Journal of Image Processing Theory and Applications (2025)
Clausius Scientific Press, Canada

DOI: 10.23977/jipta.2025.080114
ISSN 2560-6239 Vol. 8 Num. 1

Image Semantic Segmentation Model based on AsppUNet

Qian Guo[1,a*], Yanlong Xu[1,b], Limin Sun[1,c]

[1]*School of Information and Intelligent Engineering, University of Sanya, Jiyang, Sanya, China*
[a]*qianguo@sanyau.edu.cn*, [b]*yanlongxu@sanyau.edu.cn*, [c]*liminsun@sanyau.edu.cn*
**Corresponding author*

Keywords: Aspp, UNet, Atrous Convolution, Pyramid Module, CamVid Dataset

Abstract: In this paper, we propose AsppUNet, an image semantic segmentation model based on the Atrous Spatial Pyramid Pooling(ASPP) module, to address the issue that smaller objects are prone to being overlooked during the segmentation process. Instead of using the standard pooling layers in the encoder of UNet, our model adopts a series of atrous convolution layers with progressively increasing dilation rates to reduce feature loss caused by traditional pooling operations. The ASPP module is constructed by cascading atrous convolution layers with different dilation rates, and is integrated into the decoder of UNet to aggregate multi-scale feature maps and capture multi-level contextual information. Experimental results demonstrate that AsppUNet achieves superior segmentation performance on objects of various sizes. It improves the mIoU for objects at different scales on the CamVid dataset, and effectively enhances the overall segmentation accuracy.

1. Introduction

Image semantic segmentation is an intensive prediction task that requires a combination of pixel-level accuracy and multi-scale contextual reasoning. In image semantic segmentation, the segmented objects are multi-scale, and the traditional CNN is difficult to extract the multi-scale features of the image, resulting in poor segmentation of targets at different scales. To achieve higher quality image semantic segmentation, we need to retain the image detail information as much as possible, while extracting the features of different scale targets[1]. Atrous convolution enables the CNN to obtain a larger receptive field without significantly increasing the computational complexity. By combining atrous convolution layers with different dilation rates, the network can extract multi-scale spatial features, which is beneficial for segmenting objects of varying sizes. In existing image semantic segmentation methods, pooling operations are typically employed in the encoder part of the network to reduce the spatial dimensions while expanding the local receptive field. Since semantic segmentation requires pixel-wise classification, the original image resolution is usually recovered through upsampling in the decoder. However, the ability of the decoder to reconstruct detailed image information heavily depends on the robustness and richness of the features extracted by the encoder.

The VGG16 encoder in the UNet network reduces the spatial resolution using a max pooling layer with a stride of 2×2, while simultaneously expanding the local receptive field. However, this operation often leads to significant loss of detailed image information. To mitigate such feature

degradation, we replace the max pooling layers in the VGG16 encoder with atrous convolution layers. In the encoding part of the pre-trained network, atrous convolution is employed to extract multi-scale features from local regions. This strategy allows the model to achieve a larger receptive field without increasing the number of parameters. Leveraging the effectiveness of atrous convolution, we further propose AsppUNet, a semantic segmentation model based on an encoder-decoder architecture embedded with an ASPP module, which is added after the decoder to capture and fuse multi-scale contextual features, thereby improving the overall segmentation performance.

2. Related Work

Image semantic segmentation is a challenging task that requires combining pixel-level accuracy with multi-scale contextual reasoning. FisherYu[2] proposed a atrous convolution module that aggregates multi-scale contextual information without reducing the resolution and supports insertion into existing semantic segmentation architectures at any resolution. In contrast to the pyramid-shaped architectures inherited from image classification models, the context module proposed in this paper is specifically designed for dense prediction tasks. It contains no pooling or downsampling layers, and is entirely based on atrous convolution. This design enables exponential expansion of the receptive field while preserving spatial resolution.

Atrous convolution is widely utilized in tasks such as semantic segmentation and object detection. In semantic segmentation, the classic DeepLab series and DUC[3] provide an in-depth analysis of atrous convolution. Similarly, SSD[4] and RFBNet[5] in object detection also employ atrous convolution to enhance their performance. SPPNet[6] addresses the challenge of requiring a fixed-size input image by generating a fixed-length representation regardless of the input image's size or scale. It computes feature maps from the entire image only once and then aggregates features from arbitrary regions(sub-images) to generate a fixed-length representation for training the detector. This method avoids redundant computation of convolutional features. The Spatial Pyramid Pooling(SPP) layer used in SPPNet employs a multi-level spatial pyramid structure, providing greater robustness in handling deformed objects compared to sliding window approaches that use a single window size[7]. PSPNet[8], applied in image semantic segmentation, leverages pyramid pooling to aggregate contextual information from different regions, thereby capturing global context. This approach improves segmentation accuracy by effectively integrating multi-scale contextual information.

Global prior representations have been demonstrated to effectively generate high-quality outcomes in scene parsing tasks. PSPNet provides an excellent framework for applying pyramid modules in pixel-level prediction tasks, and the proposed method achieves improved segmentation performance across a variety of datasets. Feature Pyramid Networks(FPNs)[9] are commonly used in object detection tasks. In such networks, low-level features typically contain less semantic information but provide precise localization, whereas high-level features are richer in semantics but offer coarser spatial details. Unlike conventional feature fusion approaches, FPN enhances prediction accuracy by integrating high-resolution, low-level spatial features with low-resolution, high-level semantic features, and performs predictions at each fusion stage.

3. The AsppUNet Model Architecture

3.1. AtrosUNet Structure

As shown in Figure 1, the pooling layer of the UNet encoder is replaced with an AtrousConv2D layer with a dilation rate of 2. The Atrous convolution is shown in Figure 2. This modification helps

to reduce feature loss caused by pooling operations to some extent. We refer to this modified model as AtrosUNet.

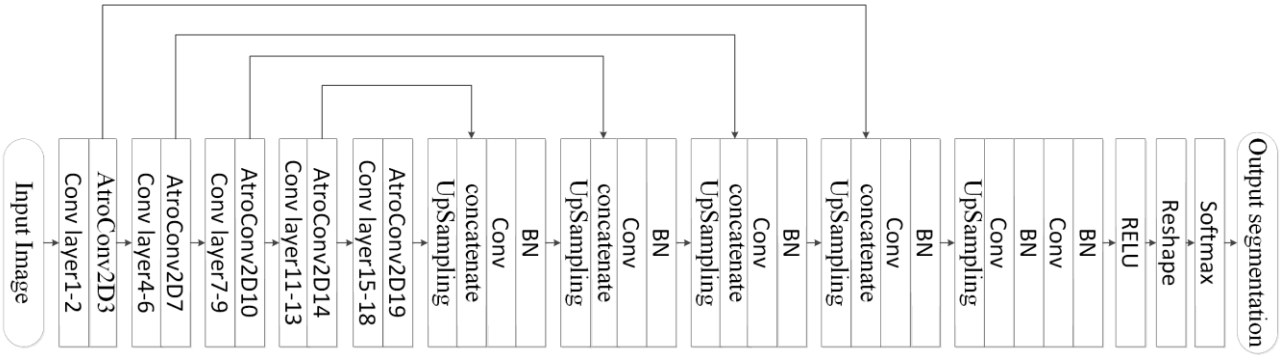

Figure 1: Network structure of AtrosUNet.

3.2. Atrous Convolution

Atrous convolution is a variant of the standard convolution operation that introduces an additional hyper-parameter known as the dilation rate, in contrast to conventional convolution. The dilation rate specifies the spacing between the elements of the convolution kernel, allowing the model to capture a larger receptive field without increasing the number of parameters or computational cost. Given a dilation rate *r*, the output *y[i]* of an dilated convolution applied to an input signal *x[i]* using a kernel *w[i]* is defined as follows:

$$y[i] = \sum_{k=1}^{m} x[i + r \cdot k] w[k] \tag{1}$$

Given that *m* is the length of *w[k]*, and *k* represents the size of the convolution kernel, the effective size of the dilated convolution kernel becomes $k' = k+(k-1)(r-1)$. As illustrated in Figure 2, consider a convolution kernel of size $3 \times 3 (k=3)$. When the dilation rate *r* is 1, the effective size of the atrous convolution kernel remains 3×3. For $r=4$, the effective size increases to 9×9, and for $r=7$, it expands to 15×15. It is evident that as the dilation rate *r* increases, more zero-padding is introduced between the elements of the convolution kernel. This introduces gaps in the receptive field, potentially leading to a loss of contextual information and spatial discontinuities in multilayer atrous convolutions using a single dilation rate. To address this issue, our proposed aspp module employs four atrous convolutions with varying dilation rates to effectively capture multi-scale target features, thereby enhancing recognition accuracy.

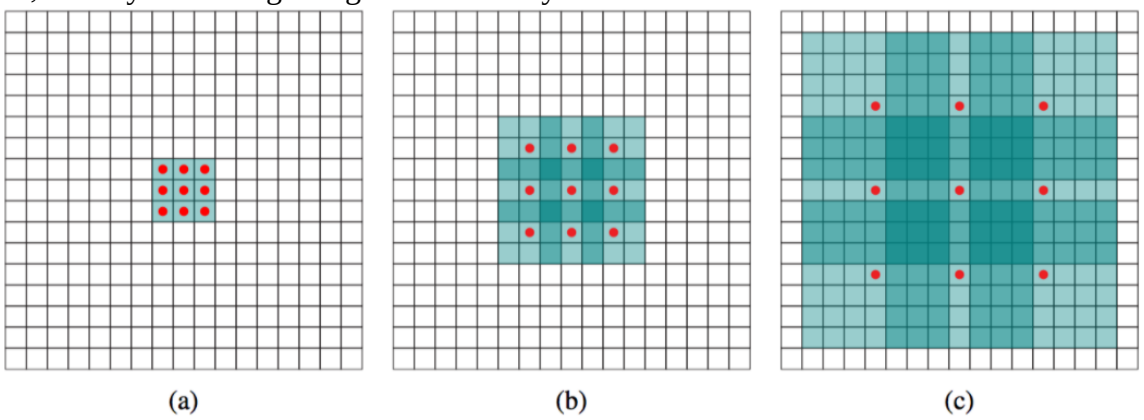

Figure 2: Atrous convolution.

The principle of atrous convolution is illustrated in Figure 2. This operation enables an exponential expansion of the receptive field without reducing the spatial resolution. Specifically, (a) shows feature map F1, obtained by applying atrous convolution with a dilation rate of 1 on F0, where each element in F1 has a receptive field of 3×3. (b) presents F2, generated from F1 using a dilation rate of 2; each element in F2 has a receptive field of 7×7. (c) depicts F3, derived from F2 via atrous convolution with a dilation rate of 4, where each element in F3 has a receptive field of 15×15. Notably, while the number of parameters increases linearly with network depth, the effective receptive field grows exponentially.

3.3. AsppUNet Structure

As illustrated in Figure 3, we insert the ASPP module behind the decoder. The ASPP module is shown in Figure 4. At this stage, the output is a feature map with a resolution of 320×320 and 64 channels, which are extracted using four parallel atrous convolution kernels with different dilation rates. Each of these kernels generates a feature map with the number of channels corresponding to the number of categories, while maintaining the same resolution. These feature maps are then concatenated and fed into the output layer for final prediction.

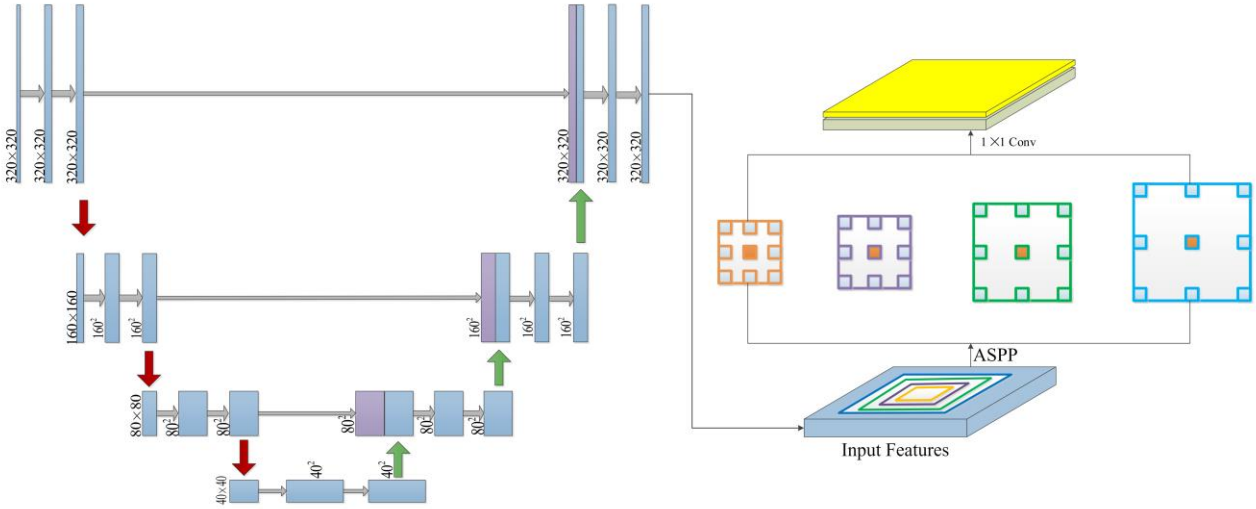

Figure 3: The AsppUNet model.

3.4. Pyramid Module

The context module is designed to enhance the performance of dense prediction architectures by aggregating multi-scale contextual information. It takes C input feature maps and generates C output feature maps, maintaining the same spatial dimensions and channel count. This consistent input-output structure allows the module to be seamlessly integrated into existing semantic segmentation networks, enabling feature maps to pass through multiple layers that expose contextual information and thereby improve prediction accuracy. The aspp module processes the input in parallel using atrous convolutions with different dilation rates. This approach is equivalent to capturing image-level contextual information at multiple scales. As illustrated in Figure 4, to classify the center pixel (highlighted in orange), the module employs four parallel atrous convolution kernels with distinct dilation rates to extract multi-scale features. These feature maps are then concatenated and transformed through a 1×1 convolution operation to produce the final output feature maps.

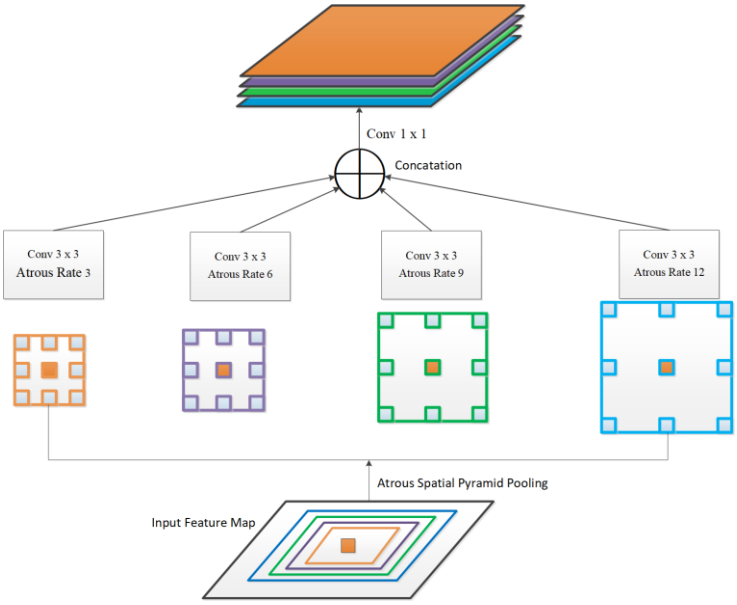

Figure 4: The ASPP module.

4. Experimental Results and Analysis

This paper conducts experiments based on the two improvement strategies proposed above. First, it verifies that replacing standard pooling with atrous convolution in the image semantic segmentation model yields better performance. Second, an aspp module is added after the encoder-decoder architecture to extract multi-scale contextual information, which improves the mIoU and enhances the overall segmentation performance. In the network, the standard pooling layers are replaced with atrous convolutions using a dilation rate of *2*. Two sets of dilation rates are defined for the pyramid module: *(6, 12, 18, 24)* and *(3, 6, 9, 12)*, respectively.

4.1. AtrousUNet Experiment

Experiment 1 corresponds to the model illustrated in Figure 1. Compared to the standard UNet model, its segmentation performance shows significant improvement. The specific experimental results are analyzed as follows:

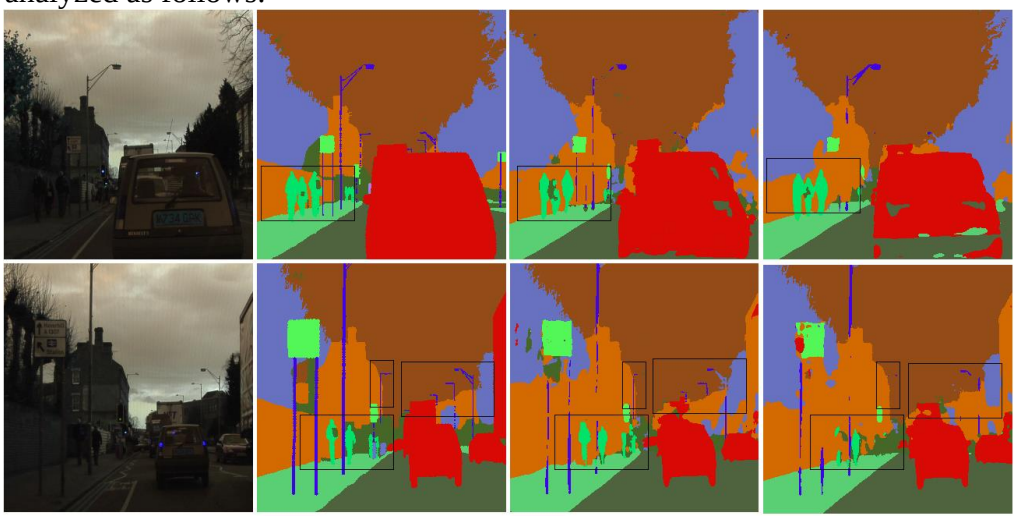

Figure 5: Segmentation results comparison between AtrosUNet and UNet on the testset.

As shown in Figure 5, the figure contains two groups of images. The first column displays the original images, the second column shows the corresponding labels, the third column presents the segmentation results from the AtrosUNet model, and the fourth column shows the segmentation results from the UNet model. As can be seen from the figure, the segmentation of light poles, billboards, and cars in the third column(AtrosUNet) of the first group is clearer compared to the fourth column(UNet). Additionally, in the second group, the outlines of pedestrians and light poles within the black box in the third column are not only segmented more clearly but also exhibit a higher degree of completeness compared to those in the fourth column. These observations indicate that the AtrosUNet model achieves superior segmentation performance compared to the UNet model. Furthermore, atrous convolution proves to be more suitable for dense prediction tasks such as image semantic segmentation.

Table 1: Segmentation results of AtrosUNet, CBAMUNet and UNet model on the CamVid testset.

Algorithms	UNet	CBAMUNet[10]	AtrosUNet
Trees	51.2	85.9	66.4
Sky	94.1	93.2	93.4
Buildings	84.8	65.3	65.0
Cars	82.5	89.8	85.0
Lighs Poles	32.4	41.3	41.5
Roads	96.9	95.8	88.8
Sidewalks	89.9	90.2	73.4
Pedestrian	26.9	30.7	50.6
Fences	20.2	16.9	29.1
Traffic Lighs	62.2	59.5	68.0
Cyclist	26.3	20.8	48.9
mIoU	60.6	62.7	64.5
PA	89.0	89.2	90.5

As shown in Table 1, the average mIoU for pedestrians, fences, and traffic lights is significantly improved in the AtrousUNet model compared to the CBAMUNet model[10]. Specifically, light poles, which are considered larger targets, and pedestrians, which are smaller targets, both show enhanced segmentation performance. The mIoUs for segmenting light poles and pedestrians using the AtrousUNet model are higher than those obtained with the CBAMUNet model. The atrous convolution operation helps reduce feature loss to a certain extent, thereby effectively capturing targets of various scales.

4.2. AsppUNet Experiment

Experiment 2 corresponds to the AsppUNet model shown in Figure 3, where four atrous convolution layers simultaneously perform feature extraction on the input feature maps, and then concatenate the feature maps at different scales, in order to obtain a more efficient combination of atrous convolution rates, this experiment sets two atrous convolution rates, namely *(6,12,18,24)* and *(3,6,9,12)*.

The experimental results indicate that better performance is achieved when the atrous convolution coefficients are set to (3,6,9,12), as shown in Figure 6. In this figure, from left to right, the columns represent: the original image, the ground truth label, the segmentation result of the proposed AsppUNet model and CBAMUNet model. Within this set of images, the segmentation performance in the region of interest(highlighted by the black box) demonstrates a clear advantage of the AsppUNet model over the CBAMUNet model. Specifically, in the black box of the first row,

the purple object corresponds to a cyclist. The AsppUNet model successfully segments the outline and accurately identifies the target, whereas the CBAMUNet model fails to capture it clearly in the fourth column. In the third row, the orange object within the black box represents a building. The AsppUNet model achieves complete and accurate segmentation of the building structure, while the CBAMUNet model exhibits fragmented or incomplete segmentation in the corresponding area. These results effectively validate the enhanced segmentation capability of the AsppUNet model on this dataset.

As shown in Table 2-1 and Table 2-2, the AsppUNet model achieves a significantly higher mIoU on the CamVid dataset compared to other models. Its segmentation performance is notably improved for object categories such as trees, cars, roads, and traffic lights. In particular, the proposed model in this paper attains a segmentation accuracy of 65.1% on the "cyclist" class, which substantially outperforms previous models, as illustrated in Figure 6. While existing models demonstrate good performance on certain categories but poor results on others, the proposed model not only improves segmentation accuracy significantly for several classes, but also maintains stable performance across other categories without notable degradation.

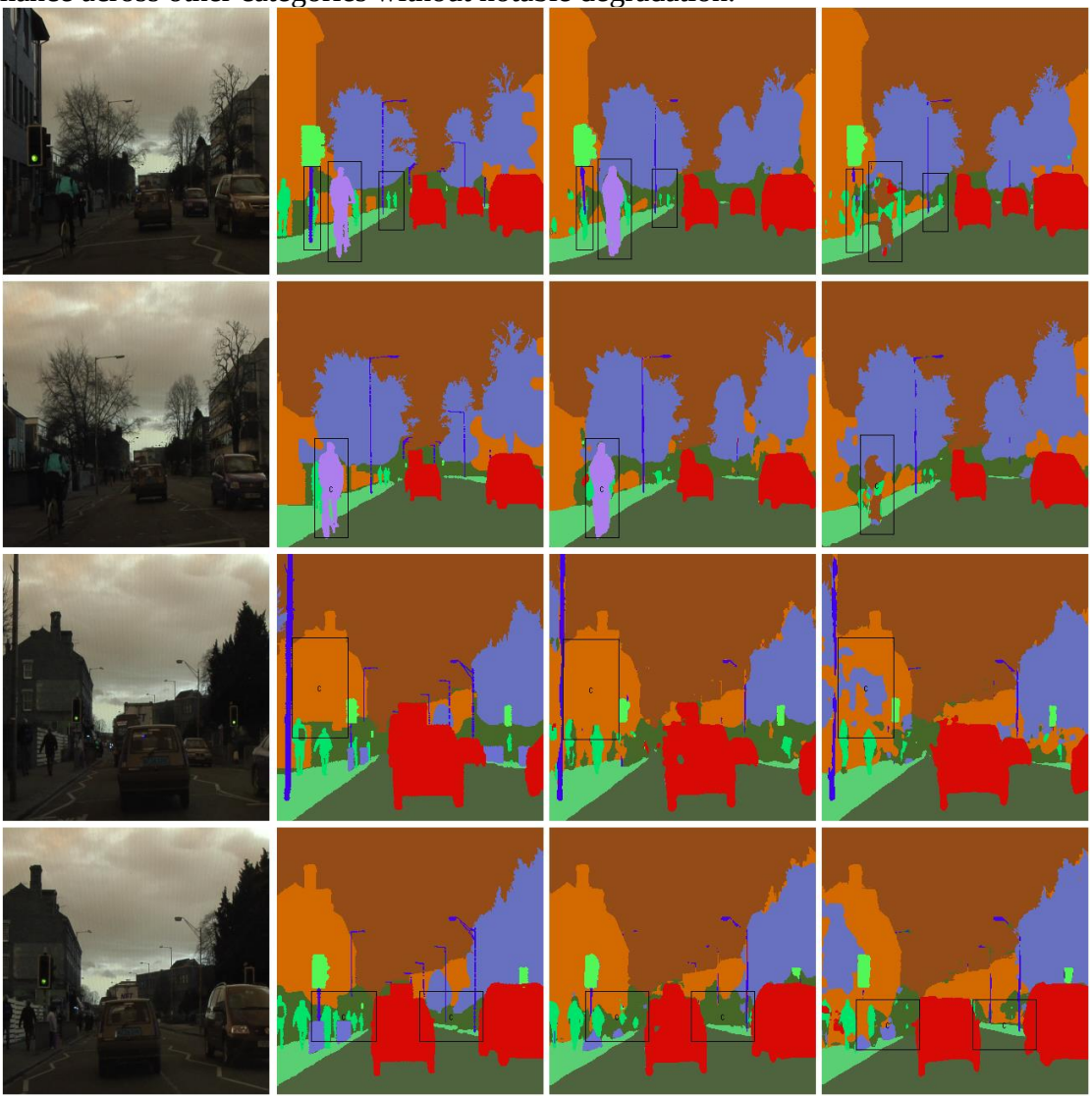

Figure 6: Segmentation results comparison between AsppUNet and CBAMUNet on the testset.

Table 2-1: Segmentation results of UNet, AsppUNet and other models for each category on the testset.

Algorithms	Trees	Sky	Buildings	Cars	Light poles	Roads	Sidewalks
FCN-8	71.0	88.7	77.8	76.1	19.9	91.2	72.7
DeconvNet[11]	-	-	-	-	-	-	-
ReSeg[12]	-	-	-	-	-	-	-
DeepLab-LFov[13]	74.6	89.0	81.5	82.2	14.3	92.2	75.4
SegNet	52.0	87.0	68.7	58.5	16.0	86.2	60.5
ENet[14]	77.8	95.1	74.7	82.4	35.4	95.1	86.7
LRN[15]	73.6	76.4	78.6	75.2	30.4	91.7	80.1
UNet	51.2	94.1	84.8	82.5	32.4	96.9	89.9
AsppUNet	81.1	94.5	73.6	93.3	30.1	97.3	85.7

Table 2-2: Segmentation results of UNet, AsppUNet and other models for each category on the testset.

Algorithms	Pedestrian	Fences	Traffic Lights	Cyclist	Backgrounds	mIoU	PA
FCN-8	41.7	24.4	32.7	31.0	-	52.0	88.0
DeconvNet[11]	-	-	-	-	-	48.9	85.9
ReSeg[12]	-	-	-	-	-	58.8	88.7
DeepLab-LFov[13]	48.4	27.2	42.3	50.1	-	61.6	-
SegNet	25.3	17.9	24.8	30.7	-	50.2	88.6
ENet[14]	67.2	51.7	51.0	42.5	-	51.3	-
LRN[15]	43.5	41.0	40.1	48.1	-	61.7	
UNet	26.9	20.2	62.2	26.3	-	60.6	89
AsppUNet	41	29.1	64.9	65.1	36.6	68.7	91.2

All models in this paper were trained using 12 segmentation categories, including the background class. In contrast, the other models listed in the table did not report mIoU values for the background category. We believe that evaluating the performance on the background class is also important for a comprehensive assessment of segmentation accuracy. As previously mentioned, two different atrous convolution configurations were tested in this study. Due to time constraints, we only compared the segmentation performance of these two settings, as shown in Figure 7. The results indicate that the combination of *(3,6,9,12)* achieves better segmentation performance than *(6,12,18,24)*. It can be observed that the optimal choice of atrous convolution rates may depend on both the dataset and the model architecture. Determining the best configuration typically requires extensive experimentation. In this work, we only compared two combinations to demonstrate the flexibility of such parameter settings. It is possible that the *(3,6,9,12)* combination is not yet optimal, but further exploration was limited by objective constraints.

As shown in Table 3, we designate the model with an atrous convolution rate of *(6,12,18,24)* as Model A, and the model with a rate of *(3, 6, 9, 12)* as Model B. According to the table, Model B exhibits superior segmentation performance compared to Model A on categories such as sky, buildings, roads, and pedestrians. Specifically, the mean mIoU for the atrous convolution rate of *(3,6,9,12)* is significantly higher than that of Model A. These findings suggest that the atrous convolution rate combination of *(3,6,9,12)* is more suitable for this task or, more specifically, better suited for the AsppUNet model.

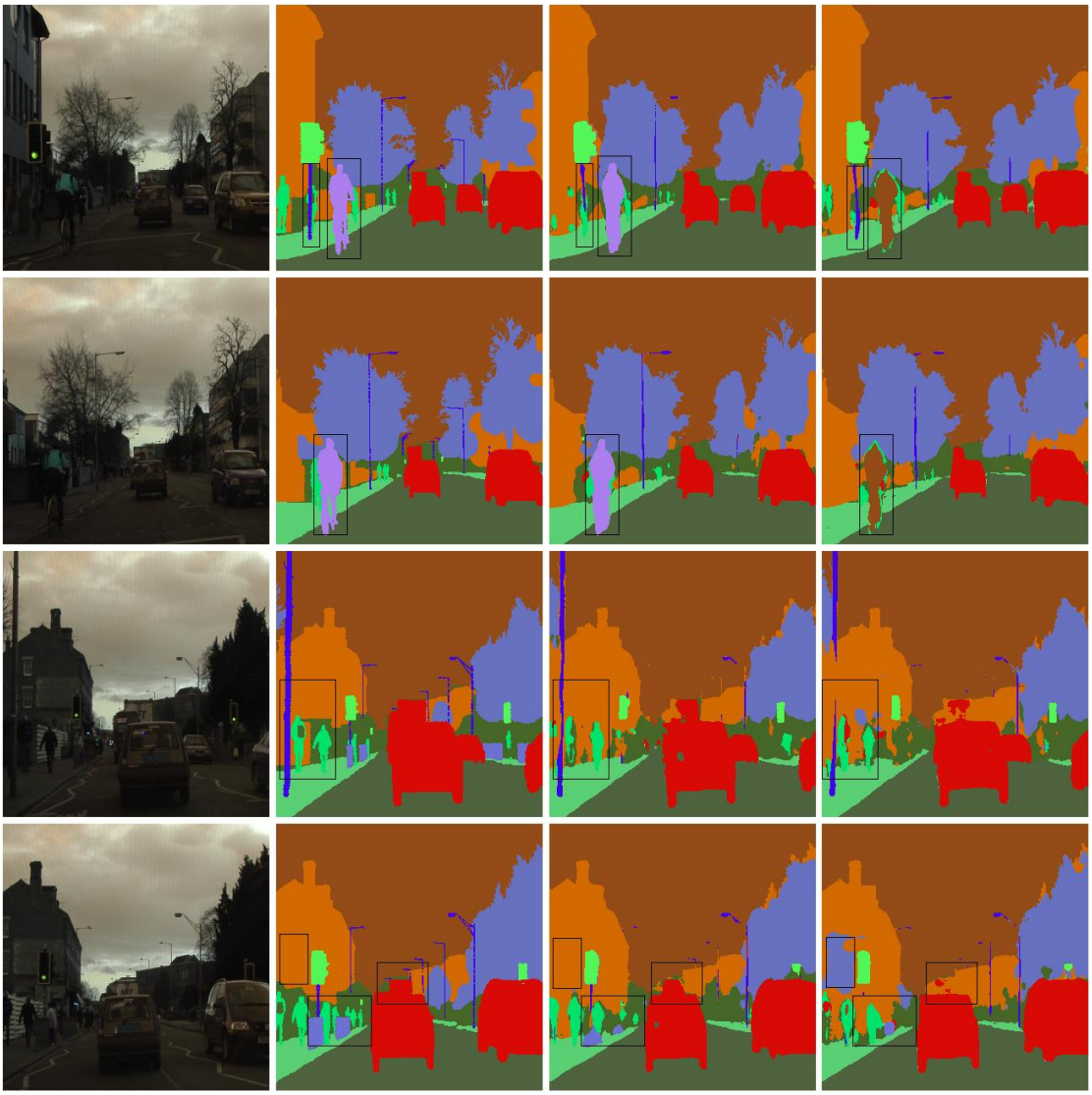

Figure 7: Segmentation results of AsppUNet with different atrous convolution rates, from left to right, the original image, the ground truth annotation, the segmentation result with dilation rates (3,6,9,12), and with dilation rates (6,12,18,24).

Table 3: Effect of different atrous convolution rates on AsppUNet performance.

Atrous convolution rate	Sky	Buildings	Roads	Pedestrians	mIoU
model A	93.2	55.3	95.8	36.9	65.5
model B	94.5	73.6	97.3	41.0	68.7

5. Conclusions

This paper proposes two directions for model improvement. The first involves replacing the standard pooling operations in the U-Net architecture with atrous convolution to reduce feature loss during downsampling. Experimental results demonstrate that this modification leads to more complete and accurate segmentation of objects such as light poles, billboards, and cars. Consequently, the mIoU scores for these categories are also improved to some extent, indicating that the atrous convolution-based pooling layer achieves better performance in dense prediction tasks compared to conventional pooling layers. Based on these findings, the second improvement

introduces the AsppUNet model, which incorporates an encoder-decoder pyramid module with atrous convolution at the end of the decoder in the network. This design enables the aggregation of multi-scale feature maps and allows the model to derive a combination of atrous rates better suited to the CamVid dataset through experimental comparison. As shown in the experiments, AsppUNet is capable of accurately recognizing challenging targets such as cyclists, segmenting their outlines clearly, and providing more complete segmentation for objects of varying sizes, including pedestrians and light poles. Overall, the AsppUNet model achieves a better balance in mIoU across different object sizes and improves the overall segmentation performance of the encoder-decoder framework.

Acknowledgements

This work was supported in part by University-level research projects of Sanya University. The research project is titled: Research on multimodal customer portrait in automobile precision sales (Project Number: USYJSPY22-42).

References

[1] Mnih V, Heess N, Graves A, et al. Recurrent Models of Visual Attention [J]. Advances in Neural Information Processing Systems, 2014,3.
[2] Yu F, Koltun V. Multi-scale context aggregation by atrous convolutions [J]. arXiv preprint arXiv:1511.07122, 2015.
[3] Wang P, Chen P, Yuan Y, et al. Understanding Convolution for Semantic Segmentation [C]. In 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), 2018: 1451–1460.
[4] Liu W, Anguelov D, Erhan D, et al. Ssd: Single shot multibox detector [C]. In European conference on computer vision, 2016: 21–37.
[5] Deng L, Yang M, Li T, et al. RFBNet: Deep Multimodal Networks with Residual Fusion Blocks for RGB-D Semantic Segmentation [J]. CoRR, 2019, abs/1907.00135.
[6] He K, Zhang X, Ren S, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition [J]. IEEE transactions on pattern analysis and machine intelligence, 2015, 37 (9): 1904–1916.
[7] Lazebnik S, Schmid C, Ponce J. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories [C]. In 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), 2006: 2169–2178.
[8] Lin T Y , Dollar P , Girshick R ,et al.Feature Pyramid Networks for Object Detection[J].IEEE Computer Society, 2017.DOI:10.1109/CVPR.2017.106.
[9] H. Zhao, J. Shi, X. Qi, X. Wang and J. Jia, "Pyramid Scene Parsing Network," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 2017, pp. 6230-6239, doi: 10.1109/CVPR.2017.660.
[10] Qian G ,Yanlong X .Image semantic segmentation model based on CBAMUNet[C]//University of Sanya (China),2024:
[11] Noh H, Hong S, Han B. Learning deconvolution network for semantic segmentation [C]. In Proceedings of the IEEE international conference on computer vision, 2015: 1520–1528.
[12] Visin F, Ciccone M, Romero A, et al. Reseg: A recurrent neural network-based model for semantic segmentation [C]. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2016: 41–48.
[13] Chen L C, Papandreou G, Kokkinos I, et al. Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs [J]. Computer Science, 2014 (4): 357–361.
[14] Yu F, Koltun V. Multi-scale context aggregation by dilated convolutions [J]. arXiv preprintarXiv:1511.07122, 2015.
[15] He K, Zhang X, Ren S, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition [J]. IEEE transactions on pattern analysis and machine intelligence, 2015, 37 (9): 1904–1916.