Research on Optimization Strategy of Adaptive Spatial-Depth Convolutional Neural Network Based on Computer Vision Tasks

Jiabao Sean XIAO

Shenzhen Middle School

Abstract: With the continuous development of technology, long-term effective data accumulation has gradually enabled models to overcome the dependency on data caused by overfitting problems. However, the vast amount of data makes it impossible for models to fully acquire effective features of data in limited time, resulting in significant differences in performance between training and testing sets. Atrous convolution enhances the feature extraction ability of models by expanding the receptive field of convolution kernels. However, incorrect atrous values can lead to a decrease in convolution efficiency. To address this issue, this paper proposes an adaptive atrous convolutional neural network based on online inference strategy, which enables convolution kernels to adjust their atrous values based on different contents at the pixel level of images. Experiments show that the proposed adaptive atrous convolutional neural network can effectively improve the feature extraction ability of models and can be flexibly embedded into various convolutional neural networks and applied to various computer vision tasks.

Keywords: atrous convolution; convolutional neural network; computer vision

1. Introduction

With the continuous development of deep learning technology, deep learning related technologies have been integrated into various applications in life and have achieved good results. Convolutional neural networks, as a very important network technology in deep learning, were initially proposed to solve image classification problems. Convolutional networks not only improved the classification of whole images, but also made progress in structured output tasks, such as video surveillance and object detection. Convolution kernels are a key component of convolutional neural networks and have become an indispensable part of convolutional neural network design. Their ability depends on the performance of layer-wise representation of spatial features on the input region, which is usually called the receptive field. The lower-level convolution kernels are relatively close to the input image, so they are more sensitive to the local receptive field of the input image. Many convolutional neural networks can achieve good performance in various visual tasks by continuously stacking convolution kernels, i.e., increasing the depth of convolutional neural networks, to enable higher-level convolution kernels to obtain larger receptive fields and obtain complete features of images.

Simply stacking convolution layers can lead to many additional problems, such as gradient disappearance and feature sparsity, which can reduce the efficiency of convolution and prevent effective improvement of model performance. Therefore, researchers have begun to study dynamic adjustment of the computational method of convolution kernels to enable networks to adaptively adjust the receptive field of convolution kernels based on input data. This research direction focuses on constructing dynamic internal structures through data-driven methods to better utilize spatial variations from input.

Some techniques tend to develop differentiable approximations for traditional image-adaptive filters and integrate them into convolutional neural networks for end-to-end training. For example, some studies have introduced bilateral filters as independent components into convolutional neural networks, simplified parameters, and derived corresponding gradient descent algorithms, enabling filter parameters to be learned from data. These methods to some extent realize dynamic adjustment of convolution methods based on input sample features. However, most of them rely on adding learnable hidden modules to achieve adaptive adjustment of convolution kernels, which cannot be extended to more general convolutional neural networks.

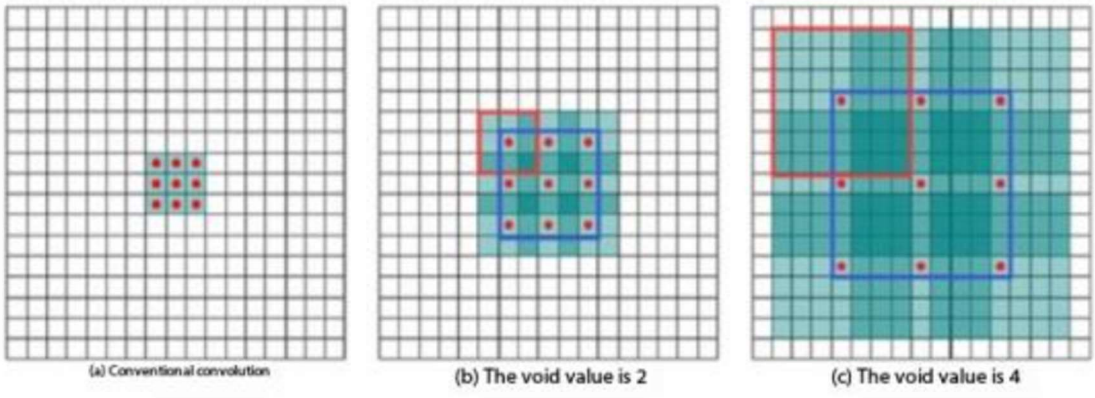


Figure 1-1 Dilated Convolutional Kernel

The introduction of Dilated Convolutional Networks (DCN) provided a new approach to enlarging the receptive field of convolutional kernels. DCN introduces a new parameter called dilation value, which defines the distance between points on the convolutional kernel when processing feature maps. The schematic diagram of the dilated convolutional kernel is shown in Figure 1-1, where (a) represents the conventional convolution, (b) represents the dilated convolution with a dilation value of 2, and (c) represents the dilated convolution with a dilation value of 4. It is a convolutional approach proposed for image segmentation problems, which allows for the use of dilated convolution to enlarge the receptive field of the kernel without downsampling or losing information. With the addition of dilation, the 3x3 convolutional kernel can have a receptive field of 5x5 or larger with the same number of parameters and computations. Therefore, stacking dilated convolutions can effectively increase the receptive field of high-level convolutions. However, the expanded convolutional kernel is constrained

by its shared dilation, which makes it unable to perceive different spatial contents at different positions. In addition, there is no clear theoretical basis for setting the dilation value, and selecting the wrong value can cause the receptive field of the high-level convolutional kernel to exceed the actual range of the image, resulting in a decrease in the learning efficiency of dilated convolution.

To improve the convolutional efficiency of dilated convolutional neural networks, researchers have proposed various solutions, among which the most representative method is deformable convolutional networks. Deformable convolution adds additional offset values to increase the spatial sampling positions in the convolutional kernel and updates these offset values end-to-end from the data. However, due to its powerful flexibility, deformable convolutional kernels often require a fixed value, such as 1, as the upper limit for the offset values in the kernel, which means that it usually needs to be stacked with deformable convolutional layers to enlarge the receptive field to achieve better performance. However, when choosing a larger value as the upper limit for the offset values to achieve a larger receptive field, deformable convolution cannot guarantee that its effective receptive field will also increase, because increasing the upper limit of the offset values will increase its flexibility accordingly. Therefore, if the dataset is incomplete, deformable convolutional networks will focus on some local details, which means that this method is strongly dependent on experimental data, and it is only possible to avoid excessive attention to incorrect details when the experimental data is sufficient.

To further improve dilated convolutional neural networks, it is necessary to correctly address the two obvious issues that exist in most existing dilated convolutional neural network structures: fixed receptive field size and manually selected dilation value range. First, the dilation values of convolutional layers are shared among all pixels, which means that each output position has the same size receptive field. However, this contradicts the law of human observation of the world, because the size of the region of interest (ROI) usually varies greatly at different positions, so the size of the receptive field should be adjusted accordingly to encode different spatial information. Based on the above analysis, a single size receptive field is difficult to capture the diversity within and between huge samples, especially for large-scale, high-resolution image datasets. Secondly, the selected dilation value in current mainstream methods is feature-independent. For each dilated convolutional layer, researchers need to specify the dilation value before integrating it into the convolutional neural network. This usually requires strong domain knowledge about the input and output context information. Moreover, for many specific tasks, there is no clear guidance available in practice for selecting the appropriate dilation value.

This article addresses the aforementioned challenges by combining the selection of dilation values with traditional convolutional modules, merging them into a unified data-driven framework. The proposed Adaptive Dilated Convolutional Neural Network (ADCNN) is a simple yet powerful extension of the generic dilated convolution kernel, treating dilation values as learnable variables in the deep model and allowing joint optimization with other convolutional weights in an end-to-end manner. Figure 1-2 shows a comparison between different convolution kernels, where (a) represents the

conventional convolution kernel, (b) represents the conventional dilated convolution kernel, and (c) represents the schematic diagram of the adaptive dilated convolutional kernel, with different colors indicating different dilation values. In the proposed ADCNN kernel, learned dilation values change at different input positions to reflect the diversity of input spatial features, resulting in dynamic receptive fields with irregular shapes in a single layer. In the model design process, there are two main problems that need to be addressed:

(1) How to determine dilation values online

This article treats dilation values as the output results of a function with a single pixel as input. Specifically, this function samples dilation values using specific probability distributions conditional on input pixel features. To address the non-differentiability of the general sampling process, this article introduces Gumbel-Softmax as a differentiable estimator for the sampling process to ensure the end-to-end trainability of ADCNN convolutional kernel.

(2) How to initialize dilation values

For ADCNN convolutional kernel, due to the clear hierarchical structure of neural networks, this article assumes that the connection pattern between dilation values and convolutional layers is related to inter-layer connections. Therefore, the size of the receptive field at each position in the convolutional kernel is dynamically adjusted based on the information flow between corresponding inter-layer pixels during forward propagation.

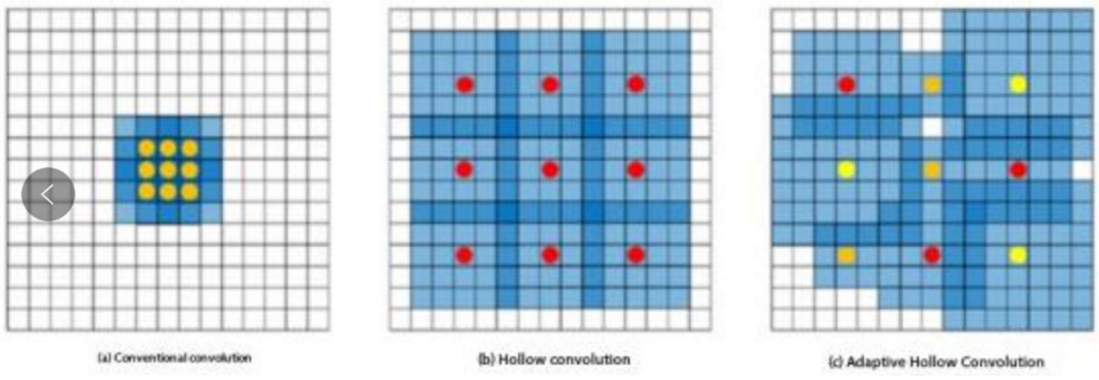


Figure 1-2 Comparison of different convolution kernels

2. Adaptive Dilated Convolutional Neural Network

By converting dilation values into the output of a function with pixels as input and implementing the function through data-driven sampling processes, this article embeds the generation process of dilation values into the deep model, which can be updated synchronously with the model update process through the introduction of hidden modules. In addition, based on the assumption that the connection pattern between dilation values and convolutional layers is related to inter-layer

connections, this article designs various information aggregation strategies to update weight information in hidden modules.

2.1 Problem Definition

Without loss of generality, this article assumes that all convolutions are two-dimensional operations. Taking the $(l-1)$th convolution layer as an example, its input is x^{l-1}, where $x^{l-1} \in R^{w^{l-1} \times h^{l-1}} w^{l-1} h^{l-1}$, representing the width and height $k_{w;d}$ of the input data (or features); $k_{w;d}$ represents the dilated convolution kernel, where d is the dilation value and w is the weight of the convolutional kernel. The relationship between the output of the convolutional layer and the input and weight can be expressed in the form of Equation (1-1).

Where s represents the size of the convolutional kernel, and$i$，j correspond to the position indices of w and h. From Equation (1-1), it can be seen that d is a constant variable independent of i，j. The purpose of the proposed method is to transform d into a function $D_{i,\ j}$ so that the output of $D_{i,\ j}$ can be sensitive to the content specific to the input position. That is, the Adaptive Dilated Convolutional Neural Network regards $D_{i,\ j}$ as an inference process, which generates dilation values by sampling from position-related hidden distributions. The basic idea of ADCNN convolutional kernel is shown in Figure 2-3.

$$y_{i,\ j}^{l} = \sum_{m=0}^{s} \sum_{n=0}^{s} w_{m,\ n} \times x_{i+dm,\ j+dn}^{l-1} \qquad (2\text{-}1)$$

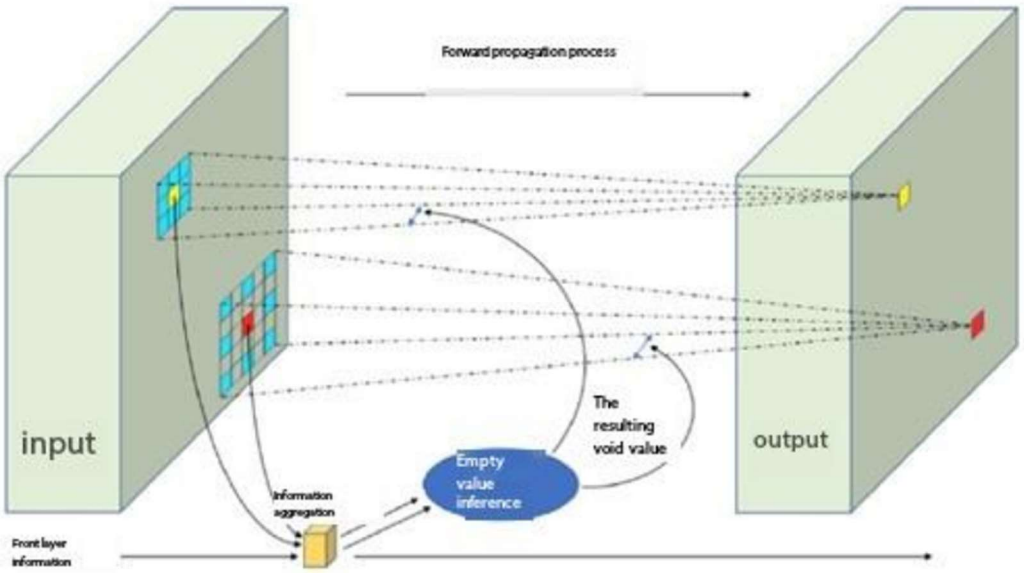


Figure 2-3 Basic idea of ADCNN convolution kernel

2.2 Online Inference Strategy for Dilation Values

Directly sampling dilation values from a categorical distribution can serve as a basic solution for generating dilation values, however, based on this dilation value generation method, gradients cannot be backpropagated through the sampling node, which affects the entire training process. Based on the approximate treatment method for the sampling process in literature^, this article uses Gumbel-Softmax as an approximate inference for discrete dilation values of $D_{i,\ j}$. Assuming that there are E effective options for dilation values and $d_{i,\ j} \in [0,\ 1]^E$ is the estimated one-hot vector for dilation values at position (i，j) the sampling process$d_{i,\ j} \sim GS\left(\hbar_{i,\ j}\right)$can be implemented using Equation (2-2).

$$d_{i,\ j} = D_{i,\ j}(\hbar) = \frac{exp\left(\left(\hbar_{i,\ j} + g_{i,\ j}\right)/\tau\right)}{\sum exp\left(\left(\hbar_{i,\ j} + g_{i,\ j}\right)/\tau\right)} \quad (2\text{-}2)$$

Where $\sum$ represents the sum of all tensor elements here, h, $h_{i_{\circ}\ j}$ is the hidden information related to content and its sub tensor $g_{i,\ j}$ is independently identically distributed and sampled from Gumbel (0,1) τ Control the degree to which GS approaches the true classification distribution.

2.3 Hidden Unit Information Aggregation Patterns

Dilation value adaptation should be controlled by the feature hierarchy structure, so this article establishes a dilation value reasoning mechanism based on inter-layer modeling to capture dependencies between abstraction levels. This article models the aggregation of inter-layer information dependence and generates hidden information h by sequentially aggregating multiple y from different layers. Suppose l represents the index of the newly added layer, and there are three aggregation options for inter-layer modeling:

(1) Recurrent Aggregation

The most basic sequential aggregation method can be expressed as shown in Equation (2-3).

$$\hbar_{i,\ j}^{l} = f\left(W_n^l \hbar_{i,\ j}^{l-1} + u_h^l y_{i,j}^{l-1}\right) \quad (2\text{-}3)$$

Where $w_{\hbar}^{l}$ and $u_{\hbar}^{l}$ are 1×1 convolution kernels, and their output channel number is E; f(.) is a non-linear activation function. Based on the recurrent aggregation mode, as the convolutional layer l deepens, $h_{i,\ j}^{l}$ continuously accumulates information from each layer, and this aggregation mode indicates that the size of the receptive field depends heavily on the connection between convolutional layers.

(2) Gate Aggregation

To more intelligently simulate inter-layer connection patterns, this article introduces a gate variable a_h^l which aggregates information from each layer in a data-driven manner to dynamically adjust inter-layer information. The entire aggregation process can be represented by Equations (2-4) and (2-5).

$$h_{i,j}^{l} = f\left(a_h^l \circ \left(W_h^l h_{i,j}^{l-1}\right) + \left(1 - a_h^l\right) \circ u_h^l y_{i,y}^{l-1}\right) \quad (2\text{-}4)$$

$$a_h^l = \sigma\left(W_a^l h_{i,h}^{l-1} + u_a^l y_{i,j}^{l-1}\right) \quad (2\text{-}5)$$

Where σ (·) is the Sigmoid activation function, $\circ$ represents element-wise multiplication. Based on the gate aggregation mode, the hidden layer does not strictly depend on the order of convolutional layers and will affect the dilation value sampling in a more complex way.

(3) Markov Aggregation

This aggregation mode is an important extreme case of recurrent aggregation mode. The Markov aggregation mode sets the convolution kernel weight w_h^l in Equation (2-3) to 0, and its specific form is shown in Equation (2-6).

$$h_{i,j}^{l} = f\left(u_h^l y_{i,j}^{l-1}\right) \quad (3\text{-}6)$$

In the Markov aggregation mode, the hidden information only depends on the input features of the current convolutional layer, and the calculation of the hidden layer information using this aggregation mode means that the size of the receptive field represented by the hidden layer information is mainly determined by the current convolutional layer and input information, without guidance from other inter-layer information.

2.4 Qualitative Analysis of the Method

The proposed adaptive dilated convolutional neural network generates dilation values through online inference strategy and applies them to the model feature extraction process, achieving the adaptive adjustment of its parameters during the training process, so as to dynamically adjust the receptive field of the convolutional kernel. To further understand the advantages of the method, this article compares and analyzes the adaptive dilated convolution with other methods of the same type from three aspects: adaptive ability of the convolution kernel, weight parameter dimensions, and spatial orthogonality.

(1) Adaptive Dilation Value vs. Adaptive Convolution Kernel

In recent years, researchers have proposed various improvement schemes for dynamically learning convolutional kernels based on different input contents[113,115]. These methods dynamically change the shape of the convolution kernel by directly manipulating the spatial properties of the convolution kernel. One typical method is Modulated Deformable Convolution, which can be expressed as shown in Equation (2-7):

$$y(p) = \sum_{k=1}^{k} w_k \,.\, x(p + p_k + \Delta p_k).\, \Delta m_k \qquad (2\text{-}7)$$

Where x represents the image sample or feature map of the input, y represents the feature map output at position p, Δpk and Δm_k are learnable offset and modulation scalars at the k-th position, respectively. This method changes the shape of the convolution kernel dynamically by using offsets and learning specific values of these offsets from the target task. Generally, adaptive convolution kernel methods tend to learn a functional relationship that makes the convolution kernel engraved with this function, that is, $w_{m,n} = F_{m,n}(x)$, where m and n respectively represent the index positions within the convolutional kernel. Compared with adaptive convolution kernel methods, ADCNN achieves convolutional kernel adaptation in a more indirect way. The dilation function D sets the target as the space of dilation values rather than the space of convolutional kernels, which includes all E possible dilation values. Therefore, the proposed adaptive dilated convolutional neural network is a special case of convolutional kernel adaptation. This method adapts to dilation values, and there are directional constraints on the modification of convolutional kernels by dilation values. This makes the learned convolutional kernels have shape constraints obtained through dilation value adaptation, while compared with adaptive convolution kernel methods, the flexibility of adaptive dilated convolution decreases, which can benefit model performance in certain situations. For example, when training data is insufficient, the convolutional neural network cannot obtain sufficient training, and the adaptive convolutional kernel method tends to focus excessively on some specific details features in the data, making the learned convolutional kernels unable to capture complete and effective features of the target object. In this case, it is necessary to impose constraints on the learning of convolutional kernels under the premise of maintaining their flexibility. Therefore, the proposed method has the potential to solve such problems.

(2) Low-dimensional Complexity vs. High-dimensional Complexity

The dimension of the expansion space of the method proposed in this paper is equal to the number of E candidate dilation values, while the convolutional kernel needs to maintain a dimension mapping relationship of $C^{L-1} \times C^{l}$on space so that it can be consistent with the channel size of the input and output. In addition, since this type of method can exponentially expand the receptive field, it is not necessary to set too many candidate dilation values in the candidate set. At the same time, the size of the channel number usually increases sharply as the convolutional neural network deepens to capture more complex and advanced abstract features. These make E significantly smaller than $C^{L-1} \times C^{l}$, which can produce a more easily learned process while not worrying about feature sparsity.

In addition, low-dimensional complexity also allows the method proposed in this paper to be deployed at a more extensive hierarchical range. Therefore, from the perspective of dimensional characteristics, the method proposed in this paper has a wider range of application scenarios.

(3) Spatial Sharing of Dilation Values vs. Spatial Orthogonality of Convolutional Kernels

Adaptive convolution kernel methods use a single function independent of the convolutional layer to generate convolutional kernels, so the generated convolutional kernels may be highly correlated with each other. Recent research has shown that a regularized space with orthogonality constraints will produce better model performance and more stable training processes. However, balancing the generation process of convolutional kernels and their spatial orthogonality is somewhat difficult. Unlike adaptive convolution kernel methods, ADCNN primarily relies on the space of dilation values, which is not only separated from the space of individual convolutional kernels but can also be shared by all convolutional layers in the convolutional neural network. This means that inter-layer aggregation patterns are easier to perform on multiple convolutional layers and can more coherently propagate information to deeper convolutional layers through the shared space of dilation values. Therefore, compared with adaptive convolution kernel methods, ADCNN can learn features of different inputs without interfering with the spatial orthogonality between convolutional kernel spaces.

3. Ablation Experiment and Analysis

The proposed adaptive dilated convolutional neural network achieves dynamically adjusting the receptive field size during the model training process, which can effectively test its effectiveness for dense prediction tasks based on computer vision. Therefore, this paper analyzes the characteristics of ADCNN in the application process through image segmentation experiments.

3.1 Experimental Settings

(1) Dataset Setting: This paper uses the PascalVOC2012 and Cityscapes datasets for image segmentation experiments, and uses the average intersection over union (mIoU) on the validation set as the evaluation metric for model performance.

(2) Parameter Setting

This paper implements various convolutional neural networks embedded with ADCNN through PyTorch deep learning. In the ablation experiment, VGG-16 is used as the backbone network for parameter initialization and model training, and Adam optimizer guides the updating of model parameters. In addition, all dilation value candidate sets of ADCNN convolutional kernels contain three candidate dilation values: {1, 2, 4} (E=3), and the default mode for information aggregation is the Markov aggregation mode.

3.2 Analysis of Convolutional Layer Adaptivity

In order to fully leverage the effectiveness of ADCNN convolutional kernels, it is necessary to clarify the relationship between the position of the convolutional kernel in the deep convolutional neural network and the model performance. This paper studies the basic backbone network VGG-16. VGG-16 consists of five convolutional layer groups, namely conv1, conv2, conv3, conv4, and conv5. Among them, the modules of conv3, conv4, and conv5 convolutional layer groups are similar and composed of three convolutional layers with different channel numbers. Therefore, the feature extraction process of these three layers is relatively similar. In this paper, conventional convolution and ADCNN convolutional kernels are replaced in the last three convolutional layer groups to study the impact of receptive fields on performance under different situations.

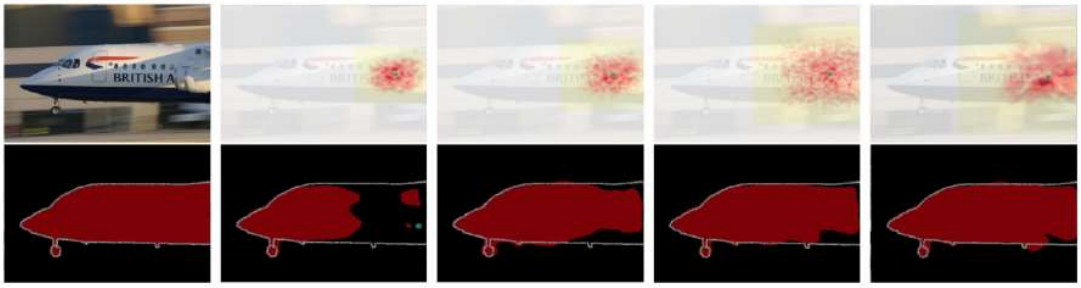
The experimental results are shown in Table 3-1. From the results in the table, it can be found that when only one convolutional layer group is modified, as the ADCNN-modified convolutional layer changes from low to high, the value of the model segmentation performance mIoU will significantly increase. This result is consistent with common experience, that replacing the convolution operation with ADCNN convolutional kernels in high-level convolutional layers performs better than replacing the bottom-level convolution with ADCNN convolutional kernels. At the same time, this result indicates that the receptive field generated by high-level convolution is larger than that of bottom-level convolution, and when replacing high-level convolution with ADCNN convolutional kernels, the spatial flexibility of the receptive field is greater, so the model can have stronger perception ability for objects of different sizes. In addition, the bottom-level convolution is closer to the input end and more sensitive to local changes, tending to focus on capturing information in smaller areas, while the high-level convolution is closer to the output distribution end, which usually relates to complex abstract information, so they are more sensitive to larger input regions.

Table 3-1 mIoU under multiple ADCNN combinations

Method	conv3	conv4	conv5	$\sigma^2\left(d_{i,\ j}\right)$	m Io U
Baseline	—	—	—	—	64.7
	✔			1.96×10^{-4}	63.9
		✔		1.84×10^{-4}	64.7
			✔	4.01×10^{-6}	66.5
ADCNN	✔	✔		2.45×10^{-4}	65.4
		✔	✔	1.24×10^{-4}	66.1
	✔	✔	✔	1.93×10^{-4}	65.9

To further understand the above findings, this paper randomly selects an image from the data set, visualizes its receptive fields and effective receptive fields (ERFs) using the trained model parameters, and puts them together with their segmentation results in Figure 3-4. The green dots represent sampling points, the red dots represent the effective receptive fields at the sampling points, and the yellow squares represent the size of the receptive field. The second row shows the segmentation results, where GT is an abbreviation for Ground Truth. To make a unified and effective comparison, these receptive fields and effective receptive fields are from the final convolutional layer of the model, i.e., the conv5-3 convolutional layer in VGG-16. Based on the visualization results, it can be found that as the ADCNN-modified convolutional layers get closer to the output layer (i.e., the higher the modified convolutional layer level), the size of the receptive field and effective receptive field corresponding to the sampling point also becomes larger. At the same time, using larger receptive fields and effective receptive fields can obtain better visual segmentation results. This result indicates that deploying ADCNN to higher convolutional layers will be beneficial for model performance in practical applications.

In addition, this article also tested more complex improvement modes by replacing more ADCNN convolution kernels with regular ones (the last three rows of Table 3-1). From the experimental results, it can be seen that stacking more ADCNNs can produce a larger receptive field. However, even with a better receptive field, stacking additional ADCNN convolution kernels will cause the model's performance to be lower than modifying a single convolution kernel. This article further studied the reasons by calculating the variance of the sampling values under different conditions. The variance statistical results are shown in the fifth column of Table 3-1, and it can be found from the statistical results that the model's performance will always decrease when combining conv5 with more ADCNN convolution kernels, and at the same time, the sampling variance will increase significantly. Meanwhile, by observing the segmentation results in Figure 3-2, it can be found that using ADCNN convolution kernels to modify high-level convolution kernels and stacking additional ADCNN convolution kernels can both expand the model's receptive field. However, by comparing the visual results of Figure 3-2 (d) and Figure 3-2 (e), it can be found that compared with the increase in the receptive field, the increase in the effective receptive field brought by stacking additional ADCNN convolution kernels is not significant. The receptive field in Figure 3-2(d) has already covered the target object, and the stacking method has produced a larger receptive field, which adds to the burden of the model recognizing positive and negative features. At the same time, considering that the segmentation result of Figure 3-2(e) covers more internal areas of the object than the segmentation result of Figure 3-2(d), there are also more ineffective segmentation areas outside the object. Therefore, this incremental increase caused by additional sampling is the reason for the performance decline, which indicates that stacking additional ADCNN convolutions may impose some additional burden on the convergence of the sampling process.



(a) Image & GT (b) conv3 only (c) conv4 only (d) conv5 only (e) conv3&4&5

Figure 3-2 The receptive fields and effective receptive fields under multiple ADCNN combinations

3.3 Analysis of Hidden Unit Information Aggregation Modes

In this paper, we propose different information aggregation modes between hierarchical layers to model the weight information of hidden units. To analyze the impact of different information aggregation modes on model performance, we design multiple sets of comparative experiments to compare the performance improvement of the model. Based on the performance of ADCNN in different convolutional layers, we expand the conventional convolutional layer in the conv5 convolutional layer group of the VGG-16 backbone network to an ADCNN convolutional layer to avoid ineffective hole value sampling. Specifically, we expand the three convolutional layers (conv5-1, conv5-2, and conv5-3) in conv5 to ADCNN convolutional layers and connect them between layers according to each aggregation mode. In addition, to analyze the impact of residual connections on interlayer connection modes, we use the DeeplabV3+ segmentation framework based on the ResNet-101 backbone network for comparative experiments and expand the high-level convolutional layers of ResNet-101. The experimental results of the information aggregation modes are shown in Table 3-2. Based on the experimental results in the table, it can be observed that all three information aggregation strategies achieve better performance than the original network. However, compared with the other two aggregation modes, the Markov aggregation mode achieves the best performance improvement under different backbone networks. To further analyze the reasons for this result, we calculate and visualize the mathematical expectations of the hole sampling results at each pixel of each sub-convolutional layer of the expanded VGG-16 based on ADCNN. The input image that produces this result is the same as the image in Figure 3-2. The visualization results are shown in Figure 3-3, where brighter regions represent higher hole values, and darker regions represent lower hole values.

Based on the visualization results, we found that ADCNN convolutional layers with the Markov aggregation mode tend to choose larger hole values when they do not carry spatial information of the input samples, that is, the hidden information passed down from the upper layer, while the gate aggregation mode and the recurrent aggregation mode tend to adjust the size of the receptive field according to the spatial structural information of the input samples and provide guidance for selecting hole values for later convolutional layers. At the same time, the features aggregated at lower levels are

more sensitive to local information, forcing the next convolution to focus its receptive field on smaller areas to capture local feature changes. Therefore, based on the experimental results of image segmentation, the Markov aggregation mode is the best choice among the three modes, and it does not produce excessively aggregated interlayer information, making the hidden units more focused on the sampling performance of the current level. Moreover, based on this analysis, in other experiments in this paper, unless otherwise stated, the Markov aggregation mode is used as the default interlayer aggregation mode.

Table 3-2 Aggregation study on different backbones

task	image analysis	
Backbone network	VGG-16	Res Net-101
Original method	64.7	75.1
Markov aggregation model	66.5	77.2
Gated aggregation mode	65.5	76.7
Circular aggregation mode	65.3	75.6

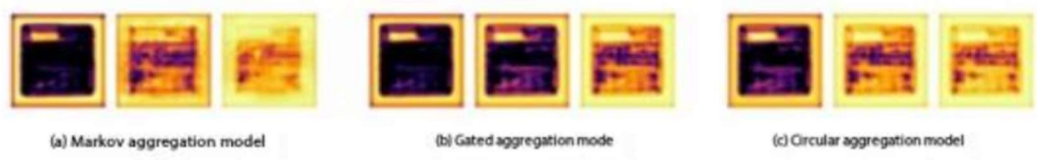


Figure 3-3 The mathematical expectation of dilation sampling at each pixel for individual sub-layers

(from left to right: conv5-1 to conv5-3)

In addition, according to the visualization results, we found that in different information aggregation modes, hole sampling tends to choose larger hole values in the boundary areas of feature maps. The main object of the image appears in the middle area of the image, and the boundary area contains less information. This indicates that ADCNN convolutional kernels tend to use larger hole values in the boundary area to achieve a larger receptive field and capture effective information. Moreover, as the convolutional kernel gradually approaches the middle area of the image, it gradually contacts the target object, and the density of effective information increases. At this time, in addition to learning the features of the object itself, it also needs to learn the boundary information of the object,

that is, to judge the foreground and background. Therefore, in the middle area, ADCNN tends to choose smaller hole values, and the visualization results are deeper.

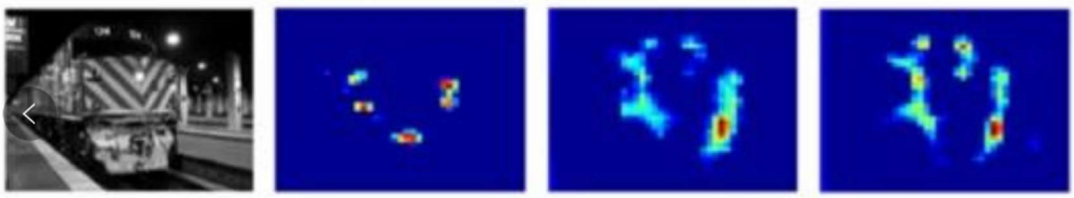
3.4 Analysis of Hole Value Inference Boundaries

To analyze the behavior patterns of hole value sampling in ADCNN convolutional kernels and explore whether there is an upper bound for the sampling space of hole values and whether there are special dependencies on certain values in hole value selection, we designed multiple sets of comparative experiments based on the VGG-16 backbone network and used ADCNN convolutional kernels with one, two, and three available hole value options to compare and analyze the changes in mIoU of image segmentation performance. In addition, based on the results of interlayer information aggregation mode analysis, we use the Markov aggregation mode to eliminate the influence of interlayer information. The other experimental configurations are the same.

Table 3-3 Dilation Boundary study on conv5 of FCN-8s

Dilation = 1	Dilation = 2	Dilation = 4	m Io U
✔			64.7
✔	✔		66.2
✔	✔	✔	66.5

Experimental results of hollow value boundary are shown in Table 3-3. According to the order of increasing available hollow value options in the ADCNN convolution kernel, this paper gradually increased the available hollow value options and compared the change in image segmentation performance mIoU. It should be noted that when the hollow value is set to 1 and there is only one hollow value option, the ADCNN convolution kernel will degrade into a regular convolution kernel. Based on the comparative experimental results, it can be found that when the number of hollow value options increases from one to two, the performance of the model has been significantly improved. However, when the number of options increases to three, the improvement in model performance is not significant. This result shows that the improvement in model performance is mainly brought by the newly added second hollow value option.



(a) Input image (b) An empty candidate (c) Two empty candidates (d) Three empty candidates

Figure 3-4 Activation maps for ADCNNs with different number of dilation options

In order to further analyze the performance differences generated by different hollow value candidates, this paper visualized the feature activation mapping of ADCNN models with different hollow value candidate sets trained on the same image. The visualization results are shown in Figure 3-4. Based on the feature visualization results, it can be found that when the hollow candidate increases from one to two, the range of the activated region in the feature is significantly improved. When the candidate increases from two to three, the increment of the activated area in the feature is not significantly improved compared to before. This indicates that more hollow value options will not further improve the model's performance. In addition, based on the online inference strategy of hollow values, ADCNN can autonomously learn and determine the upper bound of the optimal hollow value during the training process. Therefore, it is not necessary to excessively focus on setting options in the hollow value candidate set in the use of the method.

3.5 Experimental Results of Model Performance Improvement

In order to verify that the proposed ADCNN can comprehensively improve the performance of deep models in image segmentation, based on the analysis of the method's adaptability at the convolution layer level and the information aggregation mode of hidden units, this paper embeds ADCNN into more model architectures, including ResNet-101, DRN (Dilated Residual Networks), Xception, and MobileNet-v2, to verify their performance improvement. In addition, in order to verify the performance of ADCNN on more complex datasets, this paper conducts performance evaluation and comparison on the Cityscapes dataset, and compares the experimental results with current advanced image segmentation algorithms.

The comparative experimental results are shown in Table 3-4. According to the experimental results, it can be found that the proposed ADCNN can effectively improve the performance of the original network on image segmentation problems, and this improvement is stable on the PascalVOC2012 dataset and the Cityscapes dataset. Meanwhile, the method has surpassed current advanced image segmentation algorithms. This result indicates that ADCNN has strong stability and universality in improving model performance. In addition, this paper randomly selected four images from the Cityscapes dataset and used the trained original network and its ADCNN variant to visualize the segmentation results. The visualization results are shown in Figure 3-5. Among them, the first and second column images are the original images and their corresponding image segmentation truth values. The third column shows the segmentation results of the original network, and the fourth column shows the segmentation results obtained using the ADCNN proposed in this paper. Based on the visualization results, it can be seen that compared with the segmentation results of the original ResNet-101, ADCNN can correctly mark more pixel areas and provide more complete expression of details, such as vegetation and sidewalks in the image. In addition, this paper summarizes and compares the IoU (Intersection over Union) results of ResNet-101 and its ADCNN improved variants for each category on

the Cityscapes dataset and records the comparison results in Table 3-5. From the results in the table, it can be found that ADCNN can effectively improve the performance of the original backbone network on most categories.

Table 3-4 Segmentation experiment results on validation sets of VOC 2012 and Cityscapes

Pascal VOC 2012			Cityscapes		
	m Io U			m Io U	
method	conventional	ADCNN	method	conventional	ADCNN
$SSDD$ [125]	64.9	—	$Multiscale\ DEQ$ [126]	80.3	—
VGG-16+FCN-32s	62.8	65.1	$Rep\ VGG-B2$[127]	80.6	—
VGG-16+FCN-8s	64.7	66.5	$OCR(Res\ Net-101-FCN)$ [128]	80.6	—
Res Net-101+Deeplabv3+	75.1	77.2	Mobile Netv2+Deeplabv3+	70.3	71.5
Xception+Deeplabv3+	73.5	74.4	Xception+Deeplabv3+	77.5	79.0
DRN-D-54+Deeplabv3+	75.4	77.2	Res Net-101+Deeplabv3+	80.1	80.7

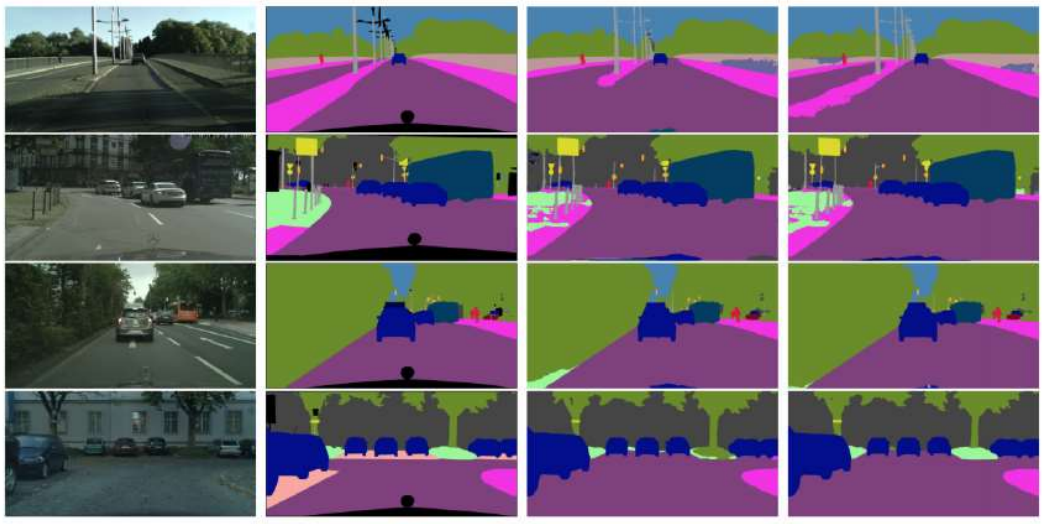


Figure 3-5 Semantic segmentation results on Cityscapes dataset

able 3-5 Performance of each class of ADCNN-ResNet-101 on the Cityscapes validation set

Backbone network	Road	Sidewalk	Building	Wall	Fence	Pole	Light
ADCNN-Res Net-101	0.984	0.867	0.934	0.610	0.654	0.668	0.737

Res Net-101	0.983	0.860	0.931	0.625	0.638	0.648	0.726

Backbone network	Sign	Vegetation	Terrain	Sky	Person	Rider	Car
ADCNN-Res Net-101	0.817	0.930	0.653	0.954	0.840	0.674	0.956
Res Net-101	0.801	0.929	0.659	0.953	0.833	0.658	0.953

Backbone network	Truck	Motorcycle	Bicycle	Bus	Train	m Io U
ADCNN-Res Net-101	0.810	0.722	0.796	0.919	0.808	0.807
Res Net-101	0.797	0.720	0.787	0.912	0.815	0.801

4.Method Applicability Experiment Results and Analysis

To validate the strong applicability of the proposed Adaptive Dilated Convolutional Neural Network (ADCNN) in this paper, which is not only applicable to dense prediction tasks such as image segmentation but can also be flexibly applied to different types of computer vision tasks and bring stable performance improvements. In this paper, three typical computer vision tasks were chosen to experiment with the proposed method, including large-scale image classification, fine-grained image classification, and optical flow estimation. The experimental results show that ADCNN can produce better performance results than conventional networks with stable performance improvement under the condition of negligible additional cost.

4.1 Large-scale Image Classification Experiment Results and Analysis

The ImageNet dataset has become an important benchmark for visual models due to its long data maintenance time, huge image collection, comprehensive category coverage, and high level of image annotation. Therefore, many computer vision tasks, such as image classification and object detection, are developed based on this dataset. However, large-scale image classification tasks on the ImageNet dataset usually require convolutional neural network models with more convolution layers or a large amount of external auxiliary data to achieve better performance. Although these methods can obtain good performance, they significantly increase the model size and training burden.

The proposed ADCNN can effectively extend the original network in a lightweight manner, bringing performance improvement to the original model while maintaining similar training efficiency. To verify this, six typical convolutional neural networks were used as backbone networks, including VGG-16, ResNet-50, ResNet-101, Wide-ResNet101-2, DRN-C-26, and MobileNet-v2, and experiments were conducted on the ImageNet dataset. Similarly to the image segmentation experiment, the Markov layer aggregation mode was used in this paper to experimentally expand the high-level convolutions of each backbone network into ADCNN convolutions.

Table 4-1 Classification accuracy of multiple backbones on ImageNet dataset

evaluating indicator	Accuracy (%)			
	Top@1		Top@5	
Backbone network	conventional	ADCNN	conventional	ADCNN
VGG-16	73.0	74.5	91.2	92.0
DRN-C-26	75.1	75.9	92.4	92.6
Mobile Net V2	71.8	72.6	91.0	90.8
Res Net-50	76.0	76.9	93.0	93.4
Res Net-101	78.3	78.8	94.0	94.2
Wide-Res Net101-2	78.8	79.1	94.3	94.4

This paper calculated the Top@1 and Top@5 classification accuracy of each original network and its ADCNN variant network on the ImageNet dataset, as shown in Table 4-1. Based on the results in the table, it can be found that the performance of ADCNN variant models on Top@1 was improved by an average of 0.8% and on Top@5 by an average of 0.25% on different network models. Given the different model complexities and convolution connection methods among these networks, these performance improvements indicate that the proposed ADCNN effectively improves the performance of the model while weakening the performance impact of complex convolution models.

Table 4-2 Model complexities of multiple backbones on ImageNet dataset

evaluating indicator	model complexity		
Backbone network	$p\#$	$\Delta\, p\#$	$(\Delta\, p\#)/(\, p\#)(\%)$
VGG-16	138M	21K	0.016
DRN-C-26	21.1M	4K	0.019
Mobile Net V2	3.5M	2.7K	0.078
Res Net-50	25.5M	1K	0.004
Res Net-101	178.2M	116.7K	0.066
Wide-Res Net101-2	507.5M	233.5K	0.046

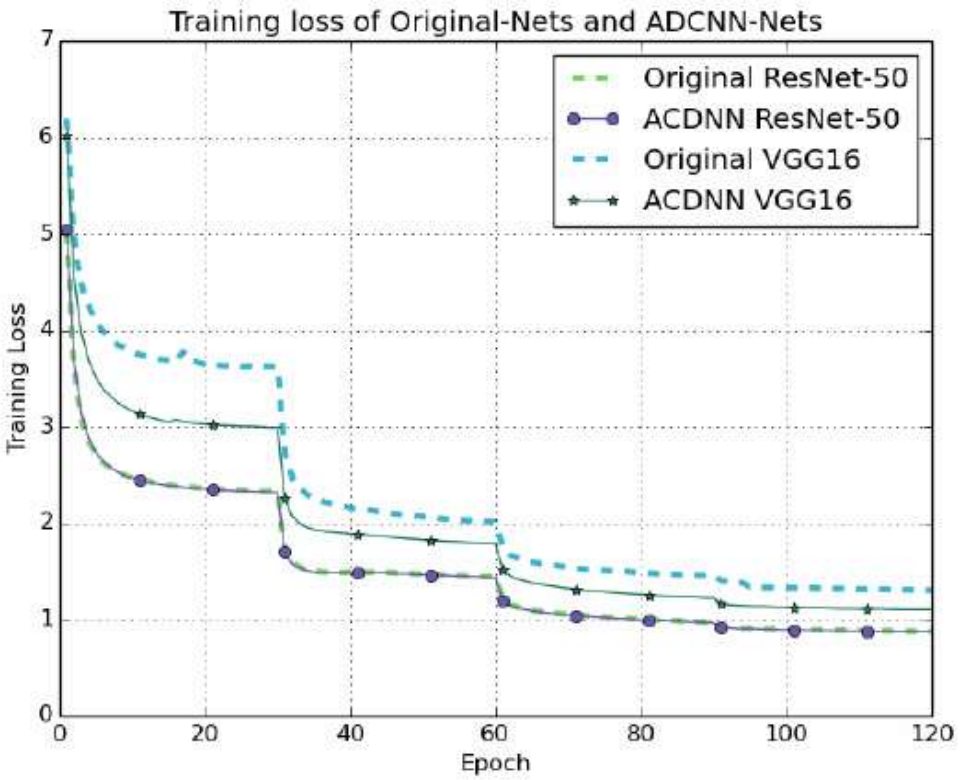


Figure 4-1 The training curve of large-scale image classification using ADCNNs based on the VGG-16 and ResNet-50

At the same time, this paper also statistically calculated the change in model complexity caused by adding ADCNN convolutions, as shown in Table 4-1. Based on the results in the table, given that each original backbone network contains nearly millions of parameters, the impact of thousands of additional weight parameters introduced by ADCNN convolutions on model complexity is relatively small. Combined with the gain in model accuracy, this result suggests that the additional weights with a size overhead of less than 0.1% bring more than ten times performance improvement. This result fully demonstrates the efficiency of ADCNN convolutions in improving performance in large-scale image classification problems, and this also indicates that the proposed ADCNN can be applied to various downstream visual tasks, such as object detection.

In addition, to verify that the model training efficiency is not affected by adding ADCNN convolutions, this paper visualized the training loss output of the original VGG-16 and ResNet-50 and their ADCNN evolved variants, as shown in Figure 4-1. Based on the change pattern of the curves in the figure, it can be found that ADCNN evolved variants have similar or even faster learning progress compared to their original networks. This result suggests that the additional weight parameters introduced by adding ADCNN convolutions do not require additional training costs to achieve convergence.

4.2 Fine-grained Image Classification Experiment Results and Analysis

Unlike general image classification problems, fine-grained image classification tasks emphasize mining subtle discriminative information in images to identify images from different subcategories. This paper designed multiple comparative experiments to verify that the proposed ADCNN can correctly handle such challenges by online selecting the dilation values for features. At the same time, this paper used a typical convolutional kernel adaptive method, deformable convolution, as the main comparative experiment for this problem. Due to the large number of parameters in the VGG-16 model, it is relatively less used in fine-grained image classification problems, so this paper conducted experiments using backbone neural network models other than VGG-16 and initialized their corresponding network structures with weight parameters pre-trained on the ImageNet dataset. The experiment was based on two typical fine-grained image classification benchmark datasets: Stanford Cars and FGVC-Aircraft. The Stanford Cars dataset contains 196 car categories, with a total of 16,185 images, including 8,144 training images and 8,041 testing images, representing variations in car manufacturer, model, and year. The FGVC-Aircraft dataset contains 10,000 images of 100 different aircraft types, of which 6,667 are used for training and 3,333 for testing. The images were proportionally resized to a resolution of 448×448.

The comparative experimental results are shown in Table 3-8. Based on the experimental results in the table, it can be found that both convolution adaptive methods improve the performance of the network model, and the proposed ADCNN can achieve better performance gains on both datasets. Combined with the performance of ADCNN in large-scale image classification problems, it can be found that the proposed ADCNN can not only be flexibly integrated into a variety of classic convolutional neural networks and achieve performance improvements on large-scale image datasets, but also effectively distinguish subtle differences between images when the training data is insufficient.

Table 4-3 Top-1 accuracy (%) for fine-grained visual classification on different databases

data set	Stanford Cars			FGVC-Aircrafts		
Backbone network	Conventional	deformable convolution	ADCNN	Conventional	deformable convolution	ADCNN
Res Net-50	89.9	92.8	93.3	87.9	89.4	90.1
Res Net-101	90.9	91.6	91.7	88.5	89.3	89.6
Wide-Res Net101-2	91.9	92.1	92.5	89.4	91.1	91.3
DRN-C-26	90.1	91.1	92.4	86.8	87.2	89.7

In addition, this paper used the trained DRN-C-26 network architecture and its convolution adaptive expansion variants to randomly select several images from the FGVC-Aircraft dataset for feature mapping extraction and visualized the results to compare the differences in feature response capabilities of different methods. The visualization results are shown in Figure 4-2, where the first column shows the original image, the second column shows the visualization results of the feature maps extracted from the original DRN-C-26 network, and the third and fourth columns show the

visualization results of the feature maps extracted from the DCN-evolved network and the ADCNN-evolved network, respectively. Based on the visualization results, it can be found that compared with the original network, DCN tends to focus on the central part of objects, while ADCNN can activate a more comprehensive range of object regions. This phenomenon indicates that when the data is insufficient, deformable convolutions tend to focus on some details in the image due to the flexibility of their convolution kernels, but ignore the overall features of the object. In contrast, the proposed ADCNN adaptively operates on the convolution kernel in the space of dilation values, and restricts the directional calculation of the convolution kernel under the premise of providing a certain flexibility to the convolution kernel. Therefore, its ability to cover all image features is effectively enhanced, which also reflects the role of ADCNN in enhancing network feature response capabilities.

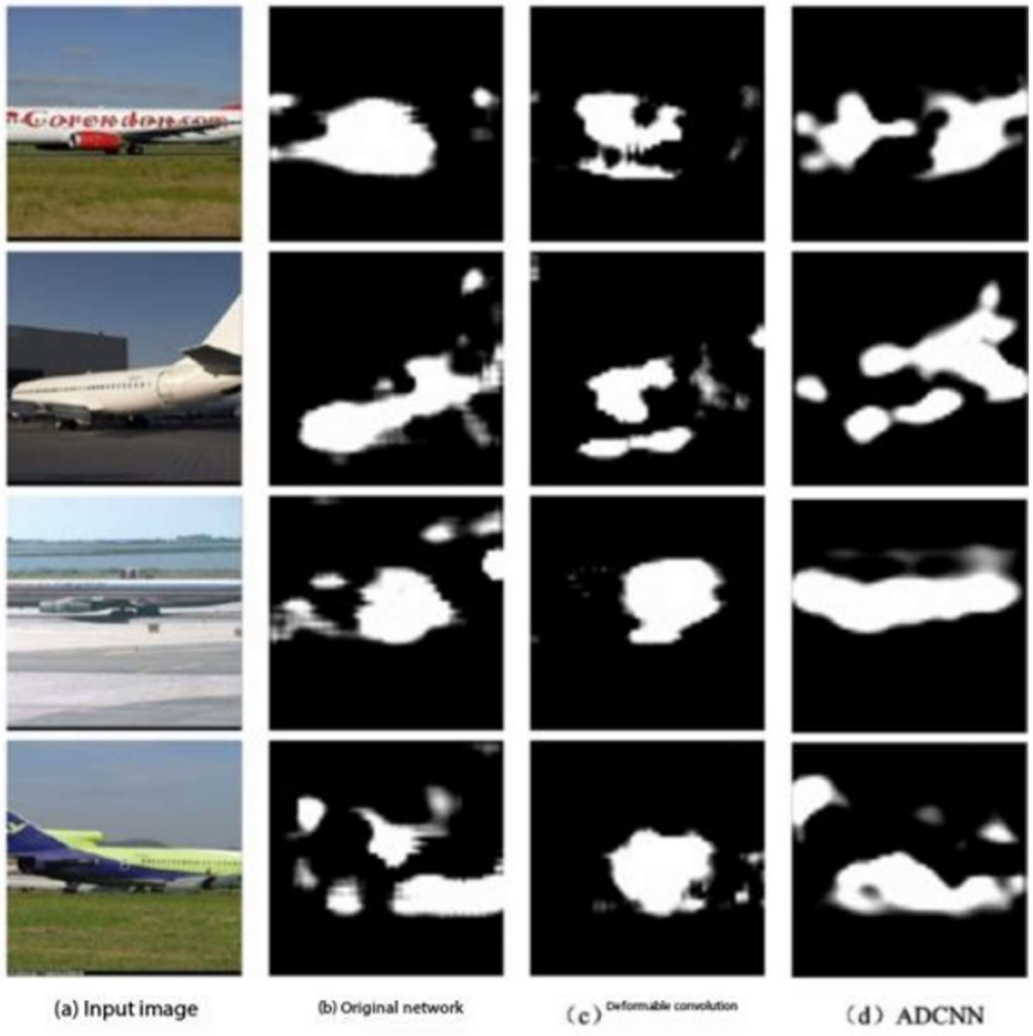


Figure 4-2 Activation maps extracted from DRN-C-26

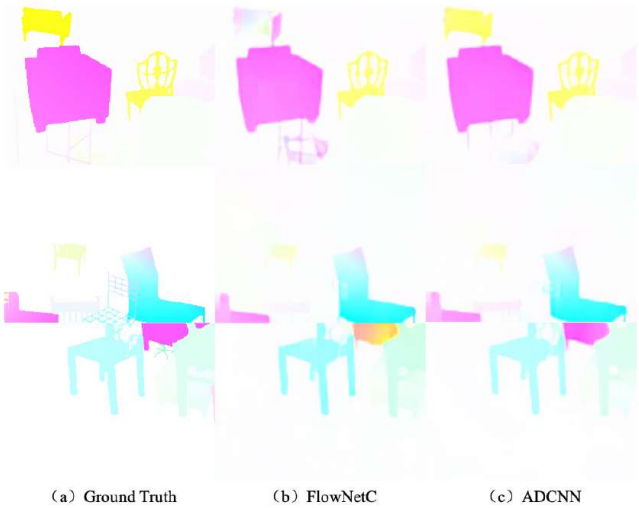
4.3 Experimental Results and Analysis for Optical Flow Estimation

Based on the experimental results, it can be demonstrated that the ADCNN proposed in this paper exhibits good performance in static visual tasks. To analyze whether ADCNN can be applied to more

extensive computer vision tasks, such as motion-based prediction tasks, this paper selects optical flow estimation for experimentation and analysis. Optical flow estimation is an essential direction in computer vision research and plays a significant role in video understanding problems. The FlyingChairs dataset was used in this paper for optical flow estimation experiments, consisting of 22,872 image pairs and corresponding flow fields. In addition, two variants of FlowNet, FlowNetS and FlowNetC, were used as the base backbone neural network models for comparative experiments.

Table 4-4 Average End-Point-Error (aEPE) on FlyingChair dataset

Task	optical flow estimation
Model Method	a EPE
Flow Net S[130] 2.78	
Flow Net S+Seg Aware[131] 2.36	
Flow Net S+LS-DFN,s = 7[132] 2.34	
Flow Net S + ADCNN 1.84	
Flow Net C[130] 2.19	
Flow Net C+LS-DFN, s = 7[132] 2.11	
Flow Net C+LS-DFN, s = 9[132] 2.06	
Flow Net C+ADCNN 1.71	



（a）Ground Truth （b）FlowNetC （c）ADCNN

Figure 4-3 Optical Flow Estimation results on FlyingChairs dataset.

The Average End-Point Error (aEPE) was used in this paper to quantitatively measure the performance of optical flow estimation. The aEPE is calculated by the Euclidean distance between the estimated flow vector and the true flow vector. Comparative experimental results are shown in Table 4-4, indicating that the model modified by the ADCNN convolution kernel reduces aEPE and surpasses the current state-of-the-art methods. Furthermore, using the trained original FlownetC network and its ADCNN evolution variant, three randomly selected predicted results were visualized in the dataset, as shown in Figure 4-3. The first column represents the sample's predicted true value, the second column represents the original FlownetC's generated result, and the third column represents the ADCNN evolutionary variant's generated result. The color represents the direction of the optical flow, and the brightness represents the magnitude of the optical flow. The comparison results show that ADCNN provides more reasonable optical flow estimation results than conventional models.

Dilated Convolutional Neural Networks (DCNN) can effectively expand the receptive field of convolutional kernels, enhance the model's feature extraction ability, and improve the model's robustness on standard data benchmarks. However, the parameter that determines its expansion ability, i.e., the dilation value, lacks clear guidance in the experimental setting process. Incorrect choice of the dilation value can cause the convolution efficiency to drop, leading to a decrease in model performance. To solve this problem, this paper proposes an Adaptive Dilated Convolutional Neural Network based on online inference strategies, designing an online inference strategy for dilatation values, introducing the Gumbel Softmax function to approximate the dilatation value sampling process, and transforming the sampling process through hidden units differentiably. In addition, multiple inter-layer information aggregation modes are proposed to further model the information transmission mode of hidden units and analyzed for performance comparison through experiments. The experimental results show that the proposed Adaptive Dilated Convolutional Neural Network can be flexibly embedded in various convolutional neural networks and brings stable performance improvement. Furthermore, this method can be widely applied to various computer vision tasks, such as image-based dense prediction tasks like image segmentation, which can improve the model's performance on large-scale datasets, accurately extract global features from images with insufficient data, and demonstrate the method's effectiveness in motion-oriented visual tasks.

5. Conclusion

With the continuous advancement of scientific research, effective data accumulation has gradually overcome the traditional overfitting problem's dependence on data. However, the vast amount of data increases data complexity sharply, presenting higher demands for the model's feature extraction ability. Traditional neural networks are limited in their feature extraction ability due to the inflexible adaptation of convolution kernel receptive fields based on image content. In contrast, Dilated Convolutional Neural Networks enhance the receptive fields of convolution kernels by sparse convolution, allowing the same size convolution kernels to process larger image areas with varying receptive field sizes through different dilation values. Furthermore, stacking dilated convolutions can

greatly improve the model's receptive field, effectively enhancing its feature extraction ability. However, traditional Dilated Convolutional Neural Networks lack effective guidance in setting dilation values, and incorrect choices can lead to a decrease in convolutional operation efficiency and impact model performance.

To solve this issue, this paper proposes an Adaptive Dilated Convolutional Neural Network based on online inference strategies. The online inference strategy designs the dilation value as an output of a pixel-based function via a data-driven sampling process. Additionally, to embed the discrete sampling process into the deep learning model update process, Gumbel Softmax is used as an approximation estimation for the dilation value inference process. Moreover, the inference process is affected by inter-layer information, and to accurately model the information transmission process between layers, a hidden unit is designed to control the inference process, and multiple inter-layer information aggregation modes are proposed to update the hidden unit parameters. This paper applies the Adaptive Dilated Convolutional Neural Network to several typical computer vision tasks and compares it with recent advanced methods. According to the comparative experimental results, the proposed method can be flexibly embedded in various convolutional neural networks and effectively enhances the model's feature extraction ability with minimal additional consumption.