Hindawi
Wireless Communications and Mobile Computing
Volume 2021, Article ID 4948067, 11 pages
https://doi.org/10.1155/2021/4948067

Research Article

A Novel Approach of Intelligent Computing for Multiperson Pose Estimation with Deep High Spatial Resolution and Multiscale Features

Haiquan Wang,[1] Xiangyang Wang,[2] Yijie Shi,[2] Yanping Li,[2] Chunhua Qian,[3] and Rui Wang[2]

[1]*Department of General Surgery, Shanghai General Hospital of Shanghai Jiaotong University, Shanghai 200080, China*
[2]*Key Laboratory of Specialty Fiber Optics and Optical Access Networks, Joint International Research Laboratory of Specialty Fiber Optics and Advanced Communication, School of Communication and Information Engineering, Shanghai University, 200444 Shanghai, China*
[3]*Department of Endocrinology and Metabolism, Shanghai Tenth People's Hospital, Tongji University, School of Medicine, Shanghai 200072, China*

Correspondence should be addressed to Yanping Li; yanpingli@shu.edu.cn, Chunhua Qian; chqian2003@126.com, and Rui Wang; rwang@shu.edu.cn

Received 5 April 2021; Revised 30 May 2021; Accepted 22 June 2021; Published 7 July 2021

Academic Editor: Xiangjie Kong

Currently, human pose estimation (HPE) methods mainly rely on the design framework of Convolutional Neural Networks (CNNs). These CNNs typically consist of high-to-low-resolution subnetworks (encoder) to learn semantic information and low-to-high subnetworks (decoder) to raise the resolution for keypoint localization. Because too low-resolution feature maps in encoder will inevitably lose some spatial information, which cannot be recovered in the upsampling stages, keeping high spatial resolution features is critical for human pose estimation. On the other hand, due to scale variation of human body parts, multiscale features are also very important for human pose estimation. In this paper, a novel backbone network is proposed specifically for HPE, named High Spatial Resolution and Multiscale Networks (HSR-MSNet), which maintain high spatial resolution features in deeper layers of the encoder and meanwhile construct multiscale features within one single residual block via subgroup splitting and fusion of feature maps. Experiments show that our approach outperforms other state-of-the-art methods with more accurate keypoint locations on COCO dataset.

1. Introduction

Human pose estimation (HPE) is one of the most fundamental tasks in computer vision—which is aimed at predicting the locations of body joints from input images. Recently, the human pose estimation methods based on Convolutional Neural Networks (CNNs) have achieved a great breakthrough [1–6], since CNNs have the powerful ability to learn rich convolutional feature representations [7]. For example, for single-person pose estimation, the state-of-the-art models have improved the performance from less than 50% PCKh@0.5 to more than 90% PCKh@0.5 [8–12] on the MPII benchmark [13]. However, multiperson pose estimation still faces two main challenges:

(1) There may be occlusion between different people, which will cause ambiguities of joints

(2) Some invisible joints are hard to be predicted

In order to solve these challenges, existing methods, such as CPN [14] and SimpleBaseline [15], employ ResNet [16] as the backbone to obtain feature maps with large downsampling. However, too large downsampling will cause image spatial information loss [17], leading to difficulties for joint

context recovery. On the other hand, due to camera view change or foreshortening, scales of different body parts may still be inconsistent, even if training images are warped to the same scale [9]. Therefore, scale variation of human body parts is also one of the main challenges. Previous works [9, 10, 18] have shown that multiscale or pyramid features are beneficial for solving the problems caused by scale changes.

In this paper, a novel backbone network is proposed specifically for HPE, named High Spatial Resolution and Multiscale Networks (HSR-MSNet). The network could maintain high spatial resolution features in deeper layers while keeping large receptive fields and construct multiscale features within one single residual block by channel split and fusion. Experiments on COCO keypoint detection dataset demonstrate the effectiveness of HSR-MSNet. At the same time, the network architecture of HSR-MSNet is very lightweight, which means that it will be possible to implement functions similar to MobileNet [19] on Internet-of-Things (IoT) devices.

2. Related Works

2.1. Single-Person Pose Estimation. DeepPose [5] is the first human pose estimation method based on deep learning, which treats the body joint as a CNN-based regression problem, and its backbone consists of a softmax classifier, five convolution layers, and two fully connected layers. Subsequent methods mostly apply heatmaps that could characterize the probability of each keypoint at different locations for pose estimation [20]. The accurate location of a keypoint is further estimated by selecting the maximum value in the aggregation heatmaps. Convolutional Pose Machine (CPM) [21] is a multistage architecture where the belief maps and image features generated in the previous stage are served as input for the next stage [22]. For CPM, large receptive fields are used to learn long-range spatial relationships and the gradient vanishing problem is eliminated by using intermediate supervision. The features of stacked hourglass network [23] (Hourglass) are processed across all scales and consolidated to best capture the various spatial relationships associated with the body [23]. The above two models (CPM and Hourglass) achieve state-of-the-art performance, which all adopt intermediate supervision to generate detailed heatmaps for the joint locations.

2.2. Multiperson Pose Estimation. Multiperson pose estimation is a more challenging problem than single-person pose estimation for many computer vision applications. Due to occlusion and complex background, it is difficult to obtain accurate location results for multiperson pose estimation. A common approach is bottom-up, which detects human joints throughout the image region and then makes the groups of joint candidates for each person. The main problem of the bottom-up approach is to model the joint-to-individual associations [24]. Cao et al. [25] proposed a novel model to detect the 2D pose of several people in an image, which uses a nonparametric representation (Part Affinity Fields (PAFs)) to associate body parts with individuals in the given image. The architecture is aimed at learning part positions and their association jointly by two branches of the same sequential prediction process.

Another pipeline to multiperson pose estimation is top-down [14, 15, 26, 27], which first detects each person in the image and then conducts single-person pose estimation for each single person. This top-down approach is not suitable when crowds are in close proximity, because it will result in significant overlap between bounding box regions of people. Fang et al. [27] proposed a novel regional multiperson pose estimation (RMPE) framework, which applies SSD [28] or Faster RCNN [29] to locate persons in an image and uses Hourglass [23] to predict pose of each people. Wei et al. [21] proposed the Cascaded Pyramid Network (CPN) for multiperson pose estimation, which uses Mask RCNN [25] to detect persons and then designs CPN to predict each person's pose.

2.2.1. High Spatial Resolution Features. Some state-of-the-art human pose estimation architecture, such as Hourglass [23], CPN [14], and SimpleBaseline [15], are shown in Figure 1. These CNN-based methods are typically encoder-decoder architecture, which consists of high-to-low resolution subnetworks (encoder) to learn semantic information and low-to-high subnetworks (decoder) to raise the resolution for the keypoint locations. Hourglass stacks multiple encoder-decoder subnetworks together to get progressively refined heatmaps. The "RefineNet" of CPN plays the role to explore the context information of "hard" keypoints to further improve the performance. SimpleBaseline simply takes ResNet as its backbone and adds additional deconvolution layers to raise the resolution of feature maps to predict keypoint heatmaps.

Because too low-resolution feature maps in encoder will inevitably lose some spatial information, which cannot be recovered in the upsampling stages, keeping high spatial resolution features is critical to improve the performance of human pose estimation.

In [30], DetNet maintains high spatial resolution in deeper layers to deal with the problem that large downsampling factors may compromise the location capability. Sun et al. [18] proposed the High-Resolution Net (HRNet) that consists of parallel high-to-low resolution subnetworks with multiscale feature fusion, which learns reliable high-resolution features by maintaining high-resolution representations through the whole networks. The information is exchanged repeatedly across multiresolution subnetworks, each of which receives information from other parallel ones.

2.2.2. Multiscale Features. Multiscale features are very important for pose estimation due to scale variation of human body parts. Most existing methods [14, 29] represent the multiscale features in a layer-wise manner fusing different level (scale) features together.

PyraNet proposed by Pishchulin et al. [8] and Res2Net proposed by Gao et al. [31] both construct multiscale features within one single residual block. By means of extending residual block to multiscale pyramids, PyraNet [8] designs Pyramid Residual Modules (PRMs) to enhance the invariance in scales. Multiscale features are obtained by applying

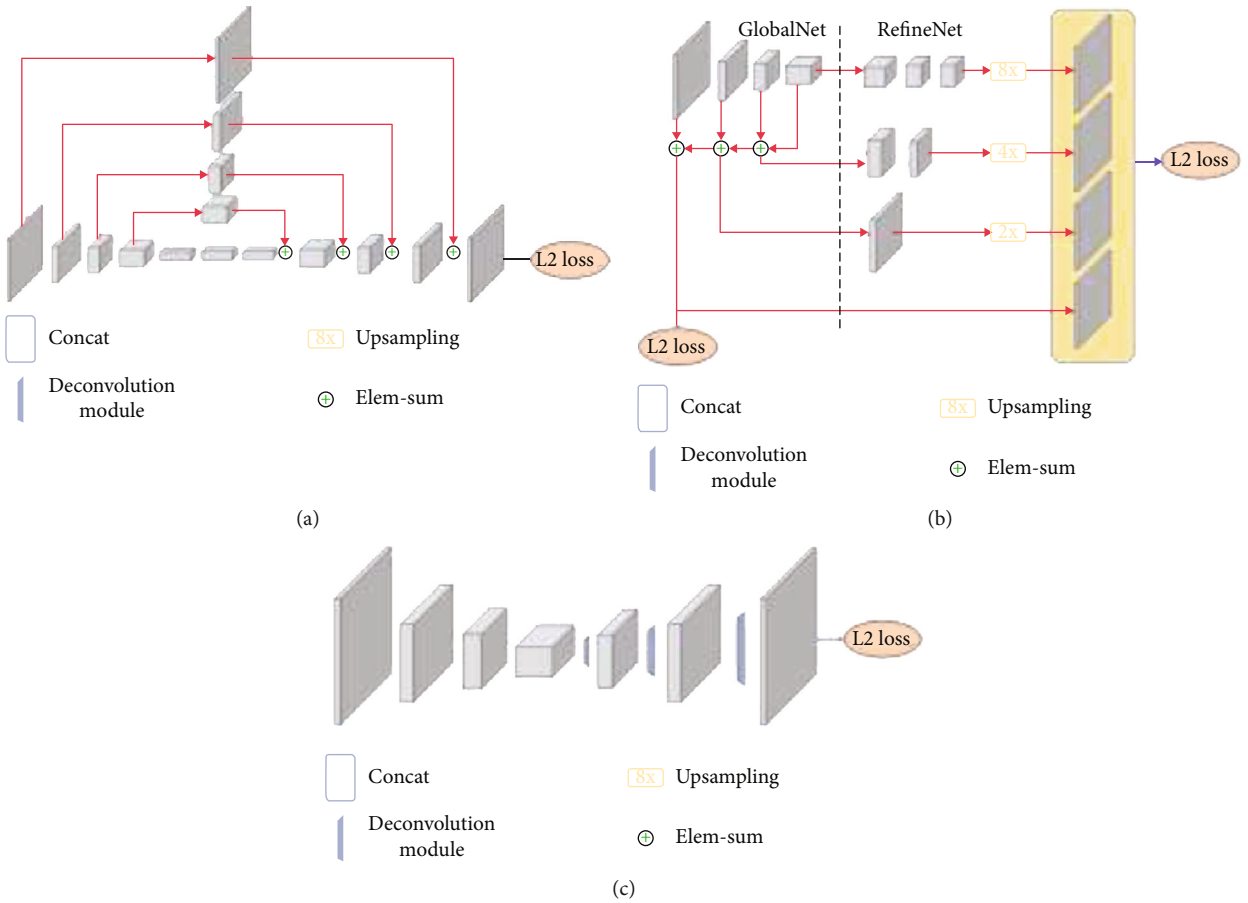

Figure 1: Some state-of-the-art pose estimation architecture: (a) Hourglass; (b) CPN; (c) SimpleBaseline.

different subsampling on input features in a multibranch residual block. Res2Net [31] represents multiscale features at a granular level and increases the range of receptive fields for each network layer. Specifically, Res2Net implements multiscale features via splitting channels of feature maps into subgroups and fusing these channel groups hierarchically. It has been proven that Res2Net can boost many backbone networks in some vision tasks, including object detection, semantic segmentation [32, 33], and salient object detection [29].

3. Our Approach

Similar to the previous works [14, 15, 25], we adopt the top-down pipeline for multiperson pose estimation, as illustrated in Figure 2. The whole framework consists of two parts: human detection and human pose estimation. First, a human detector is used to find all persons in the input image and generate a set of human bounding boxes. Then, a human pose estimation approach is applied to predict the keypoints for each single person by dealing with those human bounding boxes.

3.1. Human Detection. The state-of-the-art object detection method, YOLOv3 [34], is utilized for human detection. All eighty categories from the COCO dataset [35] are utilized to train YOLOv3, but only human bounding boxes are used for our model. In the network of YOLOv3, 53 convolution layers with some shortcut connections are used for image feature extraction and the size of feature map can be adjusted through the convolution stride. Drawing on the idea of feature pyramid networks, YOLOv3 uses multiple scales to detect objections with different sizes, the finer the grid cell, the finer the object can be detected. In addition, the softmax is replaced with logistic classifier; in this way, multilabel object detection can be supported when detecting objects. More detail about YOLOv3 could be found in [34].

3.2. Human Pose Estimation with High Spatial Resolution Features

3.2.1. Motivation. Backbone networks play an important role in human pose estimation because of their abilities to extract effective features from the input images, which is critical for classification and keypoint localization. ResNet [16], as a traditional backbone network, has been widely used [36] and achieved outstanding performance in many state-of-art networks for human pose estimation such as SimpleBaseline [15] and CPN [14]. Accordingly, ResNet has high efficiency for image feature extraction. However, there still exist the following shortcomings when ResNet is used as the backbone network for human pose estimation.

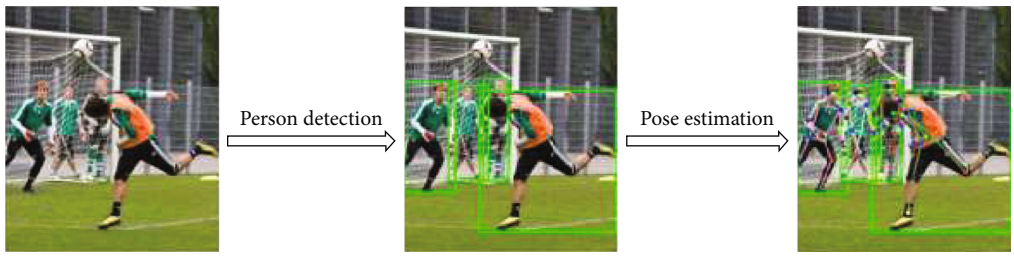

Figure 2: Top-down pipeline for multiperson pose estimation.

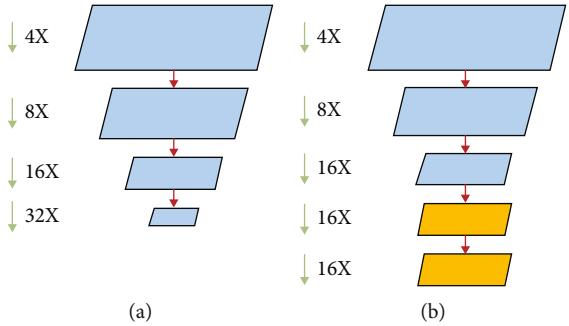

Figure 3: Comparisons of different backbones used in human pose estimation: (a) ResNet backbone; (b) HSRNet backbone.

(1) Poor accuracy of keypoint localization. There are five stages in ResNet, and in each stage, the feature maps are sampled down. Compared to the input image, the feature maps have strides of 32 with strong semantic information and have large valid receptive fields, which bring a great performance in classification task. However, the downsampling with 32 strides may lead to the loss of local information and further affect the accuracy of keypoint localization in human pose estimation

(2) Invisibility of small joints. Another drawback of large stride is the missing of small keypoints. For some occluded joints which contain less information, the spatial resolution of input image is greatly reduced when the large stride feature maps are extracted

(3) Ambiguities. In the case of occlusion between multiple persons, one human bounding box may contain the keypoints of other persons. There exists the loss of the context information while the input image is converted into feature maps with large strides. It is hard to distinguish which keypoints belong to the right person without the context information

To solve these problems, inspired by DetNet [30], we reserve the first four stages of ResNet and replace the fifth stage with two new stages, as shown in Figure 3(b), named as HSRNet (High Spatial Resolution Network). In these two stages, the feature maps are no longer sampled down and the valid receptive fields are expanded. Thus, we can not only ensure the classification of each keypoint but also reserve more semantic information of the feature maps, which will be helpful to improve the accuracy of keypoint localization.

3.2.2. Our Model for Keypoint Localization. A simple network structure is adopted in our model, as shown in Figure 4(b). Firstly, we use HSRNet as backbone network to generate feature maps with semantic information. Then, a few deconvolutional layers with batch normalization and ReLu activation are applied to generate heatmaps from the low-resolution feature maps. Finally, Mean Squared Error (MSE) is used as the loss between the predicted heatmaps and target heatmaps. The target heatmaps for keypoints are generated by applying a Gaussian centered at the ground-truth location of keypoints. Compared with SimpleBaseline [15], HSRNet is adopted to replace the original ResNet. Since the size of output feature maps of HSRNet and ResNet is different, the deconvolutional layer is reduced to keep the same size of feature map.

As shown in Figure 5(b), HSRNet is our backbone network with two new stages based on the existing ResNet [16]. As shown in Figure 6, in the two new stages, original bottlenecks A and B are slightly altered into two new bottlenecks C and D. Original3×3convolution layer is converted into a convolution layer with dilation of 2. And the bottleneck C does not have a downsampling, which ensures that the size of feature maps will not change during these two new stages. Similar to ResNet, the stack method is applied, and original bottlenecks A and B are replaced by bottlenecks C and D, as shown in Figure 5. Since the fifth stage in ResNet has a downsampling with a stride of 2, if only a new stage is used, it will reduce the valid receptive field, leading to a negative effect for human pose estimation. Therefore, we utilize two new stages to gain feature maps of higher spatial resolution, simultaneously without sacrificing valid receptive field.

3.3. Human Pose Estimation with Multiscale Features. Based on Res2Net [31], we design a new multiscale module

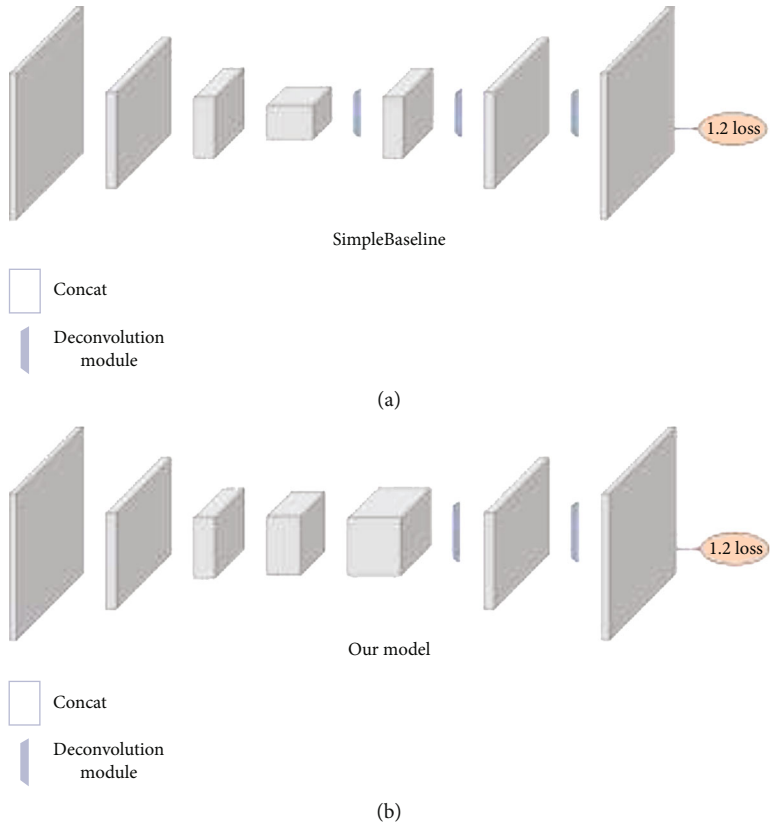

FIGURE 4: Illustration of network architecture for human pose estimation: (a) SimpleBaseline; (b) our model.

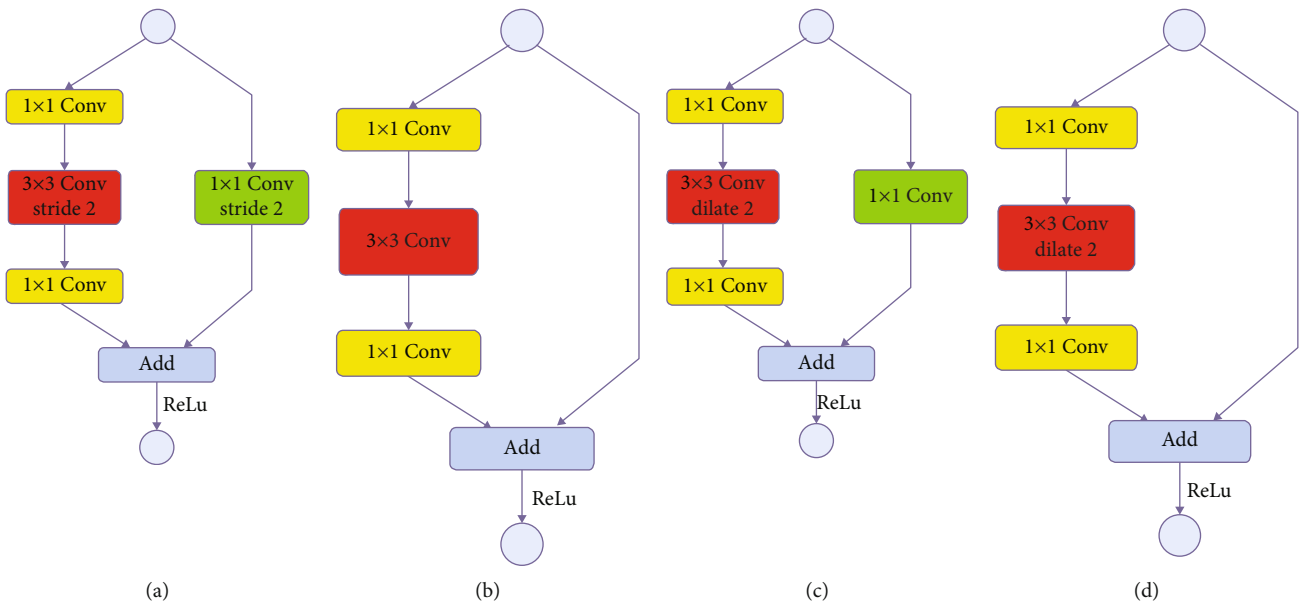

FIGURE 5: Detailed structures of different bottlenecks: (a) original bottleneck with 1×1 Conv; (b) original bottleneck; (c) dilated bottleneck with 1×1 Conv; (d) dilated bottleneck.

(MSNet) (as shown in Figure 7(b)) specifically and further integrate it into HSRNet to learn multiscale features. Our entire framework is named as HSR-MSNet.

The structure of Res2Net [31] is shown in Figure 7(a). Res2Net uses hierarchical groups of 3×3 convolution filters to extract multiscale features. Specifically, Res2Net

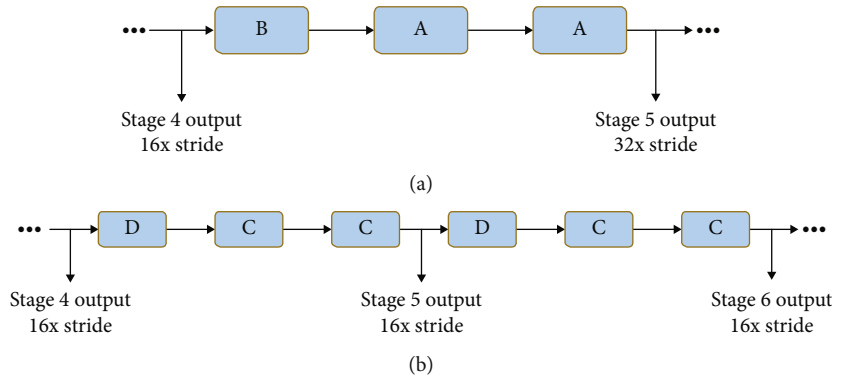

FIGURE 6: Detailed structures of ResNet and HSRNet. A, B, C, and D are bottlenecks illustrated in Figure 5. HSRNet follows the same design as ResNet before stage 4, while keeping spatial size after stage 6. (a) The fifth stage of ResNet; (b) two new stages of HSRNet.

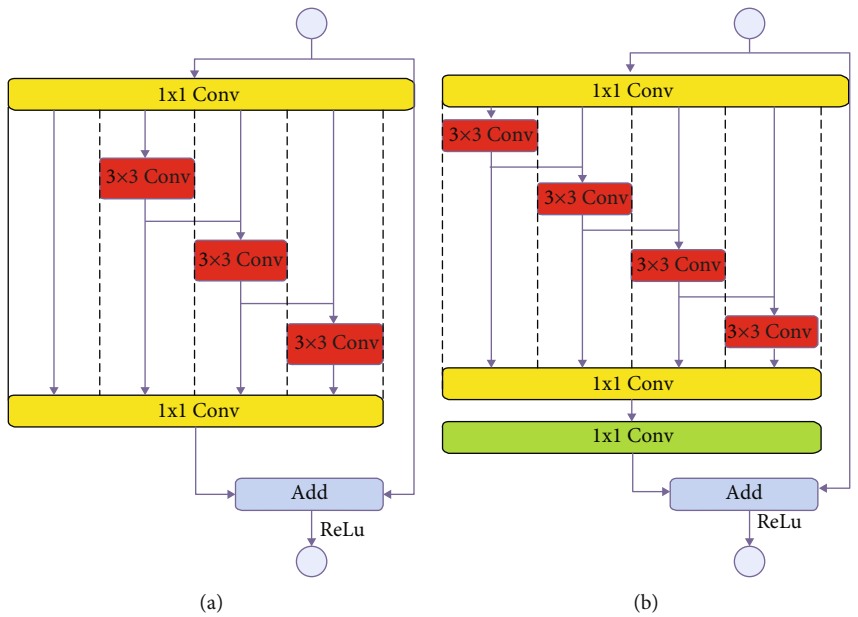

FIGURE 7: Comparison of two multiscale building blocks: (a) original Res2Net block; (b) our proposed MSNet.

implements multiscale features by splitting channels of feature maps into subgroups (3 groups as shown in Figure 7(a)) and fuses these channel groups hierarchically. Since we have retained the first four stages of ResNet, Res2Net module can be easily integrated into the first four stages of our network backbone. As shown in Figure 7(b), we add a 3×3 convolution layer compared with Res2Net, in order to increase the receptive fields. In addition, we use multiple smaller scale 3×3 convolution layers to replace the 3×3 convolution layer with stride 2. There are two advantages in this network structure. Firstly, multiple smaller 3×3 convolution layers can learn more context information of keypoints compared with one convolution layer with stride 2, especially for small-scale persons. Since the output feature maps of stage 4 are 16x strides with respect to input image, the convolution layer with stride 2 will be more difficult to extract semantic features for small-scale persons in detail. Secondly, the MSNet module can extract deep semantic features in multiscale style, while keeping large receptive fields. We denote the output feature maps of the 3×3 convolutional layer $Conv()$ as y_i, $i \in \{1, 2, \cdots, n\}$, where n is the total numbers of subgroups that the feature maps are split into evenly. Then, y_i could be expressed as

$$y_i = \begin{cases} Conv(x_i), & i = 1, \\ Conv(x_i + y_{i-1}), & 1 < i \leq n, \end{cases} \tag{1}$$

where x_i denotes the results of input feature maps x after 1 $\times$ 1 convolution and split evenly.

Among the multiscale features, not all the features are equally valid for human pose estimation. In order to balance the relationship among channel features, we add a SE (Squeeze-and-Excitation) block [37] before the residual connections (Figure 7(b)), which can learn the importance of each feature channel and promote important features while

Table 1: Comparisons on the COCO val2017 dataset.

Method	Backbone	Input size	Deconv. layers	Deconv. kernel size	AP
a	ResNet-50	256×192	3	4	70.4
b	ResNet-101	256×192	3	4	71.4
c	ResNet-152	256×192	3	4	72.0
d	ResNet-50	256×192	3	2	70.1
e	ResNet-50	$256x192$	3	3	70.3
f	ResNet-50	384×288	3	4	72.7
g	HSRNet-59	256×192	2	4	71.4
h	HSRNet-110	256×192	2	4	71.8
i	HSRNet-161	256×192	2	4	72.6
j	HSRNet-59	256×192	2	2	70.9
k	HSRNet-59	256×192	2	3	71.2
l	HSRNet-59	384×288	2	4	73.2

suppressing the less useful features for the current vision task. As a result, the final formula of our MSNet module could be written as

$$f = \mathrm{Re\ Lu}(x + K(y)), \tag{2}$$

where y represents the concatenation of $y_i (i \in \{1, 2, \cdots, n\})$, f represents the final output feature maps, ReLu is the activation layer we used, and $K()$ represents the SE block and 1×1 convolution layer.

4. Experiments

4.1. Datasets. We evaluate our model on popular MS COCO benchmark [35]. There are more than 200K images and 250K person instances labeled with keypoints in the COCO dataset which contain train set, validation set, and test set. 150K person instances are publicly available for training and validation. Our models are only trained on COCO train2017 dataset (includes 57K images and 150K person instances), no extra data involved, and ablation studies are conducted on the COCO val2017 dataset. Finally, we report the results on COCO test-dev2017 set to make a fair comparison with the public state-of-the-art methods.

4.2. Model Training and Inference. Our models are trained on the NVIDIA Tesla P100 GPU using PyTorch and optimized by Adam algorithm with a batch size of 32 for 140 epochs. The learning rate is initialized as 0.0001 and decreased by a factor of 0.1 at 90th and 120th epoch. The ground-truth human box is made to a fixed aspect ratio, which is height : width = 4 : 3 by extending the box in height or width. It is then cropped from the image and resized to a fixed resolution. The default resolution is 256×192, which is the same as the state-of-the-art methods [14, 15] for a fair comparison. During the model training, we use a pretrained model trained on ImageNet classification task [1].

We test HSR-MSNet with 59, 110, and 161 layers, named HSR-MSNet-59, HSR-MSNet-110, and HSR-MSNet-161, respectively. HSR-MSNet-59 is derived from [18], which can be download at https://github.com/guoruoqian/DetNet_pytorch. For HSR-MSNet-110 and HSR-MSNet-161, we adopt to initial the parameters of the first four stages from ResNet pretrained models [16].

The standard COCO metrics [35] are used to evaluate our approach, including AP (averaged precision), OKS (object keypoint similarity) thresholds, AP_{50} and AP_{75} (AP at different IoU (Intersection over Union) thresholds), AP_m and AP_l (AP at different scales: middle and large), and AR (average recall). The OKS plays the same role as the IoU in object detection, which is calculated from the distance between predicted keypoints and ground-truth keypoints normalized by scale of the person.

4.3. Quantitative Results. The experiments with HSRNet and HSR-MSNet are implemented to investigate the effectiveness of keeping high spatial resolution features and multiscale features for human pose estimation, respectively.

4.3.1. High Spatial Resolution Features. For the experiments with HSRNet, a human detector is introduced with AP 56.4 on COCO val2017, and the performance of ResNet and HSRNet with various options is listed in Table 1. Since the feature maps of HSRNet have higher spatial resolution than ResNet, two deconvolution layers are utilized to maintain the same size of output heatmaps. Methods a, b, c, d, e, g, h, i, j, and k with 256×192 input size eventually generate 64×48 heatmaps, and methods f and l with 384×288 input size generate 96×72 heatmaps.

(1) Size-varied backbone. Methods a, b, and c and g, h, and i compare the results of HSRNet and ResNet by size-varied backbones, which illustrate that HSRNet is better than ResNet by 0.4 AP at least in comparable size of backbones. Similar to ResNet, the larger the size of HSRNet backbone, the better the performance. As can be seen from Table 1, AP increases 0.4 from

Table 2: Comparisons on the COCO val2017 dataset.

Method	Backbone	Input size	Deconv. layers	Deconv. kernel size	AP
a	HSRNet-59	256 × 192	2	4	71.4
b	HSRNet-59	384 × 288	2	4	73.2
c	HSR-MSNet-59	256 × 192	2	4	71.5
d	HSR-MSNet-59	384 × 288	2	4	73.3

Table 3: Comparisons with Hourglass, CPN, and SimpleBaseline on the COCO val2017 dataset.

Method	Backbone	Input size	OHKM	AP
8-stage Hourglass	—	256 × 192	N	66.9
8-stage Hourglass	—	256 × 256	N	67.1
CPN	ResNet-50	256 × 192	N	68.6
CPN	ResNet-50	384 × 288	N	70.6
CPN	ResNet-50	256 × 192	Y	69.4
CPN	ResNet-50	384 × 288	Y	71.6
SimpleBaseline	ResNet-50	256 × 192	N	70.4
SimpleBaseline	ResNet-50	384 × 288	N	72.2
Ours	HSRNet-59	256 × 192	N	71.4
Ours	HSRNet-59	384 × 288	N	73.2

HSRNet-59 to HSRNet-110 and 1.2 from HSRNet-59 to HSRNet-161

(2) Various kernel sizes of deconvolution layers. Methods a, d, and e and j, i, and k prove that HSRNet also outperforms ResNet by at least 0.8 AP with various kernel sizes of deconvolution layers. AP increases 0.3 from kernel size 2 to 3 and 0.5 from kernel size 2 to 4 in HSRNet-59

(3) Different sizes of input images. a and f and g and l methods illustrate that the higher the resolution of input images, the better the performance. In addition, HSRNet improves 1 AP compared with ResNet

Since HSRNet has more parameters than ResNet in comparable size of backbone, it may be hard to demonstrate that HSRNet structure is better than ResNet. However, it is worth noting that AP of methods b and g are the same. The performance is comparable between HSRNet-59 and ResNet-101, which implies the ability of high spatial resolution to improve the performance of human pose estimation significantly.

4.3.2. Multiscale Features. The experiments are implemented with HSR-MSNet and compared with HSRNet to show the importance of multiscale features for human pose estimation, as shown in Table 2.

A human detector is adopted with AP 56.4 on COCO val2017. Due to the fact that HSR-MSNet can extract semantic features with multiscale and assign weights to these features by the SE module, HSR-MSNet has a better performance than HSRNet with the same parameters which is improved by 0.1AP. Obviously, multiscale features play an important role in human pose estimation, because it requires an understanding of large-scale features for the classification of keypoints, as well as small-scale features for localization of keypoints. SE block is also indispensable since it can assign weights to different features and make the most effective features prominent among others. The combination of multiscale features and weight distribution can further improve human pose estimation.

4.3.3. Comparisons with Other State-of-the-Art Methods. The results of HSRNet and other state-of-the-art models including Hourglass [23], CPN [14], and SimpleBaseline [15] on the COCO val2017 dataset are shown in Table 3. For fair comparison, a human detector provided by SimpleBaseline is introduced with 56.4 AP, which is comparable to Hourglass and CPN with 55.3 AP.

Our method exceeds Hourglass by 4.5 AP in the same input size of 256 × 192. Compared with CPN, our method outperforms CPN without OHKM by more than 2.6 AP and CPN with OHKM by 1.6 AP. Although our model is based on SimpleBaseline, it has been proved in Table 1 that our method performs better than SimpleBaseline. It is obvious that these performance gains are benefited from keeping high spatial resolution in deeper layers of the backbone encoder networks.

To gain a better performance, a human detector of 60.9 AP is applied to obtain human bounding boxes on COCO test2017 dataset. For reference, CPN uses a human detector with person detection AP of 62.9 on COCO minival split dataset and SimpleBaseline uses a human detector of 60.9 AP on COCO std-dev split dataset. As shown in Table 4, CPN uses the ResNet-Inception and SimpleBaseline uses

Table 4: Comparison experiments on the COCO test2017 dataset. Results of compared methods are cited from [15].

Method	Backbone	Input size	AP	AP_{50}	AP_{75}	AP_m	AP_l	AR
G-RMI [38]	ResNet-50	256 × 192	64.9	85.5	71.3	62.3	70.0	69.7
CPN	ResNet-Inception	384 × 288	72.1	91.4	80.0	68.7	77.2	78.5
FAIR* [37]	ResNeXt-101-FPN	—	69.2	90.4	77.0	64.9	76.3	75.2
G-RMI*	ResNet-152	384 × 288	71.0	87.9	77.7	69.0	75.2	70.6
oks* [37]	—	—	72.0	90.3	79.7	67.6	78.4	77.2
Bangbanggren∗ [37]	ResNet-101	—	72.8	89.4	79.6	68.6	80.0	78.7
CPN*	ResNet-Inception	384 × 288	73.0	91.7	80.9	69.5	78.1	79.0
SimpleBaseline	ResNet-152	384 × 288	73.7	91.9	81.1	70.3	80.0	79.0
Ours	HSRNet-59	384 × 288	72.6	91.2	79.9	69.0	78.9	77.8
Ours	HSR-MSNet-59	384 × 288	72.8	91.2	80.5	69.6	79.0	78.3

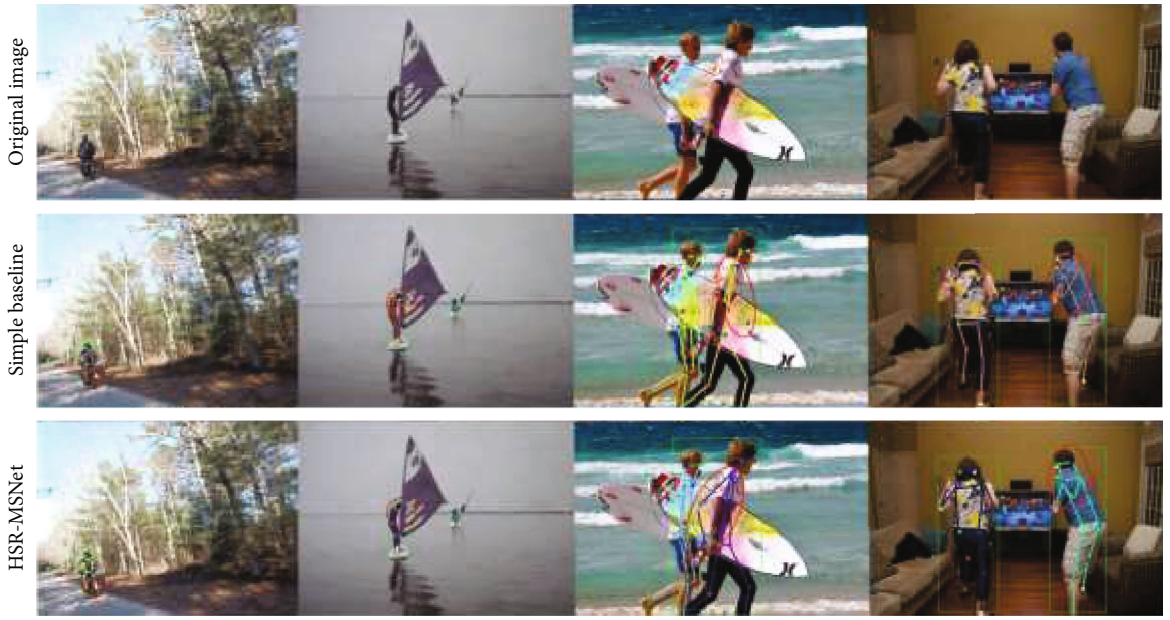

Figure 8: Qualitative comparisons on the COCO test2017 dataset.

ResNet-152 for human pose estimation. Our method only utilized a small backbone but outperforms some state-of-the-art models including G-RMI [38] and FAIR* [37] and achieves a potential performance of 72.6 AP and 72.8 AP.

Since our model is based on SimpleBaseline, the changes we made are very slight, which have a negligible impact on the overall parameters of the model. In other words, compared to SimpleBaseline, the computational complexity of our model is almost the same.

4.4. Qualitative Comparisons. Some qualitative comparison results on the COCO test2017 dataset are shown in Figure 8. We compare our approach with SimpleBaseline [15] and use comparable model to predict the keypoints for fair comparison. SimpleBaseline utilizes ResNet-50 as backbone, and our method is HSR-MSNet-59. As shown in Figure 8, the first row contains some original images of the COCO test2017 dataset, the second row is the results predicted by SimpleBaseline [15], and the third row is our results.

It is obvious that our method can better predict the keypoints of partially occluded or small people by utilizing higher resolution and multiscale feature maps. As the first column of Figure 8 has shown, our method can perceive and predict the keypoints more accurately for small people. For complex backgrounds, the second column illustrates that our method can separate the person from the background easily with high-resolution feature maps to avoid misidentification. In addition, our method can accurately predict the "hard" joints that are occluded or invisible in the third and fourth columns of Figure 8. Especially in the third column, HSR-MSNet makes it easier to detect the left arm which is heavily occluded in this image and gives a more accurate prediction. It is attributed to the ability of HSR-MSNet to mine more context information of "hard" joints.

5. Conclusion

In this paper, a novel backbone network, named HSR-MSNet, is proposed specifically for human pose estimation, which maintains high spatial resolution features in deeper layers of the encoder while still keeping large receptive fields. We also design a building module to learn multiscale features within one single residual block by splitting channels of feature maps into subgroups and then fuse these channel groups hierarchically. For multiperson pose estimation, our model can learn efficient context information of "hard" joints, due to partial occlusion or small scale of persons. Experiments on the COCO keypoint detection dataset show that our model outperforms other state-of-the-art methods, such as Hourglass [23], CPN [14], and SimpleBaseline [15] with respect to standard COCO metrics. At next steps, we will focus on evaluating our model on other human datasets, such as MPII dataset, and then further experiments on lightweight devices.

Data Availability

The data that support the findings of this study are openly available in COCO at doi:10.1007/978-3-319-10602-1_48, reference number [29].

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by the National Natural Science Foundation of China grant 61771299.

References

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[2] N. Kalchbrenner, E. Grefenstette, and P. Blunsom, "A convolutional neural network for modelling sentences," 2014, https://arxiv.org/abs/1404.2188.

[3] Y. Kim, "Convolutional neural networks for sentence classification," 2014, https://arxiv.org/abs/1408.5882.

[4] S. Ji, W. Xu, M. Yang, and K. Yu, "3d convolutional neural networks for human action recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 221–231, 2012.

[5] A. Toshev and C. Szegedy, "*DeepPose: human pose estimation via deep neural networks*," 2013.

[6] Z. Bi, L. Yu, H. Gao, P. Zhou, and H. Yao, "Improved VGG model-based efficient traffic sign recognition for safe driving in 5G scenarios," *International Journal of Machine Learning and Cybernetics*, vol. 3, 2020.

[7] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[8] L. Pishchulin, M. Andriluka, P. Gehler, and B. Schiele, "Strong appearance and expressive spatial models for human pose estimation," in *2013 IEEE International Conference on Computer Vision*, pp. 3487–3494, Sydney, NSW, Australia, 2014.

[9] W. Yang, S. Li, W. Ouyang, H. Li, and X. Wang, "Learning feature pyramids for human pose estimation," in *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 1281–1290, Venice, Italy, 2017.

[10] L. Ke, M. C. Chang, H. Qi, and S. Lyu, "Multi-scale structure-aware network for human pose estimation," in *2018 25th IEEE International Conference on Image Processing (ICIP)*, pp. 713–728, Athens, Greece, 2018.

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

[12] H. Zhang, H. Ouyang, S. Liu et al., "Human pose estimation with spatial contextual information," 2019, https://arxiv.org/abs/1901.01760.

[13] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele, "2D Human pose estimation: new benchmark and state of the art analysis," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3686–3693, Columbus, OH, USA, 2014.

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

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

[16] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, Las Vegas, NV, USA, 2016.

[17] J. Chen, H. Ying, X. Liu et al., "A transfer learning based super-resolution microscopy for biopsy slice images: the joint methods perspective," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 18, no. 1, pp. 103–113, 2020.

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

[19] A. Howard, M. Sandler, and G. Chu, "Searching for MobileNetv3," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1314–1324, Seoul, Korea, 2019.

[20] J. J. Tompson, A. Jain, Y. LeCun, and C. Bregler, "Joint training of a convolutional network and a graphical model for human pose estimation," 2014, https://arxiv.org/abs/1406.2984.

[21] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional pose machines," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4724–4732, Las Vegas, NV, USA, 2016.

[22] T. L. Munea, Y. Z. Jembre, H. T. Weldegebriel, L. Chen, C. Huang, and C. Yang, "The progress of human pose estimation: a survey and taxonomy of models applied in 2D human pose estimation," *IEEE Access*, vol. 8, pp. 133330–133348, 2020.

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

[24] B. Li, K. Liu, Y. Ji, J. Yang, and C. Liu, "Selective complementary features for multi-person pose estimation," in *2020 IEEE international conference on image processing (ICIP*, pp. 623–627, Abu Dhabi, United Arab Emirates, 2020.

[25] Z. Cao, T. Simon, S. E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7291–7299, Honolulu, HI, USA, 2016.

[26] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, , no. 2, pp. 386–397, 2020.

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

[28] W. Liu, D. Anguelov, D. Erhan et al., "SSD: Single shot multibox detector," in *European conference on computer vision*, pp. 21–37, Amsterdam, The Netherlands, 2016.

[29] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2017.

[30] Z. Li, C. Peng, G. Yu, X. Zhang, Y. Deng, and J. Sun, "DetNet: a backbone network for object detection," 2018, https://arxiv.org/abs/1804.06215.

[31] S. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang, and P. H. S. Torr, "Res2net: a new multi-scale backbone architecture," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 2, pp. 652–662, 2021.

[32] J. Xiao, H. Xu, H. Gao, M. Bian, and Y. Li, "A weakly supervised semantic segmentation network by aggregating seed cues: the multi-object proposal generation perspective," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 17, no. 1s, pp. 1–19, 2021.

[33] J. Xiao, H. Xu, W. Zhao, C. Cheng, and H. H. Gao, "A prior-mask-guided few-shot learning for skin lesion segmentation," *Computing*, pp. 1–13, 2021.

[34] J. Redmon and A. Farhadi, "YOLOv3: an incremental improvement," 2018, https://arxiv.org/abs/1804.02767.

[35] T.-Y. Lin, M. Maire, S. Belongie et al., "Microsoft COCO: common objects in context," in *Computer Vision – ECCV 2014. ECCV 2014*, pp. 740–755, Springer, 2014.

[36] Z. Cao, C. Sun, W. Wang, X. Zheng, J. Wu, and H. Gao, "Multi-modality fusion learning for the automatic diagnosis of optic neuropathy," *Pattern Recognition Letters*, vol. 142, pp. 58–64, 2021.

[37] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 8, pp. 2011–2023, 2020.

[38] G. Papandreou, T. Zhu, N. Kanazawa et al., "Towards accurate multi-person pose estimation in the wild," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4903–4911, Honolulu, HI, USA, 2017.