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

DOI: 10.23977/jipta.2025.080115
ISSN 2560-6239 Vol. 8 Num. 1

Optimization of the DeepLabv3+ Segmentation Network by Integrating Multi-Scale Spatial Information

Kui Tang[1], Lingfei Cheng[1,*], Huan Zhang[1]

[1]*School of Physics and Electronic Information, Henan Polytechnic University, Jiaozuo, Henan, China*

**Corresponding author*

Keywords: Deep Learning; Semantic Segmentation; Multi-Scale Features; Global Attention Mechanism

Abstract: Existing algorithms fail to fully utilize the rich semantic and spatial information contained in images, leading to inaccurate pixel segmentation across different categories and severe loss of detail. We propose a semantic segmentation network that focuses on multi-scale spatial information (Focusing on Multi-Scale Spatial Information for Semantic Segmentation Networks, FMSI-DeepLab). Based on the DeepLabv3+ framework, the network is improved in two main parts. In the encoder, deformable convolutions are combined with the Global Grouped Coordinate Attention (GGCA) mechanism to reconstruct the Atrous Spatial Pyramid Pooling (ASPP) module, enhancing the model's ability to capture global information across both height and width spatial dimensions, thereby enabling efficient multi-scale feature extraction. In the decoder part, “interest flow” processing is added to the low-level features, enabling them to have global connectivity at the low-level stage. Subsequently, the Multi-Scale Channel Spatial Enhanced Attention (MSEA) module is introduced to further enhance the model's focus on the edge information of the low-level features extracted by the backbone network, thereby strengthening the model's emphasis on details. Compared to the original DeepLabv3+ semantic segmentation model, the model achieves a 2.62% improvement in average intersection-over-union (mIoU) on the VOC2012 dataset, addressing issues of inaccurate image segmentation and severe loss of details.

1. Introduction

Semantic segmentation is an important research topic in the field of computer vision and plays a crucial role in many practical applications. Image semantic segmentation is a prerequisite for computer recognition and understanding of images. It is widely used in many fields such as autonomous driving [1], medical auxiliary diagnosis[2], indoor and outdoor scene analysis[3], etc., and has broad application prospects and social value .

With the increasing power of deep learning technology, semantic segmentation has entered a new phase of development. Deep learning methods can extract more and deeper feature information from images, enabling end-to-end pixel-level classification of image objects, significantly improving the accuracy and efficiency of semantic segmentation. In 2017, Chen et al.[4] improved the DeepLab method based on DeepLabv1 to create DeepLabv2. DeepLabv3[5] is based on DeepLabv2 and eliminates the corresponding cascaded convolutions, while DeepLabv3+ introduces new encoding and decoding based on DeepLabv3 and performs related cascaded processing to obtain the final features. However, the drawbacks remain evident, particularly in terms of learning long-range dependencies and spatial correlations. This limitation restricts their ability to capture global information, leading to poor performance in tasks requiring detailed analysis of complex scenes. In 2022, Azad R proposed a new image segmentation algorithm, TransDeepLab[6], based on the DeepLab and Transformer[7] architectures. This algorithm introduces an adaptive fusion mechanism for multi-scale features across contexts, achieving significant performance improvements in image segmentation tasks. In 2023, Ouyang [8] et al. proposed a new efficient multi-scale attention module (Efficient Multi-Scale Attention Module with Cross-Spatial Learning, EMA) to improve channel or spatial attention mechanisms, achieving notable results in image classification and object detection tasks. Jiao[9] et al. constructed a multi-scale dilated transformer (DilateFormer) by stacking MSDA blocks at lower stages and global multi-head self-attention blocks at higher stages, enabling more efficient interaction between domain blocks. In 2025, Si[10] et al. proposed combining channel-spatial attention mechanisms (SCSA), aiming to explore and leverage the synergistic relationship between spatial attention and channel attention to mitigate semantic differences between different feature maps.

Currently, most models struggle to fully exploit the full potential of data with rich details and multi-level structures, fail to fully uncover all potential correlations within the data, leading to issues such as blurred edge segmentation, severe loss of detail, poor segmentation performance for small and multi-scale objects, and low robustness. This paper proposes the FMSI-DeepLab model, which adopts a design more suited for multi-scale feature fusion. By leveraging spatial information in the input image and focusing more on contextual relationships, it addresses the issues of low segmentation accuracy for small objects and blurred edge segmentation, thereby improving the segmentation accuracy of multi-scale objects in images.

2. Model Overview

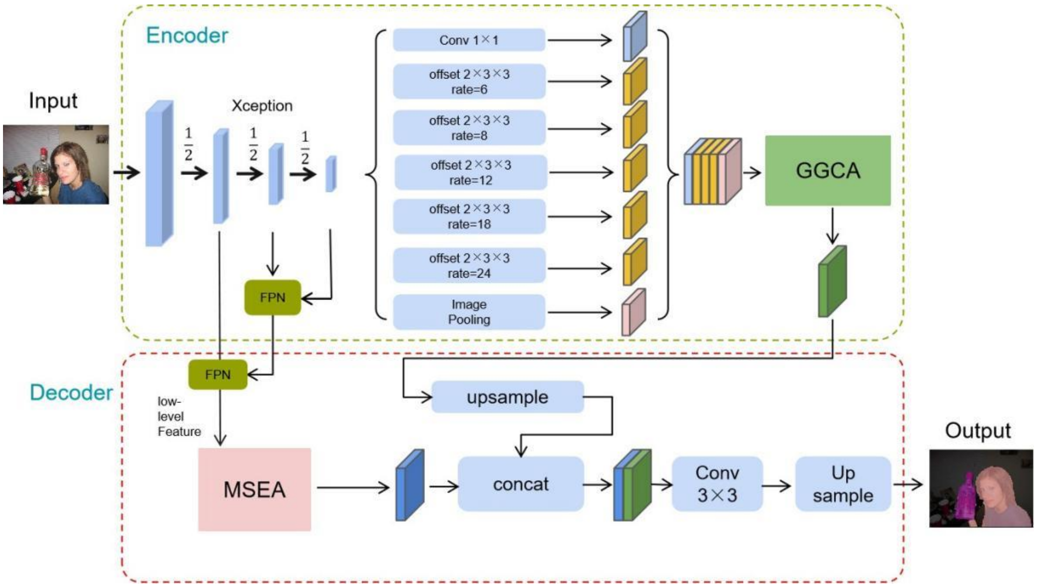

Figure 1 FMSI-DeepLab network structure

The overall structure of the FMSI-DeepLab network consists of two parts: an encoder and a decoder. The encoder uses Xception as the backbone network. Based on the original DeepLab model, it uses deformable convolution layers and global group attention blocks to encode the input image into a space with high attention to height and width. More specifically, the encoder module employs five parallel deformable convolutions with different hole rates and global average pooling to perform multi-scale sampling, capturing the feature information of the image. Then, the obtained multi-scale contextual information is fused into the decoder module using a global coordinate group attention mechanism. In the decoder, the extracted low-level semantic information is first subjected to “interest flow” processing, followed by a multi-scale spatial attention enhancement (MSEA) module to compensate for the lack of cross-channel correlation in feature extraction by the backbone network. Next, high-level multi-scale features are bilinearly upsampled, and finally, the processed low-level and high-level features are concatenated (Concat) to refine the feature representation. Through the final upsampling, pixel points are classified and restored to a segmentation map of the input image size. The overall network structure is shown in Figure 1, and the details of each component of the network will be described in detail in the following sections.

2.1 Encoder

A reconstruction of the ASPP module in the original Deeplabv3 to leverage more spatial information, specifically composed of two parts: deformable convolutions and a global coordinate-based attention mechanism. Experiments were conducted by modifying the original ASPP module's convolution dilation rate (r=6, 12, 18) to (r=6, 8, 12, 18, 24). The more densely sampled approach effectively improves segmentation accuracy. The dilation rate r controls the sampling interval, expanding the receptive field without increasing the number of parameters. For example, a 3×3 convolution with a dilation rate of 6 is equivalent to covering a 13×13 pixel region. By combining the wide-area coverage capability of dilated convolutions with the geometric adaptability of deformable convolutions, the model can simultaneously capture details and global information in complex scenes. Information about irregular targets or targets that have been rotated or deformed can be effectively collected. Finally, a global group attention mechanism is introduced to further enhance the model's ability to process multi-scale targets.

The Deformable Convolution (Def-Conv) block has two types of convolution kernels: traditional convolution kernels and convolution kernels corresponding to learned offsets. First, a standard convolution operation is used to perform preliminary feature encoding on the input image, yielding a basic feature map. The basic feature map is then input into the offset prediction branch, where a dedicated convolution layer (with 2N channels) analyzes the spatial deformation patterns of the features. Deformable convolution adds learnable offsets (Offset) to each sampling point, as shown in Equation 1:

$$y(p)=\sum_{k=1}^{K} w_k \cdot x(p+p_k+\Delta p_k) \tag{1}$$

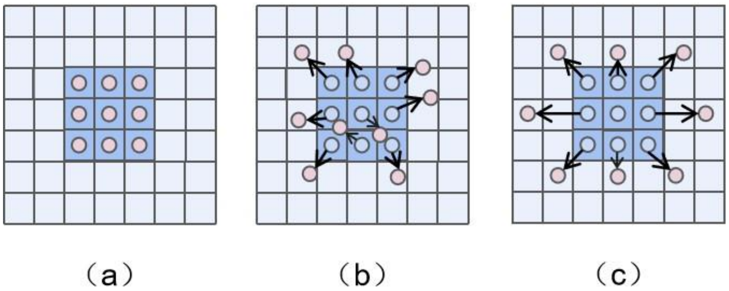

Figure 2 Principle of Deformable Convolution

Among these, the Δp_k for the Kth sampling point is determined through deformable convolution, which learns the offset to shift the sampling point toward key feature locations, dynamically optimizing the receptive field and achieving more efficient and robust feature extraction. The additional convolutional layer prediction enables the convolution kernel shape to dynamically adapt to input features. Both types of convolution kernels undergo parameter updates simultaneously via bilinear interpolation backpropagation. The module principle is shown in Figure 2, achieving end-to-end deformable adaptive learning. Among them, Figure 2 (a) shows the common 3x3 convolution kernel sampling method, (b) shows the sampling of deformable convolution, and the changes in sampling points after adding the offset, and (c) shows a special form of deformable convolution:

Traditional attention mechanisms lack the ability to simultaneously capture global information in both the height and width spatial dimensions, resulting in limited feature representation capabilities. The Global Coordinate Group Attention (GGCA) module captures multidimensional global information by performing global average pooling and max pooling in the height and width directions, respectively, thereby enhancing the comprehensiveness of feature extraction. The specific structure of this module is shown in Figure 3.First, we divide the input feature map $X \in R^{B\times H\times C\times W}$ into “G” groups according to the number of channels, with each group containing “C/G” channels. Here, “B” is the batch size, “C” is the number of channels, and ‘H’ and “W” are the height and width of the feature map, respectively. The feature map after grouping is represented as $X \in R^{B\times H\times C/G\times W}$, then we perform global average pooling and global max pooling operations on the feature map after grouping in the height and width directions, respectively, as shown in Equations (2-5):

$$X_{h,avg} = AvgPool(X) \in R^{B\times G\times \frac{C}{G}\times H\times 1} \tag{2}$$

$$X_{h,max} = MaxPool(X) \in R^{B\times G\times \frac{C}{G}\times H\times 1} \tag{3}$$

$$X_{w,avg} = AvgPool(X) \in R^{B\times G\times \frac{C}{G}\times 1\times W} \tag{4}$$

$$X_{w,max} = MaxPool(X) \in R^{B\times G\times \frac{C}{G}\times 1\times W} \tag{5}$$

For each group feature map, we apply a shared convolutional layer for feature processing. This shared convolutional layer consists of two 1×1 convolutional layers, a batch normalization layer, and a ReLU activation function, which are used to reduce and restore the channel dimension. Finally, we weight the input feature maps according to the attention weights to obtain the output feature map as shown in Equation 6:

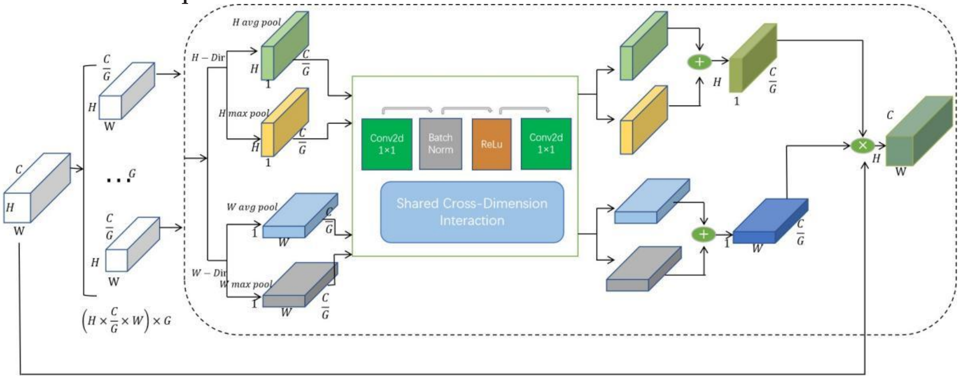

Figure 3 Global Grouped Coordinate Attention

$$O = X \times A_h \times A_w \in R^{D \times C \times H \times W} \tag{6}$$

Here, the attention weights are expanded in the height and width directions to match the size of the input feature map. The GGCA module groups the input feature maps by channel number through batch processing, reducing the computational load of each group of feature maps while maintaining the diversity and richness of feature expressions.

2.2 Decoder

For the main Xception network, channel-independent depth-separable convolutions are used to reduce the number of parameters while sacrificing cross-channel interaction capabilities. Additionally, different spatial locations may contain information of varying importance, but traditional convolution operations treat all spatial locations equally, failing to highlight critical regions. To achieve more efficient feature extraction, this paper employs “interest flows” to enhance the richness of low-level semantic information. The backbone network is divided into Entry Flow (early layers), Middle Flow (deep layers), and Exit layers. Low-level features are only derived from the early layers, resulting in a significant lack of spatial information from other parts. Therefore, we propose combining the Exit layer features with the Middle layer features after 2x upsampling, and then combining the resulting features with the Entry layer feature maps after another 2x upsampling. This process is inspired by the Feature Pyramid Network (FPN)[11]. This “interest flow” operation ensures that the low-level information extraction process incorporates more comprehensive spatial information. Additionally, a multi-scale spatial attention mechanism (MSEA) is designed, which enhances feature representation through multi-scale convolutions and dual attention mechanisms (channel attention and spatial attention). The structure is shown in Figure 4.It effectively addresses issues such as insufficient expression of multi-scale features, insufficient attention between channels, and insufficient exploration of the importance of spatial regions when convolutional neural networks process images. The overall structure is shown in Figure 4. Before merging low-level feature information in the decoder, features at different spatial scales are captured through separable convolutions of varying depths. Then, by evaluating the importance of each channel through channel attention and performing corresponding adjustments, useful feature channels are strengthened while unimportant channels are suppressed, thereby improving the quality of feature representation. The average pooling formula and max pooling are given by Equations (7-8):

$$fcl\_avg = Conv_{k=1}\left(ReLU\left(Conv_{k=1}\left(avg\_out\right)\right)\right) \tag{7}$$

$$fcl\_max = Conv_{k=1}\left(ReLU\left(Conv_{k=1}\left(max\_out\right)\right)\right) \tag{8}$$

The channel attention model is as follows: $C_a = \sigma(fcl\_avg + fcl\_max)$ Spatial attention is then introduced: focusing on the importance of specific regions in an image helps the model concentrate on processing the local regions that have the greatest impact on the final task. In the spatial attention submodule, channel-wise average pooling and max pooling operations are performed on the input features. The results of these two pooling operations are combined through a convolutional layer, and the Sigmoid activation function generates spatial attention weights to highlight important spatial regions.σ represents the sigmoid function. $AvgP(x)$ denotes average pooling, $MaxP(x)$ denotes max pooling, and spatial attention is given by equation (9):

$$S_a = \sigma\left(Conv(concat([AvgP(x), MaxP(x)], dim = 1))\right) \tag{9}$$

After generating channel attention and spatial attention, they are respectively weighted and fused with the original features. The final output combines enhanced features with channel and spatial attention information. The final output is given by Equation (10):

$$X_{final} = X_{input} * S_a + X_{input} * C_a \tag{10}$$

By optimizing the main features and using the attention mechanism, the lower-level semantic features have more spatial information before being fused with the higher-level semantic features, thereby improving the model's comprehensive grasp of the contours of the segmentation target and enhancing the segmentation accuracy.

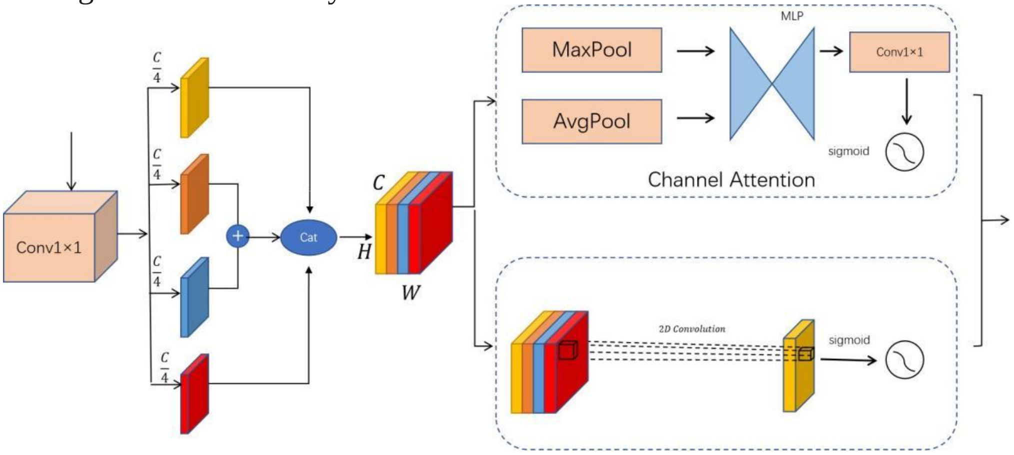

Figure 4 Multi-Scale Channel Spatial Enhanced Attention

3. Experiment results and analysis

3.1 Dataset Introduction

This paper uses the VOC 2012 dataset (PASCAL Visual Object Classes Challenge 2012, VOC 2012) to validate model performance. Initiated by the EU PASCAL network, it is a classic dataset in the field of computer vision, used for tasks such as object detection and segmentation. Data scale: Number of images: 11,530, divided into training and testing sets at a 9:1 ratio. Annotated objects: 21 categories (including background, people, vehicles, cats, dogs, airplanes, etc.).

3.2 Experimental conditions

3.2.1 Experimental environment

The experiment was based on the PyTorch deep learning framework and CUDA 11.8 library, using a Python 3.8 environment. The training parameter BatchSize was set to 4, and NVIDIA GeForce RTX 4090 was used for training. The downsampling rate of the main network was 16, the maximum learning rate was 0.007, the minimum learning rate was 7×10-5, and the total number of training epochs was 100.

3.2.2 Evaluation criteria

The experiment used the F1 score, mean intersection over union (mIoU), and mean pixel accuracy (MPA) metrics for evaluation. mIoU and MPA are commonly used evaluation metrics in image semantic segmentation tasks. mIoU measures the accuracy of the model by calculating the ratio of the intersection and union between the predicted region and the true label. It considers the overlapping region between the prediction and the true label, reflecting the quality of the segmentation boundaries between different categories. MPA is used to measure the model's prediction capability for each category of pixels. The specific formulas are (11-13):

$$F1 = \frac{2 \times P \times R}{P + R} \tag{11}$$

$$mIou = \frac{1}{k+1} \sum_{i=0}^{k} \frac{TP}{FN + FP + TP} \tag{12}$$

$$mPA = \frac{1}{k} \sum_{i=1}^{k} \frac{TP_i}{TP_i + FP_i} \tag{13}$$

3.3 Analysis of experimental results

3.3.1 Comparison and visualization of segmentation results from different models

To validate the effectiveness of the proposed algorithm in this paper, it was compared with multiple models on the VOC2012 dataset. The effectiveness of the proposed model was validated using F1-score, mIoU and MPA metrics. The models included in the comparison were U-Net, PSPNet, HRNet, Deeplabv3+ and SwinU-Net. The results of the comparison experiment are shown in Table 1. As shown in the table, the proposed model achieved 86.4%, 90.13%, and 81.65% on the F1-score, mIoU, and MPA metrics, respectively. Compared to the original Deeplabv3+ model, the average intersection-over-union (78.98%) was improved by 2.62%, and the model achieved better results in edge object and multi-scale object segmentation.

Table 1 Comparison of experimental results from different models

Model	Backbone	F1/%	MPA/%	mIoU/%
Unet[12]	VGG	83.3	72.30	60.14
PSPnet[13]	Mobilenetv2	81.5	81.11	69.93
Hrnet[14]	Hrnetv2	80.9	83.17	73.29
Deeplabv3+	Xception	83.5	85.45	78.98
SwinUnet[15]	Swin-T	-	-	67.80
Ours	Xception	86.4	90.13	81.65

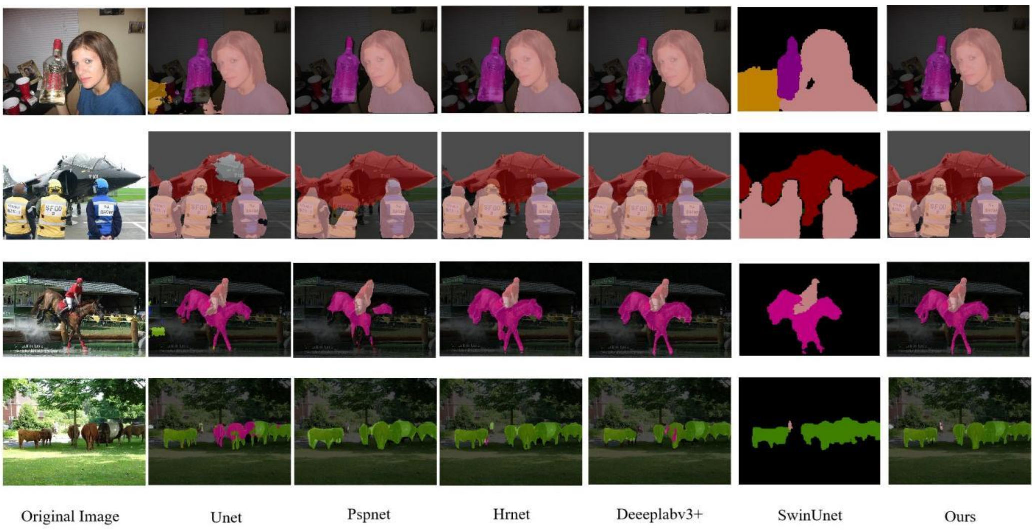

Figure 5 Visualization of Comparison Experiments from Different Models

To further illustrate the segmentation results of different models, Figure 5 presents partial visualization comparison results between the proposed network and other network models. It is evident that the proposed model performs more smoothly in extracting complex objects and captures details of small objects more effectively. Overall, the proposed model demonstrates a significant advantage over other models. For example, in the first row, the U-Net model exhibits issues such as the bottom of a bottle being missed, while in the third row, the arm of a person is not segmented, and in the fourth row, the herd of cows is not properly segmented. Although the PSPNet, Deeplabv3+, and HRNet models do not miss any objects, their segmentation results are blurry. As shown in the sixth column of the fifth comparison experiment, the SwinU-Net model exhibits severe blurring of object edges during segmentation. The model proposed in this paper achieves complete segmentation, accurate detection of small objects, and overall performance superior to other models.

3.3.2 Comparison of accuracy of different convolution hole rates in ASPP improvement

To further validate the different effects of varying hole rates on ASPP reconstruction as proposed in this paper, we conducted comparative experiments with different hole rates. Model a in the table has hole rates of 6, 12, and 18, while model b has hole rates of 6, 8, 12, 18, and 24. The original ASPP may have “blank areas” in the receptive fields of its expansion branches. Denser parallel branches can “fill in” these regions, reducing the risk of local information loss while more uniformly and continuously covering the spatial information of the feature map, thereby enhancing the capture of multi-scale features. As shown in table 2, the proposed changes in dilation rate result in a 0.89% increase in mIoU and a 0.7% increase in MPA for Model B compared to Model A. This validates the effectiveness of using different dilation rates from the original model.

Table 2 Experimental results for different void ratios

Model	Method	F1/%	MPA/%	mIoU/%
a	Deformable	84.5	88.16	79.39
b	Deformable	85.6	88.86	80.28

3.3.3 Ablation study

To validate the effectiveness of the Def-Conv convolution block combined with GGCA and MSEA attention mechanisms in improving the model's semantic segmentation performance, and to validate the accuracy of the evaluation in this paper, ablation experiments were conducted using different combinations of the four modules: SAMF, EMA, MSEA, and GGCA. The experiments were divided into five groups, and the results are shown in Table 3. Group 1 is the original Deeplabv3+ model. Group 2 replaces the original ASPP module with a parallel structure using Def-Conv convolutions, resulting in an increase in the segmentation metric mIoU to 80.28% and the metric MPA to 88.86%. Group 3 builds upon Group 2 by introducing FPN processing at the low-level semantic layer to enrich the underlying semantic information space. The segmentation metric mIoU increased to 80.65%. Group 4 introduced the GGCA attention mechanism to reconstruct the ASPP on the basis of Group 3, with metrics increasing to 89.53% and 81.06%, respectively, demonstrating the effectiveness of the GGCA attention mechanism. Group 5 is the model proposed in this paper, which adds the MSEA attention mechanism after low-level semantic extraction on the basis of Group 4. Group 5 achieved an average intersection-over-union ratio of 81.06% and an MPA of 89.53%, demonstrating a significant performance improvement and validating the correctness of the approach proposed in this paper.

Table 3: Ablation Study Segmentation Experiment Results

Xception	Def-Conv	FPN	MSEA	GGCA	MPA	mIoU
√					85.45	78.98
√	√				87.86	80.28
√	√	√			88.45	80.63
√	√	√		√	89.53	81.06
√	√	√	√	√	90.13	81.65

4. Conclusion

In summary, this paper builds upon Deeplabv3+ by introducing “interest flow” processing during the extraction of low-level semantic features, enabling the backbone network to incorporate a small amount of global spatial information when extracting low-level information. Subsequently, the MSEA attention mechanism is employed to enhance the backbone network's focus on the channel spatial dimension, thereby supplementing multi-scale spatial information. When processing high-level information, the paper introduces deformable convolutions and the GGCA attention mechanism to construct an ASPP module that focuses on more spatial information. The efficient utilization of multi-scale spatial information significantly improves the model's performance. Experimental results show that the proposed network achieves a 2.62% improvement in the mean intersection over union (mIoU) metric and a 4.68% improvement in the mean pixel accuracy (MPA) metric compared to the Deeplabv3+ model. The proposed model addresses the limitations of existing models in terms of spatial information utilization during image segmentation, resolving issues such as blurred multi-object segmentation and missing segmentation content. The overall effectiveness of the model was validated through multiple ablation experiments. Future research will focus on algorithm performance optimization and model lightweighting to further enhance the model's practicality and generalization capabilities.

References

[1] L. Wang and Y. Huang, "A Survey of 3D Point Cloud and Deep Learning-Based Approaches for Scene Understanding in Autonomous Driving," in IEEE Intelligent Transportation Systems Magazine, vol. 14, no. 6, pp. 135-154, Nov.-Dec. 2022, doi: 10.1109 MITS.2021.3109041.

[2] Chen J, Lu Y, Yu Q, et al. Transunet: Transformers make strong encoders for medical image segmentation[J]. arXiv preprint arXiv:2102.04306, 2021.

[3] Luo Z, Yang W, Yuan Y, et al. Semantic segmentation of agricultural images: A survey[J]. Information Processing in Agriculture, 2024, 11(2): 172-186.

[4] Chen L C, Papandreou G, Kokkinos I, et al. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(4):834-848.

[5] Chen, Liang-Chieh, et al. "Rethinking atrous convolution for semantic image segmentation. [J]" arXiv preprint arXiv: 1706.05587 (2017).

[6] Azad R, Heidari M, Shariatnia M, et al. Transdeeplab: Convolution-free transformer-based deeplab v3+ for medical image segmentation[C]//International Workshop on PRedictive Intelligence in MEdicine. Cham: Springer Nature Switzerland, 2022: 91-102.

[7] Han K, Xiao A, Wu E, et al. Transformer in transformer[J]. Advances in neural information processing systems, 2021, 34: 15908-15919.

[8] Ouyang D, He S, Zhang G, et al. Efficient multi-scale attention module with cross-spatial learning[C]//ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2023: 1-5.

[9] Jiao J, Tang Y M, Lin K Y, et al. Dilateformer: Multi-scale dilated transformer for visual recognition[J]. IEEE Transactions on Multimedia, 2023, 25: 8906-8919.

[10] Si Y, Xu H, Zhu X, et al. SCSA: Exploring the synergistic effects between spatial and channel attention[J]. Neurocomputing, 2025, 634: 129866.

[11] Li H, Zhang R, Pan Y, et al. Lr-fpn: Enhancing remote sensing object detection with location refined feature pyramid network[C]//2024 International Joint Conference on Neural Networks (IJCNN). IEEE, 2024: 1-8.

[12] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation[C] //Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18. Springer international publishing, 2015: 234-241.

[13] Zhao H, Shi J, Qi X, et al. Pyramid scene parsing network[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 2881-2890.

[14] Wang J, Sun K, Cheng T, et al. Deep high-resolution representation learning for visual recognition[J]. IEEE transactions on pattern analysis and machine intelligence, 2020, 43(10): 3349-3364.

[15] Cao H, Wang Y, Chen J, et al. Swin-unet: Unet-like pure transformer for medical image segmentation[C] //European conference on computer vision. Cham: Springer Nature Switzerland, 2022: 205-218.