Cascaded Hierarchical Atrous Spatial Pyramid Pooling Module for Semantic Segmentation

Lian, Xuhang; Pang, Yanwei; Han, Jungong; Pan, Jing

Published in:
Pattern Recognition

DOI:
10.1016/j.patcog.2020.107622

Publication date:
2021

Citation for published version (APA):
Lian, X., Pang, Y., Han, J., & Pan, J. (2021). Cascaded Hierarchical Atrous Spatial Pyramid Pooling Module for Semantic Segmentation. *Pattern Recognition*, *110*, Article 107622. https://doi.org/10.1016/j.patcog.2020.107622

Cascaded Hierarchical Atrous Spatial Pyramid Pooling Module for Semantic Segmentation

Xuhang Lian[a], Yanwei Pang[a,*], Jungong Han[b,*], Jing Pan[a,c]

[a] *School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, P.R. China.*
[b] *Department of Computer Science, Aberystwyth University, Aberystwyth SY23 3DB, UK*
[c] *School of Electronic Engineering, Tianjin University of Technology and Education, Tianijn 3000222, P. R. China*

Abstract

Atrous Spatial Pyramid Pooling (ASPP) is a module that can collect semantic information distributed in different scopes. However, because of the limited number of sampling ranges of ASPP, much valuable global features and contextual information cannot be sufficiently sampled, which degrades the representation ability of the segmentation network. Besides, due to the sparse distribution of the effective sampling points in the atrous convolution kernels of ASPP, large amount of local detail characteristics are easily discarded. To overcome the above two problems, a new Cascaded Hierarchical Atrous Pyramid Pooling (CHASPP) module, consisting of two cascaded components, is proposed. Each component is a hierarchical pyramid pooling structure containing two layers of atrous convolutions with the aim to densify the sampling distribution. On the foundation of such a hierarchical structure, another same structure is appended to form a cascaded module which can further enlarge the diversity of sampling ranges. Based on this cascaded module, not only rich local detail characteristics can be comprehensively presented, but also important global contextual information can be effectively exploited to improve the prediction accuracy. To demonstrate the performance of our CHASPP module, experiments

*Corresponding author
Email addresses: lianxh@tju.edu.cn (Xuhang Lian), pyw@tju.edu.cn (Yanwei Pang), jungonghan77@gmail.com (Jungong Han), jingpan23@gmail.com (Jing Pan)

on the benchmarks PASCAL VOC 2012 and Cityscape are conducted.
Keywords: Semantic Segmentation, Atrous Convolution, Atrous Spatial Pyramid Pooling(ASPP), Hierarchical Pyramid Pooling, Cascaded Module.

---

1. Introduction

Semantic segmentation is a fundamental technique in the field of computer vision [7], [3]. Its goal is to assign a label to each pixel in the image. The result of it can provide a comprehensive description of the surrounding scenes, including object categories, locations and shapes [9]. It is of broad interest for many applications, such as autonomous vehicles, robotics, and human-machine interaction [10], [6].

A challenge that exists in the task of semantic segmentation is how to tackle objects at varying scales [1], [2]. Currently, there are several predominant methods that are proved to be effective for dealing with this problem. One classical way is to rescale input images into multiple scales and put these rescaled versions into the convolutional neural networks. Through fusion of the intermediate feature maps or final score maps, features of all the versions can complement each other and thus form feature maps with more comprehensive semantic information. Though this kind of method indeed gets the performance of the networks enhanced, such improvement is achieved at the cost of dramatically lowering the efficiency of the whole system. First, the input image needs to be resampled multiple times before being imported into the networks. Second, each resampled image needs to go through the whole network once. During this process, the responses of each intermediate layer are repeatedly computed. Such inefficiency is not desirable in practical scenes, especially those requiring real-time predictions. Therefore, how to fuse the multi-scale features at the same time of not bringing about too much computation cost becomes a topic in the research of semantic segmentation.

To remedy this situation, a method based on Fully Convolutional Networks (FCN) [8] is developed, in which there are usually several skip connections be-

tween different intermediate layers of the backbone classification network. As feature maps of different intermediate layers own various semantic levels and receptive fields, the fused feature maps through the skip connections simultaneously contain contextual information collected in different scopes which can match the requirements of comprehensively describing objects with different sizes. The effect of this amounts to the method mentioned above in which the responses of the same layer are repeatedly computed when different resampled images are input.

Another line of methods aims at improving the efficiency of segmentation networks via utilizing the ASPP module. Inspired by the idea of the spatial pyramid pooling, Chen et al. [11] designed the ASPP module to more efficiently tackle this challenge. This module is composed of four parallel atrous convolution layers with different dilated rates. Features maps of each atrous convolution layer encode semantic information abstracted from different receptive fields, and all these feature maps are merged to form a final comprehensive representation of the semantic information of the whole image. The four parallel atrous convolution layers in ASPP can achieve the effect which is equivalent to probing the input image with multiple filters that own different receptive fields. With the ASPP module, the resampling process of classical methods can be avoided, and the efficiency of the whole system can be significantly improved.

Though ASPP can raise the system efficiency, it still causes another two problems at the same time when applied to practical scenes. As mentioned above, the receptive fields of the four parallel atrous convolution layers are complementary to each other, which can guarantee that the information distributing in different scopes can be sampled. With the information from different scopes, the segmentation network can have the ability to tackle objects with sizes varying in a particular range. However, such a distribution of sampling ranges cannot ensure that the information (e.g., global features and context priors), which is exploited as the key clues to predict the label of target objects, can be contained in the sampling ranges. This is caused by two reasons. One is that the diversity of the sizes of the objects in practical scenes is much larger

than that of the sampling ranges of ASPP. This means that only objects with the same sizes as those of kernels of ASPP can be sampled while information of other objects cannot be collected. The other is that distribution of objects in the image is arbitrary. Some objects are so far from the scopes covered by the convolution kernels of ASPP that the features they contain cannot be sampled. In a word, such limited variety of sampling ranges of ASPP cannot sufficiently collect information distributed in various ranges in the image.

Additionally, to avoid bringing too much extra computational load, all the points inside of the convolution kernels except for the nine effective sampling points are filled with zeroes. This measure would lead to that as the convolution kernel samples the information surrounding a particular pixel, local delicate features that correspond to the positions with zeroes in the convolution kernels cannot be collected, which may end up with the situation that these delicate features are disregarded in the final result.

To overcome the above drawbacks of ASPP, a new Cascaded Hierarchical Atrous Spatial Pyramid Pooling (CHASPP) module is proposed. Compared with the structure of the single level of the atrous convolutions in ASPP, a hierarchical structure consisting of several levels of atrous convolution layers is adopted. This structure can densify the sampling distribution and sufficiently capture important local features. Apart from this, the stacked hierarchical structures can further enlarge the diversity of the sampling ranges and expand the scopes which our module can cover.

The contributions of our paper can be summarized as follows:

1) A new hierarchical structure consisting of multiple levels of atrous convolution layers is proposed. This structure can remedy the problem of the degenerated representation of local detail features caused by the hollow kernels of ASPP via increasing the sampling density.

2) A novel Cascaded Hierarchical Atrous Spatial Pyramid Pooling (CHASPP) network which contains cascaded hierarchical structures is proposed. Through such a stacked module, the variety of receptive fields can be significantly enlarged compared with ASPP, which can guarantee that global features and contextual

information are extensively collected to make a more reasonable and accurate prediction.

3) Experimental conducted on PASCAL VOC 2012 [10] and Cityscape datasets convincingly demonstrate the performance of our network in preserving local detail features and collecting global contextual information distributing in various areas of images.

2. Related work

In this section, current segmentation methods are roughly categorized into three classes according to the measures which they adopt to diversify the sources of semantic information

Multi-scale inputs: The motivation behind this kind of methods is to make the segmentation network adapt to objects with varying scales by training the network with multiple rescaled input images [4], [5]. Though this kind of methods do not explicitly increase the number of the receptive fields, the effect of resampling input images is equivalent to simultaneously probing the input images with several kernels with different sizes. In [13], multiple images transformed by a Laplacian pyramid are respectively passed through a shared network. The shared network produces a series of feature vectors for the regions of multiple sizes centered around every pixel in the image. In [14], for encoding rich background information, each resampled version of the input image successively goes through a convolutional network and a pooling pyramid structure, and the produced feature maps of all the versions are concatenated to form the final feature maps. In [15], to capture the image long range dependencies, a recurrent network with several instances is adopted. Each instance considers as inputs both a resized RGB image and the classification attempt of the previous instance.

Intermediate features fusion: Fusing feature maps from different intermediate layers of a classification backbone network is adopted in some current methods to implicitly increase the diversity of the receptive fields. This measure

is based on the following consideration: each intermediate layer of a classification network encodes features corresponding to a particular receptive field in the input image. Merging the features from different layers can achieve the effect of aggregating information from different receptive fields. In [19], features from intermediate layers of the encoder part are combined with the features of corresponding layers in the decoder part. Long et al. [8] utilized skip connections to connect the deep layers with coarse information with the shallow layers with fine information to refine the prediction effect. Hariharan et al. [20] connected feature maps of intermediate layers and utilized the vector of pixels across the connected feature maps for the segmentation task.

Though the above methods can boost the fusion of information from different receptive fields, the predicted results of these methods are unsatisfactorily coarse. To refine the results, more sophisticated structures are placed in the skip connections between the encoder and decoder parts of the segmentation networks. In [21], a residual convolutional unit is attached to the end of each block of the backbone network. To acquire the background context from a large region, the fused feature maps undergo a chained residual pooling component with a sequence of pooling layers. Peng et al.[22] designed a global convolutional network for simultaneously maintaining the classification and localization abilities and a boundary refinement module for refining the boundary effect of the predicted results. In [23], a smooth network with channel attention blocks are connected with the intermediate blocks of the backbone network to select discriminative features. At the same time, the feature maps of intermediate blocks go through a border network to make boundaries more noticeable. Ding et al. [24] proposed a context contrasted local model to generate multi-level and multi-scale context-aware features.

Spatial Pyramid Pooling: Another method that has been widely used is to append more branches to the end of the backbone network and combine the outputs of all these appended branches. As pixels from the feature maps of a branch correspond to a particular size of receptive field, parallelizing several branches can multiply the receptive fields that the final fused feature maps

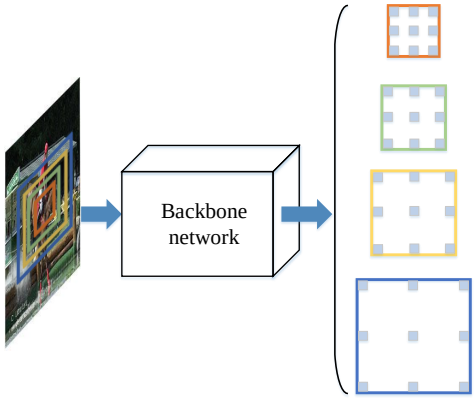

Figure 1: Atrous Spatial Pyramid Pooling (ASPP). ASPP extracts semantic information through atrous convolution layers with different dilated rates. Receptive fields that atrous convolution layers correspond to are shown in different colors.

of the pyramid pooling structures can view. In [25], three atrous layers with different dilated rates and a global pooling layer are parallelized to capture both the global and local semantic information. Similarly, Wang et al. [26] utilized parallel atrous convolution layers at the end of the backbone network. In [27], several pooling layers with different kernel sizes are utilized at the end of the backbone network to aggregate contextual information from different receptive fields. Reviewing the above methods, it can be noticed that the method utilizing spatial pyramid pooling owns both advantage and disadvantage. The advantage is that it can avoid the repeated computation of the intermediate layers of backbone networks, while the sparse sampling distribution and limited receptive fields are its disadvantages.

3. Review of atrous spatial pyramid pooling

As our method is based on Atrous Spatial Pyramid Pooling (ASPP), the problems caused by its limited variety of sampling ranges and sparse distribution of effective sampling points in ASPP will be briefly illustrated.

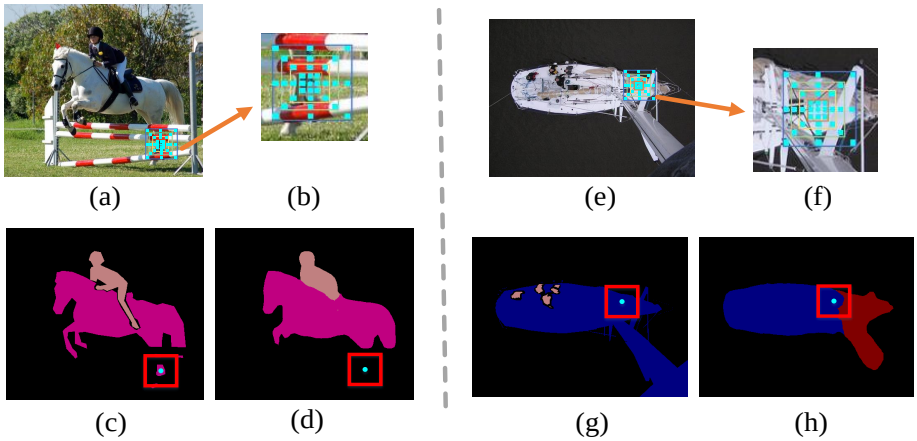

Figure 2: Illustrations of the problem caused by limited variety of sampling ranges of ASPP. (a) and (e) are Original images. (b) and (f) are Zoomed images of sampling ranges of ASPP. (c) and (g) are ground truths. (d) and (h) are predicted results of ASPP. The light blue points in the red squares of (c) and (g) are the target pixels to be classified.

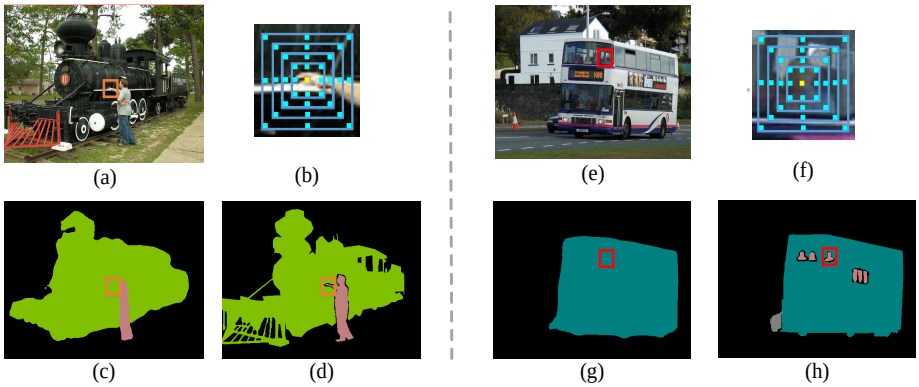

Figure 3: Illustrations of the problem caused by sparse sampling of atrous convolution kernels in ASPP. (a) and (e) are Original images. (b) and (f) are Zoomed images of sampling ranges of ASPP. (c) and (g) are predicted results of ASPP. (d) and (h) are ground truths. The yellow points in (b) and (f) are the target pixels to be classified.

3.1. Limitations of ASPP

Though ASPP can deal with objects with varying sizes to some degree, the actual challenge posed by the practical scenes is far beyond what ASPP can face. The sizes of all the objects that ASPP can collect information from only vary among four particular sizes. Nevertheless, in the open scenes, arbitrary sizes and distributions of objects bring great difficulty in sampling enough information to provide the comprehensive and accurate features. Specifically, this difficulty can be reflected in two aspects. First, due to the arbitrary sizes of the objects in practice, utilizing ASPP with limited sampling ranges can not comprehensively collect the features of objects components which are disconnected and scattered over a large scope and form a global understanding of target objects. For example, the horse legs in the red rectangle in Fig.2(a) are surrouded by the background and for from the main body of the horse. Second, the random distribution of context priors, especially those which can offer key clues for predicting the labels of local areas, always results in the rareness or intactness of the contextual information collected by ASPP. Without relatively complete context priors, the target objects are very likely to be misclassified. Such a example can be seen in Fig. 2. In Fig. 2(d), the texture of the square area surrounding the blue point is similar to that of a plane. As the sea water, which can offer the key prior to exclude the probability of a plane, can not be sufficiently sampled by ASPP, the semantic network can not discriminate the boat from the plane, which results in that the light blue point in the square of Fig. 2(f) is misclassified as plane (the red area in Fig. 2(f)).

Apart from the above problem caused by the insufficient variety of sampling ranges, the hollow structure of the atrous convolutions in ASPP also brings serious problem. As is known, to effectively expand the receptive fields of the segmentation networks at the same time of not causing too much computation cost, the convolution kernels in ASPP only partly sample the pixel values of the nine points in its sampling range and ignore the values of the other positions. In fact, much valuable local detail information which play important role in depicting the local components of objects are neglected, which results in that

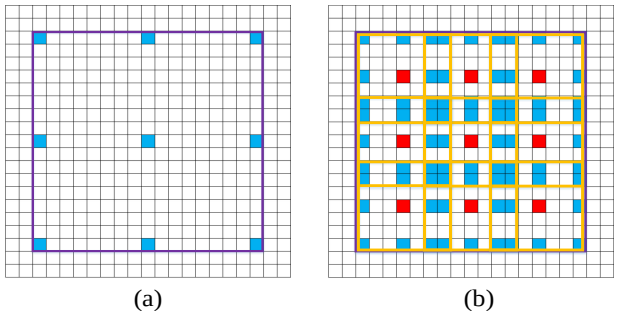

Figure 4: Comparisons between densities of effective sampling points in ASPP and our hierarchical structure. (a) Density of sampling points in ASPP. (b) Density of sampling points in our hierarchical structure. The purple squares denote the receptive fields, and the overlapping orange squares in (b) denote the finely divided sampling areas by our hierarchical structure.

many noticeable local features of objects except the main body of objects are seriously weakened or buried. Such a consequence can be clearly seen in Fig. 3. The arm of the person in Fig. 3(a) is lost in Fig. 3(c). Similarly, the passenger in Fig. 3(d) is buried by the mask of the bus (as shown in Fig. 3(f)).

For avoiding the results caused by ASPP, the CHASPP module which contains two aspects of improvements on ASPP is devised. One is to change the single level structure of ASPP into a two level hierarchical structure to densify the sampling distribution. On the foundation of such hierarchical module, the variety of sampling scopes through stacking two such structure is further enlarged. In the following section, the CHASPP will be decomposed according to the above aspects, and the principles will be respectively elaborated.

4. Cascaded hierarchical atrous spatial pyramid pooling

4.1. Hierarchical structure with muli-level atrous convolution layers

As mentioned in Sec. 3.1, the atrous convolutions adopted in many pyramid pooling structures [11], [25] is used to expand the receptive fields at the same time of not bringing expensive computation cost. To achieve this goal, all the points except the nine effective sampling points are filled with zeros to reduce the computation amount (as illustrated in Fig. 4(a)). Such an intrinsic

property determines that the atrous convolution kernel can not closely capture the characteristics of local areas that correspond to the positions of these zero values, which leads to the defective representation of local delicate structures.

For avoiding the destruction of local structures caused by the empty points of the kernels of atrous convolution, we design a hierarchical structure to increase the number of effective sampling points inside the receptive fields. In this paper, for the simplicity of statement, the atrous convolution layers in the first level is name *root layers*, and the atrous convolution layers attached to each atrous convolution layer in the first level *branch layers*. At the end of each root layer, we attach three parallel branches with different dilated rates (as shown in Fig. 10). Compared with the one-stage sampling, the sampling process is split into two stages, which are formed by one root atrous convolution layer and three branch convolution layers, with small convolution layers. The distribution of the sampling points is shown in Fig. 4(b). The squares in Fig. 4 denote the effective sampling points in the hierarchical structure. Through the sampling in each finely divided area that the branch layers pass, the unique and discriminate characteristics possessed in these areas can be elaborately collected, and finally these characteristics can be presented in the predicted results.

In fact, our idea of adopting a hierarchical strucutre with multiple atrous convolution layers is in line with those of many current semantic segmentation methods. In [12], Chen et al. developed a variation of the sampling method that the subsamplings after the last max-pooling layers of VGG [28] are skipped and changed a series of convolution layers followed by pooling layers to atrous convolutions with dilated rates 2 and 4. Besides the basic function of dense prediction, such an arrangement of the atrous convolutions can accomplish the tasks of extracting features from local areas, thus progressively raising the semantic level. Other related works [25], [26] explore the effect of stacked atrous convolution layers with different dilated rates in eliminating the gridding effect caused by many zeros padded between adjacent effective sampling points in the atrous convolution kernels, and demonstrate that through adjusting the patterns of dilated rates of serialized atrous convolution layers, the insufficient

sampling of local delicate structures can be avoided.

Following the design principle of densifying sampling distribution through stacked atrous convolution kernels, our architecture is built based on ASPP. In the building process, the original atorus convolution layer in each branch is split into a root layer and a branch layer without changing the total receptive field. On the foundation of this already formed branch layer, two additional branch layers are inserted. In our architecture, the difference among the dilated rates of these three branch layers are moderately kept. One goal of controlling the differences is to avoid that the receptive fields owned by each root layer concentrates in a narrow band which will limit the richness of the collected features. The other is to avoid redundant sampling in a particular range caused by overlapping among receptive fields belonging to different root layers. Such measure of arranging internals among receptive fields can ensure that the representative features of scopes from the near to the distant can be sampled in order and the whole collected features can form a comprehensive descriptor of the characteristics of the target objects. In Fig. 10, an implementation of our hierarchical structure is presented.

Through changing the single level of the atrous convolution layer into the new version adopting the hierarchical structure, the performance of the segmentation networks is significantly improved, especially in protecting the noticeable local characteristics of the target objects. Nonetheless, the some detail components are still coarsely represented and need to be further refined. Such a phenomenon is resulted by the fact that the features output by the backbone networks, which are the only information source of the prediction module, are lack of sufficient fine-grained information that plays important role in presenting the detail characteristics. For further refining the prediction effect of our hierarchical structure in the detail characteristics of the target object, we introduce the features of the intermediate layers of the backbone networks. As the low-level semantic information consists of highly discriminative local features, introducing such information will enlarge the differences between the intrinsic properties of two patches with the same label, which is likely to mislead the

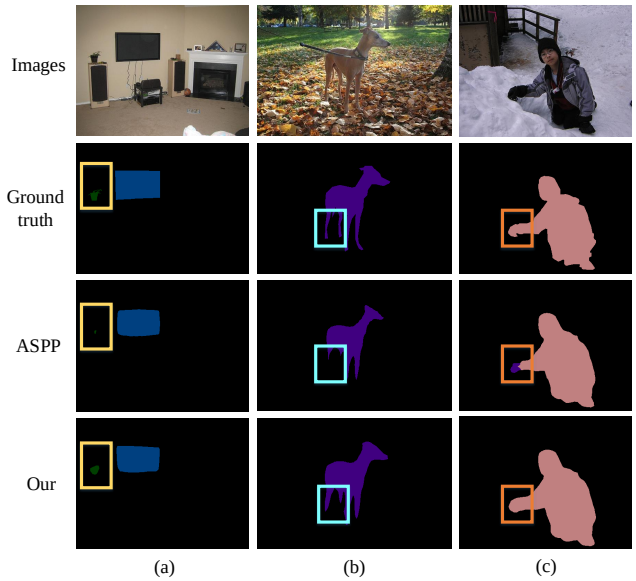

Figure 5: Illustrations of effects of changing ASPP to our hierarchical structure.

segmentation networks into making inconsistent classifications about the labels of these two patches. For avoiding such a situation, the features from blocks with high-level semantic information are utilized to ensure the consistent classification in patches with the same labels. In this paper, the features of the 16th residual unit in the third block of ResNet-101 are introduced into our hierarchical structure and respectively mix them with the features of each root layer.

The final predicted results of our hierarchical structure with introduced features of the intermediate layers of the backbone network are shown in Fig. 5. Through the comparisons in Fig. 5, it can be observed that the delicate structures in the rectangles that have been neglected by ASPP are restored to a great extent. Such an improvement in preserving local inconspicuous structures can verify that the measure of adopting the hierarchical structure can rather effectively increase the sampling density inside the receptive fields and significantly improve the capability of the segmentation networks in presenting these important local features.

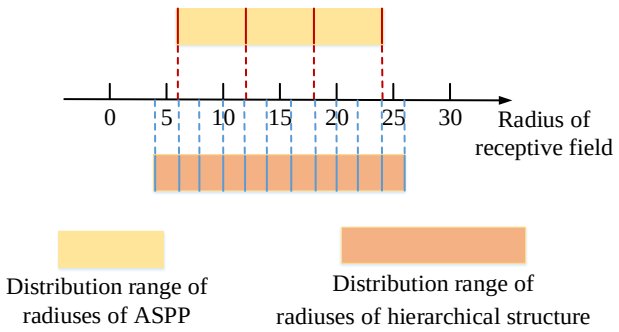

Figure 6: Illustrations of changes of distribution range and density after adopting tree-shaped structures. The light yellow band denotes the distribution range of radiuses of receptive fields in ASPP, and the red vertical lines on the this yellow band denote the exact values of receptive field radiuses. The orange band denotes the distribution ranges of radiuses of receptive fields in a single hierarchical structure, and the blue vertical lines on it denote the exact values of receptive field radiuses.

4.2. Cascaded multi-level hierarchical structures

As mentioned in the Sec. 3.1, because of the limited number of sampling scopes, ASPP is not applicable in the following situations: The first is that the components of the target objects distribute in a large scope, especially those with much disconnection. For the objects of which constituent parts are concentrated in a limited local scope without much disconnections, ASPP can extensively cover the constituent parts of the target objects and do not leave out key information. However, once the components of objects are split into fragments, the function of ASPP will be largely degraded as many valuable global information are out of the range that ASPP can reach. The second is that the contextual information, which can play an important role in providing auxiliary information to help discriminate local patches, is arbitrarily scattered in the image.

Given the above situations, our single hierarchical structure, which is utilized to achieve sufficient sampling of local features inside the receptive fields, cannot solve the problems led by the limited variety of receptive fields. This conclusion can be drawn from Figs. 6 and 7. In Fig. 7, from the view of expanding the

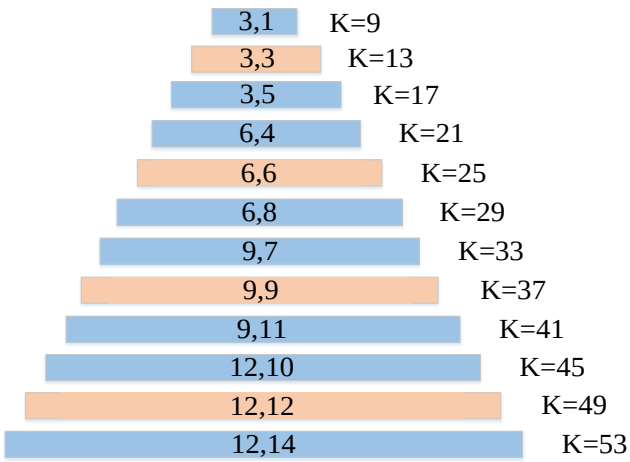

Figure 7: Illustration of the pyramid of receptive fields of our hierarchical structure. The orange band denotes the receptive fields of ASPP, and the blue bands denotes the newly produced receptive fields by our hierarchical structure. The length of each band indicates the length of the side of each receptive field, and the exact number of the side length is shown on the right side of the band.

diversity of the receptive fields outside the original scopes of ASPP, the effect of a single hierarchical structure can be considered as filling the large gap between two adjacent scopes of ASPP with additional receptive fields and not essentially increasing the variety of receptive scopes larger than the original ones of ASPP. The same situation can be observed in Fig. 6 that the receptive fields of our improved structure still stay in a fixed range and are not significantly diversified.

For significantly enlarging the diversity of the receptive fields, especially those larger than the original scopes of ASPP, on the foundation of the single hierarchical structure, a new cascade structure which contains two stacked hierarchical structures (as shown in Fig. 10) is devised Though the components of the newly added pyramid module do not vary from those of the first one, the scopes that pixels in the output feature maps can view are totally different. The primary goal of designing the first pyramid module is to avoid missing local details. Following such principle, the emphasis is on densifying the sampling distribution, not expanding the receptive fields, which results in

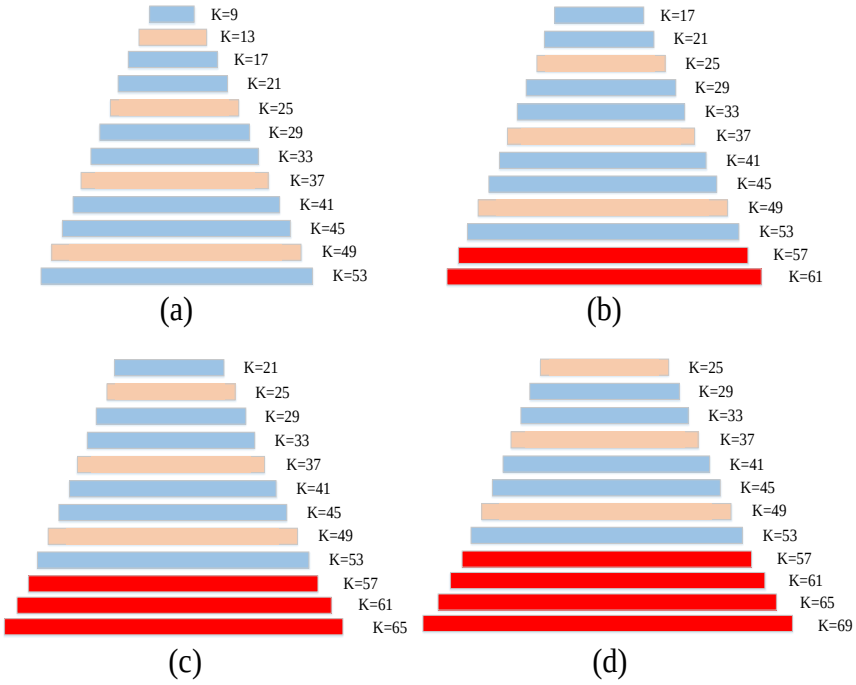

Figure 8: Examples of new receptive fields generated by the tree-shaped structures in the second pyramid module. The red bands denote the newly generated receptive fields. In this figure, the results produced by the patches with the combinations of dilated rates (3,1), (3,3) and (3,5) are respectively showed in (b), (c) and (d)

.

that the scopes pixels in the output feature maps can view are kept unchanged. Distinct from keeping to inserting the receptive fields inside the range fixed by ASPP, the second pyramid module achieves a breakthrough in increasing the sampling ranges variety outside the scopes enclosed by ASPP. As shown in Fig. 7, the first pyramid module has built a hierarchical structure of the receptive fields with uniform intervals among them. When the feature maps, containing a composite of receptive fields shown in Fig. 7, produced by the first module are passed through a particular path formed by a pair of the root layer and the branch layer in the second hierarchical structure, not only the receptive fields in the original composite will be further enlarged, but also many new ones which

are out of the range given by ASPP will be produced. As the dilated rates of the root layer and the branch layer increase, the number of the newly generated scopes will become larger. For instance, when the receptive field hierarchy with the radius array of (4,6,8,10,12,14,16,18,20,22,24,26) is passed through the path formed by the root layer and the branch layer with the combination of dilated rates of (3,1), then the radius array of the output feature maps will be (8,10,12,14,16,18,20,22,24,26,28,30) which contains two new radiuses 28 and 30. If the path is changed to the one with the dilated rates combination of (3,3), then the new update will be 28,30 and 32. Such a growing process of the number of newly generated receptive fields can be vividly illustrated in Fig. 8.

Through amplifying and diversifying effects of each pair of the root layer and branch layer in the second hierarchical structure, the total variety of the receptive fields of the whole network is significantly enriched and the visual fields of the pixels in the output feature maps become more spacious. Such changes in the receptive fields can effectively enhance the sampling and fully exploit the global features of objects and surrounding valuable contextual information. The effects of this stacked structure are shown in Fig. 9. It can be observed that relying on the key information of global features or context compensated through our stacked pyramid modules with tree-shaped structures, the misclassified parts in the rectangles of the results of ASPP can be restored, which verifies the effectiveness of our method in capturing information and classifying the labels of target objects.

4.3. Overall Framework

The overall framework of our method is shown in Fig. 10. Our segmentation network contains two stacked hierarchical structures. In each structure, there are four root layers of which each is composed of three branch layers with different dilated rates. The output features of the backbone network are passed to the first hierarchical structure of which the function is to densify the sampling distribution through four combinations of the root layers and appended three parallel branch layers. For compensating the lack of fine-grained features, the

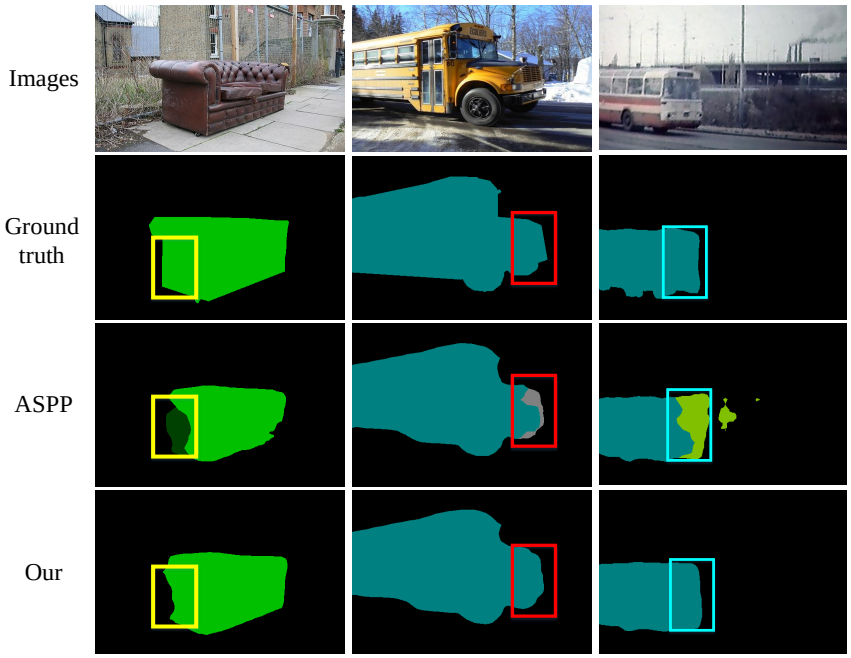

Figure 9: Illustrations of effects of enlarging the variety of receptive fields through cascaded hierarchical structures.

features of intermediate layers of the backbone network are introduced into the hierarchical module and respectively merged with the features of the root layers. Through the processing of the first module, the output features are summed and delivered to the second module to further multiply the receptive fields which aims at extensively including key global features or contextual information. The output features of the second pyramid module are finally combined, and the final predicted results are produced.

5. Experiments

In this section, the performance of our method will be evaluated on two standard benchmarks PASCAL VOC 2012 and Cityscapes.

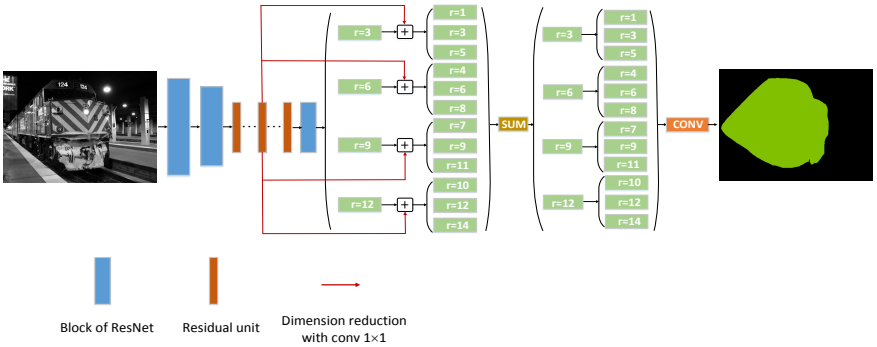

Figure 10: The overall framework of our method.

5.1. Implementation Details

Training: Our implementation is based on the platform Tensorflow. ResNet-101 [29] (pretrained on ImageNet) is used as the backbone network. As [11], learning rate poly is adopted. The learning rate is obtained by multiplying the initial learning rate with $(1 - \frac{iter}{max\_iter})^{power}$. In our experiments, the initial learning rate is set to 0.001 and the power is set to 0.9. the network is trained using the mini-batch stochastic gradient descent (SGD) [30] with the batch size of 8, weight decay 0.0001 and momentum 0.9. The mean pixel intersection-over-union (mIoU) is used as the metric to measure the prediction accuracy.

Data augmentation: For improving the robustness of our network, in the training process, input images are randomly flipped to augment the dataset in both PASCAL VOC 2012 and Cityscapes. Besides, it is found that randomly scale the input image can improve the performance. Therefore, in both PASCAL VOC 2012 and Cityscapes, the input image are rescaled with five factors $\{0.75, 1.0, 1.25, 1.5, 1, 75\}$.

5.2. PASCAL VOC 2012

In this section, the ablation studies on PASCAL VOC 2012 dataset are performed to analyze the performance of each component of our network. The PASCAL VOC 2012 benchmark contains 20 foreground classes and one background class. This dataset is composed of 1,464 images for training, 1,449 for

Table 1: Performance analysis of ASPP and our multi-level hierarchical structure. 'Stu1' is used to denote the single hierarchical structure. The numbers in each bracket denote the combination of dilated rates of all the root layers or the dilated rates of branch layers attached to each root layer. 'root1-br', 'root2-br', 'root3-br', 'root4-br' respectively correspond to the dilated rates of branch layers attached to each root layer.

Method	Stu1					mIoU
	root	root1-br	root2-br	root3-br	root4-br	
Res50	(6,12,18,24)	/	/	/	/	70.9
	(3,5,7,9)	(1,2,4)	(4,6,8)	(6,8,10)	(10,12,14)	73.2
	(4,6,8,10)	(1,3,5)	(3,5,7)	(5,7,9)	(7,9,11)	72.8
Res101	(6,12,18,24)	/	/	/	/	75.1
	(3,5,7,9)	(1,2,4)	(4,6,8)	(6,8,10)	(10,12,14)	76.8
	(4,6,8,10)	(1,3,5)	(3,5,7)	(5,7,9)	(7,9,11)	76.4

validation and 1,456 images for testing. The Semantic Boundaries Dataset is also used [31] to augment the training dataset, which results in 10,582 training images.

Ablation study for a single hierarchical structure: As stated in Sec. 4.1, for increasing the sampling density to avoid the missing of valuable information which is caused by the hollow kernels adopted in ASPP, the hierarchical structure with multiple levels of atrous convolution layers is adopted. For evaluating the effect of our multi-level structure in presenting the local noticeable detail feature, we compare the performances of ASPP and the single hierarchical structure. In our hierarchical structure, four root layers are adopted, and three parallel layers are attached at the end of each root layer. Our experiment are respectively performed on ResNet-50 and ResNet-101, and the experiment results are shown in Tab. 1. It can be seen that utilizing the hierarchical structure can significantly improve the performance compared with ASPP. For more intuitively demonstrate the effect of our hierarchical structure in preserving the local detail characteristics, the prediction results are presented in Fig. 11 It can be seen that after adopting our hierarchical structure, many important lo-

cal features which are destroyed by ASPP can be satisfyingly represented. For instance, the masks of the legs of the ostriches, which are intact in the red rectangles in the second images of column (a), are rather completely restored in the third image of column (a).

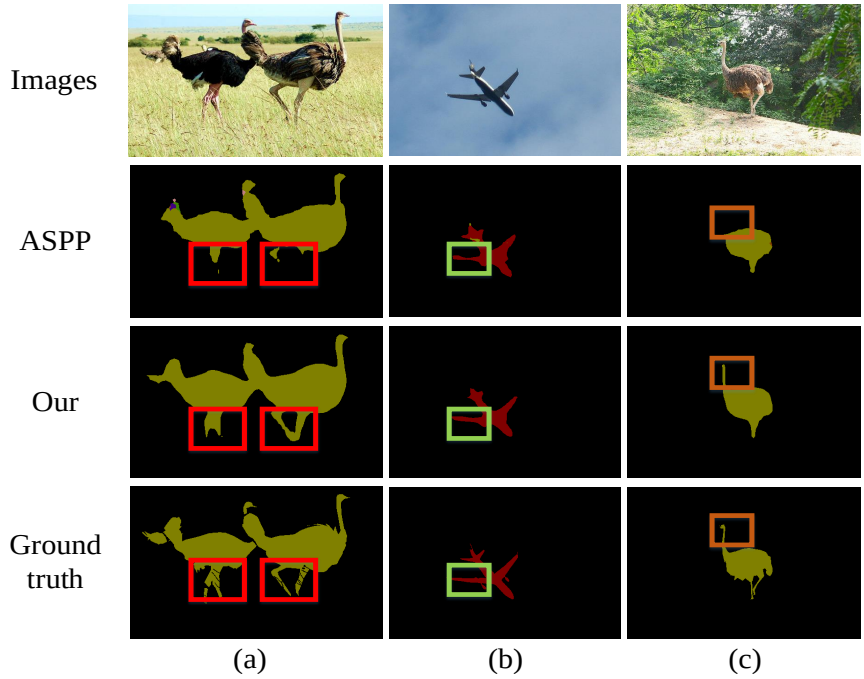

Figure 11: Comparisons between the performances of ASPP and single hierarchical structure.

Ablation study for a cascaded hierarchical structure: For demonstrating the merits of expanding the variety of sampling ranges, the performances of the single hierarchical structure and the cascaded hierarchical structure are compared. The results in Tab. 2, and the predicted results are shown in Fig. 12 It can be seen that relying on the large variety of the receptive fields of our cascaded structure, much important global information can be effectively sampled, which obviously improves the capability of the segmentation network in the discriminating local patches with the similar appearances but different labels. For example, in column (a), the properties of the legs of the cows are similar to those of the legs of the horse, which can easily result in that the

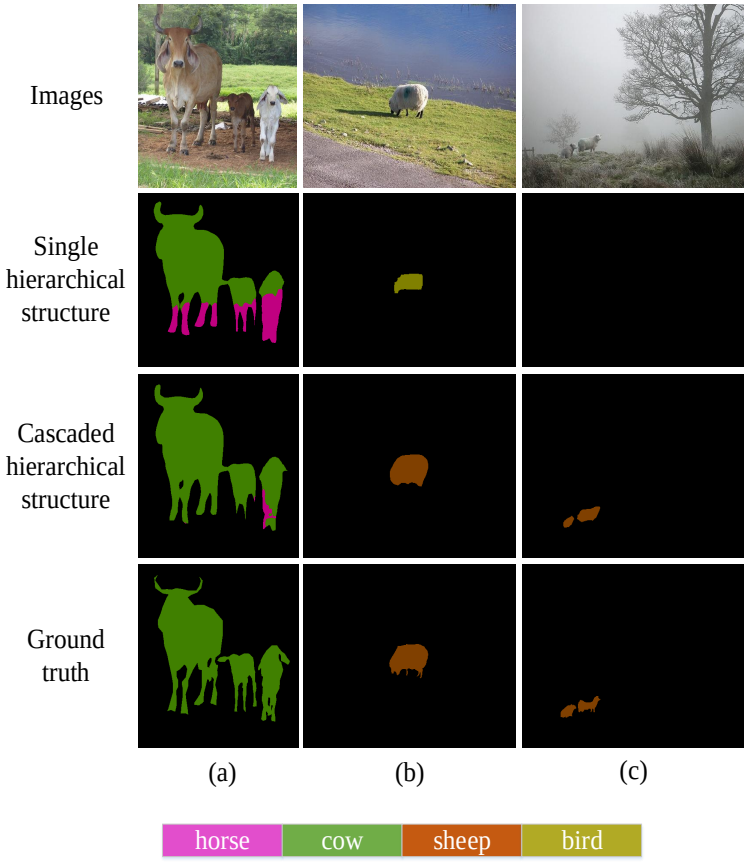

Figure 12: Comparisons between the performances of a single hierarchical structure and cascaded hierarchical structures

network mistake the cow legs for the horse legs (the pink area in the second image of column (a)). With the global information of the cows collected by the cascaded hierarchical structure, the network is not confused and can correctly distinguish these two kinds of components. Besides, with the contextual information collected by the cascaded structure with the enlarged variety of receptive fields, many misclassifications can be avoided. For example, in the second image of column (b), due to the lack of contextual information caused by the limited variety of sampling ranges, the sheep is misclassified as a bird. In contrast, as the ground information of the grass around the sheep is sufficiently collected,

Table 2: Performance analysis of single hierarchical structure and cascaded hierarchical structure. In the table, the two cascaded structures are respectively denoted as 'Stu1' and 'Stu2'. The numbers in the brackets after 'r1-r4' respectively correspond to the dilated rates of branch layers attached the root layers from the first to the fourth.

Method	Stu1		Stu2		mIoU
	root	branch	root	branch	
Res101	(3,5,7,9)	r1:(1,2,4) r2:(4,6,8) r3:(6,8,10) r4:(10,12,14)	/	/	76.8
	(3,5,7,9)	r1:(1,2,4) r2:(4,6,8) r3:(6,8,10) r4:(10,12,14)	(3,5,7,9)	r1:(1,2,4) r2:(4,6,8) r3:(6,8,10) r4:(10,12,14)	77.8
	(4,6,8,10)	r1:(1,3,5) r2:(3,5,7) r3:(5,7,9) r4:(7,9,11)	(4,6,8,10)	r1:(1,3,5) r2:(3,5,7) r3:(5,7,9) r4:(7,9,11)	77.5

the network can correctly judge the label of the sheep(as shown in the third image).

Ablation study for the number of branch layers: As mentioned in Sec. 4.1, appending branch layers to the end of root layers can increase the number of effective sampling points and boost the sampling of local delicate features, whether more effective sampling points can necessarily bring the gain in performance will be analyzed. In this part, two groups of experiments are performed. In each group, two and three branch layers are respectively added to each root layer, and their performances will be compared in order to analyze the impact when varying the number of branch layers. Our experiments are based on ResNet101, and the results are listed in Tab. 3. It can be observed that further performance gain can be achieved when the number of branch layers are increased, which again verifies that raising the sampling density can improve

Table 3: Performance analysis of the influence of using different number of branch layers.

Group	Stu1		Stu2		mIoU
	root	branch	root	branch	
1	(2,4,6,8)	r1:(2,5) r2:(4,8) r3:(7,11) r4:(10,14)	(2,4,6,8)	r1:(2,5) r2:(4,8) r3:(7,11) r4:(10,14)	77.4
	(2,4,6,8)	r1:(2,3,5) r2:(4,6,8) r3:(7,9,11) r4:(10,12,14)	(2,4,6,8)	r1:(2,3,5) r2:(4,6,8) r3:(7,9,11) r4:(10,12,14)	77.8
2	(1,3,5,7)	r1:(2,6) r2:(4,8) r3:(8,13) r4:(11,15)	(1,3,5,7)	r1:(2,6) r2:(4,8) r3:(8,13) r4:(11,15)	77.2
	(1,3,5,7)	r1:(2,4,6) r2:(4,6,8) r3:(8,11,13) r4:(11,13,15)	(1,3,5,7)	r1:(2,4,6) r2:(4,6,8) r3:(8,11,13) r4:(11,13,15)	77.7

the performance

Ablation study for the number of cascaded hierarchical structures: In the former experiment, two cascaded hierarchical structures can more comprehensively exploit the global or contextual information to improve the performances. For exploring the impact of the number of cascaded hierarchical structures on the performance, two group of experiments are performed. In each group, the dilated rates of the root layers and branch layers of a hierarchical structure are fixed, and utilize this structure as the constituent unit to form the cascaded structure with different number of hierarchical structures. In the first group, the dilated rates of root layers are set to (3,6,9,12) and the dilated rates of corresponding branch layers are set to ((1,3,5),(4,6,8),(7,9,11),(10,12,14)) While in the second group, the two combinations of dilated rates are respectively

Table 4: Performance analysis of impact of different numbers of hierarchical structures on the predicted results.

Group	Stu1	Stu2	Stu3	Stu4	Stu5	mIoU
1	✓	✓	/	/	/	78.2
	✓	✓	✓	/	/	78.3
	✓	✓	✓	✓	/	77.9
	✓	✓	✓	✓	✓	77.9
2	✓	✓	/	/	/	77.8
	✓	✓	✓	/	/	78.4
	✓	✓	✓	✓	/	78.1
	✓	✓	✓	✓	✓	78.0

(1,5,9,13) and ((1,4,7),(6,9,12),(11,14,17),(16,19,22)). The results of these two groups are listed in Tab. 4. As listed in Tab. 4, with the increasing of the number of cascaded hierarchical structures, the performances change slightly. This may be caused by that too many hierarchical structures make the receptive fields reach out of the range of the features map, which cause many useless samplings.

Results on PASCAL VOC 2012 test set: Finally, the performance of our cascaded hierarchical pyramid pooling structure on the PASCAL VOC 2012 test set is evaluated. As the annotation quality of PASCAL VOC 2012 dataset is higher than the augmented dataset [31], after training our network on the *train* and *val* sets of the augmented set, our trained network is further finetuned on *trainval* set of the original PASCAL VOC dataset. The results are listed in Tab. 5. From the results in Tab. 5, it can be seen that our method can outperform many current semantic segmentation methods, which demonstrates the high performance of our cascaded hierarchical structure. The comparisons with other methods are shown in Fig. 13.

To further improve the performance of our method, we utilize the model pretrained on COCO dataset. The comparison results are shown in Tab. 7. It can be seen that compared with the current methods, the performance of our method

Table 5: Results on PASCAL VOC 2012 test set without COCO pre-training

	DeepLabv2 [11]	DPN [32]	Piecewise [14]	Our
aeroplane	84.4	87.7	90.6	89.6
bicycle	54.5	59.4	37.6	42.7
bird	81.5	78.4	80.0	93.3
boat	63.6	64.9	67.8	71.5
bottle	65.9	70.3	70.4	80.3
bus	85.1	89.3	92.0	92.4
car	79.1	83.5	85.2	92.5
cat	83.4	86.1	86.2	93.5
chair	30.7	31.7	39.1	33.9
cow	74.1	79.9	81.2	91.8
diningtable	59.8	62.6	58.9	69.3
dog	79.0	81.9	83.8	91.8
horse	76.1	80.0	83.9	91.8
motorbike	83.2	83.5	84.3	89.8
person	80.8	82.3	84.8	86.9
pottedplant	59.7	60.5	62.1	73.8
sheep	82,2	83.2	83.2	91.3
sofa	50.4	53.4	58.2	58.1
train	73.1	77.9	80.8	86.3
tvmonitor	63.7	65.0	72.3	77.0
mIoU	71.6	74.1	75.3	81.0

is higher, which further demonstrates the superiority of our method. Besides, we replace the pyramid pooling module of DeeplabV3+ with our prediction module at the end of the backbone network. It can be observed that on top of the architecture utilized in DeeplabV3+, the performance of our method can be further improved, which proves the effectiveness of our method.

Apart from comparing the performances with current methods, we also evaluate the parameter size of our method, and summarize the parameter sizes and mIoU of different methods in Tab. 8. It can be seen that as our network utilizing

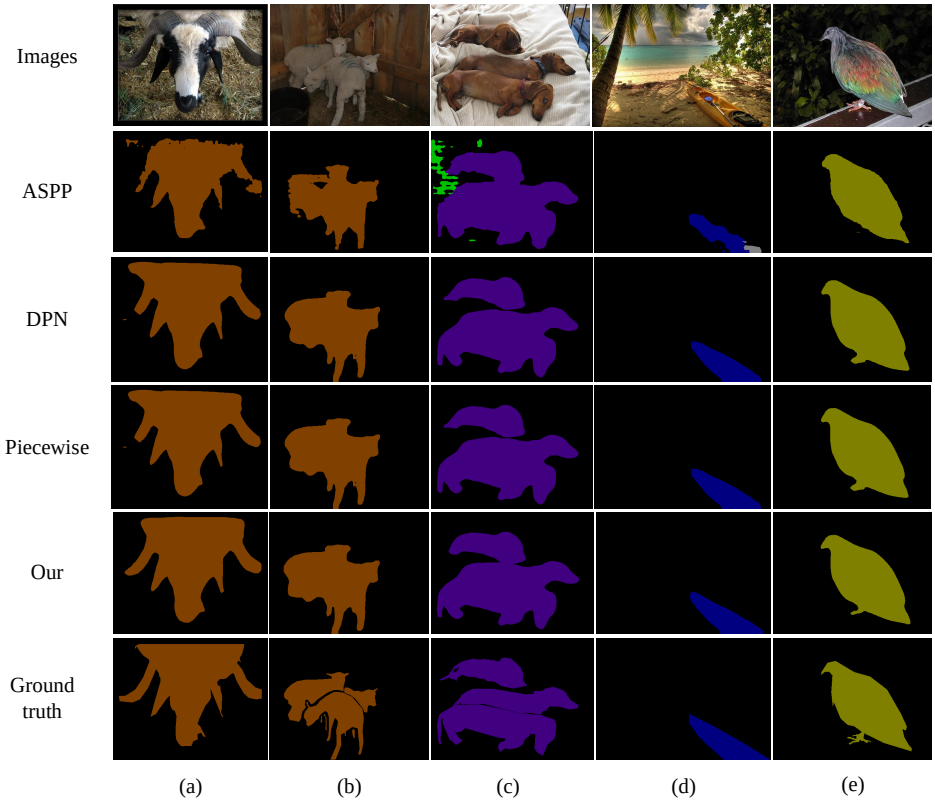

Figure 13: Comparisons between the performances of different methods on PASCAL VOC 2012 val set.

more parallel branches and cascaded structures, the parameters are apparently more than those of ASPP

5.3. Cityscapes

For further evaluating the performance of our method, our cascaded hierarchical structure is tested on the Cityscape dataset, which is a widely used benchmark for evaluating the capabilities of segmentation networks in understanding the street scene. This dataset is composed of 5,000 finely annotated street scene images of which 2,975 images are for training, 500 images for validation and 1,525 images for testing. Apart from these 5,000 finely annotated images, this dataset also contain another about 20,000 coarsely annotated im-

Table 6: Results on PASCAL VOC 2012 test set. Our model is trained on 8 Tesla V100 GPUs

Method	PSPnet [27]	DFN [23]	EncNet [36]	AAF [44]	Our
mIoU(%)	82.6	82.7	82.7	82.2	83.0

Table 7: Results on PASCAL VOC 2012 test set. Methods pre-trained on MS-COCO are marked with $^+$. Our model is trained on 8 Tesla V100 GPUs

Method	DLC$^+$ [39]	DUC$^+$ [26]	GCN$^+$ [22]	PSPnet$^+$ [27]
Mean IoU(%)	82.7	83.1	83.6	85.4

Method	DeeplabV3$^+$ [25]	Our
Mean IoU(%)	85.7	85.7

ages. The performance our method in this dataset is listed in Tab. 9, and the segmented results are shown in Fig. 14. It can be seen that our method outperforms the current segmentation methods. Furthermore, the whole architecture of Deeplabv3+ is utilized, and the pyramid module of DeeplabV3+ is replaced by our prediction module. Such a architecture on top of DeeplabV3+ can further enhance the performance of our method.

6. Conclusion

In this paper, we propose a new cascaded hierarchical atrous spatial pyramid pooling module for semantic segmentation. This module contains two aspects of improvements over ASPP: One is the single hierarchical structure which increases the sampling density through multi-level atrous convolution layers. The other one is that we cascade multiple hierarchical structures to significantly enlarge the variety of sampling ranges via extensively collecting the global and contextual information. However, the downside of our method is that the added parallel branches and the cascaded structure inevitably increase the computational costs, thus hindering our method from being applied to real-time applications.

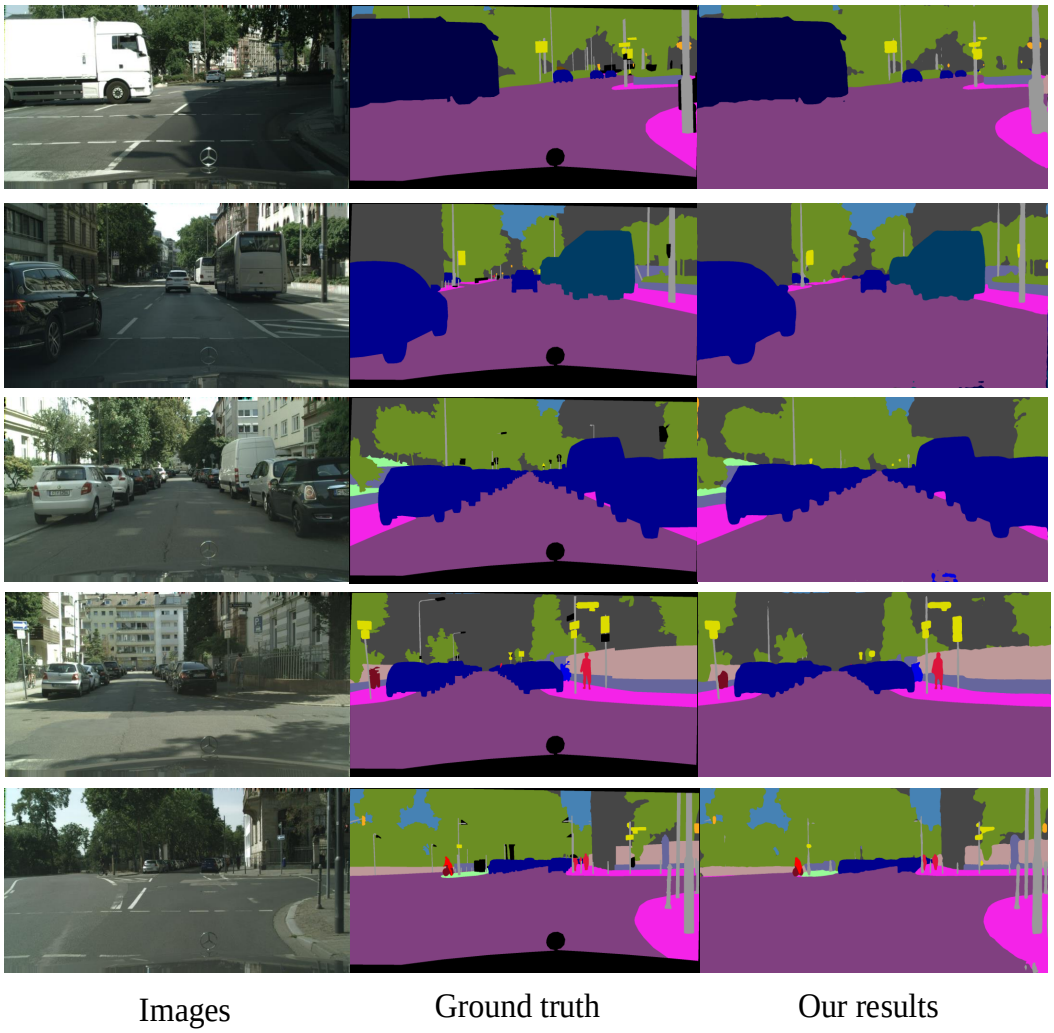

Figure 14: Comparisons between the performances of different methods on PASCAL VOC 2012 val set.

Table 8: Comparisons of parameter amounts between our method and current methods

Method	RefineNet [46]	DeepLab [11]	PSPNet [27]	Dilation-8 [34]	FCN-8s [8]	Our
# Params(M)	42.6	44.04	65.7	141.13	134.5	184.2
mIoU	82.4	79.7	85.4	75.3	67.2	85.7

Table 9: Results on Cityscapes set

Method	FCN [8]	DeepLab [11]	Piecewise [14]	AAF [44]	PSANet [42]
Mean IoU(%)	65.3	70.4	71.6	79.1	80.1

Method	DFN [23]	PSPnet [27]	DenseASPP [41]	Our
Mean IoU(%)	80.3	80.2	80.2	80.9

Our method is particularly suitable for tackling scenes in which there exist large variance among the sizes of objects, such as street scenes and road scenes. Besides, due to the dense sampling distribution, our method shows the superiority in segmenting objects which own many delicate details. Applications related to these two kinds of scenes can fully benefit from the advantages of our method.

The future work will be concentrated on two points. One is to simplify our CHASPP module in order to reach a high segmentation speed on the premise of not sacrificing too much prediction precision. The second is to develop a new segmentation network in which our simplified module can be attached to the end of each block of the backbone network and there are connections between each pair of these simplified modules. Such a network architecture can not only fully exploit the features of each block, but also realize the effective fusion of features of different blocks.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China(Gtant No. 61632018) and the National Key R&D Program of China(Grant Nos. 2018AAA0102800 and 2018AAA0102802)

References

[1] Y. Wei, X. Liang, Y. Chen, X. Shen, MM. Cheng, J. Feng, Y. Zhao, S. Yan, "STC: A simple to complex framework for weakly-supervised semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 11, pp. 2314-2320, 2017.

[2] Y. Wei, J. Feng, X. Liang, MM. Cheng, Y. Zhao, S. Yuan, "Object region mining with adversarial erasing: A simple classification to semantic segmentation approach," *in Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017.

[3] Y. Wei, X. Liang, Y. Chen, Z. Jie, Y. Xiao, Y. Zhao, S. Yan, "Learning to segment with image-level annotations," *Pattern Recognit.*, vol. 59, pp. 234-244, 2016.

[4] Z. Zhang, Y. Pang, "CGNet: cross-guidance network for semantic segmentation," *Sci. China Inf. Sci.*, vol. 63, 2020.

[5] S. Ma, Y. Pang, "Preserving details in semantic-aware context for scene parsing," *Sci. China Inf. Sci.*, vol. 63, 2020.

[6] Y. Pang, X. Bai, G. Zhang, "Special focus on deep learning for computer vision," *Sci. China Inf. Sci.*, vol. 62, 2020.

[7] P. Zhang, W. Liu, H. Wang, Y. Lei and H. Lu, "Deep gated attention networks for large-scale street-level scene segmentation," *Pattern Recognit.*, vol. 88, pp. 702-714, 2018.

[8] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," *in Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015.

[9] K. Li, W. Tao, X. Liu and L. Liu, "Iterative image segmentation with feature driven heuristic four-color labeling," *Pattern Recognit.*, vol. 76, pp. 69-79, 2018.

[10] L. Bi, J. Kim, A. Euijoon, K. Ashnil, D. Feng and F.Michael, 'Step-wise integration of deep class-specific learning for dermoscopic image segmentation," *Pattern Recognit.*, vol. 85, pp. 78-89, 2019.

[11] LC. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834-848, 2018.

[12] LC. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected crfs," *in Proc. Int. Conf. Learn. Representation.*, 2015.

[13] C. Farabet, C. Couprie, L. Majman, and Y. LeCun, "Learning hierarchical features for scene parsing," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1915-1929, 2013.

[14] G. Lin, C. Shen, A. Van Den Hengel, and I. Reid, "Efficient piecewise training of deep structured models for semantic segmentation," *in Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016.

[15] PH. Pinheiro, and R. Collobert, "Recurrent convolutional neural networks for scene parsing," [Online]. Available: https://arxiv.org/abs/1306.2795.

[16] LC. Chen, Y. Yang, J. Wang, W. Xu, and AL. Yuille, "Attention to scale: Scale-aware semantic image segmentation," *in Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016.

[17] G. Papandreous, I. Kokkinos, and PA. Savalle, "Modeling global and local deformations in deep learning: Epitomic convolution, multiple instance learning, and sliding window detection," *in Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015.

[18] I. Kokkinos, "Pushing the boundaries of boundary detection using deep learning," *in Proc. Int. Conf. Learn. Represent.*., 2016.

[19] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," *in Proc. Int. Conf. Med. Image Comput.-Assisted Intervention*, 2015.

[20] B. Hariharan, P. Arbelez, and R. Girshick, "Hypercolumns for object segmentation and fine-grained localization," *in Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015.

[21] G. Lin, A. Milan, C. Shen, and ID. Reid, " RefineNet: Multi-path refinement networks for high-resolution semantic segmentation," *in Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017.

[22] C. Peng, X. Zhang, G. Yu, G. Luo, and J. Sun, " Large kernel matters: Improve semantic segmentation by global convolutional network," *in Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017.

[23] C. Yu, J. Wang, C. Peng, C. Gao, G, Yu, and N. Sang, " Learning a discriminative feature network for semantic segmentation," *in Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018.

[24] H. Ding, X. Jiang, B. Shuai, A. Liu, and G. Wang, " Context contrasted feature and gated multi-scale aggregation for scene segmentation," *in Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018.

[25] LC. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," [Online]. Available: https://arxiv.org/abs/1706.05587.

[26] P. Wang, P. Chen, Y. Yuan, D. Liu, Z. Huang, X. Hou, G. Cottrell, "Understanding convolution for semantic segmentation," [Online]. Available: https://arxiv.org/abs/1702.08502.

[27] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, " Pyramid scene parsing network," *in Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018.

[28] K. Simonyan, and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *in Proc. Int. Conf. Learn. Representation.*, 2009.

[29] K. He, X. Zhang, S. Ren, and J. Sun, “ Deep residual learning for image recognition,” *in Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016.

[30] A. Krizhevsky, I. Sutskever, and GE. Hinton, ”Imagenet classification with deep convolutional neural networks,” *in Proc. Neural Inf. Process. Syst*, 2012.

[31] B. Hariharan, P. Arbeláez, L. Bourdev, S. Maji and J. Malik, ”Semantic contours from inverse detectors,” *in Proc. Int. Conf. Comput. Vis.*, 2011.

[32] Z. Liu, X. Li, P. Luo, C. C. Loy, and X. Tang, ”Semantic image segmentation via deep parsing network,” *in Proc. Int. Conf. Comput. Vis.*, 2015..

[33] I. Kreso, D. Causevic, J. Krapac and S. Segvic, ”Conditional scale variance for semantic segmentation,” *in Proc. German Conf. Pattern Recognit.*, 2016

[34] F. Yu and V. Koltun, ”Multi-scale context aggregation by dilated convolution,” [Online]. Available: https://arxiv.org/abs/1511.07122.

[35] G. Ghiasi and CC. Fowlkes, ”Laplacian pyramid reconstruction and refinement for semantic segmentation,” *in Pro. Eur. Conf. on Comput. Vis.*, 2016.

[36] H. Zhang, K. Dana, J. Shi, Z. Zhang, X. Wang and A. Tyagi, ”Context encoding for semantic segmentation,” *in Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018.

[37] TW. Ke, JJ. Hwang, Z. Liu and SX.Yu, ”Adaptive affinity fields for semantic segmentation,” *in Proc. Eur. Conf. Comput. Vis.*, 2018.

[38] J. Cao, Y. Pang and X. Li, ”Triply supervised decoder networks for joint detection and segmentation,” *in Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019.

[39] X. Li, Z. Liu, P. Luo, C. C. Loy and X. Tang, "Not all pixels are equal: difficulty-aware semantic segmentation via deep layer cascade," *in Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017.

[40] LC. Chen, Y. Zhu, G. Papandreou, F. Schroff and H. Adam, " Encoder-decoder with atrous separable convolution for semantic segmentation," [Online]. Available: https://arxiv.org/pdf/1802.02611v1

[41] M. Yang, K. Yu, C. Zhang, Z. Li and K. Yang, "DenseASPP for semantic segmentation in street scenes," *in Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019.

[42] H. Zhao, Y. Zhang, S. Liu, J. Shi, CC. Loy, D. Lin and J Jia, "Psanet: Pointwise spatial attention network for scene parsing," *in Proc. Eur. Conf. Comput. Vis.*, 2018.

[43] S. Mehta, M. Rastegari, A. Caspi, L. Shapiro and H. Hajishirzi, "ESPNet: Efficient spatial pyramid of dilated convolutions for semantic segmentation," *in Proc. Eur. Conf. Comput. Vis.*, 2018.

[44] TW. Ke, JJ. H, Z. Liu, SX. Yu. "Adaptive affinity fields for semantic segmentation," *in Proc. Eur. Conf. Comput. Vis.*, 2018.

[45] V. Badrinarayanan, A. Kendall, R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481-2495, 2017.

[46] G. Lin, A. Milan, C. Shen and I. Reid, "Refinenet: Multi-path refinement networks for highresolution semantic segmentation," *in Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017.