Backbone Feature Enhancement and Decoder Improvement in HRNet for Semantic Segmentation

HanLei Feng, TieGang Zhong
School of Electronic and Information Engineering, Liaoning Technical University, Huludao Liaoning, 125105, China

***Abstract*—Addressing issues such as the tendency for small-scale objects to be lost, incomplete segmentation of large-scale objects, and overall low segmentation accuracy in existing semantic segmentation models, an improved HRNet network model is proposed. Firstly, by introducing multi-branch deep stripe convolutions, features of multi-scale objects are adaptively extracted using convolutional kernels of different sizes, which not only enhances the model's ability to capture multi-scale objects but also strengthens its perception of the contextual environment.Secondly, to optimize the feature aggregation effect, the axial attention mechanism is adopted to aggregate image features along the x-axis and y-axis directions respectively, effectively capturing long-range dependencies within the global scope, and thus achieving precise positioning of objects of interest in the feature map.Finally, by implementing the progressive fusion-based upsampling strategy, it facilitates the complementary fusion of semantic information and detailed information between adjacent feature maps, thereby enhancing the model's capability to restore fine-grained details in images. Experimental results demonstrate that on the PASCAL VOC2012+SBD dataset, the mean Intersection over Union (mIoU) of the improved HRNet_S model in segmenting lower-resolution images is increased by 1.54% compared to the baseline method. Meanwhile, the improved HRNet_L model achieved a 3.05% increase in mIoU compared to the original model when handling higher-resolution image segmentation tasks on the Cityscapes dataset, and attained the highest segmentation accuracy in 15 out of the 19 different scale classification categories on this dataset.These results indicate that the proposed method not only exhibits high segmentation accuracy but also possesses strong adaptability to multi-scale objects.**

Keywords*—Semantic segmentation; HRNet; multi-branch deep strided convolution; axial attention mechanism; progressive fusion upsampling; multi-scale object adaptability

I. Introduction

Semantic segmentation aims to assign each pixel in an image to a specific semantic category, thereby enabling pixel-level understanding and segmentation of the image. In recent years, with the rapid development of fields such as autonomous driving and medical image processing, semantic segmentation technology has received widespread attention. Semantic segmentation models based on neural networks have become a research hotspot due to their high-precision segmentation capabilities. Although semantic segmentation technology has achieved remarkable application effects in multiple fields, current semantic segmentation methods still face many challenges due to the complex and variable nature of segmentation scenarios and the limitations of the segmentation models themselves. Firstly, the types of objects to be segmented are numerous, and there are significant differences in object scales. Existing semantic segmentation models have difficulties in accurately segmenting object edges and often encounter situations where small-scale objects are ignored or large-scale objects are segmented incompletely. Secondly, most existing models adopt an encoder-decoder architecture. To reduce computational complexity, multiple downsampling operations are performed during the encoder stage, which leads to the loss of a large amount of detailed information that is difficult for the decoder network to effectively recover through learning. Furthermore, existing models have not fully utilized the effective information in feature maps at various levels, making it difficult to strike a balance between shallow detailed information and deep semantic information. In response to the above issues, this paper proposes several improvements based on the HRNet [1] network framework. The main contributions are as follows:

- A Backbone Feature Enhancement Module (BFEM) is designed, which utilizes depth-wise strip convolutions of different scales to adaptively extract features from the backbone network's output. This generates appropriate weights for objects at each scale, addressing the issue of a single convolution kernel's difficulty in adapting to variations in object scales.
- A Flexible Upsampling Mechanism (FUSM) is proposed, which enhances the model's ability to restore image details by facilitating continuous information exchange between adjacent feature maps.
- To overcome the limitations of traditional spatial attention mechanisms in modeling long-distance dependencies, a novel Axial Attention Mechanism (AAM) is proposed. It independently generates attention maps along the length and width dimensions in the spatial dimension and sequentially applies these attention maps to the weighted processing of the input feature maps. While maintaining a lightweight design, it endows the output feature maps with remarkable directional sensitivity, effectively preserving the key positional information in the input features.
- For segmentation tasks involving input images with low resolution, the HRNet backbone network is combined with the Backbone Feature Enhancement Module, the Flexible Upsampling Module, and the Efficient Channel Attention Mechanism to construct the HRNet_S model.Furthermore, to address the challenge of segmenting high-resolution images, the Flexible Upsampling Module is further integrated with the AAM to optimize the decoder network of the HRNet model, resulting in the development of the HRNet_L model.

II. Related Works

To further elevate network performance, current research primarily focuses on three pivotal technological directions: aggregating multi-scale contextual information, enhancing global perception capabilities, and improving detail capture abilities.

In terms of aggregating multi-scale contextual information, multi-scale contextual aggregation networks based on dilated convolutions are widely employed. DeepLabv2 [2] achieves robust segmentation of objects across different scales by introducing atrous spatial pyramid pooling. DeepLabv3 [3] further designs a cascaded module utilizing dilated convolutions, bolstering the network's ability to detect convolutional features across multiple scales. PSPNet [4] harnesses the power of pyramid pooling modules to capture global contextual information for scene parsing tasks. DenseASPP [5] leverages an atrous spatial pyramid structure to extract a broader range of scale feature information. APCNet [6] introduces the Adaptive Pyramid Context Network, constructing multi-scale contextual representations through its meticulously designed Adaptive Context Module (ACM). SCTNet [7] enhances the efficiency of capturing multi-scale contextual information by learning semantic information alignment from Transformer to CNN.R-FPN [8] introduces learnable weights, enabling the network to adaptively utilize effective information from different scale feature maps, thereby improving the effectiveness of small object detection. The improved RCN model proposed by Zhu et al. utilizes the RFN module for multi-scale feature fusion, significantly enhancing the network's ability to recognize microscopic images [9].

To strengthen global perception and capture long-range dependencies, CCNet [10] incorporates a criss-cross attention module, collecting contextual information along criss-cross paths. OCNet [11] proposes an efficient interlaced sparse attention scheme, modeling pixel relationships through sparse relation matrices. DANet [12] adaptively integrates local and global feature dependencies, modeling semantic dependencies in both spatial and channel dimensions. Vision Transformer [13] utilizes a self-attention mechanism, dividing images into patches to construct pixel relationships and accurately capture long-distance dependencies. Building upon this, Swin Transformer [14] and Segformer [15] introduce multi-scale feature extraction and integration methods, enabling models to better understand the dependencies between elements. PIDNet [16] ingeniously utilizes ratio branches to analyze and preserve rich details in high-resolution feature maps, while leveraging integral branches to synthesize local and global contextual information, thereby capturing and processing complex long-range dependencies.

Regarding improving detail capture abilities, DeepLabv3+ [17] introduces early high-resolution feature maps during the decoding stage, enhancing segmentation accuracy at object boundaries. The UNet [18] network enhances high-resolution representations of feature maps by concatenating shallow feature maps with upsampled deep feature maps through skip connections. TransUNet [19] integrates a Transformer architecture into the UNet model, leveraging Transformers to encode image patches into sequences to capture global contextual information.

However, the improvements of most models in the past have often been confined to a specific area, which has somewhat restricted the potential for enhancing their performance [20], [21], [22]. After deeply analyzing the common characteristics of these successful cases, we successfully extracted three core elements for improving model performance. Based on this, we are committed to skillfully integrating these three aspects of improvement in order to achieve a significant leap in model performance. Therefore, we adopted a cascaded HRNet network as the basic building block, aiming to efficiently aggregate multi-scale contextual information. Subsequently, we introduced the Backbone Feature Enhancement Module and a finely designed attention mechanism to further strengthen the model's feature aggregation and representation capabilities. Finally, by introducing the Flexible Upsampling Mechanism, we significantly improved the model's ability to capture details. Compared to improvements in a single aspect, we conducted experiments to deeply explore the interaction mechanisms among different core elements and accordingly achieved an organic integration of various key improvements, thereby making the performance enhancement of the model more significant and comprehensive.

III. Proposed Method

A. Overall Model Architecture

The proposed improved semantic segmentation model based on HRNet is shown in Fig. 1. The network using the upper half of the figure as the model decoder is named HRNet_S, while the network using the lower half of the figure as the model decoder is named HRNet_L.

Firstly, the image is fed into the HRNet feature extraction network to initially obtain the deep semantic information and shallow detail information required by the model. The HRNet feature extraction network is primarily composed of four transition modules and four stage modules. The Transition module is primarily responsible for transforming the number of channels and performing downsampling operations. After processing by this module, the number of feature map channels between adjacent branches differs by a factor of two, and the length and width of the lower-resolution feature maps are half of those of the higher-resolution feature maps. Consequently, the feature maps in each connected Stage exhibit a cascaded pyramid structure. The Stage module is primarily composed of Basic Block modules, whose internal structure is identical to the residual connection modules in the ResNet network, aiming to further extract deeper-level features from the image. Within these modules, continuous skip connections occur between the branches of the cascaded pyramid feature maps to achieve the fusion of low-level and high-level features.

The HRNet_S network enriches the multi-scale characteristics of the lowest and second-lowest resolution feature map branches by introducing the Backbone Feature Enhancement Module (BFEM), which utilizes depth-wise strip convolutions with different kernel sizes. This process extends the semantic information depth of these two branches. Subsequently, the Flexible UpSampling Mechanism (FUSM) upsamples the four feature map branches in ascending order from low to high resolution. During the upsampling process, continuous information exchange occurs between different branches, achieving precise complementarity between semantic information and

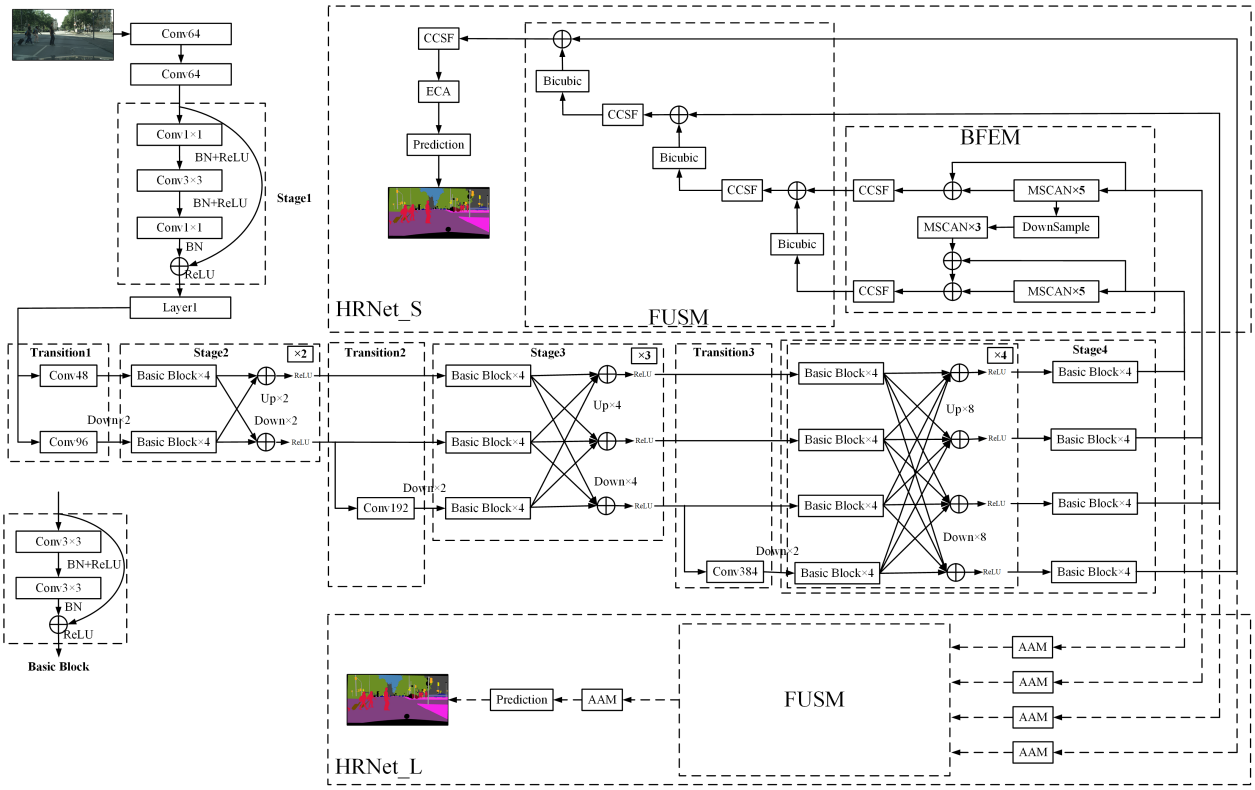

Fig. 1. Improved HRNet network architecture diagram.

detail information. Finally, the feature maps after cross-channel semantic fusion are fed into the Efficient Channel Attention (ECA) module for channel-wise weighting, enabling the network to focus on important features in a targeted manner.

In the HRNet_L network model, the feature maps output by the backbone network undergo feature aggregation along different spatial directions for each channel branch through the Axial Attention Mechanism (AAM), enhancing the model's position sensitivity to objects of interest. Immediately afterward, the feature maps of these four branches undergo upsampling in the same manner as in HRNet_S. Finally, the AAM module is utilized again to further aggregate and enhance the information in the feature maps along the spatial dimension.The HRNet_M model builds upon the HRNet_L model by adding a Cross-Channel Semantic Fusion (CCSF) module after the FUSM, while the remaining parts of the model remain consistent with the HRNet_L model.

B. Backbone Feature Enhancement Module

The structure of the Backbone Feature Enhancement Module (BFEM) is depicted in the BFEM section of Fig. 1. This module utilizes the Multi-Scale Convolutional Attention Network (MSCAN) [23] as the basic unit and strengthens the input features through step-by-step parallel connection. The final output of the sub-low-resolution feature map branch consists of two parts: one is the original feature map of the input branch, and the other is the multi-scale feature map enhanced by the original feature map after passing through five MSCAN modules. For the lowest resolution feature map branch, in addition to performing the same operations as the second-lowest resolution branch, it also concatenates with an output feature map that has undergone a $2\times$ downsampling operation and processing through three MSCAN modules. Since the concatenated branch comes from a deeper layer of the network, it contains richer and more in-depth semantic information, which is crucial for improving the accuracy of the network's category predictions. Meanwhile, the output feature maps from deeper layers have a larger receptive field, enhancing the network's ability to maintain the integrity of object segmentations and recognize large-scale objects. Finally, the semantic information from the two branches is fused through the Cross-Channel Semantic Fusion (CCSF) module. The CCSF module is composed of 1×1 convolution operations, batch normalization operations, and ReLU activation functions, enabling efficient information exchange between channels with a very low number of parameters. This operation can be represented by formula (1).

$$\mathrm{Out} = \mathrm{ReLU}\Big(\mathrm{BN}\Big(\mathrm{Conv}_{1\times1}\Big(F\Big)\Big)\Big) \tag{1}$$

Where, F represents the input feature map; Out represents the output feature map after being processed by the CCSF module.

Fig. 2(a) demonstrates the structure of the Multi-Scale Convolutional Attention (MSCA) module. Firstly, a depthwise convolution with a 5×5 kernel is employed to aggregate local information. Then, depthwise strided convolutions with sizes of 7, 11, and 21 are utilized to capture multi-scale contextual information, aiding the model in understanding contexts across

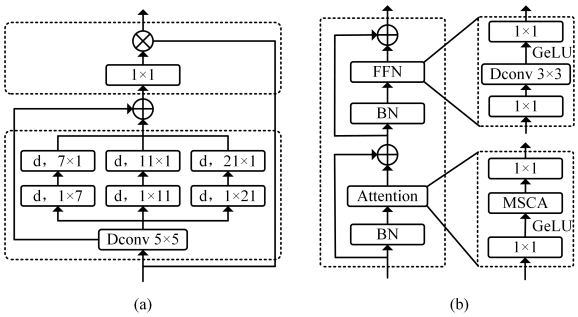

Fig. 2. (a) MSCA module and (b) MSCAN module.

different ranges [23]. Lastly, a 1×1 convolution is applied to model the relationships between different channels in the feature map, and its output is directly used to adjust the weights of the convolutional attention, achieving a weighted reconstruction of the output feature map. Compared to ordinary convolutions, the depthwise strided convolutions utilized in MSCA significantly reduce the number of model parameters. For example, a pair of 1×21 and 21×1 strided convolutions can replace a conventional 21×21 convolution, with the former requiring only 21+21=42 parameters while the latter requires 21×21=441 parameters.Mathematically, MSCA can be concisely described as a combination operation as shown in Eq. (2).

$$\begin{aligned} \text{Att} &= \text{Conv}_{1\times1}\left\{\sum_{i=0}^{3}\text{Scale}_i\Big[\text{DW-Conv}\left(F\right)\Big]\right\} \\ \text{Out} &= \text{Att}\otimes F \end{aligned} \tag{2}$$

Where: F represents the input feature map; Att denotes the attention map; DW-Conv stands for depthwise convolution; Out represents the output weighted attention feature map; $\oplus$ indicates element-wise multiplication operation in matrices; and Scale_i is the i-th branch in the MSCA diagram, where $i \in \{0, 1, 2, 3\}$.

The internal structure of MSCAN, as depicted in Fig. 2(b), primarily comprises three components: Batch Normalization (BN), Feed-Forward Neural Network (FFN), and Multi-Scale Convolutional Attention (MSCA). Among them, the FFN adopts a ResNet-like bottleneck design, aimed at reducing and expanding feature dimensions, thereby minimizing the model's parameter count and computational cost. The attention module shares a similar structure with the FFN module, with the notable difference being the replacement of the original 3×3 depthwise convolution block with the MSCA module to capture richer multi-scale contextual information. The multi-scale features and attention-weighted features output by the attention module are fused through the FFN module, generating the final feature representation.

C. Flexible Upsampling Mechanism

Upsampling operations are crucial steps indispensable for decoder networks to restore image dimensions. Traditional upsampling methods typically involve directly upsampling the deep feature maps or concatenating the upsampled deep feature maps with shallow feature maps. However, this approach has notable limitations: deep feature maps tend to lack detailed information compared to shallow feature maps, and after upsampling, there are significant differences in the distribution of semantic information between the two. This disparity and lack of information often lead to a decrease in final segmentation accuracy. To more effectively utilize the semantic and detailed information from feature maps at different scales and enhance the decoder network's ability to restore detailed information, this paper proposes a bottom-up adjacent feature map priority fusion upsampling strategy. As shown in the FUSM part of Fig. 1, this mechanism can be described by the following steps:

1) For the feature map output from the lowest resolution branch of the HRNet backbone network, a bicubic interpolation algorithm is employed for upsampling. This algorithm considers the grayscale values and their rates of change for the 16 surrounding pixels around the sampling point, generating an enlarged effect that is closer to high-resolution images, effectively preserving image details and mitigating blurring. Subsequently, the upsampled feature map is concatenated with the feature map from the next lower resolution branch along the channel dimension. Since the resolutions of the two are similar, the differences in the spatial distribution of information are relatively small. This facilitates precise alignment of key features between the feature maps, thereby promoting the retention of detailed information and reducing the introduction of noise. Following this, the concatenated feature map is processed by the CCSF module, which models the feature maps from different branches along the channel dimension, enhancing the expressive ability of important features.
2) The fused feature map is then upsampled using bicubic interpolation to match the size of the next higher resolution branch's feature map, and channel concatenation is performed with it. This process is repeated until all feature maps are upsampled to the size of the highest resolution feature map. This not only supplements the details that might be lost during the upsampling process but also enhances the semantic information in the high-resolution feature maps. Through gradual upsampling and fusion, the network can continuously interact and integrate semantic information with detailed information, thereby improving both detail recovery and category prediction accuracy. Finally, the feature map upsampled to the highest resolution will be concatenated with the feature map from the highest resolution branch, completing the entire flexible upsampling process.

Assuming F_1,F_2,F_3 and F_4 are the feature maps output by the backbone feature extraction network from the lowest to the highest resolution branches, respectively, the entire process can be expressed as Eq. (3).

$$\text{Out} = \text{CCSF}\Big(\text{Up}\Big(\text{CCSF}\Big(\text{Up}(F_1)\oplus F_2\Big)\Big)\oplus F_3\Big)\oplus F_4 \tag{3}$$

Where, $\oplus$ denotes concatenation along the channel dimension; Up represents a 2× bicubic interpolation upsampling

operation.

D. Axial Attention Mechanism

In deep learning, attention mechanisms have become a vital component, primarily comprising spatial attention and channel attention. Spatial attention mechanisms enhance focus on important regions by modeling correlations among positions on the feature map and redistributing weights to form spatial masks. In contrast, channel attention focuses on capturing the internal relationships among different channels and intelligently adjusts the importance weights of each channel's features. However, many current attention mechanisms, while fulfilling their functions, also increase the computational complexity of the model. For instance, although the Squeeze-and-Excitation (SE) [24] network can efficiently adjust weights among channels, it overlooks the significance of positional information, which is particularly crucial in tasks requiring precise spatial localization, such as semantic segmentation. The Convolutional Block Attention Module (CBAM) [25] attempts to introduce positional information in the channel dimension through global pooling, but its approach primarily relies on the multi-channel maximum and average values at each position as weight bases.This approach often captures only local information and struggles to model long-range dependencies.

To overcome this limitation, this section proposes an Axial Attention Mechanism (AAM). As shown in Fig. 3, to avoid the potential loss of positional information caused by traditional 2D global pooling, AAM decomposes the computation of spatial attention into two separate 1D feature encoding steps, performed along the length and width directions, respectively. The spatial feature vectors aggregated from these two dimensions not only contain information in their respective directions but also implicitly encode spatial position cues, which can be jointly used to efficiently integrate contextual information of spatial coordinates. The 1D pooling calculation processes along the two directions are shown in Eq. (4).

$$\begin{aligned}\text{H\_AvgPool}(F) &= \frac{1}{H}\left[\sum_{0\le j\le H} x(j,1),\cdots,\sum_{0\le j\le H} x(j,W)\right]\\ \text{W\_AvgPool}(F) &= \frac{1}{W}\left[\sum_{0\le i\le W} x(1,i),\cdots,\sum_{0\le i\le W} x(H,i)\right]\end{aligned} \tag{4}$$

Where, H_AvgPool represents the average pooling operation along the height direction, and W_AvgPool represents the average pooling operation along the width direction; H represents the height of the input image, W represents the width of the input image; F represents the input feature map, and $x(i,j)$ represents the pixel value at coordinate (i,j) in F.

Secondly, the two feature vectors rich in specific directional information are further processed to generate two attention maps. Each attention map focuses on its corresponding spatial direction and can effectively capture the long-range dependencies in that direction within the input feature map. In this way, positional information is cleverly encoded and preserved in the generated attention maps. Finally, AAM applies these two attention maps to the original input feature map through element-wise multiplication at corresponding positions, achieving the reweighting of the feature map. This operation process can be described by Eq. (5).

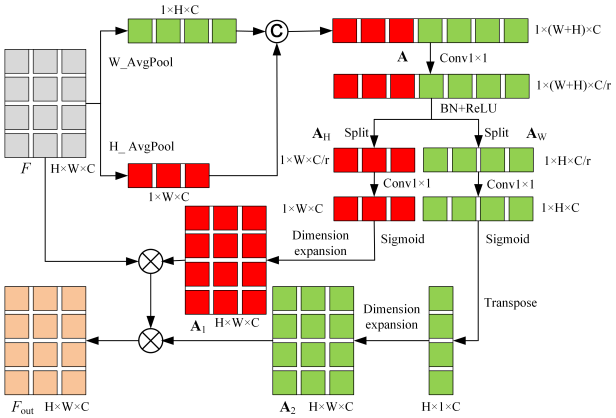

Fig. 3. Axial attention mechanism.

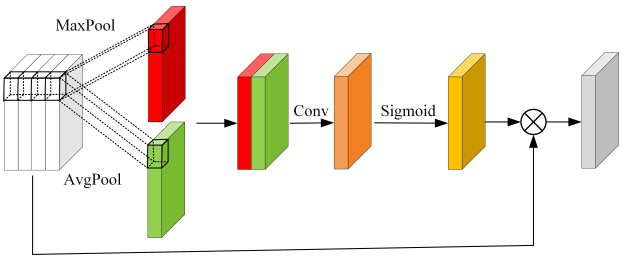

Fig. 4. Traditional spatial attention mechanism.

$$\begin{aligned}&\mathbf{A} = [\text{H\_AvgPool}(F);\text{W\_AvgPool}(F)]\\ &[\mathbf{A}_\text{H};\mathbf{A}_\text{W}]=\text{Split}\Big(\text{ReLU}\Big(\text{BN}\Big(f^{1\times1}(\mathbf{A})\Big)\Big)\Big)\\ &[\mathbf{A}_1;\mathbf{A}_2]=\text{Extend}\left[\sigma\Big(f_1^{1\times1}\Big(\mathbf{A}_\text{H}\Big)\Big);\sigma\Big(f_2^{1\times1}\Big(\mathbf{A}_\text{W}\Big)\Big)'\right]\\ &F_\text{out} = F\otimes\mathbf{A}_1\otimes\mathbf{A}_2\end{aligned} \tag{5}$$

Where, $\mathbf{A}$ represents the feature vector formed by concatenating the features from both the height and width dimensions of the input image;$\mathbf{A}_\text{H}$ and $\mathbf{A}_\text{W}$ represent the intermediate vectors obtained after performing channel dimension reduction on the features from the height and width dimensions, respectively.$\mathbf{A}_1$ and $\mathbf{A}_2$ represent the length and width directional attention matrices obtained by expanding the channel dimension and spatially expanding (replicating the single-dimensional vectors along a certain direction to match the matrix size of F) from $\mathbf{A}_\text{H}$ and $\mathbf{A}_\text{W}$, respectively.$\otimes$ denotes element-wise multiplication of corresponding positions in the matrices.

AAM not only enhances the directional sensitivity of the feature map but also retains crucial positional information, ensuring that the output results more accurately map the target regions in the image. Consequently, when performing dense prediction tasks such as semantic segmentation, AAM can significantly improve the localization accuracy of the model.

Assuming the input feature map has C channels, a height of H, and a width of W, and all convolution operations

employ 1×1 convolutional kernels. In the channel compression stage, AAM uses a compression factor r. Then, the computational complexity of the AAM attention mechanism and the traditional attention mechanism can be estimated using Eq. (6) and (7), respectively, both of which remain at the order of $O(C \cdot H \cdot W)$. Compared to traditional spatial attention mechanisms, AAM effectively captures long-range dependencies while hardly increasing computational complexity. This significant advantage makes AAM more competitive in scenarios involving high-resolution images or high real-time requirements.

$$O(\text{AAM}) = O(4 \cdot C \cdot H \cdot W) + O\Big(2 \cdot \frac{C}{\text{r}} \cdot (H+W)\Big) + O\Big(C \cdot (H+W)\Big) = O(C \cdot H \cdot W) \tag{6}$$

The computational complexity of pooling operations in both the height and width directions is $O\left(2 \cdot C \cdot H \cdot W\right)$; the computational complexity of channel compression and recovery operations is $O\left(2 \cdot \frac{C}{\text{r}} \cdot (H+W)\right)$; the complexity of the Sigmoid function can be considered as $O\left(C \cdot (H+W)\right)$; and the complexity of the two element-wise multiplications can be considered as $O\left(2 \cdot C \cdot H \cdot W\right)$. In practical applications, $\frac{C}{\text{r}}$ can be regarded as a constant, therefore, the above results can be simplified to $O\left(C \cdot H \cdot W\right)$.

The schematic diagram of the traditional spatial attention mechanism is shown in Fig. 4, and its computational complexity can be expressed as Eq. (7).

$$O\left(\cdot\right) = O\left(6 \cdot C \cdot H \cdot W\right) = O\left(C \cdot H \cdot W\right) \tag{7}$$

The computational complexity of max pooling and average pooling is $O\left(2 \cdot C \cdot H \cdot W\right)$; the computational complexity of convolution operations is $O\left(2 \cdot C \cdot H \cdot W\right)$; the computational complexity of the Sigmoid operation is $O\left(C \cdot H \cdot W\right)$; The computational complexity of element-wise multiplication operation can be considered as $O\left(C \cdot H \cdot W\right)$.

IV. Experiment

A. Experimental Dataset

The Cityscapes dataset, jointly provided by three German institutions including Daimler AG, comprises stereo vision data from over 50 cities, featuring a total of 5000 finely annotated and 20000 coarsely annotated images of urban street scenes. In the experiments, only the 5000 finely annotated images are utilized, with 2975 for training, 500 for validation, and 1525 for testing. Each image has a resolution of 1024 pixels × 2048 pixels and is densely annotated with 19 object categories. Since the labels for the test set in the Cityscapes dataset are not publicly available, this paper evaluates the models on the validation set.

The PASCAL VOC2012+SBD dataset is an extension of the PASCAL VOC2012 dataset, obtained by merging PASCAL VOC2012 with the SBD dataset. After removing duplicate images from both datasets, this augmented dataset comprises a total of 12031 annotated images covering 20 different object categories. The images are randomly split into training and validation sets at a 9:1 ratio, with 10827 images for training and 1204 images for validation. Similarly, the models are evaluated on the validation set.

TABLE I. Experimental Hardware and Software Environment

Project Detail	Specification
CPU	16 vCPU Intel(R) Xeon(R) Platinum 8350C CPU @ 2.60GHz
GPU	NVIDIA GeForce RTX 3090 GPU(24GB) × 1
RAM	42GB
CUDA	12.2
Python type	3.8.10
Operating system	Ubuntu 18.04.5 LTS
Development framework	Torch 2.2.1

B. Experimental Environment and Parameter Settings

All networks in the experiment were implemented based on the PyTorch framework, and the hardware and software environments used during the training process are detailed in Table I. The hyperparameter settings during the training process are shown in Table II. The experiments used stochastic gradient descent with momentum for gradient updates. The learning rate decay adopted a cosine annealing schedule, with the specific calculation process described in Eq. (8).

$$lr = min\_lr + 0.5 \cdot (initial\_lr - min\_lr) \cdot \left(1 + \cos\left(\pi \cdot \frac{iter}{total\_iter}\right)\right) \tag{8}$$

Where, $initial\_lr$ represents the initial learning rate; min_lr is the minimum learning rate, which is set to 1/100 of $initial\_lr$; lr represents the current learning rate; $iter$ is the current iteration number; and $total\_iter$ represents the total number of iterations.

For models trained on the Cityscapes dataset, a combination of Focal Loss and Dice Loss was adopted as the loss function. Focal Loss can effectively address the class imbalance issue present in the Cityscapes dataset, while Dice Loss enhances the accuracy of edge prediction. Combining these two losses optimizes the training effect of the model. On the PASCAL VOC2012+SBD dataset, where the class distribution is relatively balanced, a simple CE Loss (Cross-Entropy Loss) was chosen as the loss function.To optimize the training process, bilinear interpolation was used to uniformly resize the image resolution in the Cityscapes dataset to 512 pixels × 1024 pixels, while the image resolution in the PASCAL VOC2012+SBD dataset was adjusted to 512 pixels × 512 pixels. For training all models, a transfer learning strategy was employed, where pre-trained weights on the ImageNet dataset were loaded into the backbone network, and a frozen training phase of 50 epochs was first conducted. For models trained on the PASCAL VOC2012+SBD dataset, the number of training epochs was set to 200, while for models trained on the Cityscapes dataset, the number of training epochs was set to 300.

TABLE II. HYPERPARAMETER SETTINGS

Hyperparameter	PASCAL VOC2012+SBD	Cityscapes
input_pixel	512×512	512×1024
optimizer	SGD	SGD
initial_lr	0.004	0.004
lr_decay_strategy	cos	cos
momentum	0.9	0.9
freeze_iters	50	50
unfreeze_iters	250	150
total_iters	300	200
loss_function	CE Loss	Focal Loss + Dice Loss

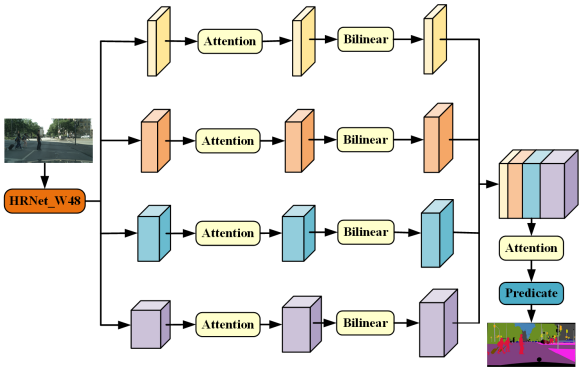

Fig. 5. Schematic diagram of attention mechanism embedding positions.

C. Experimental Results and Analysis

1) Attention selection experiment: To investigate the impact of different attention mechanisms on the segmentation performance of the model, a series of attention mechanisms were sequentially embedded into the "Attention" positions in Fig. 5 for training. These mechanisms include Strip Pooling (SP) [26], CBAM, AAM, SE, and ECA. Analyzing the experimental data in Fig. 6, it can be seen that when the test image resolution is 512 pixels×1024 pixels, the combination of the HRNet network with ECA achieves the best segmentation accuracy, with a 1.12% improvement in mean Intersection over Union (mIoU) compared to the original HRNet model. When the test image resolution is 1024 pixels×2048 pixels, the combination of HRNet with the AAM attention mechanism performs particularly well, with a 2.34% increase in mIoU compared to the original model.

ECA, AAM and CBAM, all have a positive impact on improving the segmentation accuracy of HRNet. Preliminary analysis indicates that when processing lower-resolution test images, channel attention mechanisms can significantly enhance the segmentation accuracy of the model. However, as the resolution of the test images increases, spatial attention mechanisms gradually demonstrate their unique advantage in improving model accuracy. The reason may be that when processing high-resolution images, channel attention mechanisms require deeper spatial information compression, leading to significant loss of spatial pixel information in the feature maps. Due to this information loss, channel attention weights cannot effectively measure the importance of information in the feature maps, resulting in suboptimal weighting effects on the feature maps. However, such issues do not exist when using spatial attention mechanisms.

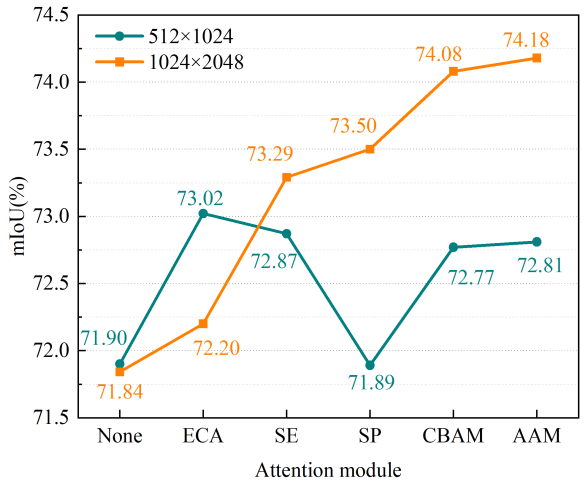

Fig. 6. Comparison of mIoU using HRNet with different attention mechanisms.

TABLE III. ABLATION EXPERIMENT RESULTS ON CITYSCAPES DATASET

BFEM	FUSM	CCSF	ECA	AAM	mIoU *(%)	mIoU ^(%)	Params /M	FLOPs */G	FLOPs ^/T
					72.34	71.79	**65.860**	**376.184**	**1.505**
✓					72.46	72.67	84.342	402.918	1.612
			✓		73.28	73.35	**65.860**	376.232	**1.505**
				✓	73.11	74.62	66.128	376.291	**1.505**
	✓				73.25	72.23	66.646	384.996	1.540
	✓	✓			73.47	73.29	67.166	419.158	1.677
	✓	✓	✓		73.70	74.37	67.166	419.205	1.677
✓	✓	✓	✓		**73.75**	73.93	85.647	445.939	1.784
	✓			✓	73.37	**74.84**	66.914	385.102	1.540
	✓	✓		✓	**73.75**	74.49	67.434	419.265	1.677

[1] "*" indicates that the image resolution used during testing is 512 pixels×1024 pixels, "^" indicates that the image resolution used during testing is 1024 pixels×2048 pixels.
[2] FLOPs stands for Floating-Point Operations, referring to the total number of basic arithmetic operations (addition, subtraction, multiplication, division, etc.) required by the model during execution, which directly reflects the computational complexity of the model.
[3] G represents billion (10^9) level of operations, and T represents trillion (10^{12}) level of operations.

2) Ablation study: To validate the effectiveness of the added modules, this section sequentially adds different (combinations of) modules to the HRNet and conducts experimental verifications. The experimental results are detailed in Table III. To more intuitively demonstrate the improvement effects of each module (and their combinations) on segmentation accuracy (mIoU), the incremental data is plotted into a bar chart as shown in Fig. 7.

When the resolution of the test images is 512 pixels×1024 pixels, the impact of each module on model performance:

FUSM: After introducing this module, the model's mIoU metric improves by 0.91%, while only introducing a small increase in parameters (0.786M) and computations (8.812G). This indicates that FUSM enhances segmentation accuracy while maintaining model efficiency, particularly excelling in detail recovery.

CCSF: Adding this module on top of FUSM further boosts

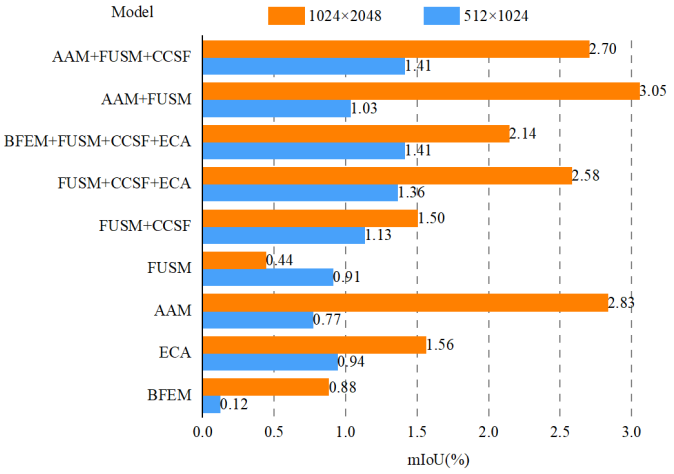

Fig. 7. The improvement in mIoU after integrating various (combinations of) modules into HRNet.

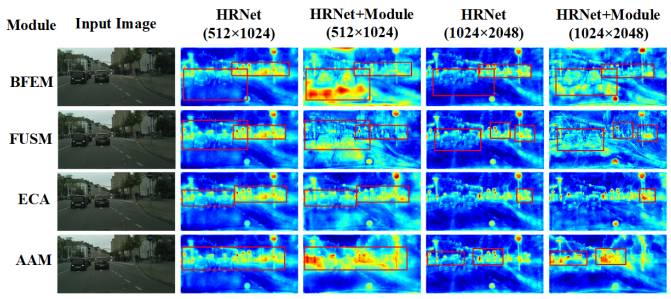

Fig. 8. Heatmap outputs before the fully connected classification layer when combining HRNet with various modules.

the mIoU metric by 0.22%. However, it significantly increases the computational cost by 34.162G. This suggests that while CCSF can improve accuracy, it comes with a relatively high computational overhead.

ECA: Incorporating this module enhances the mIoU metric by 0.27% with virtually no additional parameters or computations, demonstrating ECA's efficiency in capturing crucial information.

BFEM: Upon introduction, the mIoU metric surges to a peak of 73.75%, showcasing the module's remarkable ability to enrich multi-scale contextual information and deepen semantic understanding. However, this substantial improvement is accompanied by a notable increase in both spatial and temporal complexity of the model (an increase of 18.481M parameters and 26.734G total floating-point operations), indicating that higher computational costs are necessary to achieve this level of performance enhancement.

When the resolution of the test images is 1024 pixels$\times$2048 pixels, the impact of each module on model performance:

AAM: After introducing AAM, the mIoU significantly improves by 2.83% with only a small increase in parameters (0.268M) and almost no increase in computational complexity, demonstrating its efficient capability in processing spatial information.

FUSM: Further incorporating FUSM on top of AAM elevates the mIoU to 78.48%, achieving the optimal value. The increases in parameters and computations are relatively small, at 0.786M and 0.035T respectively, indicating that FUSM can still effectively enhance accuracy while maintaining its lightweight advantage at high resolutions.

CCSF: Adding the CCSF module to the AAM+FUSM combination unexpectedly decreases the mIoU to 74.49%, underperforming expectations, especially in high-resolution scenarios.Possible reasons for this phenomenon include: increased information redundancy or conflicts between feature maps, a high computational burden due to elevated model complexity, and complex interactions between modules that affect the model optimization process.

Fig. 8 shows heatmaps output by HRNet and its combinations with different modules, revealing the impact of each module on changes in target attention. The original HRNet heatmap primarily focuses on small targets such as bicycles and pedestrians, indicating its advantage in recognizing small-scale objects. After introducing the BFEM module, the heatmap not only continues to pay attention to small-scale targets but also starts to focus on large targets like cars and roads, suggesting that BFEM enhances the model's recognition capabilities for multi-scale targets. The addition of the FUSM module makes the temperature distribution of the heatmap more uniform and emphasizes object edges more strongly, indicating that FUSM improves the model's scene and contour perception abilities. The ECA module enables the model to more precisely focus on critical targets, contributing to enhanced feature extraction capabilities. Lastly, the AAM module clarifies the location information of objects of interest by enhancing the contrast between hot and cold regions, thereby improving the model's segmentation accuracy.

In summary, each module exhibits unique characteristics in image segmentation tasks, and their effects are significantly influenced by resolution and application scenarios. FUSM and ECA demonstrate a good balance between precision improvement and complexity control across different resolutions. While CCSF and BFEM can enhance performance, the associated increase in computational complexity cannot be overlooked. AAM is particularly efficient at high resolutions, highlighting its advantage in spatial information processing. In practical applications, the optimal combination of modules should be flexibly selected and configured based on specific requirements and environmental conditions to achieve the best balance between performance and computational efficiency.

3) Comparison with State-of-the-Art methods: To validate the performance of the proposed method in semantic segmentation, this section compares it with various state-of-the-art approaches on two standard datasets: Cityscapes and PASCAL VOC2012+SBD. The compared methods include DeepLabv3_R (based on ResNet50), DeepLabv3_X (based on Xception), PSPNet, SegFormer, SeaFormer-S, SeaFormer-B, SCTNet-S, and HRNet. The segmentation results on Cityscapes and PASCAL VOC2012+SBD datasets are presented in Tables IV and V, respectively.

Analyzing the data in Table IV, it can be observed that at a resolution of 512 pixels$\times$1024 pixels, both HRNet_S and HRNet_M achieve an optimal segmentation accuracy (mIoU)

TABLE IV. COMPARISON OF THE IMPROVED MODEL WITH STATE-OF-THE-ART METHODS ON THE CITYSCAPES DATASET

Model	mIoU *(%)	mIoU ˆ(%)	Params /M	FLOPs */G	FLOPs ˆ/T
DeepLabv3+_R[17]	70.36	71.33	**40.354**	732.532	2.930
PSPNet[4]	68.57	69.22	48.957	749.149	2.996
SegFormer[15]	72.05	72.32	47.238	**286.349**	1.145
SeaFormer-S[27]	-	70.70	-	-	**0.02**
SeaFormer-B[27]	-	72.70	-	-	0.03
SCTNet-S[7]	72.80	-	4.6	-	-
HRNet[1]	72.34	71.79	65.860	376.184	1.505
HRNet_S(ours)	**73.75**	73.93	85.647	445.939	1.784
HRNet_M(ours)	**73.75**	74.49	67.434	419.265	1.677
HRNet_L(ours)	73.37	**74.84**	66.914	385.102	1.540

TABLE V. COMPARISON OF THE IMPROVED MODEL WITH STATE-OF-THE-ART METHODS ON THE PASCAL VOC2012+SBD DATASET

Model	mIoU(%)	Params/M	FLOPs/G
DeepLabv3+_R[17]	75.68	**40.354**	366.275
DeepLabv3+_X[17]	77.82	54.714	486.773
PSPNet[4]	75.45	48.958	374.634
SegFormer[15]	**80.30**	47.239	**143.199**
HRNet[1]	77.42	65.861	188.116
HRNet_S(ours)	78.93	85.648	222.993
HRNet_M(ours)	77.62	67.435	209.674
HRNet_L(ours)	76.49	66.915	192.529

of 73.75%, surpassing the novel SegFormer model by 1.70%. Even when compared with the recent SCTNet-S model, it remains 0.95% higher. When processing images with a higher resolution of 1024 pixels×2048 pixels, the mIoU of the HRNet_L model climbs to 74.84%, outperforming SegFormer by 2.52%. Even when compared with the recent SeaFormer-B model, it remains 2.14% higher. However, in terms of parameter count and computational complexity, HRNet shows slight increases compared to SegFormer and does not have an advantage over SegFormer-B in terms of computational efficiency. Therefore, lightweight design of the model is a key focus for future research.

TABLE VI. WHEN THE TEST IMAGE RESOLUTION IS 1024 PIXELS×2048 PIXELS, THE PER-CATEGORY MIOU OF DIFFERENT MODELS ON THE CITYSCAPES DATASET (%)

Category	PSPNet	DeepLabv3+_R	SegFormer	HRNet	HRNet_L
road	91.31	92.74	92.89	93.02	**93.23**
sidewalk	72.19	71.46	73.25	72.76	**75.30**
building	86.71	87.80	87.88	88.80	**88.89**
fence	41.60	44.65	50.08	**54.12**	48.70
pole	51.69	52.68	51.79	**56.63**	56.26
traffic light	53.55	59.22	56.51	60.85	**62.84**
traffic sign	62.40	65.93	66.04	66.95	**68.78**
vegetation	66.43	70.09	70.39	71.63	**72.79**
terrain	89.55	90.01	90.38	90.52	**90.67**
sky	57.08	56.20	57.27	58.53	**59.42**
diningtable	89.00	90.17	90.32	**90.59**	90.41
person	75.28	77.64	77.53	79.15	**79.80**
rider	59.33	63.01	61.00	66.34	**68.13**
car	92.26	92.23	92.17	92.51	**93.58**
truck	55.60	60.73	70.14	63.52	**78.18**
bus	82.32	83.16	81.17	80.07	**85.34**
train	71.06	69.22	77.60	46.66	**81.89**
motorcycle	53.49	62.01	62.54	**64.28**	61.09
bicycle	71.90	73.47	72.00	73.39	**74.37**
background	61.71	64.06	65.27	65.53	**67.12**

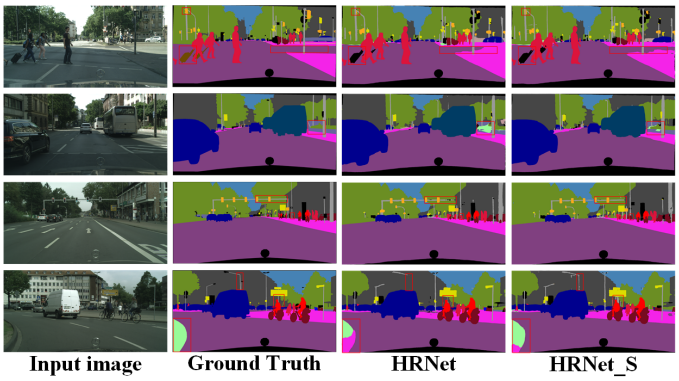

Fig. 9. The segmentation effect images of different models on the Cityscapes dataset when the test image resolution is 512 pixels×1024 pixels.

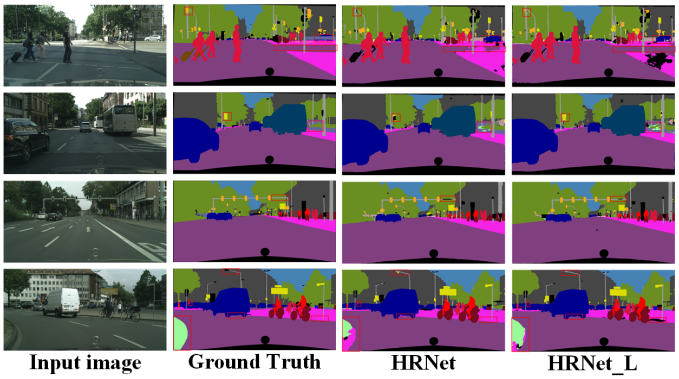

Fig. 10. The segmentation effect images of different models on the Cityscapes dataset when the test image resolution is 1024 pixels×2048 pixels.

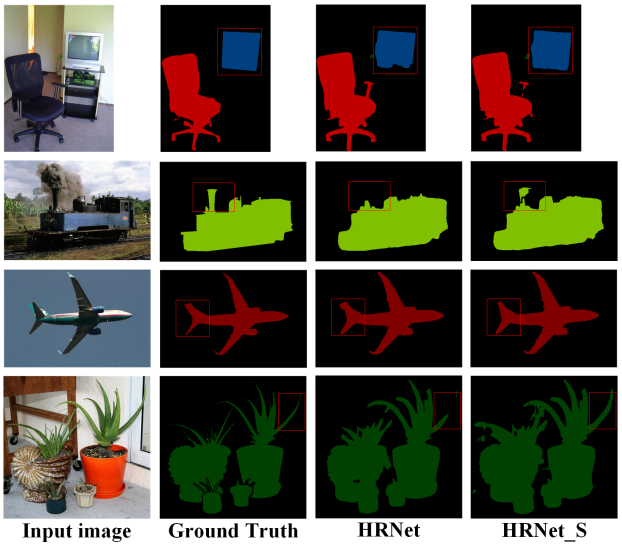

Fig. 11. Segmentation effect images of different models on the PASCAL VOC2012+SBD dataset.

As shown in Table V, on the PASCAL VOC2012+SBD dataset, for low-resolution images of 512 pixels×512 pixels, the mIoU of the HRNet_S model reaches 78.93%. Although

slightly lower in segmentation accuracy compared to the SegFormer model, HRNet_S has significantly narrowed the gap with SegFormer, demonstrating its competitiveness in low-resolution image segmentation tasks. However, from the perspective of algorithmic complexity, there is still room for further optimization of the HRNet_S model.

Combing the experimental data from Tables IV and V, from the perspectives of segmentation accuracy and model complexity, HRNet_M can be considered as a balanced choice between HRNet_S and HRNet_L. In terms of increasing computational complexity, the order is: HRNet_L, HRNet_M, HRNet_S. Additionally, the segmentation accuracy of HRNet_M also falls between that of HRNet_S and HRNet_L across different resolutions.

As shown in Table VI, which presents the category-wise mIoU data of various models on the Cityscapes dataset when tested on images with a resolution of 1024 pixels$\times$2048 pixels, it can be observed that HRNet_L significantly improves the segmentation accuracy for large-scale objects such as trucks, buses, and trains at this high resolution of 1024 pixels$\times$2048 pixels. Furthermore, HRNet_L achieves the highest accuracy in 15 out of the 19 different scale-varying categories.

To intuitively demonstrate the superiority of the proposed method in semantic segmentation tasks, this section comprehensively evaluates its segmentation effects through visual comparison experiments with the baseline method, HRNet. Fig. 9 and 10 showcase the segmentation effects of the models on Cityscapes dataset for test images with different resolutions. From these two figures, it can be observed that HRNet exhibits some shortcomings in image segmentation, particularly in road contours, traffic sign recognition, and segmentation accuracy of elongated objects such as streetlights. Additionally, it tends to confuse vegetation with terrain. However, these deficiencies are significantly improved in the proposed enhanced models, HRNet_S and HRNet_L. Fig. 11 further presents the segmentation effects of the models on the PASCAL VOC2012+SBD dataset. On this dataset, HRNet_S notably enhances the segmentation accuracy of TV and airplane wing boundaries compared to the original HRNet. Simultaneously, it exhibits more refined segmentation capabilities for complex structures such as ship funnels and vegetation leaves. This indicates that the proposed method comprehends contextual semantic information more comprehensively than the original model, validating its effectiveness in improving segmentation accuracy and detail preservation.

Based on the HRNet network, this paper introduces specific modules and successfully develops a model suitable for image segmentation across different resolutions. Specifically, the HRNet_S model demonstrates exceptional segmentation performance when dealing with lower-resolution images. However, due to its relatively high number of parameters and computational complexity, when handling higher-resolution image segmentation tasks, the significant increase in computational load becomes particularly evident, which may lead to inadequate model training and subsequently affect its final performance. Additionally, the Efficient Channel Attention Mechanism employed in HRNet_S may overly compress the valuable information in high-resolution images, adversely impacting the model's ultimate performance. In contrast, the HRNet_L model is more lightweight, thus having an advantage when dealing with higher-resolution image segmentation tasks. Its use of the AAM attention mechanism effectively aggregates valuable information in high-resolution images, which is a key factor contributing to its superior performance. However, compared to the HRNet_S model, the HRNet_L model lacks depth in semantic information, making its advantage less pronounced when handling lower-resolution image segmentation tasks.

Therefore, in practical production and life, we should select the appropriate variant of the HRNet model based on factors such as the resolution of the dataset and the application scenario, in order to achieve the best segmentation results and ensure that the model can perform optimally in image segmentation tasks across different resolutions.

V. Conclusion

Addressing the challenges faced by existing semantic segmentation networks in handling multi-scale objects, such as the tendency to lose small-scale objects, incomplete segmentation of large-scale objects, and low overall segmentation accuracy, this paper proposes an improved method based on the HRNet network. Firstly, a backbone feature enhancement module is introduced using deep stripe convolutions, which enhances the network's adaptability to multi-scale objects, overcomes the limitations of a single convolutional kernel in feature extraction, expands the model's perception range of contextual information, and enhances the network's ability to understand complex scenes. Secondly, an axial attention mechanism is employed to model the global dependency relationships within the feature maps output by the backbone network, enabling precise localization of regions of interest. Lastly, a flexible upsampling mechanism is adopted, leveraging the complementary fusion of semantic and detail information between adjacent feature maps, to effectively restore target detail information in the decoder network. Experimental results show that the proposed algorithm achieves the highest segmentation accuracy compared to other algorithms on the Cityscapes dataset. Similarly, the segmentation accuracy of the proposed algorithm on the PASCAL VOC 2012+SBD dataset is also outstanding, verifying the effectiveness of the proposed method. Further ablation studies also confirm the contributions of each improved component to enhancing the overall performance. The relevant implementation code for this paper has been uploaded to the GitHub platform (link: https://github.com/HanLeiFeng/HRNet_Series.git) for learning and exchange.

References

[1] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 5693–5703.

[2] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2017.

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

[4] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2881–2890.

[5] M. Yang, K. Yu, C. Zhang, Z. Li, and K. Yang, "Denseaspp for semantic segmentation in street scenes," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 3684–3692.

[6] J. He, Z. Deng, L. Zhou, Y. Wang, and Y. Qiao, "Adaptive pyramid context network for semantic segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 7519–7528.

[7] Z. Xu, D. Wu, C. Yu, X. Chu, N. Sang, and C. Gao, "Sctnet: Single-branch cnn with transformer semantic information for real-time segmentation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 6, 2024, pp. 6378–6386.

[8] C. Peng, M. Zhu, H. Ren, and M. Emam, "Small object detection method based on weighted feature fusion and csma attention module," *Electronics*, vol. 11, no. 16, p. 2546, 2022.

[9] M. Zhu, J. Wang, A. Wang, H. Ren, and M. Emam, "Multi-fusion approach for wood microscopic images identification based on deep transfer learning," *Applied Sciences*, vol. 11, no. 16, p. 7639, 2021.

[10] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, "Ccnet: Criss-cross attention for semantic segmentation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 603–612.

[11] Y. Yuan, L. Huang, J. Guo, C. Zhang, X. Chen, and J. Wang, "Ocnet: Object context for semantic segmentation," *International Journal of Computer Vision*, vol. 129, no. 8, pp. 2375–2398, 2021.

[12] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 3146–3154.

[13] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[14] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.

[15] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "Segformer: Simple and efficient design for semantic segmentation with transformers," *Advances in neural information processing systems*, vol. 34, pp. 12 077–12 090, 2021.

[16] J. Xu, Z. Xiong, and S. P. Bhattacharyya, "Pidnet: A real-time semantic segmentation network inspired by pid controllers," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 19 529–19 539.

[17] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 801–818.

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

[19] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, "Transunet: Transformers make strong encoders for medical image segmentation," *arXiv preprint arXiv:2102.04306*, 2021.

[20] C. Ziwen, K. Patnaik, S. Zhai, A. Wan, Z. Ren, A. Schwing, A. Colburn, and L. Fuxin, "Autofocusformer: Image segmentation off the grid," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 18 227–18 236.

[21] W. Wang, J. Dai, Z. Chen, Z. Huang, Z. Li, X. Zhu, X. Hu, T. Lu, L. Lu, H. Li *et al.*, "Internimage: Exploring large-scale vision foundation models with deformable convolutions," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 14 408–14 419.

[22] A. Hassani and H. Shi, "Dilated neighborhood attention transformer," *arXiv preprint arXiv:2209.15001*, 2022.

[23] M.-H. Guo, C.-Z. Lu, Q. Hou, Z. Liu, M.-M. Cheng, and S.-M. Hu, "Segnext: Rethinking convolutional attention design for semantic segmentation," *Advances in Neural Information Processing Systems*, vol. 35, pp. 1140–1156, 2022.

[24] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[25] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.

[26] Q. Hou, L. Zhang, M.-M. Cheng, and J. Feng, "Strip pooling: Rethinking spatial pooling for scene parsing," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 4003–4012.

[27] Q. Wan, Z. Huang, J. Lu, G. Yu, and L. Zhang, "Seaformer: Squeeze-enhanced axial transformer for mobile semantic segmentation," *arXiv preprint arXiv:2301.13156*, 2023.