Multilevel feature fusion dilated convolutional network for semantic segmentation

Tao Ku1,2, Qirui Yang1,2,3 and Hao Zhang1,2

Abstract
Recently, convolutional neural network (CNN) has led to significant improvement in the field of computer vision, especially the improvement of the accuracy and speed of semantic segmentation tasks, which greatly improved robot scene perception. In this article, we propose a multilevel feature fusion dilated convolution network (Refine-DeepLab). By improving the space pyramid pooling structure, we propose a multiscale hybrid dilated convolution module, which captures the rich context information and effectively alleviates the contradiction between the receptive field size and the dilated convolution operation. At the same time, the high-level semantic information and low-level semantic information obtained through multi-level and multi-scale feature extraction can effectively improve the capture of global information and improve the performance of large-scale target segmentation. The encoder–decoder gradually recovers spatial information while capturing high-level semantic information, resulting in sharper object boundaries. Extensive experiments verify the effectiveness of our proposed Refine-DeepLab model, evaluate our approaches thoroughly on the PASCAL VOC 2012 data set without MS COCO data set pretraining, and achieve a state-of-art result of 81.73% mean interaction-over-union in the validate set.

Keywords
Semantic segmentation, convolutional neural network, deep learning, computer vision, robot vision

Date received: 5 March 2021; accepted: 16 March 2021

Introduction
Semantic segmentation, which aims at predicting pixel-level semantic labels for an image, provides detailed semantic annotation of the surrounding environment for the robot1–3 and is a fundamental topic in computer vision. The ability to interpret scenarios is an important capability for robots interacting with the environment. Leveraging the strong capability of convolutional neural network (CNN), which have been widely and successfully applied to image classification,4–6 most state-of-the-art works have made significant progress on semantic segmentation7–13 of robot to tackle the challenges (e.g. reduced feature resolution and objects at multiple scales) in CNN-based semantic segmentation. Researchers propose various network architectures. For example, the application of atrous convolutions14–18 while keeping the number of backbone network parameters...
unchanged, and the convolution layer receptive field of each stage unchanged, the higher convolution layer can also maintain a large feature map size to facilitate small targets. The detection improves the overall performance of the model. In general, CNN often faces three challenges in image semantic segmentation: (1) reduced feature resolution, (2) the existence of multiscale objects, and (3) reduced positioning accuracy due to CNN invariance.

Dilated convolution increases the receptive field and maintains the resolution of the feature map by injecting holes into the standard convolution (max-pooling or strided convolution will reduce the feature map resolution). Compared to the original standard convolution, the dilation convolution has a hyperparameter called dilation rate, which refers to the number of intervals of the convolution kernel (e.g. standard convolution is dilatation rate = 1). Although the dilated convolution eases the contradiction between the resolution of the feature map and the size of the receptive field, it still faces many problems. For example, the neurons in the feature map have the same receptive field, which means that the semantic mask generation process only utilizes features on a single scale. However, experiences show that multiscale information would help resolve ambiguous cases and result in more robust classification. Similarly, as the resolution of feature map decreases, when the dilation rate increases, the effect of dilated convolution will be ineffective in the dilated convolution framework. This problem is defined as “gridding issue.” Since the hole convolution is a discrete sample on the feature map, there is a lack of correlation. However, in the case of a continuous stack of hole convolution, the convolution kernel is not continuous, that is, not all pixels are used for calculation. So, the way information is treated as a checkerboard here loses the continuity of the information. This is called the “gridding” problem. In this work, we propose a multiscale hybrid dilated convolution method similar to pyramid pooling to solve this problem.

In general, low-level features and high-level features complement each other; low-level feature space information is rich but lacks semantic information and high-level features are the opposite of low-level features. However, many excellent models only use the highest layer features. DeepLabV3 finely tunes the extracted highest layer features through the atrous spatial pyramid pooling (ASPP) module to obtain excellent results. DeepLabV3+ uses the lowest layer feature to improve the performance of segmentation. DFANet proposed to extract semantic features by subnetwork and substation cascade aggregation, starting with a single lightweight backbone network. So, the use of multilevel features is necessary. Therefore, we use the feature information between different levels in the backbone network to perform feature aggregation to achieve the effect of multilevel feature utilization. At the same time, we consider the resolution of each level feature map and design multiscale hybrid dilated convolution, which alleviates the “grid” problem (Figure 2) caused by dilated convolution.

Capturing multiscale context information is critical to semantic segmentation tasks. SSD uses a feature map of different stride as detection layers to detect targets at different scales, users can make plans based on the target scale of their tasks. PSPNet proposes a pyramid pooling module that can aggregate context information of different grid scales, thereby improving the ability to obtain global information; DeepLabV3+ uses dilated convolution with different dilation ratios to form a hole pyramid pooling model (called ASPP), even if rich semantic information is encoded in the final feature map, due to the pooling and striding operations of the backbone network convolution lacks the ability to extract target boundary information. Therefore, DeepLabV3+ proposes to fuse the final feature map information with the lower layer information by combining the multiscale context information coding network to obtain a more detailed segmentation effect. However, some previous works neglect the role of the backbone network. The effective use of the information of each level of the backbone network can reduce the size of the model and improve the segmentation performance. Finally, we demonstrate the effectiveness of the proposed model on the PASCAL VOC 2012 data set to achieve excellent results.

We summarize our main contributions as follows:

- We propose a multilevel extraction feature module, which greatly utilizes the feature extraction capability of the backbone network;
- A multiscale hybrid dilated convolution module has been introduced into our structure to alleviate the “gridding” problem;
- Our proposed model exceeded the results of DeepLabV3+ on PASCAL VOC 2012 without MS COCO pretraining.

In the next sections, in related works, method, and experiments, we will expand from three perspectives: multiscale hybrid dilated convolution, multiscale feature aggregation, and decoder/encoder.

**Related works**

Recently, driven by intelligent manufacturing technology and unmanned driving, semantic segmentation has significantly improved the understanding performance of robot scene. Especially, fully convolutional network (FCN)-based approaches achieve promising performance on scene parsing and semantic segmentation task. Driven by powerful deep CNNs, pixel-level prediction tasks such as scene parsing and semantic segmentation achieve great progress inspired by replacing the fully connected layer in classification with the convolution layer.

To enlarge the receptive field of CNN, use dilated convolution. In recent years, to solve computer vision...
problems more effectively, multiscale features are widely used in computer vision tasks. At the same time, multiscale features are another important factor in segmentation tasks. Based on the spatial pyramid structure,\textsuperscript{34} PSPNet\textsuperscript{18} and ASPP\textsuperscript{13} are fusing multiple multiscale feature maps to form a pyramid structure for final prediction, where PSPNet\textsuperscript{18} employs four spatial pyramid pooling (downsampling) layers in parallel to aggregate information from multiple receptive field sizes and assigns to each pixel via up-sampling. In this work, we mainly discuss about multiscale hybrid dilated convolution, multilevel feature aggregation, and the structure of the encoder–decoder.

### Multiscale hybrid dilated convolution

The multiscale feature plays a key role in semantic segmentation, especially for objects/stuff with a vast variation of scales. SegNet,\textsuperscript{25} UNet,\textsuperscript{26} and DeepLabV3\textsuperscript{10} design encoder–decoder architecture are used to fuse low-level and high-level feature maps from encoder and decoder, respectively. Spatial pyramid pool\textsuperscript{18} uses several parallel dilated convolution to extract multiscale information. ASPP\textsuperscript{13} concatenates feature maps with different dilation rates so that the output feature map aggregates multiscalar semantic information, which can improve segmentation performance. ASPP in DeepLabV2\textsuperscript{10} uses dilated convolution with four different sampling rates in the top map of features. Deeplabv\textsuperscript{13} adds batch normalization\textsuperscript{27} to ASPP. Dilated convolution with different sampling rates can effectively capture multiscale information, but it will be found that as the sampling rate increases, the effective weight of the filter (the weight is effectively applied to the feature area instead of filling 0) gradually becomes smaller. MFCAN\textsuperscript{28} uses dilated convolution instead of standard convolution, combines feature maps of different levels, and aggregates information from different receiving fields. Through multiscale information extraction, the model shows effective results on several segmentation benchmarks. However, while the PSPNet\textsuperscript{18} blindly enlarges the receptive field of neural networks, with the reduction of receptive field, as the dilation rate increases, the dilated convolution becomes more and more ineffective and gradually loses its modeling power. We use this feature to combine the resolution of each layer feature map to propose a multiscale hybrid dilated convolution module instead of the space pyramid pooling module.

### Multilevel feature aggregation

With the deeper convolutional network, the resolution of feature maps is decreasing to capture high-level semantic information, but the application of low-level information is often ignored. DeepLabV3+\textsuperscript{22} aggregates lower-level information with higher-level information for finer segmentation. DUpsampling\textsuperscript{29} spliced the highest layer of the backbone network with the second layer and achieves new state-of-the-art performance. MLFFNet\textsuperscript{11} implements feature reuse through multilevel feature fusion, which greatly reduces computational complexity. To alleviate these issues in video crowd counting, a multilevel feature fusion-based locality-constrained spatial transformer (LST) network\textsuperscript{30} is proposed, which consists of two components, namely, density map regression module and LST module. RFE\textsuperscript{31} introduces multilevel feature fusion into super-resolution reconstruction tasks and achieves excellent performance. However, the capability of feature extraction at all levels of the backbone network has not been properly applied. We have found that the segmentation effect of large target objects is not ideal. In our work, we use hierarchical aggregation to make efficient use of each level of features. Experiments have shown that the segmentation effect is effectively improved. At the same time, when using multiscale mixed dilated convolution modules for different levels, we comprehensively consider the size of the feature map and use different dilation ratios to greatly alleviate the gridding problem caused by dilated convolution.

### Encoder–decoder

To obtain more fine-grained performance, encoder–decoder networks are widely applied to many computer vision tasks and have been successfully applied in semantic segmentation tasks. The encoder is usually a pretrained classification network, such as VGG\textsuperscript{6} and ResNet.\textsuperscript{32} The decoder projects the recognition features (low resolution) learned by the encoder onto the pixel space (high resolution) semantics to obtain dense classification. DeconvNet\textsuperscript{13} uses stacked deconvolutional layers in the decoding stage to gradually recover the full-resolution prediction gradually, but it is difficult to train because of many parameters introduced by the decoder. Through the index of the pooling layer in the encoding stage and the recovery process through the index in the decoding stage, SegNet\textsuperscript{25} significantly improves the performance of semantic segmentation. RefineNet\textsuperscript{9} systematically proved the effectiveness of the encoder–decoder based structure in the semantic segmentation model. Recently, DeepLabV3++\textsuperscript{22} combined a decoder–encoder with ASPP to achieve state-of-the-art segmentation performance on a few data sets to date. Although efforts have been spent on designing a better decoder, so far almost none of them can bypass the restriction on the resolutions of the fused features and exploit better feature aggregation.

Although each model achieves good results on multiple data sets through multiscale information extraction, the “gridding” problem caused by dilated convolution is rarely mentioned. At the same time, multiscale aggregation is used in many works, but the backbone network utilization rate is insufficient. Therefore, in this work, we propose to use multilayer feature aggregation to maximize the use of backbone network features. The introduction of multiscale
hybrid dilated convolution can effectively alleviate the “gridding” problem.

**Method**

We propose a new framework that can be applied to any backbone network. In this section, we will briefly introduce three modules: multiscale dilated convolution, multilevel aggregation module, and encoder and decoder.

**Multiscale dilated convolution**

Dilated convolution is first introduced by Holschneider et al.\textsuperscript{14} and can maintain the high resolution of the feature map by replacing the max-pooling or stride convolution, and at the same time, maintain the size of the corresponding layer of the receptive field to capture multiscale information. Dilated convolution is to inject holes in the standard convolution map to increase receptive field. Compared to the original normal convolution, the dilated convolution has a hyperparameter called dilation rate, which refers to the number of kernel intervals (e.g. normal convolution is dilation rate = 1). In one dimension, dilated convolution is defined as

\[
y[i] = \sum_{k=1}^{K} x[i + d \cdot k] \cdot w[k]
\]

where dilation rate \(d\) determines the stride of the sampled input signal. We refer interested readers to see the literature\textsuperscript{20} for more details. It is worth noting that when \(d = 1\), this formula is the standard convolution. Atrous convolution is equivalent to convolving the input \(x\) with upsampled filters produced by inserting \(d - 1\) zeros between two consecutive filter values. Therefore, a large dilation rate means a large receptive field. We can adaptively modify the size of the receptive field by changing \(d\). Two-dimensional dilated convolution is constructed by inserting holes (zeros) into each pixel in the convolution kernel. However, two theoretical issues exist in the above dilated convolution framework, and we call it “gridding” issue (Figure 2) and “long-ranged information might not be a relevant issue.” As shown in Figure 2, if \(k = 3\) and \(d = 2\), there are 100 pixels in the feature map of \(21 \times 21\) for calculation, the utilization rate is 22.68\%; when \(d = 3\), 49 pixels are used for calculation, and the utilization rate is 11.11\%; when \(d = 4\), there are 25 pixels for calculation. At this time, the utilization rate is only 5.67\%. In this case, the larger the dilated ratio, the smaller the feature map is, the more serious it will be. When the dilation ratio is larger, the smaller the feature map is, the effect of convolution becomes worse. Wang et al.\textsuperscript{20} point out that at higher layers, when the dilation rate \(d\) is large, the samples from the input may be very sparse, which is not conducive to learning. We found that the convolution kernel for dilated convolution is not continuous, so this operation method will lose the continuity of information. This is severe for intensive prediction tasks at the pixel level. These disadvantages may be irrelevant when the resolution of the low-level feature map is large. The problem with the grid is that the pixels in the \(d \times d\) region of the one-layer attachment receive information from a completely different grid set, which is not good for local information consistency.

To solve this problem, we combine the dilated convolution based on the image pyramid to propose a multiscale hybrid dilated convolution module different from hybrid dilated convolution (HDC).\textsuperscript{20} At the same time, we consider the dilated convolution of different dilated ratios by considering the resolution of each layer’s feature map and balance the contradiction between the receptive field and the feature map. It is worth noting that it is necessary to avoid the use of a common factor conversion factor in the same group. Figure 2 shows that when we use the dilated ratio of 2, 3, 4 to superimpose, the number of pixels used is 157, and the utilization rate is 35.60\%. However, it can be seen that some pixels are reused because the hole ratio is 2, 4. This is what we do not want to see.

To simplify notations, we use \(C_{k,d}(X)\) to term a dilated convolution, where \(K\) is the kernel size (KS), \(d\) is the dilation rate, \(X\) is the feature input, \(AP(1)\) is adaptive pooling, and consequently, multiscale dilated convolution (MDC) is written as

\[
MDC = C_{k,1}(X) + C_{k,2}(X) + C_{k,3}(X) + C_{k,4}(X)
\]

**Multilevel aggregation module**

The structure of Refine-DeepLab is shown in Figure 1. We divide the backbone network into four blocks. We control the resolution of the feature maps of the first block and the second block to be the same, and the output feature map size is \(129 \times 129\); the resolutions of the third block and the fourth block are consistent, and the feature map size is \(64 \times 64\). The MDC dilated ratio of each layer is adjusted according to the resolution of each layer’s feature map. The output of each atrous layer is concatenated with the input feature map, and the concatenated feature map is fed into the following layer. The semantic information at all levels in the backbone network is aggregated, and the information at all levels is extracted by multiscale hybrid dilated convolution for feature aggregation of semantic information at different levels. At the same time, the use of low-level information is not only conducive to the use of features but also integrates high-resolution information, and the boundary segmentation effect is more obvious.

**Network architecture**

The whole architecture is shown in Figure 1. In general, our semantic segmentation network could be seen as an encoder–decoder structure.
Encoder: We will integrate the resolution size of the output feature graph at all levels to design dilated convolution dilation rate. The purpose of this is to alleviate the generated “gridding” effect. When the size of feature graph is not large enough, if we adopt a convolution with a large dilated rate, the consistency of local information will be destroyed. As the dilation rate increases, the data used in the input will become more and more sparse, which is not conducive to convolution learning. When the local information is lost, the correlation of the remote information will also be destroyed. We designed the $MDC_1$ dilation rate $\{3, 5, 7, 17\}$, $MDC_2$ dilation rate $\{3, 5, 7, 11\}$, $MDC_3$ dilation rate $\{3, 5, 7, 11\}$, and $MDC_4$ dilation rate $\{1, 3, 5, 7\}$. We combine the feature maps of each block.
with $1 \times 1$ convolution for channel compression and feature maps after multiscale hybrid dilated convolution. In our proposed encoder–decoder structure, we use the last feature map before being output as an encoder in the DeepLabV3. Encoder output graph contains 256 channels and rich semantic information. To simplify notations, we use $F$ to represent the result of the feature map being fusion by multiscale mixed dilated convolution and feature map without MDC operation, where $D_1$, $D_2$, $D_3$, and $D_4$ are a dilated convolution group

$$y = F_1(D_1) + F_2(D_2) + F_3(D_3) + F_4(D_4)$$  \hspace{1cm} (3)$$

**Decoder:** Our proposed decoder module is shown in Figure 1. As discussed above, the encoder is an aggregation of four aggregation modules, composed with multiscale hybrid dilated convolution and low-layer features. For inference, we do not put too much focus on designing a complicated decoder module. According to DeepLabV3+,22 we propose to fuse high-level and low-level features directly. First, $F_3$ and $F_4$ are upsampled by a factor of 2, and fuse four $F$ modules. Because our encoder is composed of four multilevel aggregation modules, we firstly fuse high-level representation from the bottom of four multilevel aggregation modules. Then, the high-level features and low-level details are added together and upsampled by a factor of 2 to make the final prediction. In the decoder module, we only implement a few convolution calculations to reduce the number of channels.

### Experiments

The proposed models are evaluated on the PASCAL VOC 2012 semantic segmentation benchmark,23 PASCAL

| Backbone       | MDC             | mIOU (%) |
|----------------|-----------------|----------|
| ResNet-101     | $\{3\}$         | 65.31    |
| ResNet-101     | $\{1, 3, 3, 3\}$| 76.42    |
| ResNet-101     | $\{1, 6, 12, 18\}$| 77.21    |
| ResNet-101     | $\{1, 3, 7, 17\}$| 77.65    |
| ResNet-101     | $\{1, 3, 5, 11\}$| 77.89    |
| ResNet-101     | $\{1, 3, 5, 7\}$| 78.57    |

mIOU: mean interaction-over-union; MDC: multiscale dilated convolution.

**Table 2.** Exploring different dilation ratio combinations to alleviate grid problems based on improving model performance.

| Backbone       | $MDC_4$ | $MDC_3$ | $MDC_2$ | $MDC_1$ | Remark column | mIOU (%) |
|----------------|---------|---------|---------|---------|----------------|----------|
| ResNet-101     | $\{1, 6, 12, 18\}$ |         |         |         | DeepLabV310    | 75.36    |
| ResNet-101     | $\{1, 6, 12, 18\}$ |         |         | $\{1\}$ | DeepLabV3+22   | 78.85    |
| ResNet-101     | $\{1, 3, 7, 11\}$ | $\{1, 3, 7, 13\}$ | $\{1, 3, 7, 17\}$ | $\{1, 5, 11, 17\}$ |               | 79.41    |
| ResNet-101     | $\{1, 3, 7, 11\}$ | $\{1, 3, 5, 11\}$ | $\{1, 3, 5, 17\}$ | $\{1, 3, 11, 17\}$ |               | 80.39    |
| ResNet-101     | $\{1, 3, 5, 7\}$ | $\{1, 3, 5, 11\}$ | $\{1, 3, 5, 17\}$ | $\{1, 3, 7, 11\}$ |               | 81.73    |

mIOU: mean interaction-over-union; MDC: multiscale dilated convolution.
$MDC_4$ feature map ($1 \times 1$ convolution used to reduce the channels of the low-level feature map from the encoder module). We will concatenate through the channel-compressed feature map $f_4$ and $MDC_4$ to get the feature map $F_4$. The $F_4$ feature map is gradually upsampled by 2 and concatenated with feature map $f_1$ after channel compression. The model structure is shown in Figure 4. Select the principle of combining different dilation ratios:

- The receptive field after using dilated convolution should not exceed the size of the feature map;
- The dilation rate combination cannot have a common divisor greater than 1.

Therefore, we first select several combinations of dilation ratios with comparative significance to verify our conclusions. We used different dilated ratio ($MDC$) and convolution KS settings to study the performance of multiscale hybrid dilated convolution modules. The results are listed in Table 1. From the table, we get the following observations. First, all multiscale hybrid dilated convolutions significantly improve performance compared to the underlying backbone network. Second, $MDC \{1, 3, 5, 7\}$ multiscale mixed dilated convolution achieves the best results, benchmarking performance increased by 13.26%. From this, we can infer that

**Table 3.** The effect of each level of feature aggregation on the performance of the model.

| MDC | mIOU (%) |
|-----|----------|
| ✓   | 62.31    |
| ✓   | 75.36    |
| ✓   | 77.31    |
| ✓   | 78.27    |
| ✓   | 79.79    |
| mIOU: mean interaction-over-union; MDC: multiscale dilated convolution.

**Table 4.** Experimental results of multilevel aggregation network on other backbone networks.

| Model  | MobileNet | DRN  | Xception | ResNet |
|--------|-----------|------|----------|--------|
| DeepLab| 70.81     | 78.87| 78.58    | 78.43  |
| Ours   | 72.59     | 80.03| 80.28    | 81.73  |

**Figure 3.** Examples of visualization on the PASCAL VOC2012 segmentation validation set. Comparison with baseline method. DeepLab produces more accurate and detailed results. (a) Image, (b) ground truth, (c) DeepLabV1, (d) DeepLabV3, and (e) ours.
Figure 4. An overview of the single-level feature dilated convolution network.

Figure 5. Experimental results of multilevel aggregation networks on the PASCAL VOC2012 validation set based on different backbone networks. The first line is based on the DeepLab model, and the second line is based on our multilevel aggregation network. (a) MobileNet, (b) DRN, (c) Xception, and (d) ResNet.
multiscale mixed dilated convolution can effectively capture multiscale features and improve model performance. In all of the following experiments, we will consider the size of the feature map and design different scale dilation ratios for different levels and convolution kernels of different sizes.

**Multilevel aggregation network**

To further validate the effectiveness of our multilevel multiscale hybrid dilated convolution network, we use ResNet-101 as our backbone network. It can be seen from Table 2 that the multilevel multiscale hybrid dilated convolution network performance is far superior to the single-level multiscale hybrid dilated convolution. The reason is that the single-level multiscale hybrid dilated convolution often ignores the low-level features. This results in poor segmentation performance for large targets. To verify the multi-level module, we designed all multiscale hybrid dilated convolution modules with a dilation ratio of \( \{1, 3, 5, 7\} \). First of all, the backbone network does not go through other special direct predictions, FCN does just that. Then, we pass the highest layer of the backbone network through the hybrid dilated convolution (MDC) process similar to the DeepLabV3 method. We gradually increase the number of feature maps participating in feature aggregation. Some feature maps need to be bilinear upsampling interpolated to be able to splice. We consider the resolution of each feature block to design a reasonable combination of dilation ratios. Here, we control the feature map \( f_1 \), the resolution of the feature map \( f_2 \) is \( 129 \times 129 \), the feature map \( f_3 \), and the feature map \( f_4 \) is \( 65 \times 65 \). We can see the improvement of the different feature blocks in Table 3. In Figure 3, the visualization example comparison between Refine-DeepLab and several Deeplab methods in the PASCAL VOC2012 segmentation validation set. Considering the size of the feature map resolution and the noise of the low-level feature information, we use the multiscale convolution of the large dilation rate to extract the low-level feature. After extracting features at each level, the same feature map with the same size as the feature map is obtained. We use the \( 1 \times 1 \) convolution to reduce the low-level feature mapping channel to process the different levels. Then, we perform bilinear upsampling on feature map \( f_3 \) and feature map \( f_4 \), and the validity of the upsampling has been verified in the literature. The four-layer features of the lower level are feature aggregated to obtain a feature map of \( 1 \times 1024 \times 129 \times 129 \) size. At this time, the high-level semantic information and the obtained low-level feature information are used for the decoding phase. The experimental results are provided in Table 2.

| Model       | mIOU (%) |
|-------------|----------|
| FCN         | 62.2     |
| Zoom-out    | 69.6     |
| DeepLabV2   | 71.6     |
| DeconvNet   | 72.5     |
| GCRNet      | 73.2     |
| DPN         | 74.1     |
| DeepLabV3   | 79.77    |
| DeepLabV3+  | 79.93    |
| Ours        | 81.73    |

Table 5. The result is the result on the PASCAL VOC2012 validation set.a

mIOU: mean interaction-over-union.
aOur method is superior to all previous advanced methods, reaching 81.73% with no pretrained MS COCO data set.

Figure 6. Examples of visualization on the PASCAL VOC2012 segmentation validation set. (a) image, (b) ground truth, and (c) Refine-DeepLab.
To evaluate the PASCAL VOC 2012 val set, we set the multiscale hybrid dilated convolution to $MDC_1 : \{1, 3, 7, 11\}; MDC_2 : \{1, 3, 5, 11\}; MDC_3 : \{1, 3, 5, 11\};$ and $MDC_4 : \{1, 3, 5, 7\}$. The deep strategy is used to train the backbone model of the enhanced training set. Specially, to verify our migration ability of the multilevel aggregation network, we will experiment in several other commonly used backbone networks. The experimental results are given in Table 4, and the visualization results of various models tested on PASCAL VOC 2012 val set are shown in Figure 5. The backbone model is ResNet101 pretrained on ImageNet. Then, we finely tune the training model on the original training. A comparison with the most advanced methods is given in Table 5. In Figure 6, examples of visualization on the PASCAL VOC2012 segmentation validation set. Obviously, our Refine-DeepLab is significantly better than the other methods on PASCAL VOC 2012.

**Summary**

Compared to DeepLabV3 and PSPNet, our method achieves better results on PASCAL VOC 2012. These results show that Refine-DeepLab can effectively use multiscale hierarchical features and multiscale spatial information. Compared with DeepLabV3, we construct multiscale context information by maximizing the multilevel features extracted by the backbone network to make the global information affinity more reasonable and higher performance.

**Conclusion**

In this article, we discuss the properties of multiscale features and propose a multiscale context representation of multilevel multiscale hybrid dilated convolution for semantic segmentation and scene analysis. Refine-DeepLab introduces multiscale expansion convolution while capturing rich context information, which can effectively alleviate the contradiction between the size of the acceptance field and the expansion convolution operation. Through the aggregation of multiple layers of semantic information, high-level semantic information and low-level semantic features are extracted to aggregate global information. Through our well-designed dilation ratios, the “gridding” problem caused by dilated convolution is alleviated, and the segmentation effect of large-size targets is solved. At the same time, our proposed Refine-DeepLab can be embedded not only in any backbone network but also in any network layer independent of the size of the input feature map. Considering the limitations of computing resources, we did not perform pretraining on MS COCO, but our experimental results show that our multilevel feature fusion dilated convolution network can be extended to other scenarios. It is worth noting that Refine-DeepLab proves that the features of each layer of the backbone network are extremely effective in improving segmentation performance. However, when we are extracting multilevel features, how to effectively extract multilevel features while ensuring that the amount of calculation remains unchanged or reduced is a problem worthy of our further solution.

**Author contribution**

The authors TK and QY contributed equally.

**Declaration of conflicting interests**

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

**Funding**

The author(s) disclosed the receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the National Key Research and Development Program of China under grant no. 2020YFB1708503.

**ORCID iD**

Qirui Yang https://orcid.org/0000-0003-2208-3268

**References**

1. He B, Yanwei S, Wang S, et al. Product environmental footprints assessment for product life cycle. *J Clean Prod* 2019; 233: 446–460.

2. He B, Liu Y, Zeng L, et al. Product carbon footprint across sustainable supply chain. *J Clean Prod* 2019; 241: 118320.

3. He B, Wang S, and Liu Y. Underactuated robotics: a review. *Int J Adv Robot Syst* 2019; 16(4): 172988141986216.

4. Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, MA, USA, 7–12 June 2015, pp. 1–9.

5. Krizhevsky A, Sutskever I, and Hinton GE. ImageNet classification with deep convolutional neural networks. In: *International conference on neural information processing systems*. December 2012.

6. Simonyan K and Zisserman A. Very deep convolutional networks for large-scale image recognition. *Comput Sci* 2014.

7. Long J, Shelhamer E, and Darrell T. Fully convolutional networks for semantic segmentation. *IEEE Trans Pattern Anal Mach Intell* 2014; 39(4): 640–651.

8. Eigen D and Fergus R. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In: *2015 IEEE international conference on computer vision (ICCV)*. Santiago, Chile, 7–13 December 2015.

9. Lin G, Milan A, Shen C, et al. RefineNet: multi-path refinement networks with identity mappings for high-resolution semantic segmentation 2016.

10. Chen LC, Papandreou G, Schroff F, et al. Rethinking atrous convolution for semantic image segmentation 2017.
11. Wang L, Xu Q, Xiong Z, et al. A multi-level feature fusion network for real-time semantic segmentation. In: 2019 11th international conference on wireless communications and signal processing (WCSP). Xi’an, China, 23–25 October 2019, pp. 1–6.

12. Junior LADO, Medeiros HR, Macêdo D, et al. Segnetres-CRF: a deep convolutional encoder-decoder architecture for semantic image segmentation. In: 2018 International joint conference on neural networks (IJCNN), Rio de Janeiro, Brazil, 8–13 July 2018, pp. 1–6.

13. Chen LC, Papandreou G, Kokkinos I, et al. DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE Trans Pattern Anal Mach Intell 2018; 40(4): 834–848.

14. Holschneider M, Kronlandmartinet R, Morlet J, et al. A real-time algorithm for signal analysis with the help of the wavelet transform. In: Wavelets time-frequency methods & phase space. Berlin: Springer-Verlag. 1989.

15. Giusti A, Dan CC, Masci J, et al. Fast image scanning with deep max-pooling convolutional neural networks. In: 2013 IEEE international conference on image processing. Melbourne, VIC, Australia, 15–18 September 2013.

16. Tompson J, Goroshin R, Jain A, et al. Efficient object localization using convolutional networks. In: 2015 IEEE conference on computer vision and pattern recognition (CVPR), 7–12 June 2015. Boston: IEEE.

17. Papandreou G, Kokkinos I, and Savalle PA. Modeling local and global deformations in deep learning: epidemic convolution, multiple instance learning, and sliding window detection. In: IEEE conference on computer vision & pattern recognition. Boston, MA, USA, 7–12 June 2015.

18. Zhao H, Shi J, Qi X, et al. Pyramid Scene Parsing Network. IEEE Computer Society 2016.

19. Huang G, Liu Z, Maaten LVD, et al. Densely connected convolutional networks. In: IEEE conference on computer vision & pattern recognition. Honolulu, HI, USA, 21–26 July 2017.

20. Wang P, Chen P, Ye Y, et al. Understanding convolution for semantic segmentation. In: IEEE winter conference on applications of computer vision. Lake Tahoe, NV, USA, 12–15 March 2018.

21. Liu W, Anguelov D, Erhan D, et al. SSD: single shot multibox detector. In: European conference on computer vision. Amsterdam, The Netherlands, 11–14 October 2016.

22. Chen LC, Zhu Y, Papandreou G, et al. Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European conference on computer vision (ECCV). Munich, Germany, 8–14 September 2018, pp. 801–818.

23. Everingham M, Zisserman A, Williams CKI, et al. The 2005 pascal visual object classes challenge. In: International conference on machine learning challenges: evaluating predictive uncertainty visual object classification. Southampton, UK, 11–13 April 2005.

24. He J, Chang SF, and Xie L. Fast kernel learning for spatial pyramid matching. In: IEEE conference on computer vision & pattern recognition. Anchorage, AK, USA, 23–28 June 2008.

25. Badrinarayanan V, Kendall A, and Cipolla R. SegNet: a deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans Pattern Anal Mach Intell 2017; 39(12): 2481–2495.

26. Ronneberger O, Fischer P, and Brox T. U-Net: convolutional networks for biomedical image segmentation. In: International conference on medical image computing & computer-assisted intervention. Munich, Germany, 5–9 October 2015.

27. Ioffe S and Szegedy C. Batch normalization: accelerating deep network training by reducing internal covariate shift. In: International conference on machine learning, Lille, France, 1 June 2015.

28. Ren D, Li J, Li L, et al. Low dose ct image denoising using multi-level feature fusion network and edge constraints. In: 2019 IEEE international conference on bioinformatics and biomedicine (BIBM). San Diego, CA, USA, 18–21 November 2019.

29. Tian Z, Shen C, He T, et al. Decoders matter for semantic segmentation: data-dependent decoding enables flexible feature aggregation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019.

30. Fang Y, Gao S, Li J, et al. Multi-level feature fusion based locality-constrained spatial transformer network for video crowd counting. Neurocomputing 2020; 392: 98–107.

31. Jin X, Xiong Q, Xiong C, et al. Single image super-resolution with multi-level feature fusion recursive network. Neurocomputing 2019; 370: 166–173.

32. He K, Zhang X, Ren S, et al. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. Las Vegas, NV, USA, 27–30 June 2016, pp. 770–778.

33. Noh H, Hong S, and Han B. Learning deconvolution network for semantic segmentation. In Proceedings of the IEEE international conference on computer vision. Santiago, Chile, 7–13 December 2015, pp. 1520–1528.

34. Mottaghi R, Chen X, Liu X, et al. The role of context for object detection and semantic segmentation in the wild. In: Proceedings of the IEEE international conference on computer vision and pattern recognition. Columbus, OH, USA, 23–28 June 2014.

35. Paszke A, Gross S, Chintala S, et al. Automatic differentiation in pytorch. In: 31st conference on neural information processing systems (NIPS 2017), Long Beach, CA, USA, 2017.

36. Liu W, Rabinovich A, and Berg AC. ParseNet: Looking Wider to See Better. ICLR 2016.

37. Chen LC, Papandreou G, Kokkinos I, et al. Semantic image segmentation with deep convolutional nets and fully connected CRFs. Comput Sci 2014; (4): 357–361.

38. Harirahan B, Arbelaez P, Bourdev LD, et al. Semantic contours from inverse detectors. In: IEEE international conference on computer vision. Barcelona, Spain, 6–13 November 2011.
39. Yu F, Koltun V, and Funkhouser T. Dilated residual networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. Honolulu, HI, USA, 21–26 July 2017, pp. 472–480.

40. Chollet F. Xception: deep learning with depthwise separable convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition. Honolulu, HI, USA, 21–26 July 2017, pp. 1251–1258.

41. Sandler M, Howard A, Zhu M, et al. Mobilenetv2: inverted residuals and linear bottlenecks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. Salt Lake City, UT, USA, 18–23 June 2018, pp. 4510–4520.

42. Mostajabi M, Yadollahpour P, and Shakhnarovich G. Feedforward semantic segmentation with zoom-out features. In: 2015 IEEE conference on computer vision and pattern recognition (CVPR), Boston, MA, USA, 7–12 June 2015.

43. Vemulapalli R, Tuzel O, Liu MY, et al. Gaussian conditional random field network for semantic segmentation. In: Proceedings of the IEEE international conference on computer vision and pattern recognition. Las Vegas, NV, USA, 27–30 June 2016.

44. Liu Z, Li X, Ping L, et al. Semantic image segmentation via deep parsing network. In: IEEE international conference on computer vision. Santiago, Chile, 7–13 December 2015.

45. Lin TY, Maire M, Belongie S, et al. Microsoft COCO: common objects in context. In: European conference on computer vision, Zurich, Switzerland, 6–12 September 2014.