2020 25th International Conference on Pattern Recognition (ICPR)
Milan, Italy, Jan 10-15, 2021

Efficient High-Resolution High-Level-Semantic Representation Learning for Human Pose Estimation

Hong Liu, Lisi Guan
Key Laborary of Machine Perception
Peking University, Shenzhen Graduate School
Email: {hongliu, guanlisi}@pku.edu.cn

***Abstract*—Ensuring that features contain both high-resolution and high-level-semantic information is important for human pose estimation, while most existing methods suffer from spatial information loss or semantic information mismatch when extracting high-resolution high-level-semantic features. To efficiently address these issues, we propose a novel Dilation Pyramid Module (DPM), which can enlarge the receptive field multiplicatively to extract high-level-semantic information as subsampling without reducing spatial resolution. DPM is composed of several consecutive dilated convolution layers of which dilation radius is specially designed to enlarge the receptive field multiplicatively and avoid the gridding issue of dilated convolution. Based on DPM, the Dilation Pyramid Net (DPN) is proposed to efficiently extract high-resolution high-level-semantic features. We experimentally demonstrate the effectiveness and efficiency of the proposed DPN with competitive performance to the state-of-the-art methods over two challenging benchmark datasets: the COCO keypoint detection dataset and the MPII Human Pose dataset.**

I. Introduction

Human pose estimation that targets at locating keypoints of the human body parts, has attracted research attention due to its potential applications in various high-level comprehension tasks, such as re-identification [1], [2], action recognition [3], [4]. To support these comprehensive high-level applications, the human pose estimator needs to effectively deal with arbitrary clothing appearances, different lights, and background clutter.

The recent advanced methods of human pose estimation aim to extract high-resolution high-level-semantic features, because high-resolution can reduce quantization error and high-level-semantic information is useful to catch global information, especially for multi-person pose estimation. While there is lacking an efficient and effective way to extract high-resolution high-level-semantic features. As shown in Fig. 1 (a) and Fig. 1 (d), some methods achieve this by directly restoring the resolution of low-resolution and high-level-semantic features extracted by a heavy backbone net. Those methods suffer from unrecoverable spatial information loss because they rely on heavy backbone nets (ResNet [5] or VGG [6]) designed for classification tasks to extract high-level-semantic features at the cost of spatial information loss. To deal with the spatial information loss, some methods fuse multi-scale features gradually like Fig. 1 (b), such as Hourglass [7] and CPN [8], while fusing multi-scale features with different semantic information suffers from the semantic information mismatch. Because those fused features of different resolution contain different level semantic information, the high-resolution low-level-semantic features may harm the semantic information of the low-resolution high-level-semantic features.

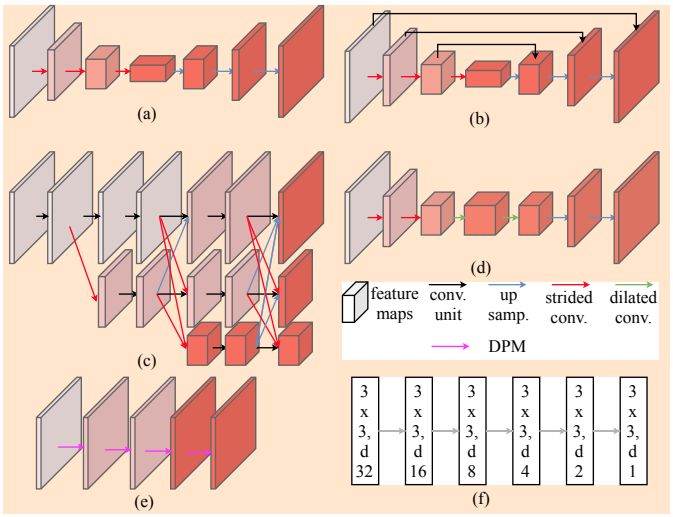

Fig. 1: Illustration of representative pose estimation methods of extracting high-resolution high-level-semantic features. Feature maps with darker color mean they contain higher level semantic information. (a) SimpleBaseline [10]:transposed convolutions for upsampling, (b) Hourglass [7], CPN [8] adopts a similar structure of (b) with an additional refine net, (c) HRNet [9], (d) Combination with dilated convolution [11], (e) is our DPN, (f) is the proposed dilation pyramid module (DPM) and d denotes the dilation radius. Bottom-right legend: conv. unit = normal convolution unit, up samp. = upsampling operation, strided conv. = strided convolution, dilated conv.= dilated convolution.

The recently proposed method, HRNet [9], keeps all features with different semantic information and resolution to extract high-resolution high-level-semantic features as shown in Fig. 1 (c). This method deals with the spatial information loss by keeping high-resolution features, while it still relies on fusing low-resolution high-level-semantic features to achieve high-level-semantic information, which also suffers from semantic information mismatch.

To extract high-resolution high-level-semantic features directly and effectively, we propose a novel dilation pyramid module (DPM), which can directly extract high-resolution high-level-semantic features without spatial information loss

978-1-7281-8808-9/20/$31.00 ©2020 IEEE

2020 25th International Conference on Pattern Recognition (ICPR) | 978-1-7281-8808-9/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICPR48806.2021.9412886

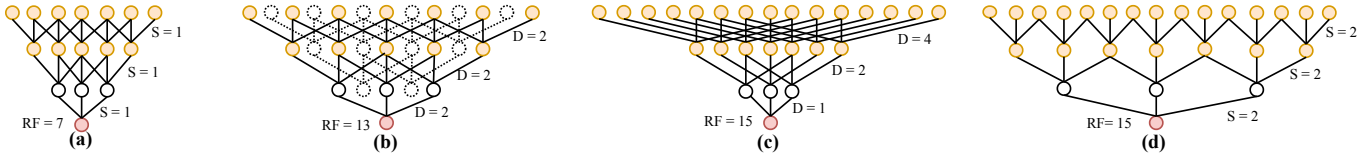

Fig. 2: The receptive fields of stacking three different convolution operations. For simplification, we illustrate their 1D counterparts. The pixels (marked in yellow) contributes to the calculation of the center pixel (marked in red) through three convolution layers with kernel size 3×3. (a) is normal 1D convolution, (b) is 1D dilated convolution with dilation radius equals to two, (c) is the proposed dilation pyramid module, and (d) is normal convolution with stride equal to two. S is stride and D is dilation radius. RF is receptive fields. It is obvious that DPM can enlarge RF multiplicatively as subsampling without spatial information loss. Compare to dilated convolution, DPM does not suffer from gridding issue of dilated convolution.

and semantic information mismatch. DPM is composed of several consecutive dilated convolution [12] layers of which dilation radius is specially designed. A simple example is shown in Fig. 2 (c).

Compared with the traditional dilation operation, DPM specially designs the dilation radius for each layer, which can enlarge receptive fields multiplicatively without spatial information loss and overcome the gridding issue of stacking dilated convolution with constant dilation radius. Based on DPM, we further propose an efficient and effective dilation pyramid network (DPN), which can achieve competitive performance to the state-of-the-art methods. As a result, with DPN, computation cost and parameter consumption can be reduced considerably.

Compared to existing widely-used networks for human pose estimation, our method has two benefits.

(i) DPN can extract high-resolution and high-level-semantic features efficiently without spatial information loss and does not need to recovery resolution by fusing features of different resolution.

(ii) DPN achieves efficiency in both parameter and computation. Compared to previous advanced methods, it contains considerable fewer parameters and lower computation cost. therefore it can be easily applied in practice.

We experimentally demonstrate the competitive performance over two benchmark datasets: the COCO keypoint detection [13] dataset and the MPII Human Pose [14] dataset.

II. Related works

Toshev et al. introduce deep neural networks into human pose estimation in [15] for the first time and demonstrate the potential promising performance of deep neural networks, methods of human pose estimation quickly shifted from classic approaches [16]–[19] to the methods based on deep neural networks [8], [10], [20], [21]. Within CNN based methods, fundamentally, there are two kinds of technical routes: regressing the coordinates of body joints [22] and estimating heatmaps of body joints followed by choosing the coordinates of the highest values as the prediction of body joints. This paper focuses on heatmap based methods.

To extract heatmaps containing both high-resolution and high-level-semantic information, different methods adopt different ways. Some methods extract low-resolution high-level-semantic features first, then restore the spatial resolution of it. For example, Xiao et al. adopt three deconvolution layers in [10] to enlarge the resolution of low-resolution high-level-semantic features by eight times. Chen et al. use a few bilinear upsampling operations in [8] to restore spatial resolution of low-resolution high-level-semantic features. Hourglass [7] and its follow-up [23]–[25] design a low-to-high process to restore spatial resolution by fuse multi-scale features gradually. Some works [11], [20] make compromise between high-level-semantic and high-resolution by replacing subsampling with dilated convolution in the last two stages in the classification net. Though replacing subsampling with dilated convolution improve the resolution of the final representations, it lower the semantic information of the features. Recent work HRNet [9] focuses on maintaining both high-resolution low-semantic features and high-resolution high-level-semantic features through the whole process for spatially precise heatmap estimation. It generates high-resolution representations through repeatedly fusing the representations produced by the high-to-low subnetworks, which reliefs the semantic mismatch problem. But it still suffers from the semantic mismatch problem caused by fusing features with different level semantic information.

To tackle this problem, we propose a novel dilation pyramid module (DPM), which can directly extract both high-resolution and high-level-semantic features without spatial information loss caused by subsampling as well as semantic information mismatch. Based on the novel dilation pyramid module (DPM), we propose the dilation pyramid neural network (DPN), which enlarges receptive fields multiplicatively as upsampling with little spatial information loss. DPN is different from most existing methods, which relies on subsampling to enlarge receptive fields multiplicatively such as [7], [8], [21]. Some related methods about human pose estimation [8], [11], [20] or segmentation [12] use dilated convolution to extract high-resolution high-level-semantic features. Those methods use dilated convolution to modify a few last stages of backbone net only for reducing spatial information loss, and they are markedly different from our method.

III. The proposed method

It is crucial to extract high-resolution high-level-semantic features for human pose estimation based on heatmap, because high-resolution can reduce quantization error and high-level-

semantic is useful to catch global information especially for multi-person pose estimation. But existing methods of extracting high-resolution high-level-semantic features are suffering from unrecoverable spatial information loss or semantic information mismatch. This paper aims to propose a method to extract high-resolution high-level-semantic features efficiently and effectively without spatial information loss and semantic information mismatch.

A. Dilation Pyramid Module

Common practices extract high-level-semantic features by enlarging the receptive fields of features. There are many ways to enlarge receptive fields such as stacking normal convolution layers, dilated convolution, and subsampling shown in Fig. 2.

The receptive fields of stacking normal convolution layers can be formulated as follow:

$$RF_{n+1} = RF_n + k - 1, \tag{1}$$

where RF_{n+1} is the receptive field of the output feature of the $(n+1)_{th}$ convolution layer. k is the kernel size of the $(n+1)_{th}$ convolution layer. It is obvious that staking normal convolution layers can only increases receptive field size linearly. A simple 1D sample of normal convolution is shown in Fig. 2 (a).

The receptive fields of stacking dilated convolution layers can be formulated as follow:

$$RF_{n+1} = RF_n + (k-1) \times d_{n+1}, \tag{2}$$

where d_{n+1} is the dilation radius of the $(n+1)_{th}$ dilated convolution layer. As shown in Fig.2 (b), the receptive field of stacking dilated convolution increases more rapidly than that of stacking normal convolution and the spatial resolution of the input feature is unchanged. But dilated convolution suffers from gridding issue, a simple example is shown in Fig.2 (b), there are only half of the pixels in the receptive field contributing to the final output. In the experiment section, we show that the gridding issue of dilated convolution is unacceptable when the dilation radius is large enough.

To increase receptive fields multiplicatively, existing methods adopt subsampling by pooling or using convolution with stride greater than one. The receptive fields improved by using stride larger than one can be formulated as follows:

$$RF_{n+1} = RF_n + (k-1) \times j_n, \tag{3}$$

$$j_{n+1} = j_n \times s_{n+1}, \tag{4}$$

where s_n is the stride of the n_{th} convolution layer, and j_n is the total stride of the first n convolution layers. As shown in Fig. 2 (d), a subsampling operation enlarges the receptive fields of input feature by four times, while the spatial resolution of the input feature is reduced by four times too and the lost spatial information can not be fully recovered. For tasks like classification do not sensitive to the spatial resolution of features, spatial information loss is not a problem. For tasks like human pose estimation, which need high spatial resolution features, spatial information loss is a major obstacle in performance improvement.

TABLE I: Architectures of proposed DPN. Building blocks are shown in brackets. The dilation radius of i_{th} 3×3 convolution in DPM is determined by 2^{6-i}. Subsampling is performed by conv2_0 with a stride of 2.

layer name	output size	DPN4_32	DPN4_64
conv0	128x96	7x7,64,stride 2	
pooling	64x48	3x3 max pool,stride 2	
conv1	64x48	[1x1, 64; 3x3, 64; 1x1,256] ×3	
conv2	32x24	1x1,128	[1x1,128; 3x3,128; 1x1,512] ×4
			1x1,256
DPM1	32x24	[1x1,32; 3x3,32; 1x1,128] ×6	[1x1,64; 3x3,64; 1x1,256] ×6
DPM2	32x24	[1x1,32; 3x3,32; 1x1,128] ×6	[1x1,64; 3x3,64; 1x1,256] ×6
DPM3	32x24	[1x1,32; 3x3,32; 1x1,128] ×6	[1x1,64; 3x3,64; 1x1,256] ×6
DPM4	32x24	[1x1,32; 3x3,32; 1x1,128] ×6	[1x1,64; 3x3,64; 1x1,256] ×6
heatmap	64x48	upsamplex2	
		1x1,17	
#Params	-	0.7 M	3.3 M

Dilation Pyramid Module (DPM) can enlarge receptive fields multiplicatively without spatial information loss and semantic information mismatch. DPM is composed of N consecutive dilated convolution layers, of which dilation radius is specially designed. DPM is defined as follow:

$$X_{out} = f_N^{d_N}(f_{N-1}^{d_N-1}(...(f_2^{d_2}(f_1^{d_1}(X_{in}))))), \tag{5}$$

where $f_i^{d_i}$ is the i_{th} dilated convolution layer and d_i is the dilation radius of it, d_i is determined by 2^{N-i}, X_{in} and X_{out} are input features and output features, respectively. The kernel size of all the dilated convolution of DPM is set to $k \times k$. The receptive fields improved by using DPM can be formulated as follows:

$$\begin{aligned} RF_{total} &= k + (k-1) * 2^1 + ... + (k-1) * 2^{N-1} \\ &= (k-1)(1 + 2^1 + ... + 2^{N-1}) + 1 \\ &= (k-1)(2^N - 1) + 1, \end{aligned} \tag{6}$$

where RF_{total} is the total receptive fields of DPM. DPM can enlarge receptive fields multiplicatively as subsampling and keep spatial resolution unchanged. By default, the kernel size of dilated convolution used in DPM is set to 3×3 for the consideration of parameter consumption and computation cost. As shown in Fig. 2 (c), the receptive field of DPM is complete. Thus DPM does not suffer from the gridding issue of dilated convolution. In the experiments section, we study the influence

of gridding issue of dilation convolution in detail. DPM keeps spatial resolution and enlarge receptive fields multiplicatively at the same time, so it is effective and efficient for tasks that need high-resolution high-level-semantic features. In the experiments section, we experimentally show the effectiveness of DPM.

B. Dilation Pyramid Neural Network

Most existing approaches of human pose estimation only focus on how to improve generalization performance, while the significant efficiency problem is put aside. This paper aims to propose an efficient and effective model. Based on DPM, the extreme lightweight Dilation Pyramid Neural Network (DPN) is proposed. Though DPM can maintain high spatial resolution and high-resolution is useful to reduce quantization error, keeping high spatial resolution usually means much more computation. To achieve a balance between performance and computation cost, we choose to keep medium-resolution. And we will show in experiments that the medium-resolution features is lightly lower in performance compared to the high-resolution feature with computation cost reduced significantly. The details of DPN are shown in Tab I. There are two proposed network DPN4_32 and DPN4_64, DPN4_32 achieves further reduction of parameters and computation cost. The proposed network DPN4_32 achieves promising performance with only 0.7 M parameters, which is only no more than one twentieth of that of the stat-of-the-art method.

IV. EXPERIMENTS AND ANALYSIS

We evaluate the performance of the proposed DPN net on two challenging benchmark datasets: the COCO keypoint detection dataset and the MPII Human Pose dataset.

A. Datasets and Settings

Datasets: The COCO dataset [13] contains about 200k images with labeled keypoints. Our model trained only on COCO train2017 dataset and is validated on COCO val2017 dataset. The standard evaluation metric is mean Average Precision (mAP) over ten object keypoint similarity (OKS) thresholds. The OKS has the same role as the IoU in object detection. OKS is calculated by the flowing formula, $OKS = \frac{\sum_i \exp(-d_i^2/2s^2k_i^2)\delta(v_i>0)}{\sum_i \delta(v_i>0)}$. Here d_i is the Euclidean Distance between the coordinates of predicted keypoint and the corresponding ground truth, s is the object scale and v_i is the visibility flag of the ground truth, k_i is a per-keypoint constant. The MPII Human Pose dataset [14] contains around 25K images with 40K subjects, there are 12K subjects for testing and the remaining subjects for the training set.

Training: Data augmentation only includes random scale ($\pm$ 30%), random rotation ($\pm$ 40°), random flip and half body data augmentation which is mentioned in HRNet [9]. Adam [26] is adopted as the optimizer. The base learning rate is 1e-3 and is dropped to 1e-4 and 1e-5 at 90th and 120th epoch, respectively. Training process is terminated within 140 epochs by default.

TABLE II: Comparison between normal convolution and DPM. DPN1 means that the model contains one DPM module. Normal means that the dilation radiuses of all the dilated convolution layers are set to one. So all the compared methods consume the same parameters and computation cost. AP is average precision.

		DPN1_64	DPN2_64	DPN3_64	DPN4_64
AP	DPM	**64.5**$_{\uparrow 4.8}$	**68.0**$_{\uparrow 4.0}$	**69.2**$_{\uparrow 4.9}$	**69.9**$_{\uparrow 5.3}$
	Normal	59.7	64.0	64.3	64.6

TABLE III: Comparison with Stacked Dilated Convolution. AP is average precision. D means that the dilation radius of all the dilated convolution layers in DPM is set to D.

	DPM	D=1	D=2	D=4	D=6	D=8
DPN1_32	**50.6**	40.1	**49.3**	48.7	43.9	38.5
DPN4_32	**62.0**	58.2	**60.6**	56.6	50.0	48.7

Testing: For the COCO dataset, we use the same person bounding boxes provided by SBL [10] to generate the cropped input image of the validation set. Following the common practice [7], [8], [10], the averaged heatmap of the original and flipped image is used to predict the joint location and then a quart offset in the direction from the highest response to the second-highest response is added to obtain final location. For the MPII dataset, the testing procedure is almost the same as that in COCO except that we adopt the standard testing strategy to use the provided person boxes instead of detected person boxes for fair comparison.

B. Ablation Studies

By default, all of the models used in the ablation study are trained on COCO train2017 dataset and are validated on COCO val2017 dataset.

Comparison with Normal Convolution: To fairly compare with normal convolution, we make the compared models possess the same parameter volume and computation cost. The normal counterpart of DPM is implemented by setting the stride and dilation radius of all the dilated convolution of DPM both to one. To rule out the effect of the resolution of feature maps, the compared models maintain the same feature map size. Results are reported in Tab II, DPM consistently outperforms its conventional convolution counterpart by a large margin, which is greater than 4 percentage in the terms of average precision (AP). Those results show the effectiveness of the proposed DPM. Because the human pose estimation task needs large receptive fields to generate high-level-semantic features and high-level-semantic information is useful to catch global clues. And DPM can achieve a much larger receptive field than its normal counterpart model and generate features with higher level semantic information than that of its conventional convolution counterpart. This is the reason that why DPM outperforms its conventional convolution counterpart.

Comparison with Stacked Dilated Convolution: Staked dilation convolution is implemented by setting the dilation radius of all the dilated convolution layers of DPM to a fixed number, D. So all of the compared models contain

TABLE IV: Comparison between high-resolution representation and medium-resolution representation. AP is average precision. DPN1 means that model contains one DPM module. Resolution is spatial resolution of input features to DPM module. The input image size is 256×192. Here the high-resolution model is trained for 210 epochs.

	Epochs	Resolution	DPN1_64	DPN2_64	DPN3_64	DPN4_64
AP	210	64×48	65.5	68.7	70.3	71.1
	140	32×24	**65.2**	**68.0**	**69.2**	**69.9**
FLOPs	210	64×48	6.0 G	7.2 G	8.4 G	9.6 G
	140	32×24	**2.1** G	**2.4** G	**2.7** G	**3.0** G

TABLE V: Comparison between reversed DPM (RDPM) and DPM. DPN1 means that the model contains one DPM module. The dilation radius of RDPM is defined by 2^{i-1}, so the dilation radius scheme of RDPM is increasing. AP is of average precision.

		DPN1_32	DPN2_32	DPN3_32	DPN4_32
AP	DPM	**50.6**$_{\uparrow 1.6}$	56.9$_{\uparrow 0.4}$	60.0$_{\uparrow 0.3}$	62.0$_{\uparrow 0.4}$
	RDPM	**49.0**	56.6	59.6	61.6

the same parameter volume and FLOPs. The results are reported in Tab III. DPM outperforms all of the stacked dilated convolution counterparts by a large margin with the same parameters and FLOPs. One reason is that DPM achieves larger receptive fields than stacked dilation convolution and the other reason is that DPM does not suffer from the gridding issue of dilated convolution. The gridding issue is shown in Tab III, the performance of stacked dilated convolution decreases when the dilation radius is larger than two. When the dilation radius is equal to 8, the performance is even lower than normal convolution ($D = 1$) by a large margin. In contrast to stacked dilated convolution, the proposed DPM module contains dilated convolution with dilation radius larger than 32 and achieves performance improvements. This experiment shows that DPM overcomes the gridding issue of dilated convolution, and simply adopting dilated convolution to human pose estimation is not applicable.

Influence of Feature's Resolution: Tab IV reports the results of high-resolution representation, 64×32, and medium-resolution representation, 32×24. Those results show that medium-resolution achieves comparable performance to high-resolution, while medium-resolution only needs about one-third of the calculation of high-resolution. Though high-resolution is effective for human pose estimation to reduce quantization error, it will bring much more computation cost and need more time to be trained well. Those experiments results show that the proposed DPN model is not sensitive to the resolution of features, it can achieve competitive performance with lower resolution.

Comparison between reversed DPM and DPM: RDPM is implemented by reversing the dilation radius setting of DPM, so RDPM has an increasing dilation radius scheme. Tab V reports the results of reversed DPM (RDPM). When the compared model only contains one module, DPM outperforms RDPM by more than one percent in terms of average precision. Though RDPM has the same receptive field with DPM, while it still suffers from gridding issue. When there is more than one module in the compared model, RDPM achieves a competitive performance with DPM, while DPM still outperforms it consistently. The reason is that stacked RDPM containing a decreasing dilation radius scheme like DPM, and this reduces the effect of the gridding issue of dilated convolution. By default, we adopt DPM as the basic module of the proposed DPN net.

TABLE VI: Compared with the state-of-the-art methods on COCO val2017 dataset. AP is average precision. Resolution of input image is 256×192.

Method	Backbone	#Params	GFLOPs	AP
OpenPose [27]	-	-	-	61.8
Mask-RCNN [28]	ResNet-50	-	-	63.1
Hourglass [7]	8-stage Hourglass	25.1 M	14.3	66.9
CPN [8]	ResNet-50	27.0 M	6.2	69.4
SBL [10]	ResNet-50	34.0 M	8.9	70.4
HRNet-w32 [9]	HRNet-w32	28.5 M	7.1	74.4
HRNet-w48 [9]	HRNet-w48	**63.6** M	**14.6**	**75.1**
DPN4_32 (ours)	DPN	**0.7** M	**1.1**	**62.0**
DPN4_64 (ours)	DPN	**3.3** M	**3.0**	**69.9**
DPN2_hrnet (ours)	HRNet-w32	29.5 M	7.8	**75.0**

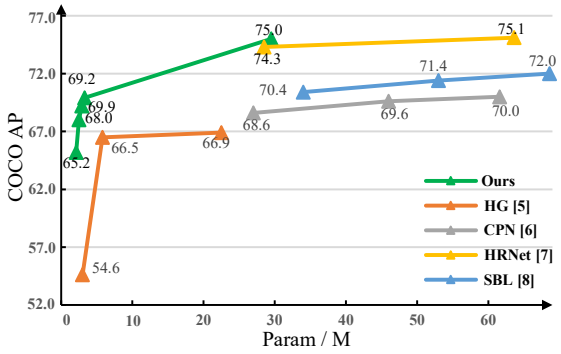

Fig. 3: Comparison with previous state-of-the-art methods about parameter efficiency on MS COCO dataset.

C. Comparison with the state-of-the-arts

COCO val2017 dataset Tab VI shows the comparison between our methods and the state-of-the-art methods. The proposed DPN4_64 achieves better performance than previous state-of-the-art methods OpenPose [27] and CPN [8] with significantly lower computation cost and much fewer parameters. The proposed lightweight model, DPN4_32, achieves better performance than OpenPose [27] with only **0.7** mega parameters. The proposed Dilation Pyramid Module (DPM) is effective to other backbone nets too. As shown in Tab VI, DPM2_HRNet has just two more DPM modules stacked to the end of the backbone net HRNet-w32 and achieves competitive performance to the heavy HRNet-w64 with only half computation and parameters of it. As shown in Fig. 3, the proposed DPN net achieves much better parameter efficiency than the previous state-of-the-art methods. As shown in Tab VI, our DPN net consumes less computation cost than previous methods achieving better or competitive performance.

TABLE VII: Compared with the state-of-the-art methods on MPII test set (PCKh@0:5). Resolution of input image is 256×256.

Method	#Patams	GFLOPs	PCKh@0.5
Deepercut [11]	42.6 M	41.2	88.5
SBL [10]	68.6 M	20.9	91.5
MCA [30]	58 M	128	91.5
JADA [31]	26.0 M	55.0	91.5
LFP [32]	28.0 M	46.0	92.0
Deeply [33]	15.5 M	**15.6**	92.3
PIL [29]	**26.0** M	63.0	92.4
DPN4_32 (ours)	**0.7** M	**1.4**	**86.0**
DPN4_64 (ours)	**3.3** M	**4.0**	**88.6**

MPII test set Tab VII shows the results on the MPII test set. Our DPN4_64 achieves more than 95% of all the previous state-of-the-art methods with parameter volume and computation cost decreased significantly. For example, compared with the PIL [29] method, our DPN4_64 achieves 95.8% of the performance of PIL [29], while our DPN4_64 only consumes 12.7% of the parameters and 6.3% of the computation consumed by PIL [29]. Those results demonstrate that our DPN net achieves better efficiency both in parameters and computation than previous methods on the MPII dataset.

V. Conclusions

In this paper, we aim to address the spatial information loss and semantic information mismatch of extracting features containing both high-level-semantic information and high-resolution. Under this consideration, we extend the traditional dilation operation with specially designed dilation radius and propose a novel dilation pyramid module, which can efficiently extract high-resolution high-level-semantic features without spatial information loss and semantic information mismatch. Base on DPM, we propose an efficient dilation pyramid net, which achieves competitive performance on two representative benchmark datasets of human pose estimation with considerable parameter and computation cost reduction. Future works will study the effectiveness of the proposed high-resolution high-level-semantic extraction method to other tasks, such as classification, segmentation, and detection.

Acknowledgment

This work is supported by National Natural Science Foundation of China (No.U1613209), National Natural Science Foundation of Shenzhen (No.JCYJ20190808182209321).

References

[1] H. Huang, W. Yang, X. Chen, X. Zhao, K. Huang, J. Lin, G. Huang, and D. Du, "Eanet: Enhancing alignment for cross-domain person re-identification," *arXiv preprint arXiv:1812.11369*, 2018.

[2] Y. Suh, J. Wang, S. Tang, T. Mei, and K. Mu Lee, "Part-aligned bilinear representations for person re-identification," in *ECCV*, 2018, pp. 402–419.

[3] M. Liu, H. Liu, and C. Chen, "Enhanced skeleton visualization for view invariant human action recognition," *Pattern Recognition*, vol. 68, pp. 346–362, 2017.

[4] J. Tu, H. Liu, F. Meng, M. Liu, and R. Ding, "Spatial-temporal data augmentation based on lstm autoencoder network for skeleton-based human action recognition," in *IEEE International Conference on Image Processing (ICIP)*. IEEE, 2018, pp. 3478–3482.

[5] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.

[6] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition"," in *ICLR*, 2015, pp. 1–14.

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

[8] Y. Chen, Z. Wang, Y. Peng, Z. Zhang, G. Yu, and J. Sun, "Cascaded pyramid network for multi-person pose estimation," *CVPR*, 2018.

[9] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," *arXiv preprint arXiv:1902.09212*, 2019.

[10] B. Xiao, H. Wu, and Y. Wei, "Simple baselines for human pose estimation and tracking," *ECCV*, 2018.

[11] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, and B. Schiele, "Deepercut: A deeper, stronger, and faster multi-person pose estimation model," in *ECCV*, 2016, pp. 34–50.

[12] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *PAMI*, vol. 40, no. 4, pp. 834–848, 2017.

[13] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *ECCV*. Springer, 2014, pp. 740–755.

[14] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele, "2d human pose estimation: New benchmark and state of the art analysis," in *CVPR*, 2014, pp. 3686–3693.

[15] A. Toshev and C. Szegedy, "Deeppose: Human pose estimation via deep neural networks," in *CVPR*, 2014, pp. 1653–1660.

[16] B. Sapp and B. Taskar, "Modec: Multimodal decomposable models for human pose estimation," in *CVPR*, 2013, pp. 3674–3681.

[17] L. Ladicky, P. H. Torr, and A. Zisserman, "Human pose estimation using a joint pixel-wise and part-wise formulation," in *CVPR*, 2013, pp. 3578–3585.

[18] P. Felzenszwalb, D. McAllester, and D. Ramanan, "A discriminatively trained, multiscale, deformable part model," in *CVPR*, 2008, pp. 1–8.

[19] L. Pishchulin, M. Andriluka, P. Gehler, and B. Schiele, "Strong appearance and expressive spatial models for human pose estimation," in *CVPR*, 2013, pp. 3487–3494.

[20] L. Pishchulin, E. Insafutdinov, S. Tang, B. Andres, M. Andriluka, P. V. Gehler, and B. Schiele, "Deepcut: Joint subset partition and labeling for multi person pose estimation," in *CVPR*, 2016, pp. 4929–4937.

[21] S.-E. W. Zhe Cao, Tomas Simon and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," *CVPR*, pp. 1302–1310, 2017.

[22] X. Sun, B. Xiao, F. Wei, S. Liang, and Y. Wei, "Integral human pose regression," in *ECCV*, 2018, pp. 536–553.

[23] X. Chu, W. Yang, W. Ouyang, C. Ma, A. L. Yuille, and X. Wang, "Multi-context attention for human pose estimation," in *CVPR*, 2017, pp. 1831–1840.

[24] L. Ke, M.-C. Chang, H. Qi, and S. Lyu, "Multi-scale structure-aware network for human pose estimation," in *ECCV*, 2018, pp. 713–728.

[25] W. Yang, S. Li, W. Ouyang, H. Li, and X. Wang, "Learning feature pyramids for human pose estimation," in *ICCV*, 2017, pp. 1281–1290.

[26] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv:1412.6980*, 2014.

[27] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *CVPR*, 2017.

[28] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *ICCV*, 2017, pp. 2980–2988.

[29] X. Nie, J. Feng, Y. Zuo, and S. Yan, "Human pose estimation with parsing induced learner," in *CVPR*, 2018, pp. 2100–2108.

[30] X. Chu, W. Yang, W. Ouyang, C. Ma, A. L. Yuille, and X. Wang, "Multi-context attention for human pose estimation," in *CVPR*, 2017, pp. 1831–1840.

[31] X. Peng, Z. Tang, F. Yang, R. S. Feris, and D. Metaxas, "Jointly optimize data augmentation and network training: Adversarial data augmentation in human pose estimation," in *CVPR*, 2018, pp. 2226–2234.

[32] W. Yang, S. Li, W. Ouyang, H. Li, and X. Wang, "Learning feature pyramids for human pose estimation," in *ICCV*, 2017, pp. 1281–1290.

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