Adaptive Dilated Convolution For Human Pose Estimation

Zhengxiong Luo[1,2,3,4], Zhicheng Wang[1], Yan Huang[3,4], Liang Wang[3,4], Tieniu Tan[3], Erjin Zhou[1]
[1] Megvii Inc [2] School of Artificial Intelligence, University of Chinese Academy of Sciences (UCAS)
[3] Center for Research on Intelligent Perception and Computing (CRIPAC)
National Laboratory of Pattern Recognition (NLPR)
[4] Institute of Automation, Chinese Academy of Sciences (CASIA)
zhengxiong.luo@cripac.ia.ac.cn {wangzhicheng, zej}@megvii.com {yhuang, wangliang, tnt}@nlpr.ia.ac.cn

Abstract—Most existing human pose estimation (HPE) methods exploit multi-scale information by fusing feature maps of four different spatial sizes, i.e. 1/4, 1/8, 1/16, and 1/32 of the input image. There are two drawbacks of this strategy: 1) feature maps of different spatial sizes may be not well spatially aligned, which potentially hurts the accuracy of keypoint location; 2) these scales are fixed and inflexible, which may restrict the generalization ability over various human sizes. Towards these issues, we propose an adaptive dilated convolution (ADC). It can generate and fuse multi-scale features of the same spatial sizes by setting different dilation rates for different channels. Specifically, it uses a regression module to adaptively generate dilation rates for different channels. This also enables ADC to adjust the fused scales according to the sizes of test persons, and thus helps ADC to have better generalization ability. ADC can be end-to-end trained and easily plugged into existing methods. Extensive experiments show that ADC can bring consistent improvements to various HPE methods. The source codes will be released for further research.

I. Introduction

Human Pose Estimation (HPE) aims to locate skeletal keypoints (e.g. ear, shoulder, elbow, etc.) of all persons in the given RGB image. It is fundamental to action recognition and has wide applications in human-computer interaction, animation, etc. This paper is interested in single-person pose estimation, which is the basis of multi-person pose estimation [1], [2].

HPE involves two sub-tasks: location (determining where the keypoints are) and classification (determining which kinds the keypoints are). The location needs plenty of local details to get pixel-level accuracy. While classification requires a relatively larger receptive field to extract discriminative semantic representations [3]. Consequently, HPE methods have to fuse multi-scale information to make a balance between these two sub-tasks [4]. Most nowadays HPE methods [5], [6], [7], [8], [9] repeatedly downscale feature maps to enlarge the receptive fields. Feature maps of different spatial sizes (i.e. 1/4, 1/8, 1/16, and 1/32 of the input image) are then resized and summed to exploit multi-scale information.

This strategy has made great achievements in HPE [7], [4], [8], but it still leaves to be desired. In this strategy, feature maps are downscaled by strided convolution (or pooling). As shown in Figure 1, during the downscaling, multiple pixels on the larger feature maps are merged into the same pixel on the smaller ones. The location information will be destroyed in this process. While during the upscaling, even if the transposed convolution [10] is used, it is hard to recover the destroyed location information. Consequently, there will be multiple possible corresponding positions on the upscaled feature maps for original single pixel. Although the final resized feature maps have the same spatial sizes, their pixels may be not well aligned. This spatial non-alignment potentially hurts the accuracy of location. Thus, it may be more preferred to fuse multi-scale features of the same spatial sizes.

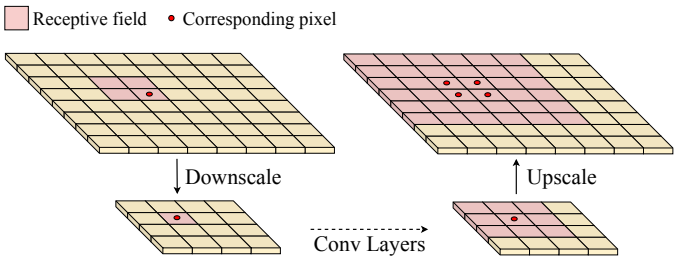


Fig. 1. The receptive fields can be easily enlarged by the downscale-conv-upscale loop. But during the upscaling, there will multiple possible postilions for the corresponding pixel. The upscaled feature maps may be not well aligned with original ones.

An alternative method is to use dilated convolution, instead of downscaling, to enlarge receptive fields. In [11], [12], multiple convolutional layers with different dilation rates are used to extract feature maps at different scales. These feature maps have the same spatial sizes and are well aligned spatially. They are concatenated and fused by 1×1 convolution to exploit multi-scale information. However, these dilation rates are still manually set and fixed, which may restrict the generalization ability over various human sizes.

Towards these issues, we propose an adaptive dilated convolution (ADC) in this paper. As shown in Figure 2, it divides channels into different dilation groups and uses a dilation-rates regression module to adaptively generate

dilation rates for these groups. Compared with previous multi-scale fusion methods, ADC has three advantages: i) Instead of using multiple independent dilated convolution layers, ADC directly assigns different dilation rates to its channels. In this way, ADC can generate and fuse multi-scale features in a single layer, which is more elegant and efficient. ii) ADC allows fractional dilation rates, which enables ADC to adjust receptive fields with finer granularity, instead of only four fixed integer scales. Thus ADC may be able to exploit richer and finer multi-scale information. iii) The dilation rates in ADC are adaptively generated, which could help ADC to generalize better to various human sizes.

ADC can be easily plugged into existing HPE methods and trained end-to-end by standard back-propagation. Our contributions can be summarized into three points:

1. We attempt to address the spatial non-alignment and inflexibility problems in nowadays multi-scale fusion methods of HPE. These problems are important to the accuracy of location and generalization ability over various human sizes.
2. We propose an adaptive dilated convolution (ADC), which could flexibly fuse well-aligned multi-scale features in a single convolutional layer by adaptively generating dilation rates for different channels.
3. The proposed ADC can be easily plugged into existing HPE methods and extensive experiments show that ADC can bring these methods consistent improvements.

II. Related Works

A. Multi-scale Fusion

Multi-scale fusion is widely adopted in many high-level vision tasks, such as detection [13], [14], [15], semantic segmentation [16], etc. On the one hand, these tasks involve both location and classification. They need multi-scale information to make a balance between these two sub-tasks. On the other hand, these tasks need to tackle objects of various sizes. They scale-invariant representations to get more stable performances. In these tasks, most methods [7], [13], [16] firstly extract a feature pyramid, which contains feature maps of different spatial sizes, and then fuse feature maps to obtain multi-scale information. However, as we have discussed above, the fused features may be not well spatially-aligned. This non-alignment may hurt the accuracy of location. For detection and segmentation, this influence could be ignored, because their evaluation metrics (IOU) are less sensitive to the accuracy of location. While HPE methods are evaluated by OKS, which will be heavily influenced by pixel-level errors. Thus the non-alignment may restrict the performance of HPE methods. In the proposed adaptive dilated convolution, multi-scale features are of the same spatial sizes, which may be more friendly to human pose estimation.

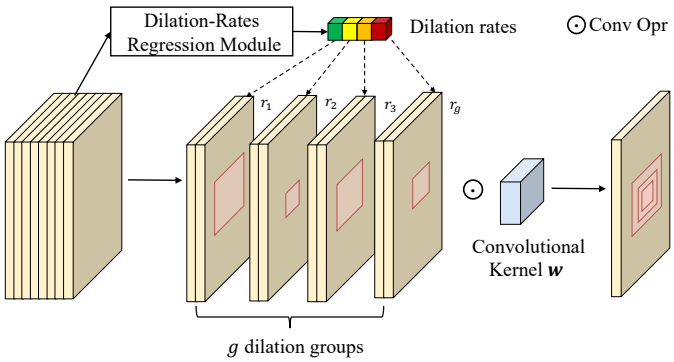


Fig. 2. Details of adaptive dilated convolution. Different dilation groups have different dilation rates, and thus have different receptive fields.

B. Dilated Convolution

The main idea of dilated convolution is to insert zeros between pixels of convolution kernels. It is widely used in segmentation [17], [18] to enlarge the receptive fields while keeping the resolutions of feature maps. As the size of its receptive field can be easily changed by adjusting its dilation rate, dilated convolution is also used to aggregate multi-scale context information. For example, in [11], the outputs of convolutional layers with different dilation rates are fused to exploit multi-scale context information. And in [12], a similar idea is adopted in an atrous spatial pyramid pooling (ASPP) module. However, these dilation rates of different layers are manually set and can only be integers, which are not flexible enough. Instead, the dilation rates in ADC can be fractional and are adaptively generated, which enables it to learn more suitable receptive fields for objects of various sizes. Besides, every dilation group in ADC can represent features at a scale, which enables ADC to fuse richer multi-scale information yet in a simpler way.

III. Adaptive Dilated Convolution

A. Constant Dilation Rates

As shown in Figure 3, original dilated convolution can be decomposed into two steps: 1) sampling according to a index set $\mathcal{I}$ over the input feature map $\mathbf{x}$; 2) matrix multiplication of the sampled values and convolutional kernel $\mathbf{w}$. The index set $\mathcal{I}$ is defined by the dilation rate r and size of kernel $k \times k$:

$$\begin{aligned} &\mathcal{I} = \{(i \cdot r, j \cdot r, c)\}, \\ &s.t. \quad \lfloor -k/2 \rfloor \leq i, j \leq \lfloor k/2 \rfloor, \quad 0 \leq c < C_{in}, \end{aligned} \tag{1}$$

where C_{in} is the number of channels in x, $\lfloor \cdot \rfloor$ denotes rounding down to the nearest integer. Specially, if $k = 3$ and $C_{in} = 1$ then

$$\begin{aligned} \mathcal{I} = &\{(-r, -r, 0), (-r, -r + 1, 0), \ldots, \\ &(r, r - 1, 0), (r, r, 0)\}. \end{aligned} \tag{2}$$

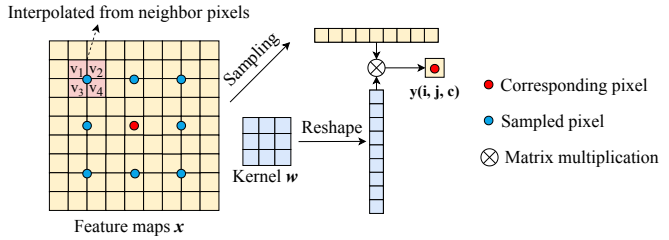


Fig. 3. Original dilated convolution can be decomposed into two steps: sampling and matrix multiplication. But in ADC, the dilation rate could be fractional (2.5 in the figure), in which case, the sampled values will be interpolated from their neighbor pixels.

For value at location (i, j, c) of the output feature map $\mathbf{y}$, we have

$$\mathbf{y}(i,j,c) = \sum_{\Delta\mathbf{p}\in\mathcal{I}} \mathbf{w}^c(\Delta\mathbf{p}) \cdot \mathbf{x}((i,j,0) + \Delta\mathbf{p}), \tag{3}$$

where $\Delta\mathbf{p}$ enumerates the indexes in $\mathcal{I}$, and $\mathbf{w}^c$ denotes the corresponding convolutional kernel for the c^{th} output channel.

The receptive field for each channel in convolutional layer is defined as the square covered by index set $\mathcal{I}$. In original dilated convolution, the receptive fields of all channels are the same. Their sizes are:

$$\begin{aligned} Area =& (\lfloor k/2 \rfloor \cdot r - \lfloor -k/2 \rfloor \cdot r)^2 \\ =& (kr - r + 1)^2. \end{aligned} \tag{4}$$

Specially, when $r = 1$, the size of receptive field is k^2.

B. Adaptive Dilation Rates

In adaptive dilated convolution, the dilation rates are no longer manually set. As shown in Figure 2, we use a dilation-rates regression module (DRM) to adaptively generate the dilation rates for different channels. DRM consists of a global average pooling layer and two fully connected layers with nonlinear activations. Suppose DRM is denoted as a function $\phi(\cdot)$, then generated dilation rate $\mathbf{r}$ is

$$\mathbf{r} = \phi(\mathbf{x}). \tag{5}$$

We divide the input channels into g dilation groups. Each group contains C_{in}/g channels. The channels in the same group shares the same dilation rate. Thus the shape of $\mathbf{r}$ is $g \times 1$. And the dilation rate of the c^{th} input channel is $\mathbf{r}^{\lfloor c/g \rfloor}$. If $g = C_{in}$, then each channel has its own dilation rate. If $g = 1$, then all channels share the same dilation rate.

Consequently, the sampling index set becomes

$$\begin{aligned} &\mathcal{I} = \{(i \cdot \mathbf{r}^{\lfloor c/g \rfloor}, j \cdot \mathbf{r}^{\lfloor c/g \rfloor}, c)\} \\ s.t. \quad &-k/2 \le i, j \le k/2, \quad 0 \le c < C_{in}. \end{aligned} \tag{6}$$

In cases where $\mathbf{r}$ is fractional, as shown in Figure 3, we use bilinear interpolation to get the sampling values. Suppose $M(\mathbf{x}, (i, j, c))$ denotes the interpolated value at (i, j, c) on $\mathbf{x}$, then we have:

$$\mathbf{y}(i,j,c) = \sum_{\Delta\mathbf{p}\in\mathcal{I}} \mathbf{w}^c(\Delta\mathbf{p}) \cdot M(\mathbf{x}, (i,j,0) + \Delta\mathbf{p}). \tag{7}$$

Similarly, in the c^{th} channel of adaptive dilated convolution, the size of receptive field is:

$$\begin{aligned} Area =& (\lfloor k/2 \rfloor \cdot \mathbf{r}^{\lfloor c/g \rfloor} - \lfloor -k/2 \rfloor \cdot \mathbf{r}^{\lfloor c/g \rfloor})^2 \\ =& (k \cdot \mathbf{r}^{\lfloor c/g \rfloor} - \mathbf{r}^{\lfloor c/g \rfloor} + 1)^2. \end{aligned} \tag{8}$$

Consequently, different dilation groups have different sizes of receptive fields. And thus ADC can fuse multi-scale information in a single layer.

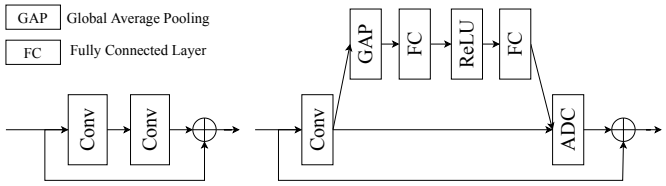
Since all involved operators are numerical differentiable [19], [?], the proposed adaptive dilated convolution can be easily plugged into existing models trained end-to-end by standard back-propagation.

C. Analysis and Discussion

Comparison with Yu et al. In [11], Yu et al. use multiple dilated convolutional layers with different dilation rates to extract features at different scales. ADC adopts a similar idea, but implements it in a simple yet efficient way. Firstly, ADC consists of only one convolutional layer. It does not use independent dilated convolutional layers or extra concatenation. Thus ADC is much more computation-economic and time-saving. Secondly, every dilation group in ADC represents features at a different scale, which enables ADC to exploit richer multi-scale information than [11]. Thirdly, the dilation rates in ADC can be fractional and are adaptively generated, instead of manually set integers. It helps ADC to generalize better to persons of various sizes.

Comparison with deformable convolution. In [19], Dai et al. propose a deformable convolutional layer, which allows the sampling index set $\mathcal{I}$ to be non-grid and irregular. It assigns an offset for each index in $\mathcal{I}$, instead of only modifying the dilation rates. Compared with adaptive dilated convolution, deformable convolution enjoys higher degrees of freedom, but it also has a much higher computational cost. More importantly, the offsets introduced in deformable convolution are completely unconstrained and independent. This may cause the input and output feature maps to lose their spatial correspondence, which also potentially hurts the accuracy of location. We also experimentally proved in Sec IV-A5 that the proposed adaptive dilated convolution is more suitable to human pose estimation than deformable convolution.

Comparison with scale-adaptive convolution. In [20], Zhang et al. propose a scale-adaptive convolution (SAC) to address inconsistent predictions of large objects and invisibility of small objects in scene parsing. SAC adaptively generates pixel-wise dilation rates to acquire flexible-size receptive fields along spatial dimensions. It works well for scene parsing, which needs to tackle objects of



(a) Original residual block (b) Residual block with ADC

Fig. 4. We replace one convolution layer in original residual block (shown in (a)) with ADC (shown in (b)).

various sizes within a single image. However, in single person pose estimation, size inconsistent across different images plays a more important role, which could be better alleviated via multi-scale fusion along channel dimension. In SAC, different pixels can have different sizes of receptive fields, but different channels share the same dilation rates. Consequently, ADC may be more suitable for single person pose estimation than SAC. In Sec IV-A5, we also experimentally prove that ADC works better than SAC in HPE methods.

D. Instantiation

We plug ADC into the backbones of frequently used HPE models, including the family of SimpleBaseline [21] and HRNet [4]. Their backbones are built up with residual blocks [22]. As shown in Figure 4, we replace one ordinary convolution layer in the original residual block by ADC. The weights of the last layer in the dilated-rates regression module are initialized as zeros and its bias are initialized as ones. Thus, the generated dilation rates in ADC are initialized are ones. The dilation groups g are set as $g = C_{in}$, in which case each group contains only one channel. Thus every channel can exploit context information at different scales, and ADC could fuse as much richer multi-scale information as it can. We also experimentally demonstrate that the performance is positively correlated to g in Sec IV-A4.

IV. Experiments

A. Experiments on COCO

Dataset. All of our experiments about human pose estimation are done on COCO dataset [23]. It contains over $200K$ persons and $250K$ persons. Our models are trained on COCO train2017 ($57K$ images), and evaluated on COCO val2017 ($5K$ images) and COCO test-dev ($20K$ images).

Evaluation metric. We use the standard evaluation metric Object Keypoint Similarity (OKS) to evaluate our models. OKS $= \frac{\sum_i \exp(-d_i^2/2s^2k_i^2)\delta(v_i>0)}{\sum_i \delta(v_i>0)}$, where d_i is the Euclidean distance between the detected keypoint and its corresponding ground-truth, v_i is the visibility flag of the ground-truth, s denotes the person scale, and k_i is a per-keypoint constant that controls falloff. We report the standard average precision (AP) and recall, including AP^{50} (AP at OKS=0.5), AP^{75}, AP (mean of AP scores from OKS=0.50 to OKS=0.95 with the increment as 0.05, AP^M (AP scores for person of medium sizes) and AP^L (AP scores for persons of large sizes).

Training. Following the setting of [4], we augment the data by random rotation ($[-30°, 30°]$), random scaling ($[0.7, 1.3]$), random translation ($[-40, 40]$), random horizontal flip and half body transform [24]. Then we crop out each single person according to their ground-truth bounding boxes. These crops are resized to 256×192 (or 384×288) and input to the HPE model.

The models are optimized by Adam [25] optimizer, and the initial learning rate is set as 1×10^{-3}. For the family of HRNet, each model is trained for 210 epochs and the learning rate decays to 1×10^{-4} and 1×10^{-5} at $170th$ and $200th$ epoch respectively. For the family of SimpleBaseline, each model is trained for 140 epochs and the learning rate decays to 1×10^{-4} and 1×10^{-5} at $90th$ and $110th$ epoch respectively. All models are trained and tested on 4 Tesla V100 GPUs. More details can be referred to the Github repository Pose[1].

Testing. During testing, we use the same person detection results provided in [21], which are widely used for many single-person HPE models [4], [3]. Single persons are cropped out according to the detection results and then resized and input to the HPE models. The flip test [4] is also performed in all experiments. Each keypoint location is predicted by adjusting the highest heatvalue location with a quarter offset in the direction from the highest response to the second-highest response [4].

1) Ablation Study: To fully demonstrate the superiority of ADC, we perform ablation studies on different models, including the family of SimpleBaseline [21] and HRNet [4]. The results are shown in Table I. As one can see, ADC can bring consistent improvement for different models. For the smallest model, i.e. SimpleBaseline-Res50, ADC brings an improvement of 1.4 on AP score. For the largest model, i.e. HRNet-W48, there is still an improvement of 0.4 on AP score. The increments decay as the AP scores increase. This may because it is harder to improve the performance of a more accurate model. From AP^M and AP^L, we can see that the improvements in medium and large persons are roughly the same. It indicates that ADC benefits equally the keypoint detection of large and medium persons.

2) Error Analysis: In this section, we use the error analysis tool in [26] to further explore how ADC help HPE models achieve better results. We mainly study four types of errors: 1) jitter: small error around the correct keypoint location; 2) missing: large localization error, the detected keypoint is not within the proximity of any body part; 3) inversion: confusion between semantically similar parts belonging to the same instance. The detection is in the proximity of the true keypoint location of the wrong body

[1]https://github.com/leoxiaobin/deep-high-resolution-net.pytorch.git

TABLE I
Results of different models with or without adaptive convolution. The input sizes are 256×192. Results are reported on COCO val2017.

Method	Backbone	ADC	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}
Simple Baseline [21]	Res50	×	70.4	88.6	78.3	67.1	77.2
		√	71.8	89.2	79.7	68.5	78.7
	Res101	×	71.4	89.3	79.3	68.1	78.1
		√	72.5	89.5	80.4	69.3	79.3
	Res152	×	72.0	89.3	79.8	68.7	78.9
		√	72.8	89.3	80.6	69.5	79.7
HRNet [4]	HRNet-W32	×	74.4	90.5	81.9	70.8	81.0
		√	75.0	90.6	82.0	71.4	81.7
	HRNet-W48	×	75.1	90.6	82.2	71.5	81.8
		√	75.5	90.8	82.3	72.3	82.5

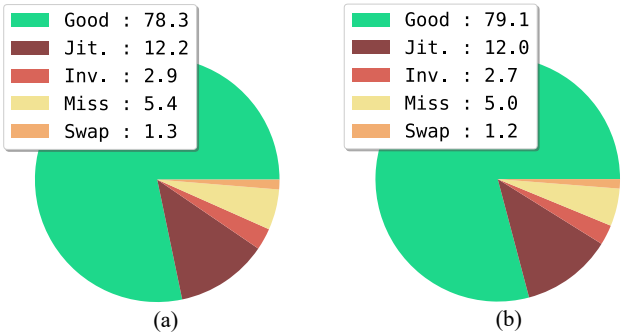


Fig. 5. Error analysis results of SimpleBaseline-Res50 (a) without ADC and (b) with ADC.

part; 4) swap: confusion between semantically similar parts of different instances. The detection is within the proximity of a body part belonging to a different person. We use SimpleBaseline-Res50 as the baseline model, and plot the error analysis results with and without ADC in Figure 5. As one can see, ADC can reduce the proportion of all four types of errors. Especially, the proportion of missing error is reduced by 0.4%. It suggests that ADC could help the model to be more robust and detect keypoints in more cases. This may be attributed to that ADC can adaptively adjust the dilation rates. The jitter error and inversion error directly indicate the accuracy of location and classification respectively. The proportions of these two errors are both reduced by 0.2%. It suggests that ADC can simultaneously benefit the location and classification of keypoints.

3) Statistical Analysis: In this section, we make a statistical analysis to further investigate how the generated dilation rates in ADC are related to the sizes of test persons. We divide the test persons in COCO val2017 into three types according to the areas of their bounding boxes. Persons whose bounding boxes have areas: 1) smaller than 32×32 are divided into the small group (53166 persons); 2) greater than 32×32 but smaller than 96×96 are divided into the medium group (25173 persons); 3) greater than 96×96 are divided into the large group (25787 persons). We still use SimpleBaseline-Res50 with ADC as the studied model. The backbone, i.e. Resnet-50, has four stages, and they contain 3, 4, 6, 3 residual blocks respectively.

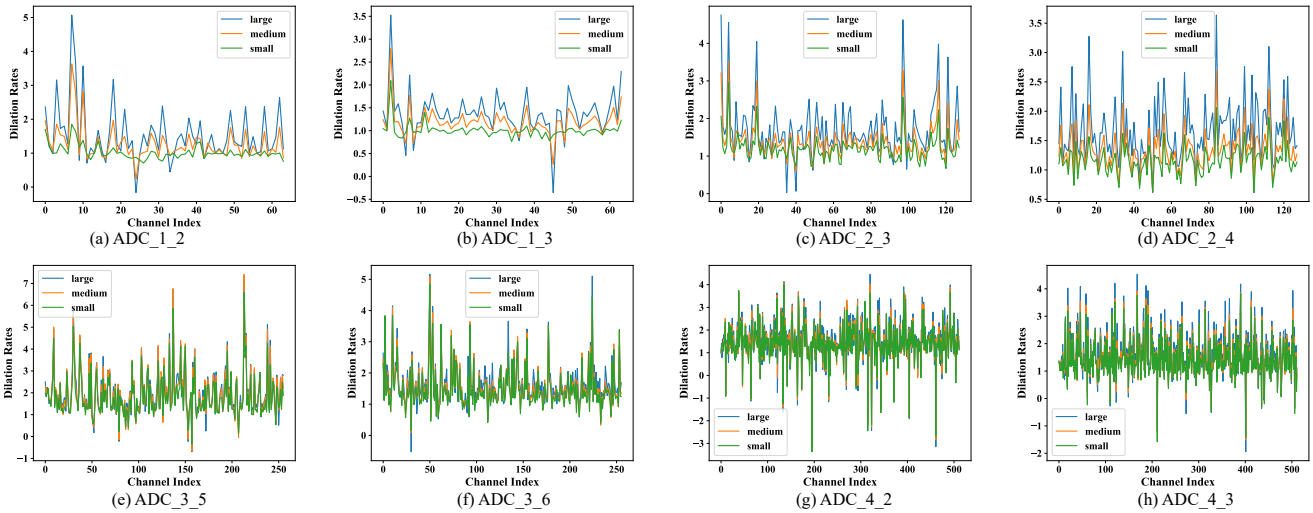
We plot the means and variations of the dilation rates of different channels in Figure 6. For example, Figure 6 (a) shows the mean dilation rates of ADC in the third block of the first stage (256 channels). Figure 6 (especially the top row) suggests that the dilation rates are closely related to the sizes of test persons. The dilation rates for larger persons are more likely to larger. It enables ADC to be more robust over various human sizes. Besides, the dilation rates for the deeper block also tend to have larger means (bottom row). It may because that deeper blocks are more concerned with semantic features and need larger dilation rates to enlarge the receptive fields. Additionally, the mean dilation rates of different channels in the same layer are quite different. The larger variance of these dilation rates also indicates that ADC can fuse rich multi-scale information via different channels.

4) Study of Dilation Groups g: We perform comparative experiments to explore the influence of dilation groups g. We use the SimpleBaseline-Res50 as the baseline. We gradually improve g from 2 to C_{in}. The results are shown in Table II. As one can see, the model performance becomes better when g increases. It suggests that the number of different dilation rates matters, which also indicates the importance of multi-scale information fusion in HPE.

TABLE II
Results of SimpleBaseline-Res50 with different dilation groups g. The input sizes are 256×192. Results are reported on COCO val2017.

Groups	2	4	8	C_{in}
AP	71.5	71.5	71.7	71.8
AP^{50}	89.1	89.3	89.1	89.2
AP^{75}	79.4	79.2	79.5	79.7
AP^{M}	68.2	68.3	68.3	68.5
AP^{L}	78.3	78.3	78.5	78.7

5) Compared with Other Methods: In this section, we experimentally prove that ADC is more suitable to human pose estimation than deformable convolution (DC) [19] and scale-adaptive convolution (SAC) [20]. Comparative experiments are performed on SimpleBaseline-Res50. As shown in Table III, although DC can bring an improvement on the baseline, its performance is inferior to that of ADC. As we have discussed in Sec III-C, the



(a) ADC_1_2 (b) ADC_1_3 (c) ADC_2_3 (d) ADC_2_4

(e) ADC_3_5 (f) ADC_3_6 (g) ADC_4_2 (h) ADC_4_3

Fig. 6. Mean dilation rates in ADC of different blocks in SimpleBaseline-Res50. The subplots are named in the format of ADC_stageID_blockID. These statistical comparisons suggest that the dilation rates for relatively larger persons are also likely to be larger.

unconstrained and independent offsets of DC may cause the input and output feature maps to lose their spatial correspondence, which potentially hurt the accuracy of location. SAC can alleviate the size inconsistent along spatial dimensions, but involves little multi-scale fusion along the channel dimension, which is more important in HPE. Consequently, the improvement of SAC is lower than both DC and ADC.

TABLE III
Results of SimpleBaseline-Res50 with deformable convolution (DC) or adaptive dilated convolution (ADC). The input sizes are 256×192. Results are reported on COCO val2017.

Method	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}
Baseline	70.4	88.6	78.3	67.1	77.2
DC [19]	71.4	89.2	79.3	67.9	78.3
SAC [20]	71.1	89.1	78.7	67.7	78.1
ADC	71.8	89.2	79.7	68.5	78.7

B. Experiments for Semantic Segmentation

Similar to human pose estimation, semantic segmentation also requires rich multi-scale information to make a balance between local and semantic features. Thus the proposed ADC should also benefit the performance of semantic segmentation models. In this section, we plug ADC into different models to demonstrate its benefits on semantic segmentation.

We use CityScapes [27] as our training (2975 images) and validation (500 images) datasets. We use FCN [28], PSANet [29], DeepLabV3 [12] and DeepLabV3+ [17] as our baseline models. The input sizes are set as 769×769. All models are trained for $40K$ iterations. More details can be referred to the Github repository mmsegmention[2]. As shown in Table IV, ADC can bring consistent improvements to different Semantic Scene Parsing models. For FCN, ADC even improves the mIOU by 4.27.

[2]https://github.com/open-mmlab/mmsegmentation.git

TABLE IV
Results (mIOU) of different models on CityScapes validation dataset. The input sizes are 769×769. All results are trained for $40K$ iterations. w/o ADC: without ADC. w/ ADC: with ADC.

Method	Backbone	w/o ADC	w/ ADC
FCN [28]	Res50	71.47	75.74
PSANet [29]	Res50	77.99	78.57
DeepLabV3 [12]	Res50	78.58	78.70
DeepLabV3+ [17]	Res50	78.97	80.11

V. Conclusion

In this paper, we mainly focus on multi-scale fusion methods in human pose estimation. Existing HPE methods usually fuse feature maps of different spatial sizes to exploit multi-scale information. However, the location information is irreversibly destroyed during the downscaling, and thus the upscaled feature maps may be not well spatially-aligned. This non-alignment potentially hurts the accuracy of keypoint location. Besides, scales of these feature maps are fixed and inflexible, which may restrict its generalization over different human sizes. In this paper, we propose an adaptive dilated convolution (ADC), which exploits multi-scale information by fusing channels with different dilation rates. In this way, each channel in ADC can represent features at a scale, and thus ADC can exploit richer multi-scale information from features of the same spatial sizes. More importantly, the dilation rates for different channels in ADC are adaptively generated, which enables ADC to adjust the scales according to the sizes of test persons. As a result, ADC can help HPE fuse better aligned and more generalized multi-scale features. Extensive experiments on both human pose estimation and semantic segmentation prove that ADC can bring consistent improvements to these methods.

References

[1] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in Proceedings of the European Conference on Computer Vision (ECCV). Springer, 2016, pp. 483–499.

[2] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional pose machines," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2016, pp. 4724–4732.

[3] Y. Cai, Z. Wang, Z. Luo, B. Yin, A. Du, H. Wang, X. Zhang, X. Zhou, E. Zhou, and J. Sun, "Learning delicate local representations for multi-person pose estimation," in Proceedings of the European Conference on Computer Vision (ECCV). Springer, 2020, pp. 455–472.

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

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

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

[7] W. Li, Z. Wang, B. Yin, Q. Peng, Y. Du, T. Xiao, G. Yu, H. Lu, Y. Wei, and J. Sun, "Rethinking on multi-stage networks for human pose estimation," arXiv preprint arXiv:1901.00148, 2019.

[8] Z. Luo, Z. Wang, Y. Cai, G. Wang, Y. Huang, L. Wang, E. Zhou, and J. Sun, "Efficient human pose estimation by learning deeply aggregated representations," arXiv preprint arXiv:2012.07033, 2020.

[9] Z. Luo, Z. Wang, Y. Huang, T. Tan, and E. Zhou, "Rethinking the heatmap regression for bottom-up human pose estimation," arXiv preprint arXiv:2012.15175, 2020.

[10] V. Dumoulin and F. Visin, "A guide to convolution arithmetic for deep learning," arXiv preprint arXiv:1603.07285, 2016.

[11] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," arXiv preprint arXiv:1511.07122, 2015.

[12] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," arXiv preprint arXiv:1706.05587, 2017.

[13] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2017, pp. 2117–2125.

[14] T. Kong, F. Sun, C. Tan, H. Liu, and W. Huang, "Deep feature pyramid reconfiguration for object detection," in Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 169–185.

[15] G. Ghiasi, T.-Y. Lin, and Q. V. Le, "Nas-fpn: Learning scalable feature pyramid architecture for object detection," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 7036–7045.

[16] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 8759–8768.

[17] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in Proceedings of the European Conference on Computer Vision (ECCV), 2018.

[18] 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," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 40, no. 4, pp. 834–848, 2017.

[19] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, "Deformable convolutional networks," in Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 764–773.

[20] R. Zhang, S. Tang, Y. Zhang, J. Li, and S. Yan, "Scale-adaptive convolutions for scene parsing," in Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2031–2039.

[21] B. Xiao, H. Wu, and Y. Wei, "Simple baselines for human pose estimation and tracking," in Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 466–481.

[22] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.

[23] 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 Proceedings of the European Conference on Computer Vision (ECCV). Springer, 2014, pp. 740–755.

[24] Z. Wang, W. Li, B. Yin, Q. Peng, T. Xiao, Y. Du, Z. Li, X. Zhang, G. Yu, and J. Sun, "Mscoco keypoints challenge 2018," in Proceedings of the European Conference on Computer Vision Workshops (ECCVW), vol. 5, 2018.

[25] D. P. Kingma and J. Ba, "Adam: a method for stochastic optimization. corr abs/1412.6980 (2014)," 2014.

[26] M. R. Ronchi and P. Perona, "Benchmarking and error diagnosis in multi-instance pose estimation," in Proceedings of the IEEE International Conference on Computer Vision, Oct 2017.

[27] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2016, pp. 3213–3223.

[28] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 4, pp. 640–651, 2017.

[29] H. Zhao, Y. Zhang, S. Liu, J. Shi, C. Change Loy, D. Lin, and J. Jia, "Psanet: Point-wise spatial attention network for scene parsing," in Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 267–283.