PSCNET: EFFICIENT RGB-D SEMANTIC SEGMENTATION PARALLEL NETWORK BASED ON SPATIAL AND CHANNEL ATTENTION

S.Q. Du[1,2], S.J. Tang[1,2,*], W.X. Wang[1,2], X.M. Li[1,2], Y.H. Lu[3], R.Z. Guo[1,2]

[1]School of Architecture and Urban Planning, Research Institute for Smart Cities, Shenzhen University,Shenzhen, P.R. China
-dusiqi2021@email.szu.edu.cn; shengjuntang@szu.edu.cn; wangwx@szu.edu.cn; lixm@szu.edu.cn; guorz@szu.edu.cn
[2]Key Laboratory of Urban Land Resources Monitoring and Simulation, Ministry of Natural Resources, Shenzhen, P.R. China
[3]School of Resource and Environmental Sciences, Wuhan University, Wuhan 430072, P.R. China- luyh@upr.cn

Commission I, WG I/7

KEY WORDS: Deep Learning, Semantic Segmentation, RGB-D Fusion, Channel Attention,Spatial Attention

ABSTRACT:

RGB-D semantic segmentation algorithm is a key technology for indoor semantic map construction. The traditional RGB-D semantic segmentation network, which always suffer from redundant parameters and modules. In this paper, an improved semantic segmentation network PSCNet is designed to reduce redundant parameters and make models easier to implement. Based on the DeepLabv3+ framework, we have improved the original model in three ways, including attention module selection, backbone simplification, and Atrous Spatial Pyramid Pooling (ASPP) module simplification. The research proposes three improvement ideas to address these issues: using spatial-channel co-attention, removing the last module from Depth Backbone, and redesigning WW-ASPP by Depthwise convolution. Compared to Deeplabv3+, the proposed PSCNet are approximately the same number of parameters, but with a 5% improvement in MIoU. Meanwhile, PSCNet achieved inference at a rate of 47 FPS on RTX3090, which is much faster than state-of-the-art semantic segmentation networks.

1. INTRODUCTION

1.1 General Instructions

Semantic maps refer to 3D scene maps with object category information. Semantic maps have wide applications in many fields such as robotics, AR/VR and autonomous driving. With the continuous development of 3D measurement technology, the construction of semantic maps has changed from traditional 2D image processing to 3D information processing, and how to achieve semantic classification of high precision 3D scenes has become a current research challenge. Deep learning provides an important basis for semantic segmentation. Traditional image-based segmentation algorithms are becoming more and more mature, but it is difficult to be directly applied to data processing in 3D space. In recent years, researchers have proposed a semantic classification algorithm framework that integrates geometric features and image features, which provides an important idea for combining 3D geometric modeling with scene semantic information recognition. In this paper, we propose a efficient RGB-D semantic segmentation framework called PSCNet,which has same computational complexity as Deeplabv3+ (Chen et al., 2018) but get higher precision.

1.2 Related Works

Semantic segmentation has been an fundamental technology in a large number of applications, such as Automatic driving, robotics, city modeling, etc. Significant works have been done in recent years based on machine learning or deep learning network. The most widely used model for semantic segmentation tasks is convolutional neural network(CNN). CNN network was first proposed by LeCun (Lecun et al., 1998), which extracts image features from the RGB imaged by convolution operation. In this work, they conclued that a local linear operation with weight sharing and translation invariance called convolution operation, which can significant contribut to the extraction of texture, location and semantic features of objects from images.. The architectural design of CNNs was then improved by , called AlexNet (Krizhevsky et al., 2012). AlexNet is the first convolutional neural network that is able to achieve better accuracy than traditional machine learning methods on the ImageNet benchmar. After that, VGG network (Simonyan and Zisserman, 2014) was then proposed by the Oxford University computer vision team, which was the first modular design network and widely used for image classification and object detection. The ResNet (He et al., 2016) is a milestone of CNN architectural design, which greatly solved the training problem through residual connections. As CNNs are usually designed as a stack of multiple layers of convolution operations FCN (Shelhamer et al., 2017) is the first convolutional neural networks applied to semantic segmentation. FCN uses the local information of the final feature map for pixel-wise classification. In details, the network adapts the fully connected layer of the VGG network to the convolution layer and compute pixel classification weight by a local linear operation. Since FCN was proposed, semantic segmentation algorithms have gradually evolved into three structure: 1.Encode-Decode networks, represented by U-Net (Ronneberger et al., 2015); 2.Multi-Scale Feature Extraction networks, represented by Deeplab (Chen et al., 2017) series; 3.Multi-Scale Encode-Decode networks represented by Deeplabv3+, which fuse the above two ideas.

The above works are mainly focused on the classification of color images. With the quick development of 3D sensors, it is more convenient to obtain both the RGB and the registered depth image. Unlike the RGB image, the Depth images contains significant geometric information and provides the possibility to learning more types of features, such as the edges, spatial

[*] Corresponding author

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-V-1-2022-129-2022 | © Author(s) 2022. CC BY 4.0 License.

location.. The geometric information of the Depth images is well complementary to the color and texture information in RGB images. A lot of works have been done onthe fusion of Depth and RGB information to improve the semantic segmentation precision.

FuseNet (Hazirbas et al., 2016) uses a parallel feature extractor for RGB-D feature extraction, and achieved SOTA (state-of-the-art) results on NYUv2 dataset (Silberman et al., 2012). Similarly, a large number of networks were proposed by using the same fusion strategy, such as RedNet (Jiang et al., 2018), ACNet (Hu et al., 2019), RDFNet (Park et al., 2017), RAFNet (Lu et al., 2020), TSNet (Zhou et al., 2020), CANet (Zhou et al., 2022), ESANet(Seichter et al., 2021), etc. The main differences between these networks is the structure of feature fusion module and the way they are decoded

In the design of the feature fusion module, FuseNet and RedNet both use a simple element wise summation to fuse RGB and Depth features. Subsequently, some researches add feature selection operation by using attention module. Among them, ACNet uses a linearly computed CAM attention module, RAFNet, ESANet uses a non-linearly computed SE attention module (Hu et al., 2018), and TSNet and CANet use a self-attention module.

Instead of conducting feature fusion in the feature extraction phase, RDFNet used Multi Model Feature Fusion Network with residuals connection for feature fusion in the decoding phase.

In summary, we can draw the following conclusions as follows. Firstly, in feature extraction phase, a large number of studies use two identical networks to extract RGB and Depth features separately. Secondly, when feature fusion is performed, most studies tend to use the attention module for feature selection in the channel direction.None of the above methods conduct feature selection in spatial direction. Two questions were raised. What happens when the spatial attention module is used for feature fusion. In this study, we found that the weight of spatial attention for some feature layers is small, which mean some part of the features is not used during the feature fusion phase. Then the second question is whether the importance of depth features in feature fusion process is consistent with the the results of Spatial Attention weight visualization. Thirdly, whether setting different weights for different feature extraction can improve the efficiency of the model

Therefore, three important questions need to be explored in this paper: a) What is the effect of spatial attention in feature fusion. b) How to define the importance of features during redundant feature rejection process c) How to design a more effective deep learning model when the features that contribute more are identified.

In this research, we first designed a parallel RGBD feature extraction structure to test the effect of different attention modules on feature extraction. Second, we visualize the weight of spatial attention as a way to analyze the mechanism of spatial attention during feature fusion. Third, we analyzed the relationship between spatial attention weights and feature importance by ablation experiments, and streamline the network based on feature importance and computational cost. Fourth, we redesigned the ASPP module in Deeplabv3+ network, substantially compressed the computation while keeping the accuracy unchanged. Finally, an improced RGBD semantic segmentation network, PSCNet, is designed for sematic segmentation based on Deeplabv3+. The experimental results show that our proposed network, PSCNet is able to achieve better segmenttion accuracy and lower computational cost with comparable number of parameters.

1.3 Contribution

The contribution of this paper can be summarized as follows:

1. We found that the spatial attention module can effectively improve the classification accuracy with the addition of a small number of parameters. We integratedSA Module and SE Module and used them for feature filtering.

2. We found that using the full ResNet for feature extraction on Depth is not the most efficient approach. Removing the last residual block results in a higher efficiency preparameters. In addition we found that spatial attention weight map standard deviation is correlated with feature information volume.

3. Our work proposes an efficient multi-scale feature extractor called WW-ASPP. WW-ASPP module reduces 30% computation cost and 8 Million parameters when compared to ASPP in Deeplabv3+.

4. We construct PSCNet, a more efficient RGBD semantic segmentation network which achieve 47FPS FP32 inference speed on RTX3090 and 5% MIoU improvement comparing with the original Deeplabv3+.

2. METHOD

In this research, we propose an new RGB-D semantic segmentation network, PSCNet, which can significantly reduce the parameters redundancy and improve the effectiveness of the framework. We first designed a RGB-D parallel feature extraction backbone to analysis the effects of three attention modules, including Spatial Attention (Woo et al., 2018), SE (Hu et al., 2018), and CAM (Hu et al., 2019), on feature fusion process. Then, by analyzing the contribution of SA, the weights of SA is determined with a well designed weight model. Based on Deeplabv3+ framework, we streamline the network by filter those parameters with low parameter efficiency and proposed a Lite-ASPP module by introducing depth-wise convolution and increasing the number of output channels. Summarizing all improvement,we generated a lightweight model, PSCNet.

This section proceeds with a introduction of the RGB-D parallel feature extraction network, followed by a detailed description of the selection of the attention module. Finally the depth-wise-wide-ASPP is introduced. The backbone, depth-wise-Wide-ASPP and the classifier were integrated for the construction of PSCNet.

2.1 Parallel RGB-D Feature Extraction Backbone

The parallel feature extraction backbone is similar to the structure used in previous studies, based on the FuseNet idea. As shown in **Figure 1**, the backbone contains two ResNet50 networks (He et al., 2016), which is used for the extraction of RGB and Depth features individually. ResNet is an image classification network using shortcut, and its basic structure is a Residual Layer, which contains multiple Residual Blocks.

The backbone in this framework removes the down-sampling operation of the third and fourth residual layers, and uses dilation convolution to keep the receptive field. The dilation parameter of the third residual layer and the fourth residual layer is 2 and 4 respectively. Element-wise summation is used for fusion of RGB and Depth features in the network before each residual layer. Also, the attention module is added in the subsequent study is in this process.

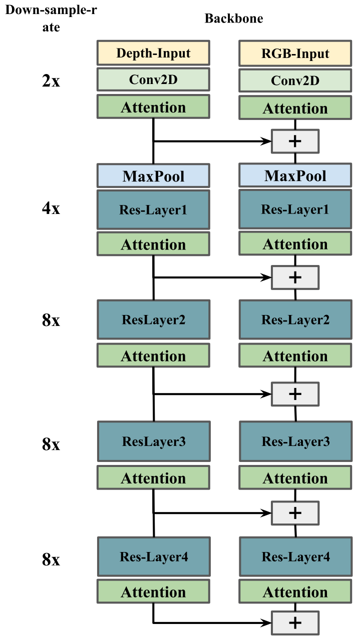


Figure 1. Backbone of PSCNet.

2.2 Selection of the Attention Module

The weight of the features are generally calculated by the attention mode. The attention model enables the network to rthe features that are useful to the trainning task and rejects those features with less contribution to the task. Therefore, the attention mechanism allows the model to select better features when training.

The core of the attention module is the learnable module. In terms of the learnable module, it is affected by gradient back propagation and can be classified into two types: linear and nonlinear computational modules. Basically, the learnable module is always learning a combination rule of features and generating the weights of those features. In addition, the attention module is divided into spatial attention module and channel attention module according to the dimension of the feature' weights. The spatial attention module weights the feature map in the two-dimensional direction. In addition to the first two dimensions, channel direction is also used to weight the feature map during the channel attention processing.

In order to explore the effect of different attention modules on the model and to obtain the optimal attention modules, three types including the local spatial attention (SA) module, the Squeeze Excitation (SE) and the Channel Atten module (CAM) are tested as follows.

The SA module (Woo et al., 2018) uses 7x7 convolution as a learnable module, and generates attention weights on max-pooling and average-pooling of feature map.The SA module is calculated as follows.

$$R_s = \sigma[Conv_{7\times7}(cat(Mean(F), Max(F)))]$$
$$\begin{aligned} where\ R_s &= \text{result of spatial attention},\ F = \text{Feature Map} \\ Mean &= \text{pixel mean value} - \text{channel direction} \\ Max &= \text{pixel max value} - \text{channel direction} \\ Conv_{7\times7} &= \text{2D convolution with kernel size } 7 \end{aligned} \tag{1}$$

The module structure is as follows:

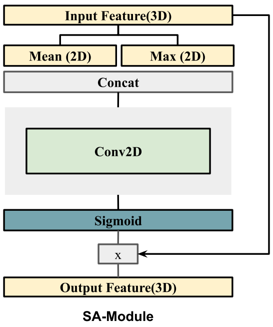


Figure 2.. SA Module.

The Squeeze-and-Excitation (SE) module (Hu et al., 2018) is a channel attention module that first abstracts the global features of each channel using global adaptive pooling, and the learnable module generates the importance weights of each layer using a nonlinear Multi-layer perception.The module structure is shown in Figure3.The SE module is calculated as follows:

$$R_c = \sigma[MLP_{2-Layers}(AdaptAvgPooling(F))]$$
$$\begin{aligned} where\ R_s &= \text{result of channel attention},\ \text{F} = \text{Feature Map} \\ MLP_{2-Layers} &= \text{Multi} - \text{Layer Preception} \end{aligned} \tag{2}$$

The CAM (Hu et al., 2019) module is a channel attention module that first abstracts the global features of each channel using global adaptive pooling, and the learnable module uses a linear 1x1 convolution to generate the importance weights of each layer.The module structure is shown in Figure4.The CAM module is calculated as follows:

$$R_c = \sigma[Conv_{1\times1}(AdaptAvgPooling(F))]$$
$$\begin{aligned} where\ R_s &= \text{result of channel attention},\ \text{F} = \text{Feature Map} \\ Conv_{1\times1} &= \text{2D convolution with kernel size } 1 \end{aligned} \tag{3}$$

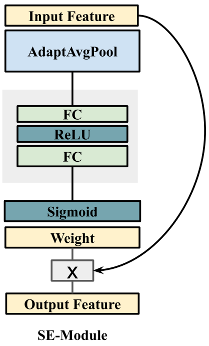


Figure 3. SE Module

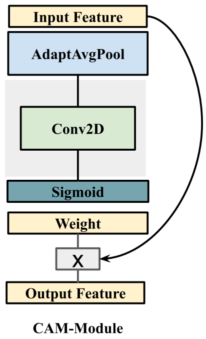


Figure 4.CAM Module

2.3 Depth-wise-Wide-Atrous Spatial Pyramid Pooling

In order to compress the model parameters and reduce the computational cost, we redesigned the Atrous Spatial Pyramid Pooling (ASPP) module based on the framework of Deeplabv3+ (Chen et al., 2018), by using depth-wise separable convolution.

In convolution modules, a single convolution kernel needs to compute all feature maps. Depth-wise separable convolution will be conducted on multiple convolution operation, in which each convolution kernel only computes on one layer of the feature map and greatly reduces the convolution computational cost.. However, as demonstracted in MobileNetv2 (Sandler et al., 2018), after depth-wise separable convolution, a large amount of information will be lost when channels is too small and use ReLU as activation function. Therefore, this study increases the output channels of the dilation convolution in ASPP module from 256 to 512 to reduce information loss. The structure of the depth-wise-Wide-ASPP module in this paper is shown in **Figure** ..

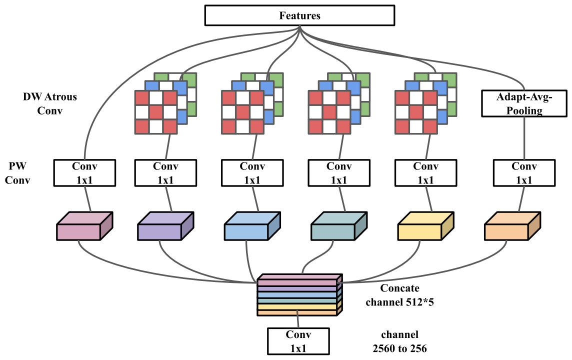


Figure 5. Depth-Wise-Wide-ASPP Module.

2.4 PSCNet

To achieve a more efficient RGB-D semantic segmentation network, PSCNet was constructed by a streamlining backbone and redesigning ASPP module..The structure of PSCNet is as followed:

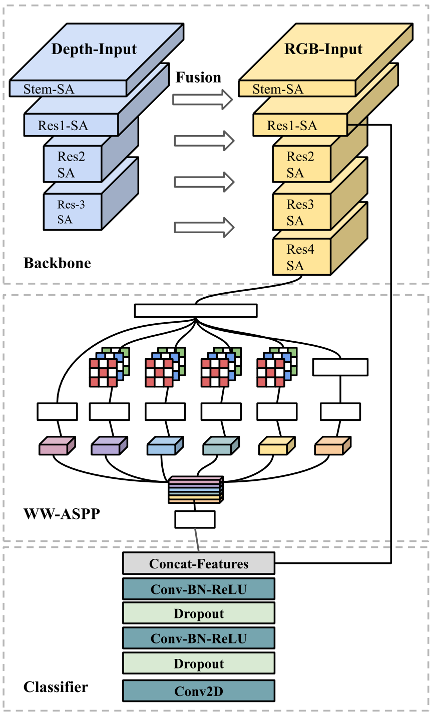


Figure 6. Framework of PSCNet

Similar with Deeplabv3+, PSCNet uses an encode-decode structure and uses the streamlined ResNet50x2-Spatial Attention backbone in the encoding side, and number of the parameters is 8millons more than that of the original ResNet50. Depth-wise-wide-ASPP module in Section 2.3 is used for the extraction of multi-scale feature, and the number of parameters is 8 millons less than the original ASPP module. Through the optimization of the method, the proposed framework in this paper has approximately the same parameters as the original Deeplabv3+.

3. RESULT AND DISCUSSION

3.1 Dataset and Training Details

In this study, we used the NYUv2 dataset for training purpose. NYUv2 contains 1449 images with semantic annotations and corresponding depth maps. The image resolution of the benchmark is 640*480 pixels. The training set consists of 795 RGB images and corresponding Depth images, and the validation set consists of 659 images. The NYUv2-40 classes are used as the training label. The background portion of the label is removed for training, prediction, and computational accuracy. RandomHSV, RandomScale and CenterCorp operations are used for data enhancement of the training images, and RGB and depth data are normalized separately.

The hardware platform for experimental analysis is AMD EPYC 7462 CPU, Nvidia RTX3090 GPU, and 80 GB RAM. Software versions used Pytorch 1.1.10, CUDNN 8.3.

The backbone was initialized using ImageNet pre-training weights. The loss function is cross-entropy loss. We use an auxiliary loss calculated by RGB-Res-Layer3 feature and FCN Classifier. AMP (Automatic Mixing Precision) is used for training and stochastic gradient descent (SGD) for optimization. We set SGD momentum to 0.9, weight decay to 0.002 and initial learning rate to 0.002. We use poly learning rate scheduler with parameter 0.9 and batch size to 8. Total epoch is 500, 50000 iterations.

3.2 Ablation Experiments of Attention Module

The Attention module is an efficient module that improved segmentation precision by increasing a few parameters.In this experiment, we expect to verify which attention module or their combinations is more beneficial for the improvement of classification accuracy. To analyze the effect of spatial attention on accuracy, this paper first compares the classification accuracy of Deeplabv3+ network on NYUv2 using the backbone proposed in 2.1. We inserted four attention modules before element-wise summation. The result is shown in Table 1:

Attention	MIoU	Improvement	AP	Parameters
SA	46.4%	+0.8%	73.9%	64M
SE	46.1%	+0.5%	74.0%	66M
CAM	44.7%	-0.9%	73.7%	**75M**
SA+SE	**46.8%**	**+1.2%**	**74.3%**	66M
Baseline	45.6%	+0%	73.8%	64M

Table 1. Comparison accuracy of different attention module on NYUv2. Baseline:Non-Attention, SA:Spatial Attention,SE:Squeeze-and-Excitation, CAM:Channel attention Module

First, the experimental results shows that the use of single spatial Spatial Attention Module can effectively improve the accuracy of the model. Using the SA Module alone results in higher accuracy than using SE Module, CAM, and Baseline. The results indicate that feature selection in the spatial direction is superior to that in the channel direction.

Secondly, the research found that CAM model has the lowest accuracy, but the most model parameters. It is means that the linear computation module is not the most optimal choice, and that a large number of parameters causes a large amount of over-fitting.

The results, shows taht the SA Module can work in synergy with the SE Module. Th best classification accuracy can be obtained by using both spatial attention and channel attention. The overall improvement in accuracy is more than one percent, basically equal to the sum the improvement using SE and SA separately, which proves that the two modules are not contradictory.

In this paper, we speculate that the SA enhances the complementarity of RGB and Depth features by improving contrast of feature map. SA uses the Sigmoid function to calculate the feature importance weights. The sigmoid function is a marginal distribution function in the Bernoulli distribution (0-1 distribution), and a high weight means the feature has a higher probability of being 1.

Therefore, the Sigmoid function can ignore some unimportant features and enhance the separation between foreground and background of the feature map. There is a certain complementarity between the information contained in RGB and Depth images. Spatial attention enhances the complementary components of RGB and depth, which in turn improves the effect of feature fusion.

3.3 Ablation Experiments of Feature Fusion

This study discusses the redundancy problem during extracting features on depth images. Using a backbone with the same size as a RGB feature may result in over-extraction problem and then lead to an increase in redundant computation. This is mainly due to the fact that color images contain more information than depth images. In this paragraph, we hope to find the cost-benefit balance when performing feature extraction on depth images.

Prior to discussing the redundancy of the network, we need to define how to evaluate the redundancy. Those operations that significantly increase the computation while contributing little to accuracy are referred to as redundant operations. The redundancy of the network can be evaluated by the Precision Pre-Million-Parameters (referred to as P/P) is used to define and evaluate the redundancy of the network as follows.

$$\mathrm{P/P} = \frac{MIoU}{Parameters} \tag{4}$$

Furthermore, we use a step-by-step approach to increase the number of modules and carry out an ablation experiment on the fusion times of depth features. This experiment aims to analyze the cost and benefit by adding new modules. The experiment result is shown in the following Table 2.

	Baseline	M1	M2	M3	M4	M5
Stem		√	√	√	√	√
Layer1			√	√	√	√
Layer2				√	√	√
Layer3					√	√
Layer4						√
MIoU (%)	40.4	44.4	44.6	44.8	45.7	**46.4**
Parameters	40	40.4	40.6	41.8	48.9	**63**
FLOPS (G)	214	214	218	225	259	**331**
Efficiency	1.01	**1.10**	**1.10**	1.07	0.93	0.74

Table 2. Ablation Experiment of Used Feature Fusion Layers

The result has shown that the redundancy of the network is positively related to the number of modules used for depth feature exrtaction. Fusing the third and fourth residual layers achieve higher accuracy. However, more parameters are used and computational efficiency is reduced. In comparison with Deeplabv3+, the network has a higher efficiency than Deeplabv3+ when only fusing shallow modules, but gradually decreases to 70% of Deeplabv3+ as the fusion proceeds. It is obvious that the network can be made to achieve high accuracy by stacking parameters, but such a network is difficult to be applied in a practical environment.

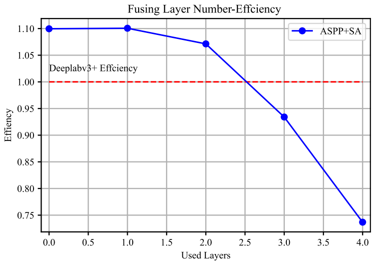


Figure 7. Efficiency of Used Feature Fusion Layers.

In summary,the balance between efficiency and cost is the case of fusing the first two residual layers.The balance point means that fusion strategy not lead to high computation cost and get significant precision improvement.

3.4 Spatial Attention Applicability Analysis

The ablation experiment is an effective method for quantifying the contribution of features. However, ablation experiment is a very time consuming process, and the complete ablation experiment of this research took hundreds GPU hours. Our second question, therefore, is whether we can analyze the redundancy of feature extraction operations without performing an ablation experiment.

For this question, we speculate that the feature importance can be estimated by SA module weight. SA weights are derived from the M5 network, which can be viewed as a global reflection of the features found in that layer. To analyze the relationship between this weight and feature importance, we first visualized the SA weights of each layer, and the results are shown in Figure10.

Figure 8. Example RGB Data

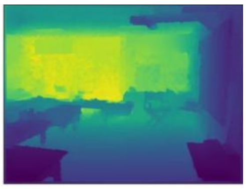


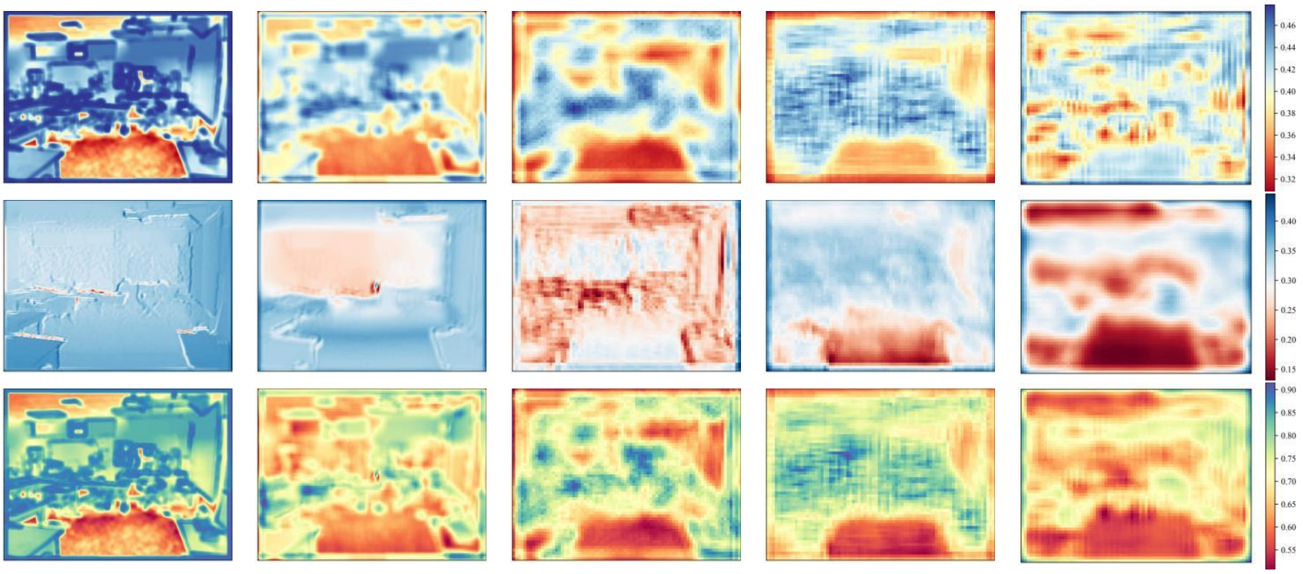
Figure 9. Example Depth Data

It can be found that the contrast of the SA weight map in shallow layers is lower, whereas the contrast of deeper layers is higher. By analyzing the change of accuracy after fusion of each layer, we found that deeper features contribute more to the accuracy. This seems to imply that the contrast of SA weight map contrast is positively correlated with features contribution. We performed a correlation analysis (Spearman's rank correlation coefficient) for the standard deviation(Std) of the SA weight map, which can quantify the contrast, and the contribution of each layer. The results showed that the two were statistically related.

Correlation	P-Value
0.872	0.054 (<0.1)

Table 3. Spearman's Rank Correlation Coefficient of SA std and Layer Precision Contribution

Therefore, we believe that it is feasible to estimate the feature importance using SA weights. Especially in the case of limited computing resources. A model containing all fused features needs to be trained, and the significance of each feature is estimated by examining the output of SA Std.



(a) Stem-Layer-weight **(b)** Res-Layer1-weight **(c)** Res-Layer2-weight **(d)** Res-Layer3-weight **(e)** Res-Layer4-weight

Figure 10. Weight Map of Spatial Attention Module in Different Layer.
From top to bottom, the image shows SA Map of RGB feature, SA Map of Depth Features and the element summation of two SA Map, which represent the SA Map of fusion feature.

We speculate that the primary reason for the large contribution of deep-layer features is that deep-layer features are enriched with semantic information extracted from the Depth image, which can complement the semantic information in the RGB image. The depth image does not have a lot of texture and color information, which makes CNN fail to extract more information from the shallow layer.

In summary, we only need to train the network once to estimate the importance of the features in it by SA Std. As a result, the estimation of feature importance, or the explanation of network principles, can be influenced to some extent by the weights of SA in future studies.

3.5 Ablation Experiment of Deepwise-Wide-ASPP

The above paper discussed how to analyze the efficiency of the network and how to simplify the backbone. In order to further improve the efficiency of the network, we propose the Deepwise-Wide-ASPP (WW-ASPP), a module based on depthwise separable convolution, which reduces the computation cost of the original ASPP module. In this part, the efficiency of our proposed WW-ASPP module will be analyzed based on the evaluation method in 3.4.

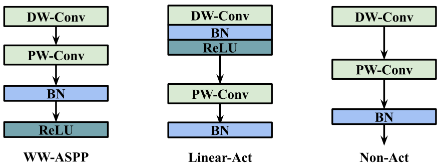


Figure 11. Different DW-PW Convolution Block Design

First, we discuss the design of the WW-ASPP module to ensure the optimal model of our proposed network. WW-ASPP employs depth-wise separable convolution to achieve the equivalent of the original 3x3 convolution using two convolutions of DW-PW. The DW-PW convolutions were combined using Xception's strategy, i.e., no module was added between the DW and PW convolutions, and a batch normalization layer and a ReLU activation function were applied after the PW convolution.

However, besides Xception, some other studies use different strategies. For example, MobileNetv2(Sandler et al., 2018) adds ReLU activation function after DW convolution, but no activation after PW convolution(Linear-Act); ConvNext(Liu et al., 2022) proposes that using fewer activation functions can improve the network accuracy(Non-Act). In this paper, we verify the three strategies respectively using the following results, which are shown in the following Figure 11.

	WW-ASPP	Linear-Act	Non-Act	Baseline
MIoU	45.4%	44.4%	43.5%	45.7%
GFLOPS	89	89	89	259
Parameters	40	40	40	49

Table 4. Precision of Different DW-PW Block Design

The results demonstrate that the WW-ASPP module, is the one with the least accuracy degradation compared to the original ASPP module. Subsequently, we analyzed the efficiency of using the WW-ASPP module. The results are shown in the Figure 12.

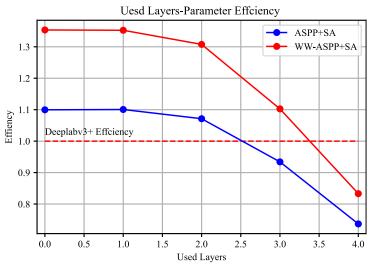
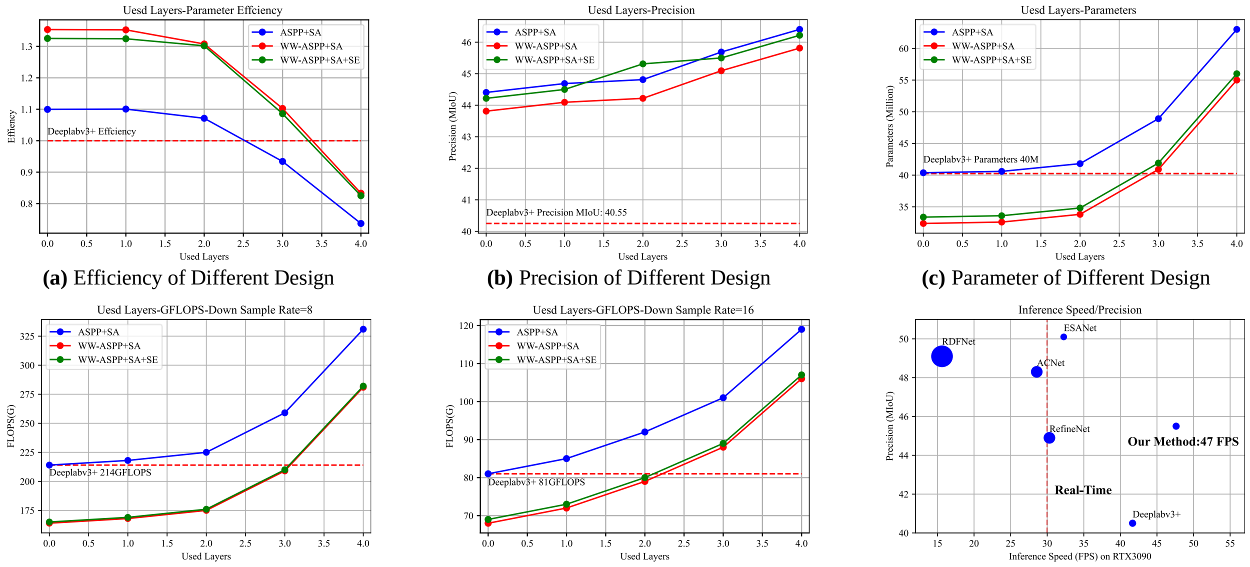


Figure 12. Efficiency Comparison of ASPP&WW-ASPP

The results show that the network efficiency of each parameter is greatly improved by using the WW-ASPP module, and the computation cost is only 30% of that before the improvement.



(a) Efficiency of Different Design

(b) Precision of Different Design

(c) Parameter of Different Design

(d) FLOPS of Different Design (stride 8)

(e) FLOPS of Different Design (stride 16)

(f) Comparison of Different Model

Figure 13. Comparison Result.
The subplot (a)-(e) shows comparison result of different module and layer of backbone used in PSCNet.
Subplot (f) shows comparison result of different Model.

3.6 Overall Efficiency Evaluation

PSCNet is the optimal structure selected among several models based on the above discussion. PSCNet uses Deeplabv3+ as a benchmark, and we select the model from a range of models based on the layer used to extract depth features. The result with the closest parameter efficiency to the Deeplabv3+ network and the highest accuracy is chosen.

Figure13 (a) illustrates that the WWASPP module makes a significant contribution to the improvement of model efficiency. Figure13 (b) illustrates that the accuracy of the network has been significantly improved by integrating the above improvements compared to Deeplabv3+. By comparing Figure (d) with Figure (e), we found that the reduction in computation caused by WW-ASPP is more apparent in the network with higher resolution feature maps.

In summary, we found that the network has the same efficiency per parameter as Deeplabv3+ when using SE, SA, and WW-ASPP with Depth's first three residual layers. At this point, we choose this structure as the final structure of the PSCNet. The final PSCNet network parametric number is 40m, MIoU is 45.4%, and 89 GFLOPS. The improved network MIoU is 5% higher and the computational effort is 60% lower than Deeplabv3+ with a downsampling rate of 8 and an accuracy of 40.5%.

3.7 Comparison with Existing Network

In conclusion, we compare the proposed PSCNet with other semantic segmentation networks shown in Table 5 in terms of accuracy, efficiency, and computational speed.

The absolute accuracy obtained by the network proposed in this paper is comparable to that obtained by other methods. It should be noted that it has achieved second-high efficiency pre-parameters that are only marginally lower than ESANet. In terms of FP32 inference speed, PSCNet is significantly faster than traditional methods, averaging 47 frames per second(FPS) on RTX3090.

The results indicate that increasing network accuracy regardless of cost may result in a huge number of parameters, which dramatically increases the training cost of the network. RDFNet contains 450 million parameters and ACNet contains 120 million parameters, which contain a large number of redundant operations in terms of per-parameter efficiency analysis. In the real world of engineering, each computational resource has a cost, and under no circumstances should you use a large amount of resources to optimize redundant parameters.

Model	MIoU	Parameters	Speed	Efficiency
Deeplabv3+ (Baseline)	40.5%	40M	42 FPS	100%
RDFNet	49.1%	450M	15 FPS	10%
ACNet	48.3%	118M	28 FPS	37%
TSNet	46.1%	-	-	-
RefineNet	44.7%	118M	30 FPS	35%
LW-RefineNet	43.6%	46M	-	95%
RAFNet	47.5%	-	-	-
ESANet	50.1%	34M	32 FPS	**134%**
PSCNet-T	45.4%	40M	**47 FPS**	106%
PSCNet-L	46.2%	52M	30FPS	89%

Table 5. Precision and Other Details of Different Model

Furthermore, the results indicate that although the use of DW convolution can greatly reduce the amount of parameters and computation needed, DW convolution is a low-throughput computational module and can significantly slow down inference speed. Theoretically ESANet has fewer parameters and lower computational demand compared to PSCNet, but because of its heavy use of DW convolution, it makes the inference speed severely slowed down.

4. CONCLUSION

In this paper, based on a detailed analysis and discussion of the different modules of the segmentation network, we proposed an efficient semantic segmentation network, PSCNet, which uses an encode-decode structure and WW-ASPP module for the extraction of multi-scale feature. Based on the experimental analysis, we can draw the following conclusions.

1) the research demonstrates that the conbination of spatial attention and channel attention can signifucantly improved the segmentation accuracy. After integrating of spatial-channel

attention, the MIoU of the network increased by 1.2% compared to Baseline.

2) Since the depth image and the visual image contain different information, using the same backbone for feature extraction can easily lead to parameter redundancy and model efficiency reduction. The standard deviation of SA Weight maps is used to determin the importance of the features.

3) AWW-ASPP module is proposed based on depthwise convolution. The experimental result shows that WW-ASPP significantly reduces parameters and computation of the network with only a small impact on accuracy.

Finally, comparison test proves PSCNet is more balanced in precision and efficiency. The PSCNet improves the MIoU by 5% compared to the Deeplabv3+. Also, the PSCNet has the highest inference speed (47 FPS) and the second highest efficiency per parameter among the networks involved in the comparison.

ACKNOWLEDGEMENTS

This work was supported in part by the National Key Research and Development Program of China (Projects Nos. 2019YFB210310, 2019YFB2103104) and in part by a Research Program of Shenzhen S and T Innovation Committee grant (Projects Nos. JCYJ20210324093012033, JCYJ20210324093600002), the Natural Science Foundation of Guangdong Province grant (Projects No. 2121A1515012574), the Open Fund of Key Laboratory of Urban Land Resources Monitoring and Simulation, MNR(Nos. KF-2021-06-125,KF-2019-04-014), the National Natural Science Foundation of China grant (Projects Nos. 71901147, 41901329, 41971354, 41971341) and the Foshan City to promote scientific and technological achievements of universities to serve industrial development support projects(Projects No. 2020DZXX04).

REFERENCES

Chen, L.-C., Papandreou, G., Schroff, F., and Adam, H.: Rethinking atrous convolution for semantic image segmentation, arXiv preprint arXiv:1706.05587, 2017.

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., and Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation, Proceedings of the European conference on computer vision (ECCV), 801-818,

Hazirbas, C., Ma, L., Domokos, C., and Cremers, D.: Fusenet: Incorporating depth into semantic segmentation via fusion-based cnn architecture, Asian conference on computer vision, 213-228,

He, K., Zhang, X., Ren, S., and Sun, J.: Deep Residual Learning for Image Recognition, 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 27-30 June 2016, 770-778, 10.1109/CVPR.2016.90,

Hu, J., Shen, L., and Sun, G.: Squeeze-and-excitation networks, Proceedings of the IEEE conference on computer vision and pattern recognition, 7132-7141,

Hu, X., Yang, K., Fei, L., and Wang, K.: ACNET: Attention Based Network to Exploit Complementary Features for RGBD Semantic Segmentation, 2019 IEEE International Conference on Image Processing (ICIP), 22-25 Sept. 2019, 1440-1444, 10.1109/ICIP.2019.8803025,

Jiang, J., Zheng, L., Luo, F., and Zhang, Z.: Rednet: Residual encoder-decoder network for indoor rgb-d semantic segmentation, arXiv preprint arXiv:1806.01054, 2018.

Krizhevsky, A., Sutskever, I., and Hinton, G. E.: Imagenet classification with deep convolutional neural networks, Advances in neural information processing systems, 25, 2012.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P.: Gradient-based learning applied to document recognition, Proceedings of the IEEE, 86, 2278-2324, 1998.

Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., and Xie, S.: A ConvNet for the 2020s, arXiv preprint arXiv:2201.03545, 2022.

Lu, R., Chen, B., Cheng, Z., and Wang, P.: RAFnet: Recurrent attention fusion network of hyperspectral and multispectral images, Signal Processing, 177, 107737, 2020.

Park, S.-J., Hong, K.-S., and Lee, S.: Rdfnet: Rgb-d multi-level residual feature fusion for indoor semantic segmentation, Proceedings of the IEEE international conference on computer vision, 4980-4989,

Ronneberger, O., Fischer, P., and Brox, T.: U-net: Convolutional networks for biomedical image segmentation, International Conference on Medical image computing and computer-assisted intervention, 234-241,

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., and Chen, L.-C.: Mobilenetv2: Inverted residuals and linear bottlenecks, Proceedings of the IEEE conference on computer vision and pattern recognition, 4510-4520,

Seichter, D., Köhler, M., Lewandowski, B., Wengefeld, T., and Gross, H.-M.: Efficient rgb-d semantic segmentation for indoor scene analysis, 2021 IEEE International Conference on Robotics and Automation (ICRA), 13525-13531,

Shelhamer, E., Long, J., and Darrell, T.: Fully Convolutional Networks for Semantic Segmentation, IEEE Transactions on Pattern Analysis and Machine Intelligence, 39, 640-651, 10.1109/TPAMI.2016.2572683, 2017.

Silberman, N., Hoiem, D., Kohli, P., and Fergus, R.: Indoor segmentation and support inference from rgbd images, European conference on computer vision, 746-760,

Simonyan, K. and Zisserman, A.: Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556, 2014.

Woo, S., Park, J., Lee, J.-Y., and Kweon, I. S.: Cbam: Convolutional block attention module, Proceedings of the European conference on computer vision (ECCV), 3-19,

Zhou, H., Qi, L., Huang, H., Yang, X., Wan, Z., and Wen, X.: CANet: Co-attention network for RGB-D semantic segmentation, Pattern Recognition, 124, 108468, 2022.

Zhou, W., Yuan, J., Lei, J., and Luo, T.: TSNet: Three-stream self-attention network for RGB-D indoor semantic segmentation, IEEE intelligent systems, 36, 73-78, 2020.