35] learned in the fully annotated data set is refined into a field image with only 2D pose annotations. And [36] proposed a weakly supervised transfer learning method that uses mixed 2D and 3D label data for recognition. In contrast, Reference [37] [38] [5] uses the CNN framework for credit, and Dense Pose [37] converts the 2D image into a 3D human body model for real-time recognition, that is, the pixels of the human surface in the 2D image project onto the 3D human body surface. Using the Kinect device [38], the 2D and 3D joint positions are regressed in real-time using a fully convolutional posture formula. Real-time detection of 2D multi-person poses [5], significantly increasing the speed while maintaining detection accuracy.

In this paper, we use edge devices for real-time human pose detection. The data is in a real scene with multiple people. Simultaneously, considering the edge device's performance, we use a 2D multi-person to estimate the human body pose.

B. Model Designs

1) DEEP NEURAL NETWORKS ON EDGE DEVICES

More and more edge devices need matching deep neural networks, which require less computing power and improve recognition speed. Google Net [39] uses a modular structure design to reduce computational complexity. Inception [40][41][42] series aims to improve the expressive ability without increasing the calculation. InceptionV2[40] uses the convolutional solution method, replacing two 3x3 convolutional layers with a 5x5 convolutional layer. InceptionV3 [41] is further improved, decomposing 7x7 into two one-dimensional convolutions to increase the nonlinearity of the network. Inception V4[42] combines the Inception module with the residual connection, which dramatically improves the training speed. Res Net [43] [44] uses a bottleneck structure, a residual structure to enhance network performance, and has become a backbone feature extraction in the computer vision.

2) GROUP CONVOLUTIONS

Group convolution was first proposed by Alex Net [45]and improved based on the GPU allocation model. ResNet [46] ultimately demonstrates the effectiveness of group convolution. The deep separable convolution model is currently the state-of-the-art group convolution, which is exploited in a lightweight application framework. MobileNetV2 [21] improves the MobileNetV1[20] series, using Inverted Residuals and linear activation methods to improve accuracy. MobileNetV3[22] combines the advantages of V1 and V2. Meanwhile, to minimize edge devices' resource consumption, the SE module and h-switch activation function are added. Under the framework of MobileNetV3[22], Ghost Net [26] uses Ghost bottleneck, which uses fewer parameters to generate more feature maps, replacing the bottleneck structure.

3) MULTI-INFORMATION FUSION

The multi-resolution fusion method [3][47] decomposes the source object into resolutions of different scales and then feeds the multi-resolution aggregation into multiple networks. Reference [3][4][48] combine low features into high-level feature resolution through jumpers. The cascaded pyramid network [4] base on the idea of shortcut, which gradually combines the feature elements generated from high to low resolution. Reference [23] Based on the concept of deep fusion, the multi-scale resolution is repeatedly fused. Simultaneously, multi-information fusion approach [23] is to cross-fuse different groups of information to enhance data information sharing. Shuffle Net [18] proposes the Channel shuffle method to exchange information between groups. This information is the feature map after group convolution, which ensures the accuracy of the model. Cross Info Net [25] first generates two sets of feature sub-branches, sharing useful supplementary information, and then cross-cascades the updated feature branches. Our approach, inspired by multi-information task fusion, aims to use a lightweight network to reduce computer consumption and cross-fusion of shared information to improve accuracy.

III. Method

Our method adopts a broad bottom-up pipeline to predict the human body's key points based on real-time multi-person recognition. As shown in Fig 1, the first unit is the feature extraction module. Its method uses the most advanced lightweight network, which consumes less, has fewer parameters, and can produce more effective feature maps. The second part is the basic cross-information mixing module. This module first decomposes the feature map into two branches: key points heatmaps and pairwise relations (part affinity fields, pafs), merging the two units. The last part is the strengthened information module, fed from the baseline feature map to the network for cross fusion again, and finally, the key points extract.

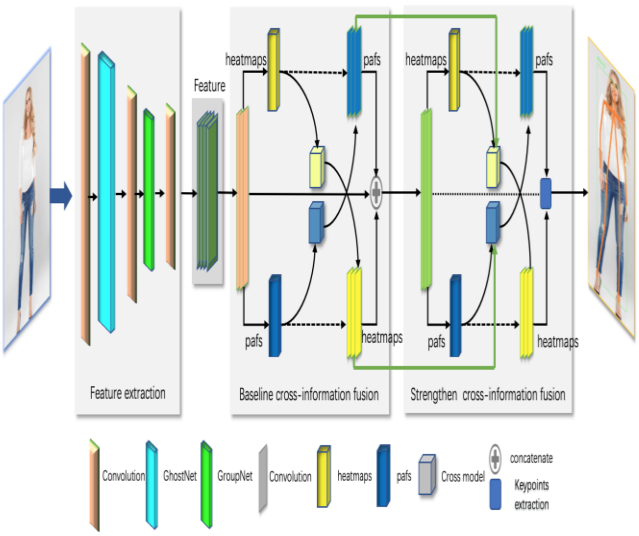

FIGURE 1. **Overall network architecture with lightweight cross-fusion sharing setup**

A. Lightweight feature extraction

As the initial feature extraction stage, we use Resnet18 and Resnet50[43]as the backbone network. However, the evaluation results are not satisfactory, and the average accuracy has not been improved based on the increase of parameters. To ensure the accuracy of using fewer parameters, we design a lightweight feature extraction network, as shown in Fig 2. Firstly, we apply Ghost bottleneck and Ghost bottleneck down-sampling [26]to generate an efficient feature map. Define the input image as I of the size $3´ W´ H$, the convolution kernel size $3*3$, and the output $512´ W'´ H'$. Then perform refined feature extraction through group convolution and group convolution down-sampling [33], the convolution kernel size $3*3$ and $1*1$, feature extraction image F size $128´ W"´ H"$, which feeds the follow-up refine network module.

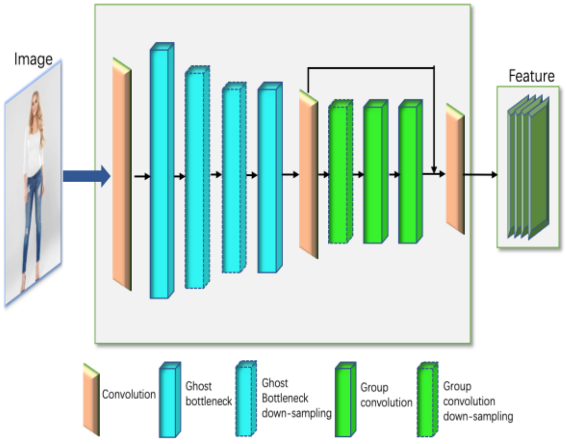
Image
Feature
Convolution
Ghost bottleneck
Ghost Bottleneck down-sampling
Group convolution
Group convolution down-sampling

FIGURE 2. **The initial lightweight feature extraction module**

B. Baseline cross-information fusion architecture

The standard pose estimation method designs the human body pose into binary tree branches: key points heatmaps, and pafs (part affinity fields), and then optimize them separately. Although pafs are composed of key points heatmap pairs, these two parts' eigenvalues are output independently. They can neither remove refine feature maps nor enhance data information. To better extract forceful shared knowledge and reduce fusions, we propose a novel key points acceptable extraction model based on cross-information fusion. The model compartmentalizes two parts: baseline cross-information fusion architecture and strengthens cross-information fusion architecture. First, we illustrate baseline cross-information information, as shown in Fig3.

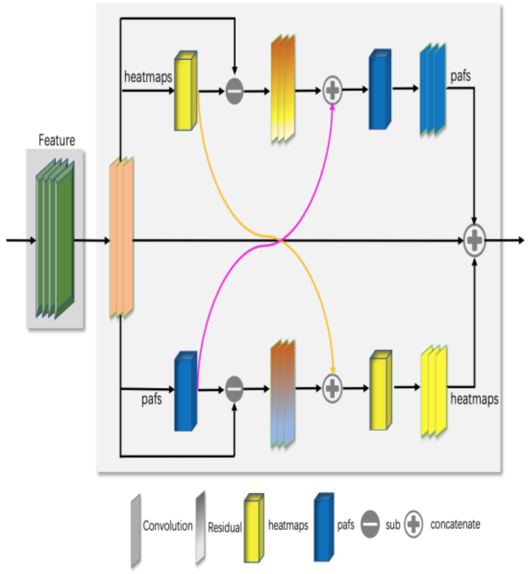

FIGURE 3. **Baseline cross-information fusion module**

The classic network structure first extracts the feature maps F captured from the initial feature and then feeds the feature maps to the convolutional layer to pull the local feature branch of the heatmaps H and pafs P. Finally, the updated feature maps F' obtained by cascading the heatmaps feature H and the pafs feature P, and the feature maps F poured into the next detailed convolution module, which repeats 2-5 times. Nevertheless, this method does not concern the shared information between the branches except for the cascading input of features at each stage. However, the relationship between the pafs feature branch and the feature branch's heatmaps feature may be inclusive or mutually exclusive. For example, the pafs feature branch includes some local characteristics of the heatmaps, which may enhance the pafs quality or noise. Still, it is useful for feature extraction of the heatmaps itself, and vice versa. We attempt to utilize a cross-information fusion network to enhance data and acquire more effective feature maps, take sufficient advantage of shared information, and eliminate potential noise.

The detailed network diagram is shown in Fig 3, the initial feature map F^0, which includes global heatmaps feature and global pafs feature information. The feature branch of the local heatmaps F_h^1 and the local pafs F_p^1 are respectively obtained by formula 1.

$$F_h^1 = F^0 * h\ ;\ F_p^1 = F^0 * p \tag{1}$$

Secondly, in formula 2, the global feature maps F^0 subtracts the local heatmaps feature F_h^1 and the local pafs feature F_p^1 respectively to obtain $\overline{F_h^1}$ and $\overline{F_p^1}$, thereby reducing noise interference and enhancing local information.

$$\overline{F_h^1} = F^0 - F_h^1\ ;\ \overline{F_p^1} = F^0 - F_p^1 \tag{2}$$

Thirdly, cross-cascade $F_h^1 \oplus \overline{F_p^1}$ and $F_p^1 \oplus \overline{F_h^1}$ update F_h^1 and F_p^1 respectively to generate enhanced heatmaps branch F_h^1 and pafs branch F_p^1, as shown in formula 3.

$$F_h^1 = (F_h^1 \oplus \overline{F_p^1}) * h\ ;\ F_p^1 = (F_p^1 \oplus \overline{F_h^1}) * p \tag{3}$$

Finally, in formula 4, the global feature maps F^0, enhanced F_h^1 and enhanced F_p^1 cascade to produce a rich global feature map F^1.

$$F^1 = F^0 \oplus F_h^1 \oplus F_p^1 \tag{4}$$

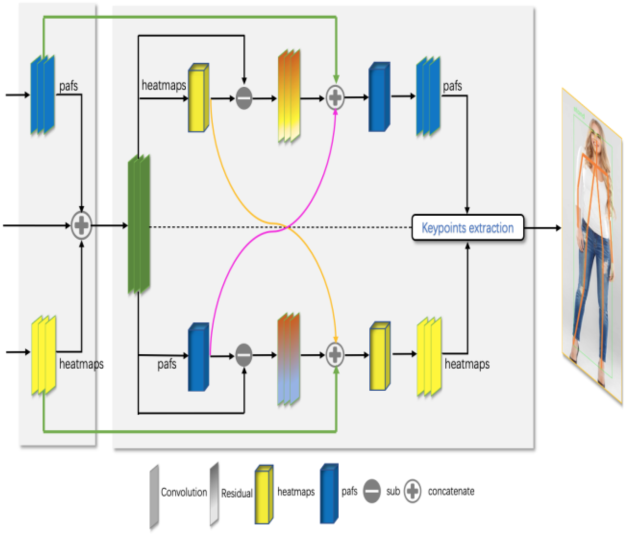

FIGURE 4. **Strengthen cross-information fusion module**

C. Strengthen cross-information fusion architecture

The feature maps F^1 produced in the baseline cross-information fusion stage contains both global features and fine local features related to heatmaps and pafs. As shown in Figure 4, this stage performs enhanced feature extraction on F^1. Equations 5 and 6 are similar to equations 1 and 2.

$$F_h^2 = F^1 * h\ ;\ F_p^2 = F^1 * p \tag{5}$$

$$\overline{F_h^2} = F^1 - F_h^2\ ;\ \overline{F_p^2} = F^1 - F_p^2 \tag{6}$$

$$F_h^2 = (F_h^2 \oplus \overline{F_p^2} \oplus F_h^1) * h\ ;\ F_p^2 = (F_p^2 \oplus \overline{F_h^2} \oplus F_p^1) * p \tag{7}$$

It can be seen from the figure that the local heatmaps feature F_h^2, the residual heatmaps feature $\overline{F_p^2}$ and the global heatmaps feature branch F_h^1 from the previous stage cascaded together to extract more sophisticated local features, similar to pafs part. Formula 7 adds international units F_h^1 and F_p^1 are based on cross-cascade, enhances the data's practical information, and helps extract more high-precision joint features.

Based on the above formula, the cross-information fusion structure describes in Algorithm 1. Among them, the number of feature extraction modules are controlled by k. The input value of each stage is the feature value F^k, and the output feature F^{k+1} update after crossover and cascade operations. Finally, the heatmaps feature F_h^{k+1} and pafs feature F_p^{k+1} from F^{k+1} are extracted for key points recognition and pose estimation. Let $k=1$, the key points of the human body are obtained by F_h^1 and F_p^1 feature maps.

```
Algorithm cross-information fusion
Input:
    Feature $F$; Stage $k$
    heats maps feature extraction: $h$
    pafs feature extraction: $p$
    convolution operator: $*$
    concatenate: $\oplus$
Initialize:
    $F^0 = F$; episode
    $F_h^0 = 0\ ;\ F_p^0 = 0$
$k \leftarrow 0$
while episode is not terminated do
    $F_h^{k+1} = F^k * h\ ;\ F_p^{k+1} = F^k * p$
    $\overline{F_h^{k+1}} = F^k - F_h^{k+1}\ ;\ \overline{F_p^{k+1}} = F^k - F_p^{k+1}$
    $F_h^{k+1} = (F_h^{k+1} \oplus \overline{F_p^{k+1}} \oplus F_h^k) * h\ ;\ F_p^{k+1} = (F_p^{k+1} \oplus \overline{F_h^{k+1}} \oplus F_p^k) * p$
    $F^{k+1} = F^k \oplus F_h^{k+1} \oplus F_p^{k+1}$
    $k \leftarrow k+1$
end while
key points extraction according to $F_h^{k+1}$, $F_p^{k+1}$
Output: key points
```

IV. Experimental Results on CoCo

A. COCO Key points Detection

1) DATASET

We also evaluate our method on MS COCO2017 dataset [49], which contains more than 200,000 pictures and 250,000 individual instances labeled with 17 key points. Among them, 2017 Train images [118K/18G] for training, 2017 Val image [5K/1G] for test verification, the corresponding annotation information of the picture 2017 Train/Val annotations [241MB].

2) EVALUATION METRIC

We evaluate the target detection results by the COCO dataset standard evaluation metric Object Key points Similarity (OKS). The object prediction key points have the same format as ground truth: $[x_1, y_1, v_1, \ldots, x_k, y_k, v_k]$, where x_k, y_k are the coordinates of the Key points, and v_k is the visible sign, v is 0,1,2, which means unlabeled, occluded, and visual, respectively. OKS is defined as for formula 8.

$$OKS = \Sigma_i[\exp(-d_i^2/2s^2k_i^2)\delta(v_i>0)] \Big/ \Sigma_i \delta(v_i>0) \tag{8}$$

Here d_i is the Euclidean distance between the detected key points and the corresponding ground truth, sk_i is the standard deviation. We also reflect standard average precision and recall scores: AP(IoU=0.50:0.95); AP50and AP75 (IoU=0.50, 0.75); APM for medium objects and APL for large objects (IoU=0.50:0.95), and AR at IoU=0.50:0.95.

3) TRAINING DETAILS

We preprocess the image and reset the image size to 368*368. Data enhancement includes random rotation ([-40, 40]) and random cropping ([0.5, 1.1]). Using Ghost Net[26] as the Backbone, using ImageNet pre-training weights for initialization, at the same time, choosing Adam algorithm to train the model, the initial learning rate is 4e-5, and weight decay is 5e-4.

Our network execution used PyTorch, two devices RTX 2080Ti GPU and the training time is 1 day. NVIDIA TX2 1080 GPU training time is 3 days while deleting extra programs to provide enough memory.

4) **RESULTS ON THE VALIDATION SET**

We reflect the results of our modus and other state-of-the-art methods in Table 1. Our approach of combining Ghost Net [26] and Cross-fusion achieved both state-of-art AP and inference speed with the respective GFLOPs counts. Table 1 shows the comparison of GFLOPs to AP performance on the validation set. (i) Compared to Ghost Net [26]. Both Retina Net [51] and Faster R-CNN [52] choose Ghost Net [26] as the skeleton, and our method also used the Ghost Net [26] skeleton framework. After cross information processing, our AP is much higher than the AP of Retina Net [51] and Faster R-CNN [52], while the AP of Retina Net [51] is similar to that of Faster R-CNN [52]. (ii)Compared to lightweight Open Pose [33]. Lightweight Open Pose [33] uses light networks, such as MobileNetV1[20], Dilated MobileNetV1[50], and Dilated MobileNetV2[21][50]. The highest AP value is 43.2 from the Dilated MobileNetV1[50] method, and our process reaches 44.4, which exceeds the optimal strategy in the lightweight Open Pose [33]. (iii) Compared to Open Pose

[53]. Both the Open Pose [53] and the lightweight Open Pose [33] methods use a bottom-up approach, and the light Open Pose [33] method uses two stages of Refinement extraction. The methodology also uses a bottom-up, and two-layer progressive cross-information sharing refine extraction method. Although, our AP value with the lightweight crossover method is lower than Open Pose's two-stage AP value of 46.2, the GFLOPs of Open Pose [53] are indeed as high as 80.3, while the GFLOPs of our approach is only 18.02.

TABLE I
ACCURACY VERSUS COMPLEXITY OF PROPOSED NETWORK ON COCO VALIDATION SET

Method	Backbone	AP	GFLOPs
Open Pose Refinement2	VGG-16	46.2	80.30
Lightweight Open Pose	Mobile Net V1	37.9	23.30
	Dilated Mobile Net V1	43.2	31.30
	Dilated Mobile Net V2	39.6	27.20
Retina Net	Ghost Net	26.6	-
Faster R-CNN		26.9	-
Lightweight Cross		44.4	18.02

B. Self-comparisons

Since Mobile Net [20] proposed, more and more lightweight network framework with similar or improved series was designed. Res Net [43], primarily for essential feature extraction in the visual field, is currently the most widely applied CNN feature extraction network. Its residual system can maintain a substantial increase in accuracy with increasing depth. We first evaluated networks from the ResNet [43] family to replace the Mobile Net [20]and started from ResNet-18.

We reported the results of our method in Table 2. Above all, we have compared ResNet-18 and ResNet-50 as skeletons. ResNet-18 has higher AP and AR than ResNet-50. Secondly, using our method for delicate feature extraction, the AP and AR of ResNet-18 and ResNet-50 are improved, and the AP value of ResNet-18 exceeded the lightweight Open Pose [33] way with Dilated MobileNetV1 [20] [50]. Finally, we utilized the Ghost Net [26] skeleton network combined with the cross-fusion method to obtain the highest AP value of 44.4.

TABLE II
SELF-COMPARISON RESULTS ON BACKBONE SELECTION

Method	Backbone	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}	AR
Lightweight Open Pose	ResNet-18	38.4	61.6	39.8	34.5	44.8	41.3
	ResNet-50	35.2	59.7	34.7	32.4	41.0	38.8
Ours	ResNet-18	43.5	64.0	46.0	39.2	50.3	46.6
	ResNet-50	39.9	62.6	40.2	37.1	45.4	43.8
	Ghost Net	44.4	64.5	46.7	41.1	50.0	47.7

C. Loss function comparison

We conducted COCO data training on the network in a supervised manner. The mean square residual value (MSE) is the minimum value of the estimated key points and the ground truth as the training loss function.

1) HEATSMAP LOSS

We use the heatmaps as a restriction to guide the network to obtain better global feature extraction, and the heatmaps loss function detection at each stage show in Equation 8.

$$L_H^k = \Sigma_{j=1}^{J} \| H_j^k - \hat{H}_j \|_2^2 \tag{8}$$

Where k is the value of the extraction stage, J is 19, the body's key points, and the background H_j^k and $\hat{H}_j$ are the estimated heatmaps and the ground truth, respectively. The heatmaps resolution is 46*46 px, the value is the Gaussian distribution heat map generated by the joint sample points, and the offset is 4 px. The overall heatmaps loss function is 9:

$$L_H = \text{å}_{k=0}^{K} a_u L_{H^u}^k + \text{å}_{k=0}^{K} a_v L_{H^v}^k \tag{9}$$

We set $K = 1$, $L_{H^u}^k$ is the heatmaps loss function before crossover, and the corresponding balance factor is a_u value of 0.01, and $L_{H^v}^k$ is the heatmaps loss function after crossover, and the balance factor a_v is 1.

2) PAFS LOSS

Like the heatmaps loss function, the pafs loss function (10) and (11) are defined as follows.

$$L_P^k = \Sigma_{i=1}^{I} \| H_i^k - \hat{H}_i \|_2^2 \tag{10}$$

$$L_P = \text{å}_{k=0}^{K} b_u L_{P^u}^k + \text{å}_{k=0}^{K} b_v L_{P^v}^k \tag{11}$$

Where pafs is the key points pairing of the human body, so the value I is 38, the stage value K is 1, and the balance factors b_u and b_v are also 0.01 and 1, respectively.

3) LOSS FUNCTION AND COMPARISON

Based on the above definition, (12) is the overall loss function. a is the balance factor value of the loss function of the entire heat map of 0.6? Similar to a , the amount of b is 0.4.

$$L = a L_H + b L_P \tag{12}$$

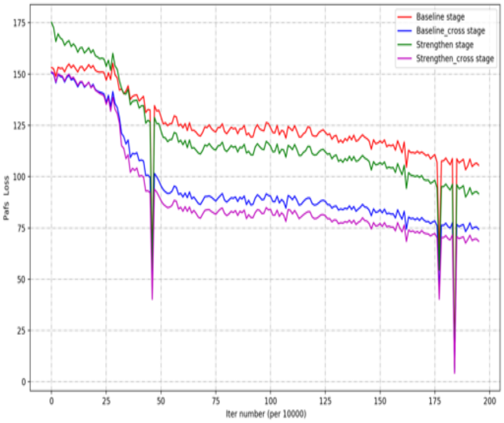

FIGURE 5. **Pafs loss function comparison**

Fig 5 compared the pafs loss function value, which divided into a baseline phase and an enhancement phase, partition into two parts before and after the crossover. Initially, Image Net pre-training weights used for initialization. The baseline stage loss value is lower than the strength stage loss value, and the gap between them gradually decreased. In contrast, the baseline-cross stage loss, which almost coincides with the strengthen-cross stage, starts to approach the baseline stage's and then slowly expands. Subsequently, due to the increase in the number of iterations, all loss values dropped rapidly, and the loss values in the strength stage were progressively lower than the baseline stage values. Eventually, the fluctuation of the loss value stabilized. The strengthen-cross loss is much lower than the uncrossed loss value, and the enhancement phase loss is lower than the baseline loss value.

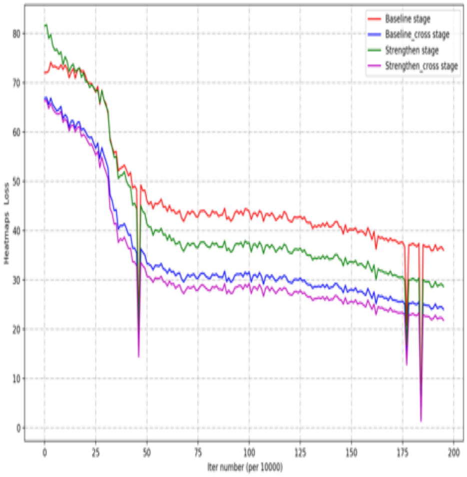

FIGURE 6. **Heatmaps loss function comparison**

Fig 6 shows a comparison of heatmaps losses, similar to pafs loss values. During the rapid decline of the loss value, the baseline stage's loss values, and the strengthening stage coincide. The strengthen-cross loss value is lower than the baseline-cross stage loss value throughout the training iteration process.

V. Conclusion

In this paper, we proposed a lightweight cross-fusion neural architecture for human pose estimation, which can generate accurate key-point heatmaps and deploy them on edge devices. Our method showed that the network can run on edge devices, and Ghost Net is a powerful and efficient backbone for human pose estimation. Furthermore, lightweight pose estimation network, the backbone Ghost Net matching cross-information and fusion framework have achieved excellent results using the MS COCO 2017 data set. In future work, we plan to study lightweight top-down methods to improve recognition accuracy and generalization capabilities.

REFERENCES

[1] E. Shelhamer, J. Long, T. Darrell, "Fully Convolutional Networks for Semantic Segmentation," IEEE Transactions on Pattern Analysis and Machine Intelligence., vol. 39, no. 4, pp. 640–651, April. 2017.

[2] S. E. Wei, V. Ramakrishna, T. Kanade, et al. Convolutional Pose Machines. In CVPR, 2016, pp. 4724–4732.

[3] A. Newell, K. Yang, J. Deng. Stacked hourglass networks for human pose estimation. In ECCV, 2016, pp. 483–499.

[4] W. Yang, S. Li, W. Ouyang, et al. Learning feature pyramids for human pose estimation. In ICCV, 2017, pp. 1290–1299.

[5] Z. Cao, T. Simon, S. E. Wei, et al. Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. In CVPR, 2017, pp. 1302–1310.

[6] E. Insafutdinov, L. Pishchulin, B. Andres, et al. DeeperCut: A Deeper, Stronger, and Faster Multi-Person Pose Estimation Model. In ECCV, 2016, pp.34–50.

[7] T. Sekii. Pose proposal networks. In ECCV, 2018, pp.8–14.

[8] X. Nie, J. Feng, J. Xing, et al. Pose Partition Networks for Multi-Person Pose Estimation. In ECCV, 2018, pp.705–720.

[9] G. Papandreou, T. Zhu, L. C. Chen, et al. PersonLab: Person Pose Estimation and Instance Segmentation with a Bottom-Up, Part-Based, Geometric Embedding Model. In ECCV, 2018, pp.282–299.

[10] T. Pfister, J. Charles, A. Zisserman. Flowing ConvNets for Human Pose Estimation in Videos. In ICCV, 2015, pp.1913–1921.

[11] B. Xiao, H. Wu, Y. Wei. Simple Baselines for Human Pose Estimation and Tracking. In ECCV, 2018, pp.472–487.

[12] S. Han, Mao H, W. J. Dally. Deep Compression: Compressing Deep Neural Networks with Pruning, Trained Quantization and Huffman Coding. arXiv: 1510.00149v5, 2016.

[13] J. H. Luo, J. Wu, W. Lin. ThiNet: A Filter Level Pruning Method for Deep Neural Network Compression.In ICCV, 2017, pp.5058–5066.

[14] M. Rastegari, V. Ordonez, J. Redmon, et al. XNOR-Net: ImageNet Classification Using Binary Convolutional Neural Networks. In ECCV. Springer, 2016, pp.525–542.

[15] B. Jacob, S. Kligys, B. Chen, et al. Quantization and Training of Neural Networks for Efficient Integer-Arithmetic-Only Inference. In CVPR, 2018, pp. 2704–2713.

[16] G. Hinton, O. Vinyals, J. Dean. Distilling the Knowledge in a Neural Network. Machine Learning., to be published. DOI: 10.4140/TCP.n.2015.249.

[17] S. You, C. Xu, C. Xu, et al. Learning from Multiple Teacher Networks. Acm Sigkdd International Conference on Knowledge Discovery & Data Mining., to be published. DOI: 10.1145/3097983.3098135.

[18] X. Zhang, X. Zhou, M. Lin, et al. ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile Devices. In CVPR, 2018, pp. 6848–6856.

[19] N. Ma, X. Zhang, H. T. Zheng, et al. ShuffleNet V2: Practical Guidelines for Efficient CNN Architecture Design. In ECCV, 2018, pp. 122–138.

[20] A. G. Howard, M. Zhu, B. Chen, et al. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. arXiv:1704.04861, 2017.

[21] M. Sandler, A. Howard, M. Zhu, et al. Mobilenetv2: Inverted residuals and linear bottlenecks. In CVPR, 2018, pp. 4510–4520.

[22] A. Howard, M. Sandler, G. Chu, et al. Searching for MobileNetV3. Searching for mobilenetv3. In ICCV., to be published. DOI: 10.1109/ICCV.2019.00140.

[23] K. Sun, B. Xiao, D. Liu, et al. Deep High-Resolution Representation Learning for Human Pose Estimation. In CVPR, 2019, pp. 5693–5703.

[24] S. Ruder. An overview of multi-task learning in deep neural networks. arXiv:1706.05098, 2017.

[25] K. Du, X. Lin, Y. Sun, et al. CrossInfoNet: Multi-Task Information Sharin g Based Hand Pose Estimation. In CVPR, 2019, pp. 9986-9905.

[26] K. Han, Y. Wang, Q. Tian, et al. GhostNet: More Features From Cheap Operations. In CVPR, 2020, pp. 1580-1589.

[27] T. Y. Lin, M. Maire, S. Belongie, et al. Microsoft COCO: Common Objects in Context. In ECCV, 2014, pp. 740–755.

[28] L. Pishchulin, E. Insafutdinov, S. Tang, et al. DeepCut: Joint Subset Partition and Labeling for Multi Person Pose Estimation. In CVPR., to be published. DOI: 10.1109/CVPR.2016.533.

[29] E. Insafutdinov, L. Pishchulin, B. Andres, et al. DeeperCut: A Deeper, Stronger, and Faster Multi-Person Pose Estimation Mode. In ECCV., to be published. DOI: 10.1007/978-3-319-46466-4_3.

[30] E. Insafutdinov, M. Andriluka, L. Pishchulin, et al. ArtTrack: Articulated Multi-person Tracking in the Wild. In CVPR, 2017, pp. 6457–6465.

[31] G. Papandreou, T. Zhu, N. Kanazawa, et al. Towards Accurate Multi-Person Pose Estimation in the Wild. In CVPR, 2017, pp. 4903–4911.

[32] H. S. Fang, S. Xie, Y.W. Tai, et al. RMPE: Regional Multi-person Pose Estimation. In ICCV, 2018, pp. 4321–4330.

[33] D. Osokin. Real-time 2D Multi-Person Pose Estimation on CPU: Lightweight OpenPose. In ICPRAM., to be published. DOI: 10.5220/0007555407440748.

[34] A. Newell, Z. Huang, J. Deng. Associative embedding: End-to-end learning for joint detection and grouping. In Advances in Neural Information Processing Systems, 2017, pp. 2274–2284.

[35] W. Yang, W. Ouyang, X. Wang, et al. 3D Human Pose Estimation in the Wild by Adversarial Learning. In CVPR., to be published. DOI: 10.1109/CVPR.2018.00551.

[36] X. Zhou, Q. Huang, X. Sun, et al. Towards 3D Human Pose Estimation in the Wild: a Weakly-supervised Approach. In ICCV, 2017, pp. 398–407.

[37] R. A. Güler, N. Neverova, I. Kokkinos, et al. DensePose: Dense Human Pose Estimation In The Wild. In CVPR, 2018, pp. 7297–7306.

[38] D. Mehta, S. Sridhar, O. Sotnychenko, et al. VNect: Real-time 3D Human Pose Estimation with a Single RGB Camera, Acm Transactions on Graphics., to be published. DOI: 10.1145/3072959.3073596.

[39] C. Szegedy, W. Liu, Y. Jia, et al. Going deeper with convolutions. In CVPR, 2015, pp. 1–9.

[40] S. Ioffe, C. Szegedy. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. arXiv:1502.03167v3, 2015.

[41] C. Szegedy, V. Vanhoucke, S. Ioffe, et al. Rethinking the inception architecture for computer vision. In CVPR, 2016, pp. 2818–2826.

[42] C. Szegedy, S. Ioffe, V. Vanhoucke, et al. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. In AAAI, 2017, pp. 4278–4284.

[43] K. He, X. Zhang, S. Ren, et al. Deep Residual Learning for Image Recognition. In CVPR, 2016, pp. 770–778.

[44] K. He, X. Zhang, S. Ren, et al. Identity Mappings in Deep Residual Networks. In ECCV, 2016, pp. 630–645.

[45] A. Krizhevsky, I. Sutskever, G. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In Advances in neural information processing systems, 2012, pp. 1097-1105.

[46] S. Xie, R. Girshick, P. Dollár, et al. Aggregated Residual Transformations for Deep Neural Networks. In CVPR., to be published. DOI: 10.1109/CVPR.2017.634.

[47] Y. Chen, Z. Wang, Y. Peng, et al. Cascaded Pyramid Network for Multi-Person Pose Estimation. In CVPR., to be published. DOI: 10.1109/CVPR.2018.00742.

[48] L. Ke, M. C. Chang, H. Qi, et al. Multi-Scale Structure-Aware Network for Human Pose Estimation. In ECCV, 2018, pp. 731–746.

[49] W. Havard, L. Besacier, O. Rosec. SPEECH-COCO: 600k Visually Grounded Spoken Captions Aligned to MSCOCO Data Set. In GLU., to be published. DOI: 10.21437/GLU.2017-9.

[50] F. Yu, V. Koltun, T. Funkhouser. Dilated Residual Networks. In CVPR., to be published. DOI: 10.1109/CVPR.2017.75.

[51] T. Y. Lin, P. Goyal, R. Girshick, et al. Focal Loss for Dense Object Detection. In ICCV, 2017, pp. 2980–2988.

[52] S. Ren, K. He, R. Girshick, et al. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence., vol. 39, no. 6, pp. 1137–1149, June. 2017.

[53] OpenPose:https://github.com/CMU-Perceptual-Computing-Lab/openpose .

FIRST A. AUTHOR (F'82) received the B.S. degree in Computer Science and Technology from Nanjing Normal University, Jiangsu China, in 2005 and the M.S. degree in Computer Application Technology from Nanjing Normal University, Jiangsu China, in 2009. She is currently pursuing the Ph.D. degree in Computer Science and Technology at Hohai University, College of Computer and Information, Jiangsu China.

Since 2012, she was a Lecturer with the Department of Computer Science, Zijin College, Nanjing University of Science and Technology, Nanjing, Jiangsu, China. Her research interest includes developing the human pose estimation, IoT artificial intelligence using Edge computing, and point cloud registration.

SECOND B. AUTHOR (M'57) received the B.S. degree in Nanjing University, the M.S. degree in Southeast University, Jiangsu China, and the Ph.D. degree in mechanical engineering from the Hong Kong Polytechnic University, Hong Kong, China.

He is currently a professor and doctoral supervisor in the School of Computer and Information, Hohai University. He has published many academic papers in international and domestic authoritative academic journals (such as Neural Computation, IEEE Transactions on Neural Networks, Science in China-Series F, etc.). In recent years, his research interests mainly include machine learning, artificial neural networks, machine vision, pattern recognition, graphic grammar, and information visualization.

Dr. Author He is a member of the Machine Learning Technical Committee of the IEEE SMC Society, an associate editor of IEEE Transactions on Systems, Man, and Cybernetics-Part B. A reviewer of multiple international academic journals (such as EEE Transactions on Neural Networks, Information Science, Neurocomputing, etc.). He is a member of the program committee of several international academic conferences.

THIRD C. AUTHOR, JR. (M'83) received the B.S. degree in Computer Science and Technology from Nanjing Normal University, Jiangsu China, in 2006 and the M.S. degree in Computer Application Technology from Nanjing Normal University, Jiangsu China, in 2009. He is currently pursuing the Ph.D. degree in Computer Science and Technology at Nanjing University, Department of Computer Science and Technology, Jiangsu China.

Since 2017, he has been an Assistant Professor with Nanjing Institute of Tourism and Hospitality. He is the author of five books, more than 40 articles, and more than 4 inventions. His research interests include Intelligent optimization, evolutionary computing and computer vision.