Neurocomputing 453 (2021) 853–864

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Adaptive deformable convolutional network

Feng Chen [a], Fei Wu [b,*], Jing Xu [c], Guangwei Gao [b,f], Qi Ge [d], Xiao-Yuan Jing [b,e]

[a] School of Computer Science, Nanjing University of Posts and Telecommunications, Nanjing, China
[b] College of Automation, Nanjing University of Posts and Telecommunications, Nanjing, China
[c] School of Law, Hohai University, Nanjing, China
[d] College of Telecommunications and Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing, China
[e] School of Computer, Wuhan University, Wuhan, China
[f] Institute of Advanced Technology, Nanjing University of Posts and Telecommunications, Nanjing, China

ARTICLE INFO

Article history:
Received 3 January 2020
Revised 15 June 2020
Accepted 22 June 2020
Available online 8 September 2020
Communicated by Wenguan Wang

Keywords:
Deformable convolution
Semantic segmentation
Object detection
Geometric transformation

ABSTRACT

Deformable Convolutional Networks (DCNs) are proposed to solve the inherent limited geometric transformation in CNNs, showing outstanding performance on sophisticated computer vision tasks. Though they can rule out irrelevant image content and focus on region of interest to some degree, the adaptive learning of the deformation is still limited. In this paper, we delve it from the aspects of deformable modules and deformable organizations to extend the scope of deformation ability. Concretely, on the one hand, we reformulate the deformable convolution and RoIpooling by reconsidering spatial-wise attention, channel-wise attention and spatial-channel interdependency, to improve the single convolution's ability to focus on pertinent image contents. On the other hand, an empirical study is conducted on various and general arrangements of deformable convolutions (e.g., connection type) in DCNs. Especially on semantic segmentation, the study yields significant findings for a proper combination of deformable convolutions. To verify the effectiveness and superiority of our proposed deformable modules, we also provide extensive ablation study for them and compare them with other previous versions. With the proposed contribution, our refined Deformable ConvNets achieve state-of-the-art performance on two semantic segmentation benchmarks (PASCAL VOC 2012 and Cityscapes) and an object detection benchmark (MS COCO).

1. Introduction

Due to the object scale, pose, viewpoint, and part deformation, accommodating these geometric variations is still challenging in sophisticated computer vision tasks [1–6]. Different from the manner of depending on large data with sufficient variations or that of using transformation-invariant features and algorithms [1,7,8], Deformable ConvNet (DCNv1) [9] was proposed to learn a 2D spatial offset to let the grid sampling locations swim with respect to the proceeding feature maps. However, since the adaptive learning of the filter deformation is limited, DCNv1 still suffers from the problem of the irrelevant image content. Then, Zhu et al. [10] revised it and proposed Deformable ConvNet v2 (DCNv2). They made a reformulation of deformable convolution that includes channel-wise attention named modulation mechanism, and then they stacked more such deformable convolutions into networks to intensify the control of sampling over a broader range of feature levels [8]. The sampling locations finally surround the distinct object content and the channel weights are activated by modulated mechanism to judge the impact of sampling point. Ideally, if the sampling points locate in irrelevant content, the channel weight would be punished by multiplying a small factor to alleviate its influence. Therefore, both the spatial and channel attention should share the same function to be object-sensitive. However, due to proceeding the same input feature map by separate convolutions without spatial-channel interaction, these two attention modules may be unknown to each other and hard to adapt synchronously. This limitation makes the modulation mechanism a simple learnable feature amplitude that doesn't powerfully constraint the channel weight exactly, so that the spatial attention weights do not correlate well with feature importance measures. Besides, DCNv2 replaces 13 standard convolutions with deformable convolutions in ResNet101 [11]. This choice is suboptimal to enhance the adaptive geometric modeling ability.

Deformable convolution still has potentials to be excavated. In this paper, inspired by the observation that spatial offset and modulated offset share an internal relationship [12,13], we mainly from

* Corresponding author.
E-mail address: wufei_8888@126.com (F. Wu).

https://doi.org/10.1016/j.neucom.2020.06.128
0925-2312/

two aspects, deformable modules and deformable organization, to refine the deformable convolutional network, which is named Adaptive Deformable ConvNet (A-DCN).

For the deformable modules, we propose to leverage an inter-module information to connect spatial and channel attention modules. We first reformulate the deformable convolution and deformable RoIpooling by correlating spatial attention with channel attention. As shown in Fig. 1, in spatial attention, an adaptive dilation factor is introduced to initialize the sample locations and to decompose the displacement of them, aiming to strengthen the process of offset learning. Second, the adaptive dilation could be multiplied in modulated mechanism. Thus, the spatial and channel attention could be interpolated. We use this information as a constraint added into the channel-wise attention (modulation mechanism) to correlate it with the tendency or choice of spatial attention. Considering the refinement and interdependency of both spatial and channel attention, the geometric transformation ability in manipulating spatial support regions is further improved.

From the aspect of deformable organizations, both DCNv1 and DCNv2 term deformable convolution as a powerful counterpart of regular convolution. They replace the plain counterpart in ResNet [11] with deformable convolutions. However, the factors of influencing enhanced geometric modeling ability are not comprehensively exploited in these works. In this paper, we develop deformable convolution into feature aggregation module as an independent congregation based on high-level feature maps. And then we can make an empirical study of much more general and various settings of three kinds of deformable convolutions (i.e., standard, modulated, and our adaptive deformable convolutions). The study yields some significant findings for the property of deformable convolutions.

Existing applications [12,9,10] on object detection are much more than that on semantic segmentation. And some of their findings on these two tasks are different. Therefore, we make a thorough analysis to delve the performance of deformable convolutions on semantic segmentation. Besides, we propose a method named Adaptive Deformable ConvNet (A-DCN) for semantic segmentation [14,15] which achieves state-of-the-art performance on PASCAL VOC 2012 [16] and Cityscapes [17]. To further assess the generalization of deformable convolution, our A-DCN is also evaluated on object detection benchmark, achieving outstanding gains over the original model on COCO [18].

Our contributions can be summarized as follows:

(1) We reformulate the deformable module to strengthen the adaptive transformation ability. The refinement in spatial, channel attention, and spatial-channel interdependency allows the deformation to be more powerful with minimal cost.
(2) An empirical study is conducted on the deformation organization, which concludes the factors of cooperation within deformable convolutions. These experiments will give some hints for the further works using deformable convolution.
(3) More ablation study is implemented on semantic segmentation to analyze the capability and property of previous and our proposed deformable convolutions. These experiments provide extensive understanding of dense prediction.
(4) We propose Adaptive Deformable ConvNet (A-DCN) which could be incorporate into state-of-the-art CNNs [14,15,19,20] of computer vision. And our methods also achieve more superior performance than original architecture on two benchmarks of semantic segmentation (PASCAL VOC 2012 [16] and Cityscapes [17]) and one benchmark of object detection (COCO [18]). Our code now is available on https://github.com/Chenfeng1271/Adaptive-deformable-convolution.

2. Related work

Attention mechanism enables a neural network to focus on relevant content and exclude redundant context. It first showed superior advancement in the field of natural language processing (NLP) [21–23], such as the landmark Transformer attention module [24]. Then the success of attention is adopted into computer vision tasks [25,26] to capture long-range dependencies and contextual information. On semantic segmentation, inspired by non-local module [27], the works of [28–30] model pixel-to-pixel relation to exploit object-wise context to update the representation for

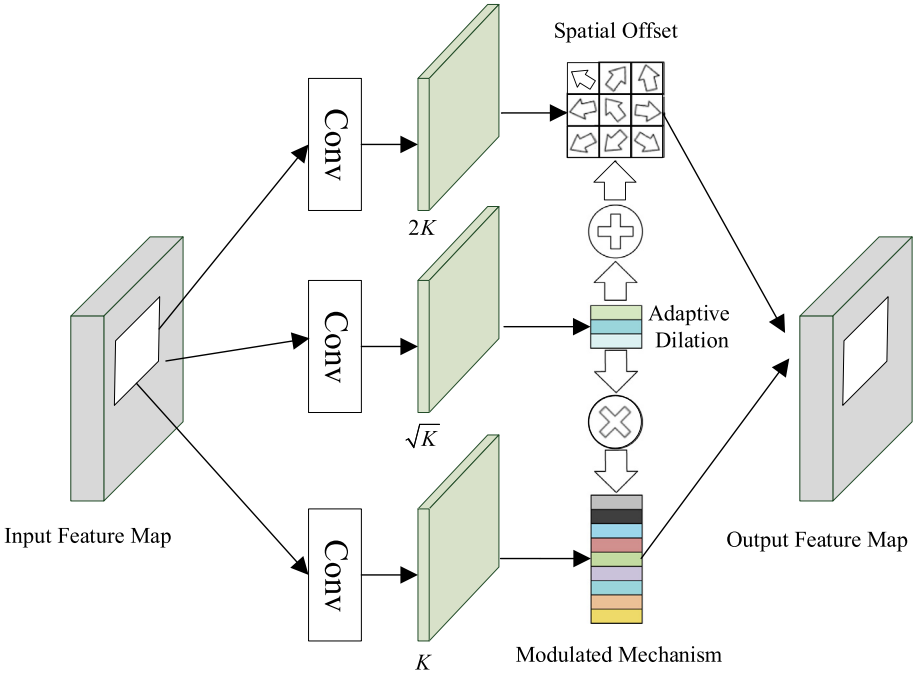

Fig. 1. Scheme of adaptive deformable convolution. K is the kernel size of it (e.g., 3×3).

each pixel. [31] designs CO-attention Siamese Network to address unsupervised video object segmentation task. Inspired by the observation of human attention and primary object judgement, [32] constructs two sub-task models to simulate human dynamic attention mechanism. On object detection, RelationNet [33] introduces object-wise correlation to model the relation between objects. [25] proposes domain attention which is composed by few SEblocks[34]. [35] proposes a hierarchical pyramid attention module to highlight and capture salient object edge. ASNet [36] uses a hierarchy of convLSTMs to sequentially refine the saliency features over multiple steps. Recently, GNNs attract more researchers' attention to reason visual relationship. Working as a more flexible and basic attention network, GNNs could focus on input graph locally and globally, which bring them reasoning ability as meeting partial observation [37]. Besides, recent works focus on the theoretical explanation of attention mechanism. [38] makes an empirical study on spatial attention that decomposes each attention module in terms of key, query and relative position. [39,40] delve the relationship between self-attention and convolutional layers by reformulating these attention modules.

Deformable ConvNet v1 [9] proposes a learnable offset to augment the spatial sampling locations with respect to preceding feature maps. Given a convolutional kernel of K sampling locations, let w_k, p_k and Δp_k refer to the weight, handpicked offset and learnable offset for the k-th location, respectively. The deformation process can be formulated as:

$$y(p) = \sum_{k=1}^{K} w_k \cdot x(p + p_k + \Delta p_k), \tag{1}$$

where $x(p)$ and $y(p)$ denote the feature representations at location p from the input feature maps x and from output feature maps y. Standard deformable convolution only possesses top branch of Fig. 1. In practical learning, an extra convolution is applied to model the offset function, allowing the deformation to condition on the input in a local, dense and learnable manner [41,42].

Additionally, the authors observed that the pixels near the center of receptive field have much larger impact and the effective receptive field shares a Gaussian distribution [9,43,44]. This observation allows them to employ a metric named 'effective dilation' to interpret the internal mechanism of deformable convolution. The effective dilation, which measures the distances between all adjacent pairs of sampling locations in the filter, inspires us to active the predefined offset p_k to facilitate the learning process.

Deformable ConvNet v2 [10] introduces the modulation mechanism into the standard deformable module [9] to strengthen the capability in manipulating spatial support regions. Zhu et.al., [10] reformulated the modulated deformable convolution as:

$$y(p) = \sum_{k=1}^{K} w_k \cdot x(p + p_k + \Delta p_k) \cdot \Delta m_k, \tag{2}$$

where Δm_k is the learnable modulation scalar for the kth location. The modulated deformable convolution can be divided into spatial attention with offset Δp_k and channel-wise attention with modulation Δm_k. Spatial and channel attention refers to top and bottom branches of Fig. 1. Both the offset Δp_k and modulation Δm_k are obtained via a separate convolutional layer applied over the same input feature maps x with $2K$ and K output channels respectively. However, the single separate convolution with only input obstructs the correlation of attention weight between spatial and channel attention. In this paper, we refine the modulated deformable convolution and make a well design of two attention mechanisms and spatial-channel interdependency that interacts the tendency of them. Besides, DCNv2 replaces 10 more plain counterpart than the setting of DCNv1 in the ResNet [11] with deformable convolution, to considerably enhance the model's ability of geometric transformation. This simple replacement focuses on the deformation brought by stacking deformable convolutions in backbone and may ignore the cooperation of deformable convolutions. In Section 5, in terms of three kinds of deformable convolution, we delve the keys of the cooperation that contribute to better performance.

3. Adaptive deformable modules

3.1. Adaptive deformable convolution

To further facilitate the ability of adaptive learning of deformable convolution, we refine it from the aspects of spatial, channel attention and their interdependency. Naturally, the new version is named adaptive deformable convolution (a-dconv). We assume that images share the linearity of feature maps and the pixels at the same distance from receptive field center should possess comparable impact [9,43,26]. It means that the local image content changes gradually. And according to the distance to the activation unit, the attention weight of pixels that deformable convolution focuses on has obvious phases. Thus, we introduce the idea of adaptive dilation factor to model this obvious phase. As illustrated in Fig. 2(a) and (b), the moving of each point in offset could be decomposed. Take the three points in the middle slice as an example, the moving of them is divided into $s_k \cdot p_k$ and Δp. Specifically, $s_k \cdot p_k$ takes relative long step and is same in three points' moving, which could generally describe the distance of three points to the center. The final condition is like Fig. 2(c): the points of each slice are under single phase. The adaptive dilation factor s_k could locally represent the distance of phase to the center, which directly indicates the impact of pixels. Therefore, we name s_k as phase distance. Then, the channel weight m_k could be further activated by this distance information. In general, the channel weight of far sampling point is depressed and that of the close one is aggravated.

With a convolutional kernel size of N which has $K = N^2$ sampling locations (e.g., 9 sampling locations in a 3×3 convolution), we let $w_k, p_k, \Delta p_k = \{\Delta a_{ij}, \Delta b_{ij}\}$ and Δm_k refer to the weight, handpicked offset, learnable offset and modulation scalar for the kth location. $\{a_{ij}, b_{ij}\}$ locates the position of sampling points in sampling grid or spatial dimension. The refined deformable convolution can be formulated as:

$$y(p) = \sum_{k=1}^{K} w_k \cdot x(p + s_k \cdot p_k + \Delta p_k) \cdot ((1 - s_k) \cdot \Delta m_k), \tag{3}$$

where $s_k \in \mathbb{R}^N$ is the adaptive dilation factor that contains the general distance information of sampling locations. $x(p)$ and $y(p)$ denote the feature representations at location $p = \{a'_{ij}, b'_{ij}\}$ from the input feature maps x and that from output feature maps y. The final location in the input $(p + s_k \cdot p_k + \Delta p_k) = \{a_{ij}, b_{ij}\}$ after adding offset can be denoted as:

$$\begin{aligned} a_{ij} &= a'_{ij} + s_{ij} \cdot d_{ij} + \Delta a_{ij}, \\ b_{ij} &= b'_{ij} + s_{ij} \cdot d_{ij} + \Delta b_{ij}, \end{aligned} \tag{4}$$

where $i \in [-N/2, N/2] \cap Z, j \in [-N/2, N/2] \cap Z$ locate the integral coordinate in the kernel grid, and $s_{ij} \in \{s_k\}_{k=1}^{K}$ is an adaptive dilation vector for the position (i,j). d_{ij} is the predefined dilation rate for position (i,j). Since s_k and Δp_k are fractional, $x(p)$ is computed by bilinear interpolation. Following DCNv1 and DCNv2 [9,10], we let $s, \Delta p_k$ and Δm_k come from a separate convolution with $3K + \sqrt{K}$ output channels. The 2 K output channels model the spatial offset $\{\Delta p_k\}_{k=1}^{K}$, and the consecutive K output channels correspond to $\{\Delta m_k\}_{k=1}^{K} \in [0,1]$ which is activated by sigmoid function. $\{s_k\}_{k=1}^{K} \in [0,1]$ that is modeled by the remaining $\sqrt{K}$ channels is a

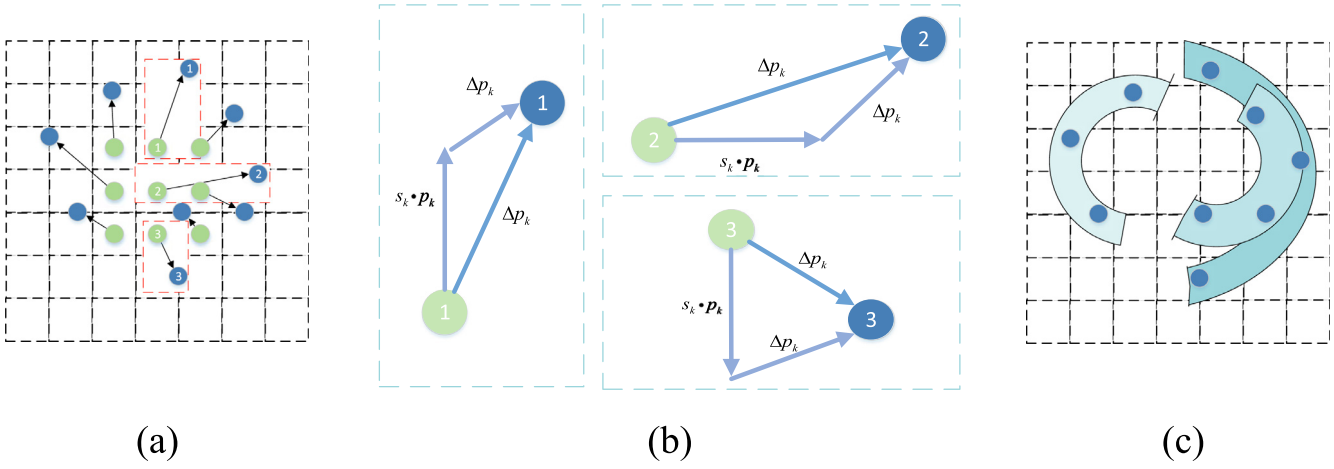

Fig. 2. Illustration of displacement of sampling locations in adaptive deformable convolution. The green points are predefined sampling locations and the blue points are the obtained sampling locations after updating the offset. $s_k \cdot p_k$ intentionally refers to a large step and Δp_k refers to relative small step. (a) is the final moving in offset; (b) is an example of decomposition of displacement of three points of middle slice in (a); and (c) is the illustration of modeling phase distance. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

tensor cooperating with Δp_k at spatial dimension. And the learning rate for these three parts is set to 0.1 times of the corresponding layers.

Standard and modulated deformable convolutions [9,10] apply the task of offset learning to Δp_k. This adaptive dilation factor aims to disentangle the displacement of sampling locations into a large step $s_k \cdot p_k$ and a tiny step Δp_k as shown in Fig. 2. The large step is determined by adaptive dilation factor, interacting the gradients at the locations that share similar distance to the kernel center. This decomposed manner advances the whole training of Deformable ConvNet with negligible computation, because that adaptive dilation factor only takes $\sqrt{K}$ channels. Besides, the adaptive dilation factor can be used in initializing the sampling locations, which is a much more flexible way than traditional dilation. The range of spatial offset is unrestricted while the modulation scalar is within $[0, 1]$. We adjust the s_k with $[0, 1]$ for magnitude correlation. In this paper, we use 3×3 deformable convolution in our Adaptive Deformable ConvNet where $\Delta p_k, s_k$ and Δm_k are defaulted to 0, 1 and 1 respectively.

In modulated deformable convolution [10], the spatial attention (spatial offset) and channel attention (modulation scalar) mechanisms are independent, since both of them are connected in the parallel way. When the sampling locations move, the learning process of modulation may not correspondingly adjust so as to become inexact this setting. In our work, since the pixels near the active units have larger impact, we use adaptive dilation factor which contains the distance information, to interact the spatial and channel attention parts. Therefore, the modulation module is also sensitive to the distance. In the proceeding of feature, spatial and channel parts could cooperatively learn the relevant image content and exclude the irreverent content.

3.2. Adaptive deformable RoIpooling

As an aligned transformer [20,45] which converts an input feature map with arbitrary size into the fixed size one, RoIpooling is widely used in object detection. We also introduce a corresponding adaptive deformable RoIpooling which uses spatial-channel interdependency to enhance the geometric modeling ability.

Given an input RoI, RoIpooling is taken to divide the RoI into K spatial bins. The cells belonging to each bin, are aggregated to compute the bin output. Denote $p_k, \Delta p_k, s_k$ and Δm_k as predefined offset, learnable offset, adaptive dilation factor and modulation scalar for the kth bin, respectively. In the adaptive deformable RoIpooling, the output $y(k)$ of kth bin can be denoted as:

$$y(k) = \sum_{j=1}^{n_k} x(p_{kj} + s_k \cdot p_k + \Delta p_k) \cdot ((1 - s_k) \cdot \Delta m_k)/n_k, \tag{5}$$

where n_k is the number of sampled bin cells of the kth bin, and $x(p)$ is the feature at location p using bilinear interpolation to compute the offset movement. p_{kj} is the sampling location at j-th gird cell of kth bin. The practical usage follows [10] that employs two *fc* layers of 1024-D and an additional *fc* layer with $3K + \sqrt{K}$ channels. As adaptive deformable convolution, the first 2 K channels are normalized learnable offset $\{\Delta p_k\}_{k=1}^{K}$ and K channels are normalized modulation scalar $\{\Delta m_k\}_{k=1}^{K}$ using sigmoid function. The remained $\sqrt{K}$ channels are used to produce adaptive dilation factor $\{s_k\}_{k=1}^{K}$. The learning rates of these additional *fc* layers are the same as those of existing layers.

3.3. Relationship of elements

To facilitate the understanding of adaptive deformable convolution, we reformulate three kinds of deformable convolutions in terms of self-attention mechanisms [39].

First, to exploit the elements (i.e., query, key and relative position between key and query content) of attention module in computer vision tasks, we follow [38,39] to give a generalized attention formulation. Let k represent a key element with content x_k and let q index a query element with content z_q. In multi-head self-attention, the output feature y_q can be formulated as follows:

$$y_q = \sum_{m=1}^{M} W_m \left[\sum_{k \in \Omega_q} A_m(q, k, z_q, x_k) \odot W'_m x_k \right], \tag{6}$$

where M represents the attention head, $A_m(q, k, z_q, x_k)$ is the attention weights in mth attention head. Ω_q is support region to compute the corresponding output query. In the case of convolution, Ω_q is the convolutional window. W_m and W'_m are the learnable weights.

For deformable convolution, the learnable weights are updated based on query content and relative position. Thus, standard

deformable convolution (s-dconv) could be represented as a special instantiation:

$$y_q^{\text{s-dconv}} = \sum_{m=1}^{M} W_m \left[\sum_{k \in \Omega_q} A_m^{\text{s-dconv}}(q,k,x_q) \odot W'_m x_k \right], \tag{7}$$

$$A_m^{\text{s-dconv}}(q,k,x_q) = G(k, q + p_m + w_m^{\top} x_q), \tag{8}$$

where x_q denotes the query content, $G(\cdot)$ denotes the bilinear interpolation kernel. p_m acts equally as predefined offset p_k in Eq. 1. $w_m^{\top}$ is a projection metric that projects query content x_q to deformable offset dimension. $A_m^{\text{s-dconv}}$ is the spatial attention weight based on query content and relative position.

For modulated deformable convolution (m-dconv), it can be reformulated as:

$$y_q^{\text{m-dconv}} = \sum_{m=1}^{M} W_m \left[\sum_{k \in \Omega_q} A_m^{\text{s-dconv}}(q,k,x_q) \cdot C(q,k,x_q) \odot W'_m x_k \right], \tag{9}$$

where $C(q,k,x_q)$ is a kind of channel attention named modulated module, which only uses query content. In the proceeding of feature, the spatial position of each realigned pixel is identical to that of input pixel in modulated module. Therefore, this modulated module works independently with spatial attention and realigns the channel weight of each pixel. In the case of modulated deformable convolution, the position difference between spatial and channel attention may make the cooperation of them inconsistent.

In our adaptive deformable convolution, the output $y_q^{\text{a-dconv}}$ is computed as follows:

$$y_q^{\text{a-dconv}} = \sum_{m=1}^{M} W_m \left[\sum_{k \in \Omega_q} G(k, q + p_m + a_m w_m^{\top} x_q) \cdot C(q,k,x_q,a_m) \odot W'_m x_k \right], \tag{10}$$

where a_m is the adaptive factor that models the relationship between the position of spatial attention and that of channel attention. Therefore, taking the relationships of two attention modules into consideration, adaptive deformable convolution could measure the compatibility of key-query pair more exactly.

4. Adaptive deformable convolutional network (A-DCN)

4.1. A-DCN for semantic segmentation

We use PASCAL VOC [16] and Cityscapes [17] to train our model. PASCAL VOC 2012 obtains 20 semantic categories. Following the protocols in [14,15], we augment the *SBD* dataset [46] into training. Finally, the total dataset has 10582 annotated images for training, 1449 annotated images for evaluation and 1456 annotated images for testing. For Cityscapes, it is a roadway dataset containing 19 semantic categories. And following [28], train set has 2975 images and validation set has 500 images and test set has 1525 images.

We train our model with no bells and whistles. The code of adaptive deformable convolution is revised based on official code of DCNv1 [9] and mm-detection [47]. In all models of semantic segmentation, e.g., DeepLabv2 [48], DeepLabV3 [14] and DenseASPP [49], regular convolutions are replaced with deformable convolutions. We train our model uses SGD with mini-batch 8 on 8 GPUs. The images are resized to shorter side of 360/780 for PASCAL VOC/Cityscapes, respectively. The total epoches are set to 50/240 for these two datasets using random crop. We use the ploy learning rate policy where the initial learning rate 4×10^{-3} is multiplied by $1 - \left(\frac{iter}{max\_iter}\right)^{power}$ with power 0.9. The momentum is set to 0.9 and the weight decay is 10^{-4}.

For evaluation of ablation study, the results are implemented on validation set with single scale input. Besides, to examine the generalization of our adaptive deformable convolution, the final results of different models in Section 6 are based on validation set with multi-scale input. We use mean interaction-over-union (mIoU) over image pixels as our metric of semantic segmentation. Following DCNv1 and DCNv2, we also use mIoU@V and mIoU@C for PASCAL VOC and Cityscapes respectively to denote the results of two datasets.

4.2. A-DCN for object detection

We use COCO [18] to train our models. Following the protocol of [20,19] of using MS COCO 2017, we use 118 K trainval set for training and 5 K images of validation set for evaluation.

We incorporate the adaptive deformable modules into Faster R-CNN [20], Mask R-CNN [45] and Cascade Mask R-CNN [50]. For these models, 256 RoIs are sampled for the regional proposal. The scale of input is resized to 800 pixels in shorter side. The implementations of these networks are 35 K and 240 K iterations for PASCAL VOC and COCO on 8 GPUs respectively. The learning rates are set to 0.02 with momentum 0.9 and weight decay 0.0001. For other hyper-parameters, we employ the default setting of these networks in mm-detection [47]. Besides, we use ResNet-101-FPN pretrained on ImageNet as the backbone of the models without feature mimicking. For evaluation, the average precision (AP) is used for our metric.

5. Well organization for deformable ConvNet

In this section, we discuss the suitable organizations and settings of deformable convolutions to maximize the geometric transformation modeling ability. Due to that previous works have explored much more detailed implementation of deformable modules on object detection than that on semantic segmentation, in this work, we pay more attention on the ablation study on semantic segmentation. The feature aggregation between encoder and decoder has more flexible and deliberated variants. In these variants, all factors that impact on the performance of deformable convolution could be included. Therefore, we replace the plain counterpart with deformable convolution in feature aggregations (e.g., ASPP [15]) which follows the backbone, on semantic segmentation.

Connection Type: We arrange three kinds of deformable convolutions in the way of stacked residual connection, parallel connection and dense connection separately. These connection types construct the basic CNNs. Considering the intuition that deformable convolution works better on high-level feature maps, we place the part containing deformable convolutions after the backbone, acting as feature aggregation for ablation. In each model, ResNet101 is employed as backbone and DeepLabv3's decoder is used for upsampling. As shown in Table 1, different versions of deformable convolutions almost make a gain over using regular convolution via various connection types, except in deformable DenseASPP [49]. In three-kind connection types, the residual connection achieves the most obvious promotion than others, while the performance of it may not be the best one. In dense connection type (i.e., DenseASPP and deformable DenseASPP), different from that standard and modulated deformable convolutions lead a performance degradation, our adaptive deformable convolution achieves an improvement. We believe that multiple skip connections cause the standard and modulated weight learning to be vulnerable. This degradation is more severe as using the modulated

Table 1
The effect of using different connection types of deformable convolutions. s-dconv, m-dconv and a-dconv represent the standard, modulated and adaptive deformable convolutions respectively. ASPP and DenseASPP are the feature aggregations of DeepLabv3+ and DenseASPP respectively. The stacked residual connection type is the same as residual block of ResNet [11].

Method	Connection	Convolution	mIoU@V	mIoU@C
Stacked residual convolutions	Stacked residual	Regular	76.94	74.89
Deformable stacked convolutions	Stacked residual	s-dconv	77.39	75.24
		m-dconv	77.83	75.60
		a-dconv	**78.06**	**76.11**
ASPP	parallel	regular	78.08	75.32
deformable ASPP	parallel	s-dconv	78.23	75.83
		m-dconv	78.22	75.91
		a-dconv	**78.70**	**76.43**
DenseASPP	dense	regular	77.35	76.85
deformable DenseASPP	dDense	s-dconv	75.65	76.55
		m-donv	75.42	76.49
		a-dconv	**77.82**	**77.09**

deformable convolution. The unrelated position between spatial and channel attention in modulated deformable convolution aggravates the vulnerable condition.

Basically, the learnable offset is a relative position between key and query in term of self-attention. And the supporting key is restricted by a local window centered at query location. For the parallel connection (i.e., deformable ASPP [15]), it could be regarded as a multi-head attention module (other connection types have more series connections), including long-range dependency and horizontally-symmetric inter-dependencies and leading to flexibly adjust the deformation learning.

More Deformable Convolutions: The most direct way of enhancing spatial adaptive transformation is adding more deformable convolutions [10]. Ablation study of where deformable convolution should be placed could be divided to backbone and feature aggregation. For the deformable convolution in backbone, we follow [9] that applys more deformable convolutions and replaces their counterparts in Res5 block of ResNet-101 serially. As shown in Table 2, continuously increasing deformable convolutions indeed allows Deformable ConvNet to achieve better performances when DCNs use 1 to 3 deformable convolutions. In FCN [51], using standard, modulated, and our adaptive deformable convolutions respectively achieve 71.86%, 72.36%, 72.54% mIoU as the best performance. However, when most replacement happens at low-level stages without guide, such as feature mimicking, the gain brought by adding more deformable convolutions may not become obvious or even fall down. When using 6 deformable convolutions in DeepLabv2 [48], the results of using three deformable convolutions separately decrease 0.44%, 0.09%, 0.15%. This phenomenon is different from the observation that using more deformable convolutions gives better results in object detection [9].

Table 2
More deformable convolutions in the last 1, 2, 3, and 6 convolutional layers (of 3×3 filter) in backbone of ResNet101. s-dconv, m-dconv and a-dconv denote standard, modulated, adaptive deformable convolutions respectively. Results (mIoU, %) are reported on PASCAL VOC 2012 validation set.

Usage of deformable convolution (#layers)	Version	FCN	DeepLabv2
res5c(1)	s-dconv	69.57	73.56
res5b,c(2)		70.32	74.51
res5a,b,c(3,default)		**71.86**	**74.92**
res5&res4b22,b21,b20(6)		71.58	74.48
res5c(1)	m-dconv	69.78	73.80
res5b,c(2)		70.65	74.95
res5a,b,c(3,default)		**72.36**	**75.19**
res5&res4b22,b21,b20(6)		72.30	75.10
res5c(1)	a-dconv	69.70	73.96
res5b,c(2)		70.87	75.21
res5a,b,c(3,default)		**72.54**	**75.36**
res5&res4b22,b21,b20(6)		72.50	75.21

For adding more deformable convolutions in feature aggregation, we use different stages/branches in deformable DenseASPP and deformable ASPP where each 3 × 3 regular convolution is replaced by deformable convolution. As shown in Fig. 3, three kinds of deformable convolutions improve final results in deformable ASPP module. Compared with the setting using regular convolutions, that of standard or modulated deformable convolutions has about 0.5% promotion and that of adaptive deformable convolution has further about 0.5% improvement. However, this improvement doesn't repeat in deformable DenseASPP module. For standard and modulated deformable convolutions, the curves

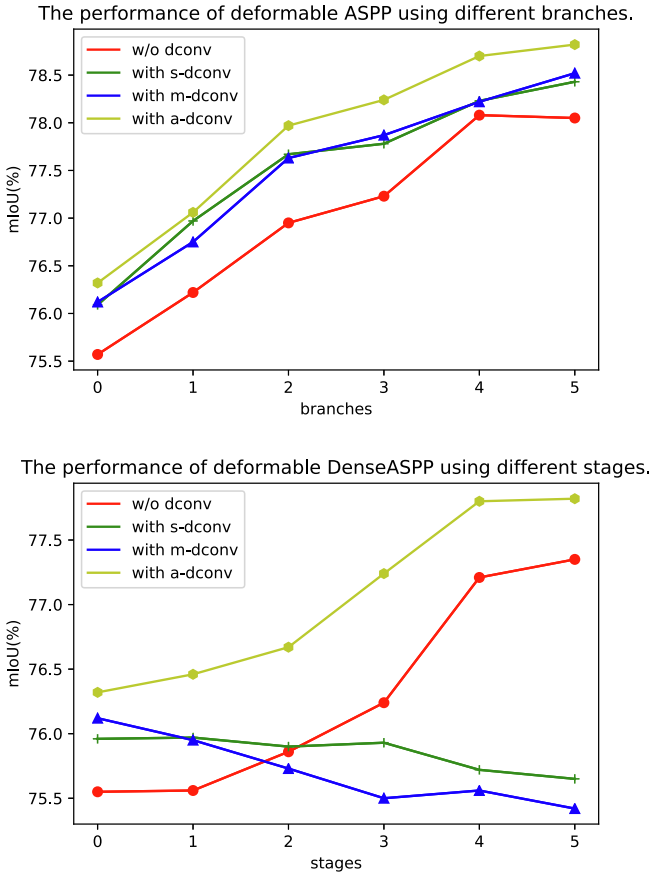

Fig. 3. Adding more deformable convolutions in the branches or stages of feature aggregation. Results are reported on PASCAL VOC 2012 validation set using ResNet101. s-dconv, m-dconv and a-dconv denote standard, modulated, adaptive deformable convolutions respectively.

fall down as adding more convolutions. In the same condition, only adaptive deformable and regular convolutions could increase their results. This phenomenon also verifies the superiority of our proposed method.

Image Resolution and Output Stride: DCNv1 and DCNv2 [9,10] delve the influence of image resolution to the performance of Deformable ConvNet on object detection. In this work, we find that Deformable ConvNets are sensitive to image resolution and output stride of backbone on semantic segmentation. Fig. 4 indicates the results of applying regular ConvNets and three kinds of Deformable ConvNets with different image resolutions and evaluation output strides (eval OS). We use image resolutions of $\{200, 300, 400, 500, 600, 700\}$ for PASCAL VOC 2012 val set in the case of evaluation output stride of 8 and 16. When the eval OS is 8, the curves have similar tendency. The curves of regular convolution and three kinds of deformable convolutions all achieve their summits as image resolution is 500. However, in the case of eval OS 16, the curves become smoother from 400 to 700 than that of eval OS 8. In addition, in two cases of eval OS, the models using deformable convolutions outperform that using regular convolution, which advocates the effectiveness of enhancing geometric transformation by deformable convolution. Compared with other curves, the curve of our adaptive deformable convolution achieves the best performance in various image resolutions.

Higher-level Input: We use ResNet50, ResNet101, DRN [52] as backbone of Deformable ConvNet where deformable convolution is inserted in feature aggregation. Table 3 shows that deformable convolution does depend on high-level input. The average margin between using ResNet50 and ResNet101 is about 2.5%, while the average gap between using DRN and ResNet101 is just about 0.4%. Though deformable convolution could work better depended on high-level feature input, the performance of deformable convolution may not further improve too much by too sophisticated backbone.

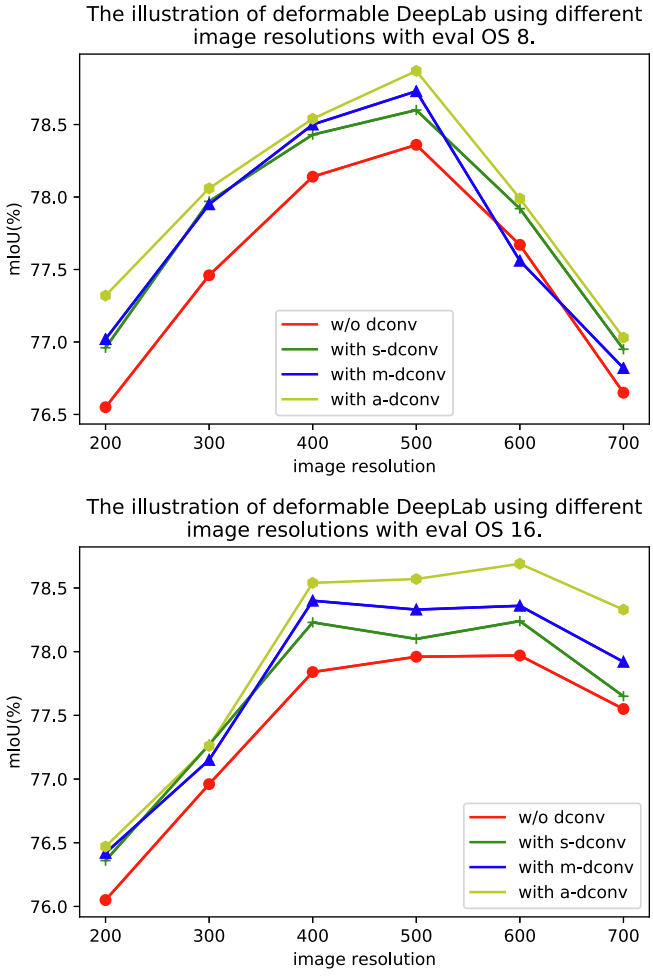

Fig. 4. Ablation study on various image resolutions. For each kind of deformable convolutions, three convolutions are added in ResNet101. The results are reported on PASCAL VOC 2012 val set.

6. Experiments

In this section, we insert our adaptive deformable modules into existing state-of-the-art models [20,45,50,48,51], in the fields of semantic segmentation and object detection.

6.1. Adaptive deformable ConvNet for semantic segmentation

To verify the effectiveness of our method, we add our deformable convolution into existing methods, such as DeepLabv2 [48], DenseASPP [49] and DeepLabv3 [14]. We use adaptive deformable convolution in backbone and feature aggregation. For details, in the fifth stage of ResNet, each 3×3 regular convolution in residual blocks, which are denoted as res5a, res5b and res5c sequentially, is replaced by a 3×3 deformable convolution. In the feature aggregation part, we keep the basic structures of ASPP, DenseASPP, etc., and then replace all 3×3 regular convolutions of them with 3×3 deformable convolutions. The difference of designing semantic networks of using our adaptive deformable convolution with that using previous works can be summarized as threefold: First, considering that deformable convolutions work better on high-level feature map, we choose to insert deformable convolution into feature aggregation, instead of placing it in the low-level stage of backbone. Second, the sensitivity of deformable convolution with different connection types is examined. Different from only using residual connection in previous works, leveraging effective connection to reinforce the geometric modeling ability may be more optimal, especially in flexible feature aggregation, which also indicates the general ability of convolutional function. Third, we focus on balancing the additional cost and reinforced deformable modeling ability. Comparing with standard and modulated deformable convolutions, our adaptive convolution effectively improves its convolution-wise geometric ability by inter-module connection with trivial extra cost. The positions of deformable convolutions at high level stage also enable each convolution to proceed low-resolution feature map, costing relative less memory. As illustrated in Table 4, this organization could further enhance the deformable modeling ability. In all models with ResNet50 or ResNet101, using adaptive deformable convolutions could gain more 1.5% improvement than the same original models. Besides, we compare our methods with some strong baseline models, i.e., DANet [28], EncNet [53] and Dilated FCN [54]. As shown in Table 5, the performance of our refined methods from classical models, e.g., DenseASPP, are comparable to that of current powerful baseline models.

We visualize the effective sampling locations in Fig. 5. The points of adaptive deformable convolution surround the activation unit (green point) and model of that concentrates more on the relevant content. The visual results of deformable ASPP on PASCAL VOC 2012 and Cityscapes are shown in Fig. 6 and Fig. 7, which indicates the outstanding qualitative performance of the proposed adaptive deformable convolution.

6.2. Adaptive deformable ConvNet for object detection

We incorporate our adaptive deformable modules and previous deformable modules into existing object detection models which

Table 3
The performance (mIoU, %) of using higher level feature as input. All backbones use regular convolutions. Experiments are conducted on PASCAL VOC 2012 validation set.

Backbone	Deformable stack residual			Deformable ASPP			Deformable DenseASPP		
	s-dconv	m-dconv	a-dconv	s-dconv	m-dconv	a-dconv	s-dconv	m-dconv	a-dconv
ResNet50	74.56	74.97	75.68	75.26	75.37	76.08	72.87	72.45	75.04
ResNet101	77.39	77.83	78.06	78.23	78.22	78.70	75.65	75.42	77.82
DRN-D-105 [52]	77.80	78.05	78.43	78.43	78.39	78.92	76.02	75.90	78.12

Table 4
The results (mIoU, %) of Adaptive Deformable ConvNets for semantic segmentation on PASCAL VOC 2012 validation set. 'dcn' indicates the deformable convolution, and 'a-dcn' indicates our adaptive deformable convolution.

Method	Backbone	w/o dcn	With a-dcn
DeepLabv2	ResNet50	72.37	74.65
DeepLabv2	ResNet101	74.30	76.21
DenseASPP	ResNet50	75.66	77.02
enseASPP	ResNet101	77.35	78.96
DeepLabv3	ResNet101	78.08	79.04

Table 5
Performance comparison of our methods with strong baselines on PASCAL VOC 2012 validation set.

Baseline	mIoU (%)
Dilated FCN	77.3
DANet (PAM + CAM)	79.0
EncNet (Encoding + SE-loss)	78.4
Deformable DenseASPP (ours)	79.0
Deformable DeepLabv3 (ours)	79.1

have achieved state-of-the-art performance. Each 3×3 regular convolution in the fifth stage of ResNet is replaced as the setting in semantic segmentation, and the RoIplooing is replaced by deformable RoIpooling. Because of additional cost of adding more convolutions in backbone and strict structure of the detection networks, we follow the same designing strategy as previous works [9,10]. As shown in Table 6–8 in Appendix, the enhanced adaptive deformable modeling is examined. First, we provide baseline for Faster R-CNN [20], Mask R-CNN [45] and Cascade Mask R-CNN [50]. The results of the baseline model using setting of aligned RoIpooling [9] are 34.7% AP score for Faster R-CNN. The same setting achieves AP^{bbox}/AP^{mask} scores of 40.4%/35.3% for Mask R-CNN and AP^{bbox}/AP^{mask} scores of 42.3%/36.3% for Cascade Mask R-CNN. When more regular convolutions are replaced by deformable counterparts, i.e., dconv@c3-c5, all these three models obtain 2% to 3% improvement. Besides, we make a comparison between original, modulated and adaptive deformable convolutions. In the same setting of replacing three counterparts from c3 to c5, the networks using our adaptive deformable convolution perform better than those using original and modulated deformable convolutions. In addition, the experiment also indicates the usefulness of adaptive deformable RoIpooling that we propose. The setting of three adaptive deformable convolution and RoIpooling leads to the highest score in three networks. The AP score of Faster R-CNN [20] is 41.7% which is about 7% higher than the baseline model. This improvement represents in other two networks with 7.4% AP^{bbox} improvement and 5.7% AP^{mask} improvement on Mask R-CNN, and with 6.2% AP^{bbox} improvement and 5.6% AP^{mask} improvement on Cascade Mask R-CNN. Besides, in the second row of Fig. 5, effective sampling points of our method could focus more on target, which indicates its superiority.

6.3. Computational efficiency

We further assess the computational efficiency of adaptive deformable convolution. First, we analyze the spatial property, implementation cost and complexity of regular convolution, dynamic convolution and three kinds of deformable convolutions, as shown in Table 6. Compared with the implementation cost of standard and modulated deformable convolutions, adaptive deformable convolution only adds trivial cost which could be ignored. And our adaptive deformable convolution owns spare and global spatial property whose complexity is same as regular convolution, s-dcn and m-dcn.

Fig. 5. The visualization of effective sampling locations. Each triplet of images denotes the spatial support of the method using standard, modulated or adaptive modules respectively, with using deformable convolution@conv3-conv5. The images in second row are produced by the methods which use corresponding deformable RoIpooling.

Fig. 6. Qualitative results of deformable DeepLabv3 on PASCAL VOC 2012. Three rows denote the original image, prediction of deformable DeepLabv3 and ground truth respectively.

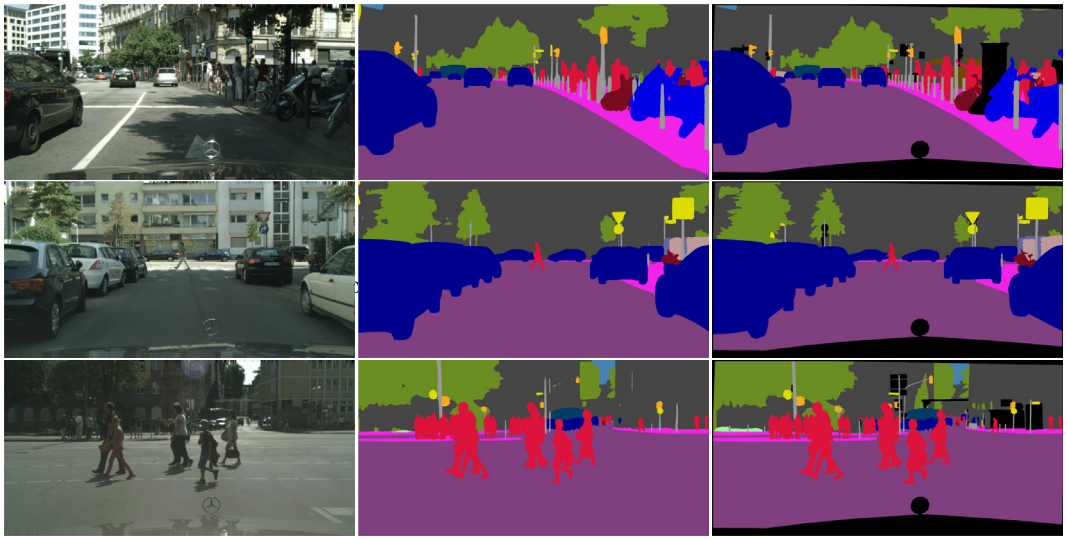

Fig. 7. Qualitative results of deformable DeepLabv3 on Cityscapes.

Furthermore, Table 7 shows the comparison of FLOPs and memory cost of A-DCN and previous works. The experiments are made on semantic segmentation and object detection tasks. Dense ASPP and DeepLabv2 use the same setting as Table 4. Faster RCNN and Mask RCNN use the same setting (adconv@(c3-c5 + adpool)) as Table 8. It is obvious that A-DCN would improve the original performance with trivial extra cost.

7. Conclusion

In this work, we propose new deformable convolutional networks which achieve obvious improvement comparing to their original networks. A new adaptive deformable convolution and RoIpooling further enhance the geometric modeling ability by connecting the spatial positions between spatial attention and channel attention. Besides, we delve the factors that influence the enriched

Table 6
The computational efficiency of regular convolution, dynamic convolution and three kinds of deformable convolutions. N_s is the number of spatial elements, i.e., width by height for image; C is the channel representation dimension; N_k is the kernel size; N_g denotes the number of feature groups in dynamic attention; N_m is the kernel size of separate convolution in deformable convolution.

Convolution	Spatial property	Implantation	Complexity
Regular		convolution	Spare, local
$N_k C^2$		$O(N_s^2 C^2 N_k)$	Dynamic convolution
Spare, local	–	$O(N_s C N_g N_k + N_s C^2)$	
s-dcn	Spare, global	$N_k C^2 + 2N_k N_m C$	$O(N_s^2 C^2 N_k)$
m-dcn	Spare, global	$N_k C^2 + 3N_k N_m C$	$O(N_s^2 C^2 N_k)$
a-dcn	Spare, global	$N_k C^2 + 3N_k N_m C + \sqrt{N_k} N_m C$	$O(N_s^2 C^2 N_k)$

Table 7
Comparison of inference time and memory of standard, modulated, adaptive deformable convolutions.

Model	Inference time (fps)			Memory(GB)		
	s-dcn	m-dcn	a-dcn	s-dcn	m-dcn	a-dcn
Dense ASPP	5.4	5.0	4.9	4.7	4.9	4.9
DeepLabv2	8.4	7.8	7.6	4.2	4.3	4.3
Faster RCNN	9.6	9.3	9.2	3.9	3.7	4.0
Mask RCNN	7.0	6.9	6.6	4.5	4.5	4.6

Table 8
Ablation study (AP, %) on enriched deformable modeling. 'dconv', 'mdconv' and 'adconv' represent standard, modulated, adaptive deformable convolutions. 'dpool', 'mdpool' and 'adpool' denote standard, modulated, adaptive deformable RoIpooling. Besides, '@(c3-c5)' stands for that the positions of deformable convolutions are at stages conv3 to conv5.

Method	Setting	Faster R-CNN	Mask R-CNN		Cascade Mask R-CNN	
		AP	AP^{bbox}	AP^{mask}	AP^{bbox}	AP^{mask}
Baseline	Regular (RoIpooling)	32.0	-	-	-	-
	Regular (aligned RoIpooling)	34.7	36.6	32.2	38.4	32.4
	dconv@(c5) + dpool(DCNv1)	38.0	40.4	35.3	42.3	36.3
Enahanced deformation	dcnv@(c5)	37.2	39.9	35.0	42.2	36.3
	dconv@(c4-c5)	39.8	41.5	35.7	43.1	36.8
	dconv@(c3-c5)	40.0	41.7	36.4	44.0	37.6
	dconv@(c3-c5)+dpool	40.6	42.0	36.5	44.3	**38.1**
	mdconv@(c3-c5)	40.3	42.6	36.9	44.2	37.0
	mdconv@(c3-c5)+mdpool(DCNv2)	41.0	43.1	37.0	44.5	38.0
	adconv@(c3-c5)	40.8	43.2	37.3	44.4	37.8
	adconv@(c3-c5)+adpool	**41.7**	**44.0**	**37.9**	**44.6**	38.0

deformation on semantic segmentation, including connection type, adding more deformable convolutions, image resolution and higher-level input. All these experiments covering three kinds of deformable convolutions give a comprehensive understanding for designing Deformable ConvNets. Following these understanding, we propose Adaptive Deformable ConvNets for semantic segmentation that improve the original performance with a large margin. In addition, to verify the effectiveness of our deformable modules, we insert them into existing models, e.g., Faster R-CNN, and these models also outperform the original models.

For future work, we notice that even with a learnable convolution, the optimization of the larger step in offset decomposition is under the indirect restriction of the whole offset and modulation. We believe a clearer and straightforward constraint of modeling inter-module correlation for this larger step is necessary, to ensure the step to consist with the true distance of a group of points to the center.

CRediT authorship contribution statement

Feng Chen: Writing - original draft, Data curation, Software, Resources. **Fei Wu:** Conceptualization, Writing - review & editing, Supervision. **Jing Xu:** Visualization, Investigation, Data curation. **Guangwei Gao:** Methodology, Software, Validation. **Qi Ge:** Writing - review & editing, Formal analysis. **Xiao-Yuan Jing:** Methodology, Supervision.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

The work described in this paper was fully supported by the National Natural Science Foundation of China (Nos. 61702280, 62076139, 61972212, 61972213), Natural Science Foundation of Jiangsu Province (Nos. BK20170900, BK20190089), National Postdoctoral Program for Innovative Talents (No. BX20180146), China Postdoctoral Science Foundation (No. 2019M661901), Jiangsu Planned Projects for Postdoctoral Research Funds (No. 2019K024), CCF-Tencent Open Fund WeBank Special Funding (No. CCF-WebankRAGR20190104), and Scientific Research Starting Foundation for Introduced Talents in NJUPT (NUPTSF, No. NY217009).

References

[1] D.G. Lowe, Distinctive image features from scale-invariant keypoints, Int. J. Comput. Vis. 60 (2) (2004) 91–110.
[2] M. Jaderberg, K. Simonyan, A. Zisserman, et al., Spatial transformer networks, in: Advances in Neural Information Processing Systems, 2015, pp. 2017–2025.
[3] Y. Jeon, J. Kim, Active convolution: learning the shape of convolution for image classification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4201–4209.
[4] J. Gu, H. Hu, L. Wang, Y. Wei, J. Dai, Learning region features for object detection, in: Proceedings of the European Conference on Computer Vision, 2018, pp. 381–395.
[5] L. Sifre, S. Mallat, Rotation, scaling and deformation invariant scattering for texture discrimination, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 1233–1240.
[6] R. Girshick, F. Iandola, T. Darrell, J. Malik, Deformable part models are convolutional neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 437–446.
[7] R. Girshick, J. Donahue, T. Darrell, J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 580–587.
[8] R. Zhang, Making convolutional networks shift-invariant again, arXiv preprint arXiv:1904.11486.
[9] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, Y. Wei, Deformable convolutional networks, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 764–773.
[10] X. Zhu, H. Hu, S. Lin, J. Dai, Deformable convnets v2: More deformable, better results, arXiv preprint arXiv:1811.11168.
[11] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[12] Y. Xiong, M. Ren, R. Liao, K. Wong, R. Urtasun, Deformable filter convolution for point cloud reasoning, arXiv preprint arXiv:1907.13079.
[13] W. Luo, Y. Li, R. Urtasun, R. Zemel, Understanding the effective receptive field in deep convolutional neural networks, in: Advances in Neural Information Processing Systems, 2016, pp. 4898–4906.

[14] L.-C. Chen, G. Papandreou, F. Schroff, H. Adam, Rethinking atrous convolution for semantic image segmentation, arXiv preprint arXiv:1706.05587.
[15] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, H. Adam, Encoder-decoder with atrous separable convolution for semantic image segmentation, in: Proceedings of the European Conference on Computer Vision, 2018, pp. 801–818.
[16] M. Everingham, S.A. Eslami, L. Van Gool, C.K. Williams, J. Winn, A. Zisserman, The pascal visual object classes challenge: a retrospective, Int. J. Comput. Vis. 111 (1) (2015) 98–136.
[17] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, B. Schiele, The cityscapes dataset for semantic urban scene understanding, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 3213–3223.
[18] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C.L. Zitnick, Microsoft coco: Common objects in context, in: European Conference on Computer Vision, 2014, pp. 740–755.
[19] J. Dai, Y. Li, K. He, J. Sun, R-fcn: object detection via region-based fully convolutional networks, in: Advances in Neural Information Processing Systems, 2016, pp. 379–387.
[20] S. Ren, K. He, R. Girshick, J. Sun, Faster r-cnn: towards real-time object detection with region proposal networks, in: Advances in Neural Information Processing Systems, 2015, pp. 91–99.
[21] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, E. Hovy, Hierarchical attention networks for document classification, in: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2016, pp. 1480–1489.
[22] A. M. Rush, S. Chopra, J. Weston, A neural attention model for abstractive sentence summarization, arXiv preprint arXiv:1509.00685.
[23] M.-T. Luong, H. Pham, C. D. Manning, Effective approaches to attention-based neural machine translation, arXiv preprint arXiv:1508.04025.
[24] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in Neural Information Processing Systems, 2017, pp. 5998–6008.
[25] X. Wang, Z. Cai, D. Gao, N. Vasconcelos, Towards universal object detection by domain attention, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 7289–7298.
[26] F. Wu, F. Chen, X.-Y. Jing, C.-H. Hu, Q. Ge, Y. Ji, Dynamic attention network for semantic segmentation, Neurocomputing 384 (2020) 182–191.
[27] X. Wang, R. Girshick, A. Gupta, K. He, Non-local neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7794–7803.
[28] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, H. Lu, Dual attention network for scene segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 3146–3154.
[29] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, W. Liu, Ccnet: criss-cross attention for semantic segmentation, arXiv preprint arXiv:1811.11721.
[30] H. Zhang, H. Zhang, C. Wang, J. Xie, Co-occurrent features in semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 548–557.
[31] X. Lu, W. Wang, C. Ma, J. Shen, L. Shao, F. Porikli, See more, know more: unsupervised video object segmentation with co-attention siamese networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 3623–3632.
[32] W. Wang, H. Song, S. Zhao, J. Shen, S. Zhao, S.C. Hoi, H. Ling, Learning unsupervised video object segmentation through visual attention, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 3064–3074.
[33] F. Sung, Y. Yang, L. Zhang, T. Xiang, P.H. Torr, T.M. Hospedales, Learning to compare: relation network for few-shot learning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 1199–1208.
[34] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7132–7141.
[35] W. Wang, S. Zhao, J. Shen, S.C. Hoi, A. Borji, Salient object detection with pyramid attention and salient edges, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 1448–1457.
[36] W. Wang, J. Shen, X. Dong, A. Borji, R. Yang, Inferring salient objects from human fixations, IEEE Trans. Pattern Anal. Mach. Intell.
[37] Z. Zheng, W. Wang, S. Qi, S.-C. Zhu, Reasoning visual dialogs with structural and partial observations, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 6669–6678.
[38] X. Zhu, D. Cheng, Z. Zhang, S. Lin, J. Dai, An empirical study of spatial attention mechanisms in deep networks, arXiv preprint arXiv:1904.05873.
[39] J.-B. Cordonnier, A. Loukas, M. Jaggi, On the relationship between self-attention and convolutional layers, arXiv preprint arXiv:1911.03584.
[40] S. Jain, B.C. Wallace, Attention is not explanation, arXiv preprint arXiv:1902.10186.
[41] K.-N.C. Mac, D. Joshi, R.A. Yeh, J. Xiong, R.R. Feris, M.N. Do, Locally-consistent deformable convolution networks for fine-grained action detection, arXiv preprint arXiv:1811.08815.
[42] J. Zhu, L. Fang, P. Ghamisi, Deformable convolutional neural networks for hyperspectral image classification, IEEE Geosci. Remote Sens. Lett. 15 (8) (2018) 1254–1258.
[43] M.P. Heinrich, O. Oktay, N. Bouteldja, Obelisk-net: fewer layers to solve 3d multi-organ segmentation with sparse deformable convolutions, Med. Image Anal. 54 (2019) 1–9.
[44] J.J. Koenderink, A.J. van Doorn, Representation of local geometry in the visual system, Biol. Cybern. 55 (6) (1987) 367–375.
[45] K. He, G. Gkioxari, P. Dollár, R. Girshick, Mask r-cnn, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2961–2969.
[46] B. Hariharan, P. Arbelaez, L. Bourdev, S. Maji, J. Malik, Semantic contours from inverse detectors, in: International Conference on Computer Vision, 2011.
[47] K. Chen, J. Wang, J. Pang, Y. Cao, Y. Xiong, X. Li, S. Sun, W. Feng, Z. Liu, J. Xu, Z. Zhang, D. Cheng, C. Zhu, T. Cheng, Q. Zhao, B. Li, X. Lu, R. Zhu, Y. Wu, J. Dai, J. Wang, J. Shi, W. Ouyang, C. C. Loy, D. Lin, MMDetection: open mmlab detection toolbox and benchmark, arXiv preprint arXiv:1906.07155.
[48] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A.L. Yuille, Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs, IEEE Trans. Pattern Anal. Mach. Intell. 40 (4) (2018) 834–848.
[49] M. Yang, K. Yu, C. Zhang, Z. Li, K. Yang, Denseaspp for semantic segmentation in street scenes, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 3684–3692.
[50] Z. Cai, N. Vasconcelos, Cascade r-cnn: delving into high quality object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 6154–6162.
[51] J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3431–3440.
[52] F. Yu, V. Koltun, T. Funkhouser, Dilated residual networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 472–480.
[53] H. Zhang, K. Dana, J. Shi, Z. Zhang, X. Wang, A. Tyagi, A. Agrawal, Context encoding for semantic segmentation, in: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, 2018, pp. 7151–7160.
[54] S. Gong, Z. Wang, T. Sun, Y. Zhang, C. D. Smith, L. Xu, J. Liu, Dilated fcn: Listening longer to hear better, in: 2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 2019, pp. 254–258.

Feng Chen is pursuing the Master degree in computer technology from Nanjing University of Posts and Telecommunications, Nanjing, China. His research interests include computer vision and image processing.

Fei Wu received the Ph.D. degree in computer science from Nanjing University of Posts and Telecommunications (NJUPT), China, in 2016. He is currently with the College of Automation in NJUPT. He has authored over forty scientific papers, such as TPAMI, TIP, TCYB, PR, TSE, TR, CVPR, AAAI, IJCAI and WWW. His research interests include pattern recognition, artificial intelligence, and computer vision.

Jing Xu is an undergraduate of Hohai University, Nanjing, China. Her major is criminal law and her research interests include fairness of artificial intelligence and computer vision.

Guangwei Gao received the Ph.D. degree in pattern recognition and intelligence systems from Nanjing University of Science and Technology, Nanjing, China, in 2014. Now, he is an associate professor with the Institute of Advanced Technology, Nanjing University of Posts and Telecommunications, Nanjing, China. His research mainly focuses on pattern recognition and computer vision.

Qi Ge received the Ph.D. degree in pattern recognition and intelligent systems from Nanjing University of Science and Technology in 2013. She is currently an Assistant Professor with the College of Telecommunications and Information Engineering, Nanjing University of Posts and Telecommunications. Her research interests include image segmentation, machine learning, and image restoration.

Xiao-Yuan Jing received the Doctoral degree of Pattern Recognition and Intelligent System in the Nanjing University of Science and Technology, 1998. Now he is a Professor with the School of Computer, Wuhan University, and with the College of Automation, Nanjing University of Posts and Telecommunications, China. His research interests include artificial intelligence and pattern recognition.