Hindawi
International Journal of Intelligent Systems
Volume 2023, Article ID 6358162, 15 pages
https://doi.org/10.1155/2023/6358162

Research Article

Refined Division Features Based on Transformer for Semantic Image Segmentation

Tianping Li, Yanjun Wei, Meilin Liu, Xiaolong Yang, Zhenyi Zhang, and Jun Du

School of Physics and Electronics, Shandong Normal University, Jinan, Shandong, China

Correspondence should be addressed to Jun Du; dujun@sdnu.edu.cn

Received 27 January 2023; Revised 11 June 2023; Accepted 28 July 2023; Published 19 August 2023

Academic Editor: Mohammad R. Khosravi

Transformer can build global relationships between pixels and enhance pixel representation. The existing methods only establish the context relationship from the whole image but will reduce the representation between the category areas. In addition, the existing methods based on the transformer self-attention do not combine the advantages of convolution and transformer, resulting in more calculation parameters of the model. In order to solve these two problems, this paper proposes to enhance the segmentation accuracy and performance by enhancing the relationship between image-level regions and the relationship between semantic level pixels. First, we design a refined division feature (RDF) module to enhance the channel representation and thus the same locale representation. Second, we design a transformer based on convolution (CTrans), which first computes the relationship between similar pixels and enhances the pixel representation. Then, the feature map is compressed and enriched to reduce the computational load of CTrans, and finally the relationship between pixels is established from a global perspective. We design a refined division feature module based on transformer for semantic image segmentation (RFT) model combining RDF and CTrans module. The experimental results show that the mIoU result of our method in Cityscapes test data set is 81.9%, and the model parameter is 64.6M, which is superior to other methods in terms of data. In addition, we conducted visualization experiments with Cityscapes and Pascal voc 2012 datasets with other methods, and the results showed that our method was superior to other methods.

1. Introduction

In recent years, deep learning computer vision has been rapidly developed in the fields of semantic analysis [1], image repair [2, 3], object tracking [4], super resolution reconstruction [5], and object detection [6] and has been favored by many researchers. Semantic segmentation is a fundamental technique in computer vision. The role of semantic segmentation is to predict each category for each pixel and learn the semantic and spatial information of each class, for example, the locations and classes of objects. Semantic segmentation generally deals with the issue of classifying pixels at the pixel scale, and it typically needs rich contextual semantic information. With the advancement of convolutional neural networks (CNNs), particularly a fully convolutional network (FCN) [1], many researchers have started to focus on the study of multiscale contextual features for semantic segmentation, for example, SegNet [7], DeconvNet [8], and the DeepLab series of articles [9–11]. DeepLabv3 [9] and DeepLabv3+ [10] use atrous convolution with different hole rates to extract multiscale contextual semantic information, which is not beneficial to dense segmentation. DeepLabv3 excessively uses atrous convolution, which has generated a grid effect. Similarly, PSPNet [11] uses a pyramid pooling block consisting of adaptive average pooling at different scales to extract multiscale contextual semantic information, which has the drawback of not considering the relationship between pixel dots and the neighboring pixel set. Since PSPNet pixels do not have rich upper and lower information, SA-FFNet [12] proposes VH-CAM and UC-PPM to improve the upper and lower information of pixels. However, SA-FFNet uses pyramid pooling to enter pyramid pooling of advanced feature maps, which will inevitably cause some loss of effective

information. In addition, multiple UC-PPM models will be used to compound model structures. RELAXNet [13] proposes that AGF module uses maximum pooling and average pooling, and the resulting spatial weights are not distinguishable. After being multiplied with feature maps, feature maps cannot be effectively distinguished.

EncNet [14] uses a nonadaptive method to assemble contextual semantic information and uses a homogeneous context extraction process for all pixels, which does not meet the need for different context dependencies for each pixel.

Nowadays, convolutional neural network (CNN) semantic segmentation has been widely used in various domains, such as MultiNet for autonomous vehicle driving [15], UNet3+ for medical image parsing [16], DSFPNet for object detection [17], and image classification [18].

An attention mechanism has recently been integrated into semantic segmentation networks, which counts the similarity of adjacent features to obtain contextual semantic information about pixels. For example, PSANet [19] aggregates the contextual semantic information at each position by predicting the attentional feature map. ANNet [20] improves semantic segmentation precision using long-distance asymmetric dependencies between pixels and neighboring pixels. A pixel-region relationship is calculated using OCRNet [21] to improve pixel-region representation. DANet [22] calculates pixel-to-pixel distances on feature map channels and spatial locations to improve pixels' presentation. DADCNet [23] uses the SE structure to learn the relevant information of the feature maps between the channels, so that the network attention is focused on the useful feature maps.

Based on SA-FFNet, we used group convolution, adaptive maximum pooling, and adaptive average pooling to integrate more image and texture information into the channel attention map and used the channel attention map to update each adjacent channel's feature map. In addition, we used two fully connected layers for the feature maps. The two fully connected layers not only effectively combine the linear information between the channel feature maps but also can establish the information interaction between the channel feature maps.

With the advancement of computer technology and CNNs, automatic end-to-end segmentation is now possible. Small objects, such as pedestrians, traffic signals, and traffic signs, can be segmented more efficiently through the accurate segmentation of object regions from images. A network has finer spatial information at the lower levels, but its semantic coherence is poor. Feature maps at the high-level of the network provide consistent semantic information, but their spatial information is coarse. To address this problem, literature [24] took advantage of color features and edge features to improve the face tracking reliability. DFNet [25] adopts a V-shaped structure instead of a U-shaped structure to capture multiscale contextual semantic information. An SFS and FFF module is proposed by FSN [26] to extract important features and merge them adaptively.

In a refined division feature (RDF) block, we generated a channel attention matrix to distinguish the importance of feature channels inspired by DFNet and FSN [26]. Feature maps extracted using conventional CNNs gradually decrease in resolution, and the perceptual field is limited to global information and long-range pixel dependencies.

Since the introduction of the transformer [27] in computer vision, semantic segmentation has improved dramatically. Transformer is capable of capturing global information; it can compute dynamic weights between global pixels, and it can adapt dynamically to different input images. These properties are very useful for obtaining high-level semantic information, but they are only helpful if there are sufficient data to support the transformer.

Furthermore, the transformer cannot deal with fine details in the images. This is particularly true for small targets at long distances in Cityscapes that contain multiscale targets. CMT [28] solved this problem by combining deep convolution with the transformer, where deep convolution extracted local features to compress channels and the transformer established global interdependencies between the patches. Convolution and transformer were used by UniForme [29] to extract global and local information, which effectively addresses the problems of redundancy and dependency in the learning process of the networks.

Our RDF module not only distinguishes feature map information from feature maps adaptively and removes redundant information but also establishes long-distance links to each channel's feature maps. Furthermore, we proposed a module called the CTrans block, which includes depth-separable convolution, cross-attention, and self-attention. First, RDF module processes input feature maps in parallel to establish cross-channel interactions to produce feature maps with different resolutions and depths; then, instead of using the multilayer perceptron (MLP) in the transformer, depth-separable convolution is used to extract multiscale spatial information. Finally, global feature information is captured using the transformer attention. Based on the number of parameters and accuracy of RFTNet, we compared our network with some classical networks. As shown in Figure 1, our network outperforms other classical networks not only in terms of the number of parameters but also in terms of segmentation accuracy.

For this paper, the main contributions are as follows.

(1) Because existing attention models do not consider the dependencies between feature graph channels from a macroscopic perspective, this paper proposes refined division feature (RDF) module, which can extract multiscale spatial information and establish long-distance channel dependencies. RDF is very flexible and scalable, and it can be applied to many computer vision network architectures.

(2) For existing models, transformer self-attention is used to establish long-distance dependence between pixels to improve the accuracy of semantic segmentation, which does not consider the advantages of combining convolution and transformer. From the micropoint of view, transformer based on convolution (CTrans) is proposed in this paper, which can not only enhance pixel representation and enrich

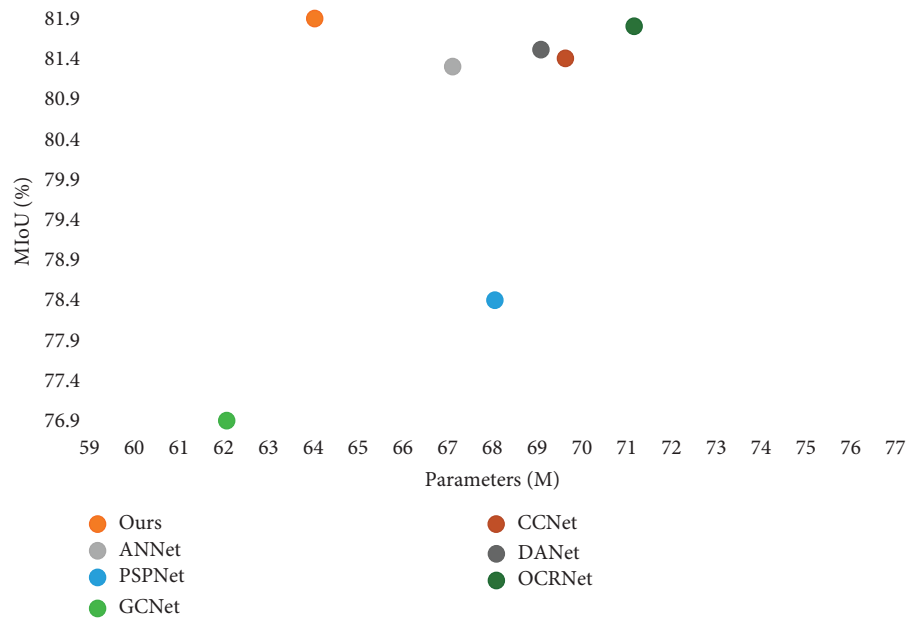


Figure 1: Comparison of the parameters and mIoU of our network and other classical networks on the dataset of Cityscapes.

and reduce model computation but also establish global relations between pixels.

(3) Combining RDF and CTrans, we proposed a semantic segmentation model of RFT to improve the accuracy and performance of semantic segmentation. The proposed framework allows this paper to achieve leading performance on segmentation-based benchmarks, including Cityscapes and PASCAL VOC 2012.

(4) For both PASCAL VOC 2012 and Cityscapes data, we obtained outstanding results. We used a single Tesla GPU V100 to develop our models. For PASCAL VOC 2012, the input image resolution was 512×512, and for Cityscapes, it was 512×1024. There were eight input batches in PASCAL VOC 2012 and only four in Cityscapes.

2. Related Work

In this section, we briefly review the related work, such as multiscale feature contextual information and convolution combined with the transformer.

2.1. Multiscale Feature Contextual Information. To combine semantic and representational information, an FCN adapts the classification network into a fully CNN, but this ignores high-resolution feature maps, which degrades the edge information.

To maximize both representation and semantic information, some subsequent studies have been conducted, which have shown improved feature information, which can broadly be divided into two categories: the first is the "Encoder–Decoder style." For example, SegNet uses an encoder–decoder structure and maximizes pooling preservation. This approach not only saves memory but also increases segmentation speed and accuracy. MCRNet [30] aims to use deep context to guide multistage fusion. Its disadvantage is that in the case of deep convolutional neural network ResNet, fusing feature maps of different stages will cause model redundancy and slow speed. When the perceptual field is too small and the object is too large, missegmentation can occur because the network ultimately does not see the object. For example, CENet [31] proposes a contextual integration network to achieve semantic segmentation. Its disadvantage is that it requires high hardware and software, and it is easy to cause overfitting.

Furthermore, suppose that the object is too small and the perceptual field is too large. In this case, the network will see additional backgrounds and redundant information, resulting in misclassification, because the network will have difficulty judging the tiny object. For example, Zhang et al. [32] proposed a method based on pyramidal consistency learning to improve the accuracy of segmentation. Its disadvantage is that it requires large computing resources, and the feature processing is not sufficient, which may lead to the loss of some detailed information. CCTseg [33] uses the prediction results obtained by DeepLabv3+. However, DeepLabv3+ uses a feature map obtained by roughly 4 times upsampling and fuses it with the feature map of the encoder, which is easy to cause resolution loss and is not conducive to semantic segmentation. To overcome this problem, DeconvNet proposes a deep convolutional network based on SegNet, where the encoders use VGG-16 convolutional layers to learn and the decoders use deconvolution with inverse pooling to upsample. RefineNet [34] and GCNet [35] combine feature maps inherent in different stages of multiscale contexts, but they lack a consistent global context.

The second style is called the "Backbone style," where DeepLabv2 uses atrous convolutions, using different sampling scales and input features, to capture target and contextual semantic information at multiple levels. To solve the

multiscale segmentation problem, DeepLabv3 developed cascaded or parallel atrous convolutions and expanded ASPP, achieving good results without requiring dense CRF postprocessing. To extract multiscale contextual semantic information, PSPNet incorporates information at different scales into the PSP module. A self-attentive method is used by OCNet [36] to learn pixel-to-pixel similarities; then, all features are aggregated using a similarity attention graph to approximate an object's context. OCRNet enriches pixel-region representations by computing pixel relationships with the regions. The spatial attention module of DANet and the channel attention module are used to capture contextual information to improve pixel representations, which improves segmentation performance.

The two methods mentioned above have two disadvantages. First, the affinity matrix is calculated by comparing pixels with other pixels, but the contextual information about a single pixel is minimal. Thus, the affinity matrix obtained is unsatisfactory. In addition, these methods focus on developing complex attention modules, which inevitably involve more computations and cannot effectively establish long-distance dependencies. To reduce the matrix computation complexity and effectively fuse global and local information, we consider improving the correlation between affinity matrices and long-distance pixel dependencies. In this study, we proposed two modules: RDF for obtaining discriminative feature maps and CTrans for integrating global and local information and establishing interpixel relationships over long distances.

2.2. Combination of Convolution and Transformer. As CNNs share weights and local fields of perception, they can effectively reduce the number of computational parameters while extracting spatial details. The translation invariance of a CNN also enhances its generalization ability. Nonetheless, the perceptual field of CNNs is limited, so they cannot capture global information and interdependencies between pixels at a distance.

On the other hand, the transformer has strong abilities to extract global information and expand the perceptual field, but it has two drawbacks: first, it is challenging to train, and second, it is not sensitive to fine details. To solve these problems, subsequent researchers combined the advantages of CNNs with those of the transformer. UniFormer uses both CNNs and transformer to effectively solve the redundancy of network learning as well as the long-distance interdependency between the pixels. AMACF [37] combines self-attention mechanism and convolutional network to extract global and local information of feature maps, respectively. TransUNet [38] and TransBTS [39] combine transformer and UNet [40] and apply them to the field of medical image segmentation and achieve very satisfactory segmentation results, but their disadvantage is that they require a large amount of training data and computing resources to train and optimize models, so it is impossible not to be used in resource-constrained and limited-number environments. TranSiam [41] proposes a method of combining depthwise separable convolution and transformer, which combines the advantages of both and reduces the amount of calculation. Its shortcoming is that it uses multihead attention, which makes the model more complex and requires more computing resources and time. Coatnet [42] proposes a combination of convolution and attention, which can adapt to processing images of various scales and improve the accuracy of segmentation. Its disadvantage is that multiple convolution kernels of different sizes are used, resulting in larger model parameters, which easily lead to model overfitting.

In addition, AMACF can adaptively distinguish the importance of feature maps according to the weights generated by the self-attention mechanism and the convolutional network, which makes the matrix operation simple.

The conformer [43] uses a CNN to extract local features and transformer to establish global relationships. Mobile-Former [44] is used to extract local features at the pixel level with efficient depthwise and pointwise convolution; by combining convolution with the transformer, global interaction is improved, and the number of randomly generated tokens is reduced. Several convolution kernels frequently lead to very computationally intensive and random image splits, so encoding patches become unstable when using several convolution kernels. Echt [45] suggested using a few convolutional kernels and transformers to encode patches to solve the problem of unstable network training and improve the segmentation effects.

The CTrans module in this study is based on the above methods, and it comprises multilayer convolution, cross-attention, and self-attention. As components in the CTrans module, convolution is used to extract local features in the early stages, cross-attention is used to improve the pixel representation in the middle stages, and self-attention is used to construct global contextual semantic information about pixels in the later stages; finally, local features are combined with global features.

Through convolution, local features are extracted further, and the number of channels is reduced, resulting in dense segmentation with rich contextual semantic information for each pixel. CTrans combines the advantages of convolution and transformer to capture global information and establish long-distance interdependencies among the pixels.

3. Method

This study presents a segmentation model based on RFTNet that combines the RDF and CTrans modules. First, we describe RFTNet in Section 3.1; then, we introduce the RDF and CTrans modules in Sections 3.2 and 3.3, respectively.

3.1. RFTNet. As shown in Figure 2, we briefly discussed refined division features based on transformer for semantic image segmentation (RFT).

First, we use ResNet101 to generate feature maps of the fifth stage. Next, we use the RDF module to identify the importance of the feature maps, and we use the RDF module to integrate the spatial information of feature maps and the channel information of feature maps into group feature maps to obtain better information interaction in global and local channel attention, which adaptively distinguishes channels according to

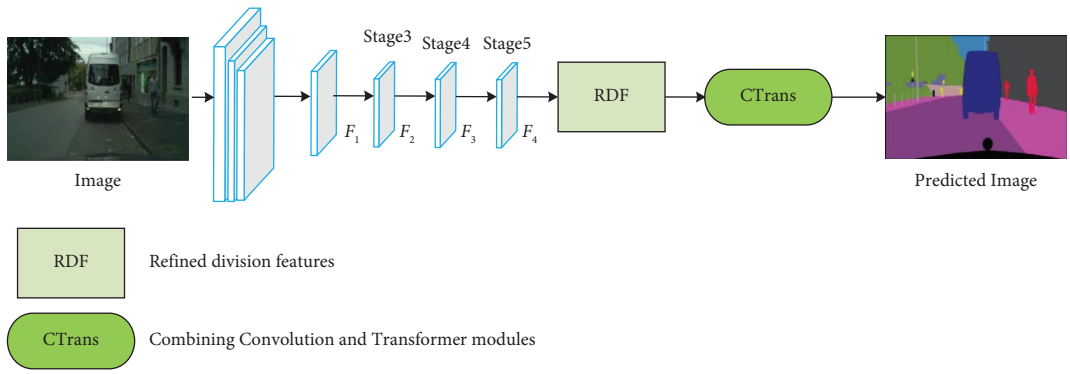


Figure 2: Overview of RFT's general frame diagram.

their importance. We use the CTrans module to extract the spatial information about the feature maps, and we use the similarity between pixel points to improve pixel representations. Finally, we use the CTrans module to extract the spatial details of feature maps, improve the pixel representation, and establish the global information relationships, and we use the FCN Head [1] upsampling method to obtain the same resolution feature maps for the multiscale feature maps.

3.2. Refined Division Features Module. According to previous research [1–14, 46, 47], fusing multiscale features can improve semantic segmentation in images with objects of different sizes. ResNet101 [48] deals with multiscale features of varying resolutions in each stage of a feature map. Low-resolution feature maps contain more semantic information than high-resolution feature maps, but high-resolution feature maps contain more detailed spatial information. In addition, large-scale objects contain weak semantic information after multiple downsampling because they have a limited perceptual field, whereas small objects contain clearer location information.

As shown in Figure 3, without increasing the computational cost, group convolution with different kernel sizes extracts multiscale feature map information with multiple branches. Therefore, feature maps of different resolutions and depths can be obtained. For each branch, group convolution can learn independent multiscale spatial information, and the RDF approach involves incorporating adaptive global average pooling (GAP) and global max pooling (GMP) modules within the framework of the RDF (residual dense feature) module. The primary objective is to effectively encode the pertinent information from the feature map into the channel attention map. This enables the RDF module to aptly discriminate and differentiate features across varying scales or dimensions, thus enhancing its ability to accommodate multiscale characteristics present within the data. A channel attention mechanism can assign different weights to each feature in a feature map, thereby generating more information. According to SAFFN, each pixel point in a feature map lacks sufficient contextual semantic information; thus, SAFFN uses irregular convolutions for channel compression, obtaining pixel points with robust contextual semantics. RELAXNet suggests that GMP can extract salient features from the feature maps.

As shown in Figure 3, in our approach, we extracted the spatial information from the input feature map using the multigroup convolution.

A feature map of RDF is given as follows:

$$f \in R^{C\times H\times W}, \tag{1}$$

where f represents the feature map output by the backbone network and C, H, and W represent the number of channels, height, and width of the feature map, respectively.

After the group convolution operation, the channel dimension of C is divided into $C/4$ in the RDF module. Through the group convolution method, input feature maps are processed simultaneously, and the resolutions of different depths are compressed, resulting in richer feature map information and effective extraction of spatial information from each channel of the feature maps. The spatial information about these channels can also be linked via group convolution for information interaction across channels.

We combine adaptive maximum pooling and adaptive global average pooling to extract feature map information. Adaptive global average pooling encodes the spatial information of a feature map into a channel attention map, whereas maximum pooling removes redundant data to reduce computation costs and alleviate the overfitting of the network. We added two fully connected layers, a nonlinear activation function (ReLU) and a softmax function, to the pooled feature map, following EPSANet [49].

Through these two fully connected layers, we can better combine the linear information between channels and improve the interaction between the channels. Next, the softmax function converts fractional values in each channel into probability values and multiplies those probability values by the feature maps in each channel so that information can be extracted efficiently. The attention weights of the RDF module channels are calculated as follows:

$$\begin{aligned} \varphi &= \mathrm{Conv}_i(g_u, k_v), \\ G_i &= ((\eta(\varphi(B_i))) + (\mu(\varphi(B_i)))), \\ \mathscr{B}_i &= \sigma(F_2\delta(F_1(G_i))), \end{aligned} \tag{2}$$

where $G_i \in R^{C/4\times H\times W}$, $\mathscr{B}_i \in R^{D_i\times 1\times 1}$, GAP stands for the adaptive global average pooling, GMP stands for the adaptive global max pooling, G_i represents the feature maps,

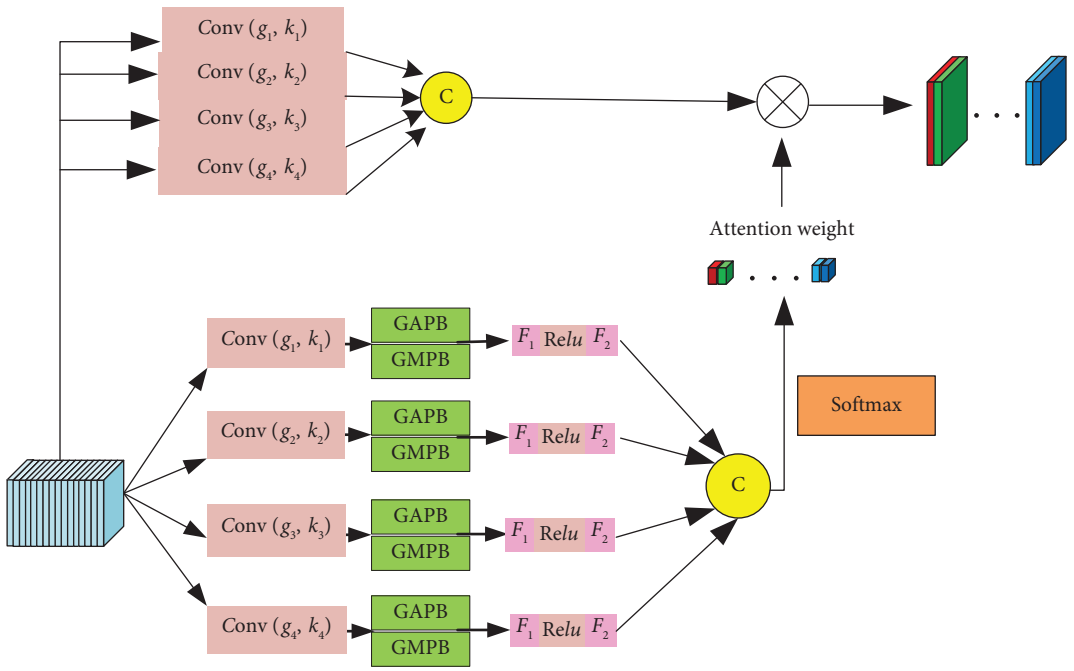


Figure 3: Detailed structure of the RDF block.

and φ stands for the phase i group convolution, where the group size is g_u and the convolution kernel size is k_v. In this block, F_1 represents the first fully connected layer, F_2 represents the second fully connected layer, δ represents the ReLU activation function, σ represents the softmax function, and $\mathscr{B}_i$ represents the channel attention map. The sigmoid and softmax activation functions are as follows:

$$\begin{aligned} \text{Sigmoid}(x_i) &= \frac{1}{1+e^{-D_i}}, \\ \text{Softmax}(x_i) &= \frac{e^{D_i}}{\sum_{i=1}^{n} e^{D_i,}} \end{aligned} \tag{3}$$

where D_i denotes the i-th channel weight value. Finally, the output feature map is computed as follows:

$$E_i = \mathscr{B}_i \varphi(B_i), \tag{4}$$

where $E_i \in R^{C/4\times H\times W}$ denotes the feature map output of the RDF module at i-th branch (i = 1, 2, 3 and 4).

3.3. Combining Convolution and Transformer Modules. This section describes the combination of convolution and transformer (CTrans) modules in detail. Figure 4 illustrates the structure of the CTrans block. It comprises depthwise convolution, cross-attention [50], and self-attention. The depth-separable convolution is used to extract spatial information from a feature map and to improve the interactions between the feature map information on each channel. Through the depth-separable convolution, the convolutional properties of the transformer are improved while reducing the computational cost. The CTrans module captures both the global and spatial details of a feature map. By replacing the transform in the input of the transformer with depth-separable convolution according to CMT, we can reduce the computational cost of encoding features while maintaining high accuracy. Therefore, the CTrans module can extract multiscale features and reduce the independence of pixels on different channels by associating pixels with each other. CTrans divides the feature extraction process into two phases: local feature extraction and long-range interdependency creation, i.e., a phase for establishing global information relationships. Through multilayer convolution, the local feature extraction phase not only extracts more pixel space information but also reduces information loss. As shown in Figure 4, the input feature map for the CTrans module is $F^* \epsilon R^{C^*\times H^*\times W^*}$. In the CTrans module, considering that the feature maps for Q_1 and K_1 branches are highly correlated, we can improve the similarities of V_1 by establishing relationships between them. We use cross-attention to calculate the similarity between Q_1 and K_1 to produce the spatial attention map; then, we apply the spatial attention map to weight V_1.

The calculation of cross-attention is as follows:

$$\text{CAttn} = \text{Softmax}\left(\frac{Q_1 {K_1}^T}{\sqrt{d_{k_1}}}\right) V_1, \tag{5}$$

where $\text{CAttn} \in R^{C^*\times H^*\times W^*}$ denotes the cross-attentive output feature map, $K_2, Q_2 \in R^{N\times d_{k_2}}$, and $N = H^* \times W^*$ denotes the number of pixels in the feature map.

In the second step, we channel-compressed the feature maps output by depth-separable convolution and extracted the spatial details. Depth-separable convolution is calculated as follows:

$$I = \text{Depth}_{\text{Conv}}(\text{CAttn}), \tag{6}$$

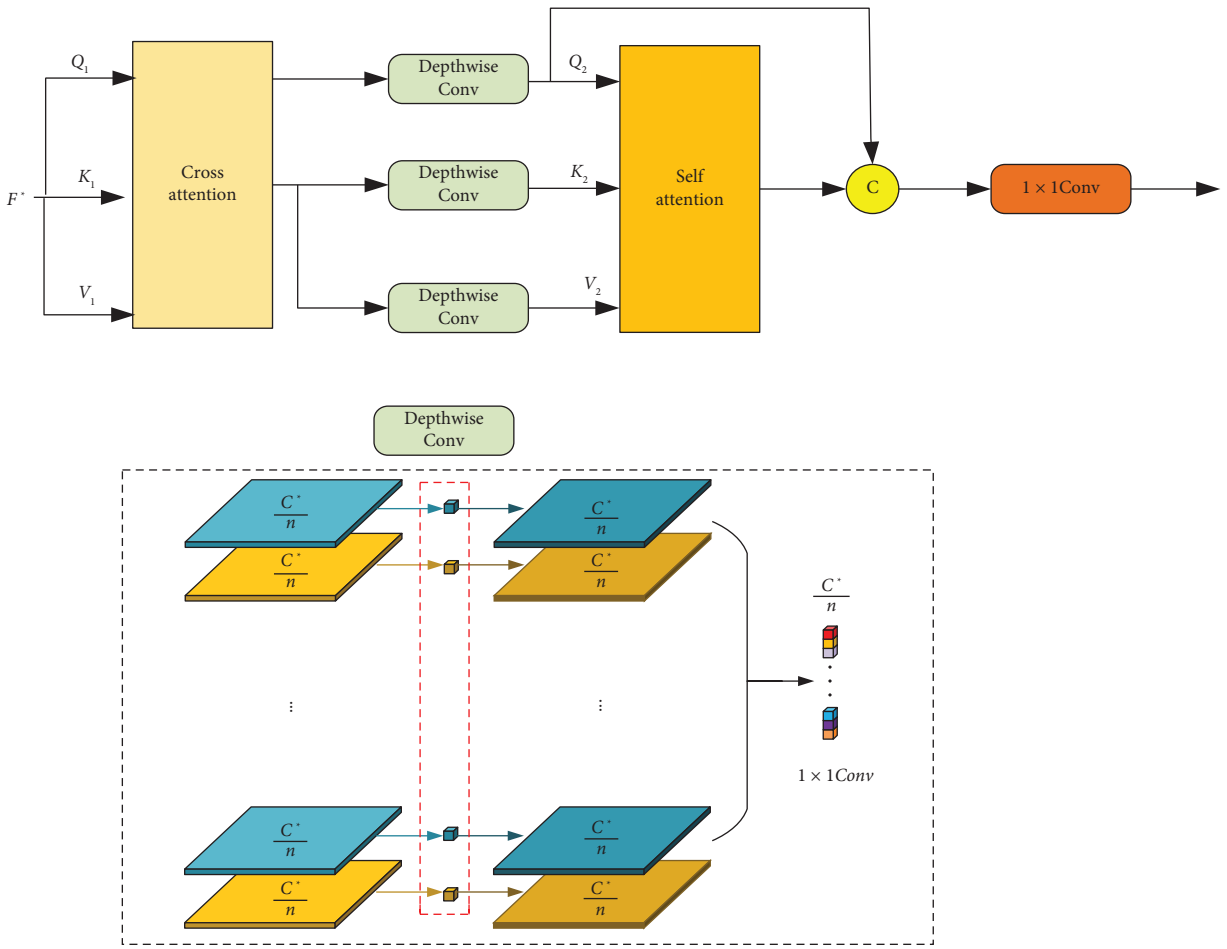


Figure 4: CTrans framework diagram comprising depth-separable convolution, cross-attention, and self-attention.

where $\text{Depth}_{\text{Conv}}$ refers to depth-separable convolution, including point-by-point and depth-by-depth convolutions, $I \in R^{C^{**}\times H^{**}\times W^{**}}$, $C^{**} = C^*/n$, $H^{**} << H^*$, and $W^{**} << W^*$. The input feature map is filtered by depth-by-depth convolution, and the input feature map channels are integrated by point-by-point convolution.

Finally, we used self-attention to establish the correlation between global pixels. CTrans combines the feature maps of high-level and low-level stages. To generate segmentation results, we fed the fused features into a classifier. We computed the inner product of Q_2 and K_2 through self-attention and multiply on the V_2 branch to establish a global pixel relationship. We then fused the output feature map with the feature map Q_2 of the lower stage of the CTrans block to refine the segmentation.

The calculation of self-attention is as follows:

$$\text{SAttn} = \text{Softmax}\left(\frac{Q_2 {K_2}^T}{\sqrt{d_{k_2}}}\right)V_2, \tag{7}$$

$$T = \text{cat}(\text{SAttn}, Q_2),$$

where $\text{SAttn} \in R^{C^{**}\times H^{**}\times W^{**}}$ denotes the self-attentive output feature map, $K_2, Q_2 \in R^{N\times d_{k_2}}$, and $N = H^{**} \times W^{**}$ denotes the number of pixels in the feature map. We stack self-attentive output feature maps and the shallow feature maps Q_2 to produce robust feature maps that contain both spatial information and contextual semantics. We stack the feature map output using the CTrans module and then input them to the category classifier.

$$Y = \text{fcn}_{\text{class}}(T), \tag{8}$$

where T represents the feature map output by the CTrans module at each stage and $\text{fcn}_{\text{class}}$ represents the classifier. $Y \in R^{K\times H\times W}$ represents the output result map, and K, H, and W represent the number of feature map channels, height, and width, respectively.

4. Experiments

4.1. Datasets. Our approach is evaluated with two primary datasets, PASCAL VOC 2012 [51] and Cityscapes [52]. PASCAL VOC 2012 is a comprehensive scene dataset containing 2,913 images with 20 categories. Of the 2,913 images, 1,464 are used for training, 1,449 for validation, and 1,456 for testing.

The Cityscapes dataset contains 5,000 high-quality pixel level annotated images of urban driving scenes, categorized into 30 categories. Of the 5,000 images, 2,975 were used for training, 500 for evaluation, and 1,525 for testing. The images were taken in 50 different cities. This dataset also contains 19,998 coarsely annotated images; here, we only used finely labeled images for 19 categories.

4.2. Implementation Details. To train the model on the Cityscapes dataset, we used stochastic gradient descent (SGD) [53] using a poly-learning rate decay strategy, where the initial learning rate is multiplied by $(1 - \text{iter}/\max\_\text{iter})^{\text{power}}$. For training and validation on the Cityscapes dataset, we used a 0.0025 learning rate, 0.9 weight decay, and 0.0005 momentum.

During the training and validation phases, we cropped the original images to 1024×512 for Cityscapes and 512×512 for PASCAL VOC 2012. The input image is randomly scaled from 0.5 to 2 and flipped horizontally during training for data augmentation. Our backbone network is ResNet101, pretrained on the ImageNet dataset [54]. For the PASCAL VOC 2012 and Cityscapes datasets, the batch sizes were 8 and 4 and the training epochs were 350 and 400, respectively. According to PSPNet, our models are optimized using two cross-entropy losses. The first loss function was applied to the output of the fourth stage of ResNet101 and the second to the model's output. Therefore, the total loss function is as follows:

$$l = \lambda l_{\text{model}} + l_{\text{backbone}_{\text{stage4}}}, \tag{9}$$

where $l_{\text{backbone}_{\text{stage4}}}$ indicates the loss function at the output of the fourth stage of the backbone and l_{model} represents the loss function at the output of our model. λ is set to 0.4.

4.2.1. Evaluation Metrics. In this paper, we use pixel accuracy (PA), intersection over union (IoU), and the mean of IoU (mIoU) as our evaluation metrics. Their calculations are as follows:

$$\begin{aligned} \text{PA} &= \frac{\sum_{i=0}^{k} P_{ii}}{\sum_{i=0}^{k} \sum_{j=0}^{k} p_{ij}}, \\ \text{IoU} &= \frac{p_{ii}}{\sum_{j=0}^{k} p_{ij} + \sum_{j=0}^{k} p_{ji} - p_{ii}}, \quad i = 0, 1, 2 \ldots k, \\ \text{mIoU} &= \frac{1}{K+1} \sum_{i=0}^{k} \frac{p_{ii}}{\sum_{j=0}^{k} p_{ij} + \sum_{j=0}^{k} p_{ji} - p_{ii}}, \end{aligned} \tag{10}$$

where PA represents the ratio of correctly identified pixels to the total number of pixels. IoU, for each class, is calculated as the intersection and concatenation of true and predicted values. mIoU is used to calculate this indicator, first calculate the IoU for each category and then calculate the average.

If there are $k+1$ classes, p_{ij} represents the number of pixels initially in class i but are predicted to be in class j. Therefore, P_{ii} is the predicted true number and p_{ij} and p_{ji} denote false positive and false negative, respectively.

4.3. Ablation Study. In this section, we conduct ablation experiments to verify the effectiveness of our method. We validate the effectiveness of the RDF and CTrans modules on the PASCAL VOC 2012 and Cityscapes datasets, respectively, through ablation experiments.

4.3.1. RDF Module. According to Figure 3, the RDF module can adaptively select features based on various channels. In the group convolution module, feature map channels are compressed to produce compressed feature maps of various spatial information scales. Adaptive maximum pooling can be used to extract salient feature information, whereas GAP can reduce the domain size constraint to preserve more information. An adaptive maximum pool can highlight the unique performance of some features, whereas an average pool can conserve more effective characteristics. As a result, combining adaptive global averaging and adaptive maximum pooling can provide a rich set of information. The sigmoid and softmax functions generate channel attention values, and the softmax function establishes long-run channel dependency and calibrates channel attention weights. In order to prove the influence of adaptive average pooling, adaptive maximum pooling, sigmoid function, and softmax function on experimental results, we successively add adaptive average pooling, adaptive maximum pooling, sigmoid function, and softmax function to the RDF module. Table 1 shows the experimental results. In the first row, GAP combined with the sigmoid function achieves an mIoU of 78.11%, but in the second row, GAP combined with the softmax function achieves an mIoU of 78.72%, which is an improvement of 0.61 compared with the combination with the sigmoid function. In the sixth row, after we included GMP, the corresponding metric reaches 78.88%, which is an improvement of 0.16% compared with the previous value of 78.72% in the second row. Accordingly, GMP improves performance by 0.16%. The Cityscapes dataset has 19 categories. Each category may appear in different scenes, so we embed information extracted from feature maps using adaptive maximum pooling and adaptive average pooling in the channel attention map. Channel attention map assigned different weights to each feature map on the channel, and improved feature differentiation can improve segmentation accuracy.

We verify the performance of the CTrans module in the network using varying convolution kernel group sizes and convolution kernel sizes. We initially set the convolution kernel group size and the convolution kernel size to 1, 2, 4, and 8 and 1, 3, 5, and 7, respectively, achieving an mIoU of 77.95%. Next, we set the convolution kernel group size and convolution kernel size to 1, 4, 8, and 16 and 1, 3, 7, and 9, respectively. This achieves an mIoU of up to 78.88%, an improvement of 0.93% compared with 77.95%. Then, this article sets the convolution kernel group size to 1, 2, 2, and 8 and 1, 4, 4, and 16, respectively, and the corresponding convolution kernel size is set to 1, 4, 4, and 16 and 1, 3, 3, and 9, respectively. The obtained mIoU is 77.76 and 77.62, respectively. Considering the above data, we selected 1, 4, 8, and 16 and 1, 3, 7, and 9 as the parameters of the model.

The metric values decrease as the group size and convolution kernel size increase, as shown in Table 2. Using multiple convolutional groups can increase the speed of

Table 1: The ablation experiments are conducted on the validation set of Cityscapes in the RDF module to verify the adaptive global average pooling (GAP), adaptive global maximum pooling (GMP), and normalization functions, sigmoid and softmax, respectively. The ablation experiments use ResNet50 as the backbone network.

GAP	GMP	Sigmoid	Softmax	mIoU (%)
✓		✓		78.11
✓			✓	78.72
	✓	✓		78.13
	✓		✓	78.02
✓	✓	✓		78.76
✓	✓		✓	78.88

Table 2: Variations in convolutional group size and convolutional kernel size in experiments.

Backbone	Kernel size	Group size	mIoU (%)
ResNet50	1, 3, 5, 7	1, 2, 4, 8	77.95
ResNet50	1, 3, 7, 9	1, 4, 8, 16	**78.88**
ResNet50	1, 3, 3, 5	1, 2, 2, 8	77.76
ResNet50	1, 3, 3, 9	1, 4, 4, 16	77.62

Bold indicates that the highest results are obtained when we choose the appropriate kernel size and group size.

model training by allowing the model to be trained in parallel simultaneously. However, training models in parallel and optimizing them with SGD can lead to slow convergence and poor accuracy depending on the input image batch size. To fully exploit multiscale information from the feature map, the group size and convolution kernel size must be appropriately increased. Table 2 shows that when we set the group size and convolution kernel size to 1, 4, 8, and 16 and 1, 3, 7, and 9, respectively, the model is optimal.

In the CTrans block, we verify the combination of convolution and transformer to improve the model performance. Different feature maps on different channels have interrelated information, so we compute the similarity of the feature maps on two branches to weight the feature maps on the third branch. In addition, we embed convolution in the transformer, enabling the CTrans module to extract spatial information as well as build global information.

We conduct ablation experiments to accurately verify the effect of embedding convolution. Table 3 shows that the experimental results using depthwise conv embedded in the transformer is 78.88%, and the experimental results using the original transformer is 78.66% with 0.22% performance improvement. The experimental results are superior to the results obtained when depthwise conv is used instead of the MLP linear projection layer. Thus, convolution can extract local information and spatial location.

4.3.2. CTrans Block. The multilayer perceptron (MLP) only has the function of linear mapping and has no feature extraction function, so it is not sensitive to spatial details. According to the experimental results, depthwise conv significantly improves the model results compared with the linear projection by MLP. We used ResNet50 as the backbone network and observed the effect of the two modules on the network gain to evaluate the effectiveness of the cross-attention and self-attention modules. Table 4 shows that when using only cross-attention and self-attention, an mIoU of 78.62% and 78.74% are obtained, respectively, indicating that self-attention is 0.12% more effective than cross-attention. Combining the two attention mechanisms achieves an mIoU of 78.88%, which is 0.14% higher than that when only self-attention is used. Multiscaling and horizontal flipping are further applied to the Cityscapes dataset, which achieved an mIoU of 79.31% and 79.56%, respectively. Therefore, the more data there are, the more effective the transformer is.

Table 3: Experimental comparison between depthwise conv and MLP embedded in the input of transformer.

Backbone	Transformer based on depthwise conv	Transformer based on MLP	mIoU (%)
ResNet50	✓		**78.88**
ResNet50		✓	78.66

Bold indicates that the highest results are obtained when we choose the transformer based on depthwise conv.

To fully demonstrate the effectiveness of the CTrans module, we compared it with OCRNet's object contextual representation (OCR) module [23]. As shown in Table 5, CTrans segmentation is 0.12% higher than OCR. We visualized the segmentation graphs for the CTrans and OCR modules in Figure 5. In the third column of the first row, the edge of the wine bottle is missing, but our segmentation result is complete. The human leg in the second row and third column is segmented into a horse, the middle of the cat in the third row and third column is segmented into different classes, and a cow in the fourth row and third column is segmented into a horse. As mentioned above, this problem is called the inconsistency problem within a class. To address this issue, we designed the RDF module to effectively handle the intraclass inconsistency issue. Furthermore, we designed the CTrans module to further mitigate intraclass inconsistency. Similarly, the fourth row's third column of cows is mispredicted because of a lack of contextual semantic.

In contrast, the above issue is nonexistent in CTrans segmentation. As shown in Table 5, despite having 6.8 more parameters than OCR, CTrans has a 0.12% higher mIoU.

(1) Combining RDF and CTrans Modules. We cascaded the CTrans and RDF modules as RFT networks to obtain superior segmentation results. The CTrans module is composed of a depth-separable convolution module, a cross-attention module, a self-attention module, and an FCN Head [1]. In addition, we replaced the transformer's MLP with depth-separable convolution so that the transformer can construct global information and convolutional features. To demonstrate the effects of the two modules, different experimental settings (Table 6) were used to show that adding the RDF and CTrans modules improved semantic segmentation. Compared with the dilated FCN, the RDF module improves mIoU by 6.36% and the CTrans module improves mIoU by 6.67%. When both modules are used, semantic segmentation yields a 78.88% improvement. The results indicate that our method improves semantic segmentation very effectively.

Table 4: Ablation experiments are conducted for self-attention and cross-attention; then, random flipping and multiscaling are applied to improve the effectiveness of the network.

Backbone	RDF and GC and DC	Cross-attention	Self-attention	MS (0.75, 1, 1.25)	FH	mIoU
ResNet50	✓	✓				78.62
ResNet50	✓		✓			78.74
ResNet50	✓	✓	✓			78.88
ResNet50	✓	✓	✓	✓		79.31
ResNet50	✓	✓	✓	✓	✓	79.56

MS denotes multiscaling, FH denotes flip horizontal, RDF denotes the refined division features, GC denotes group convolution, and DC denotes deep separation convolution.

Table 5: Comparative ablation experiments on the Cityscapes for the CTrans and OCR modules.

Method	Backbone	mIoU	Parameters (M)
OCR	ResNet50	78.76	10.5
CTrans (ours)	ResNet50	78.88	17.3

(a) (b) (c) (d)

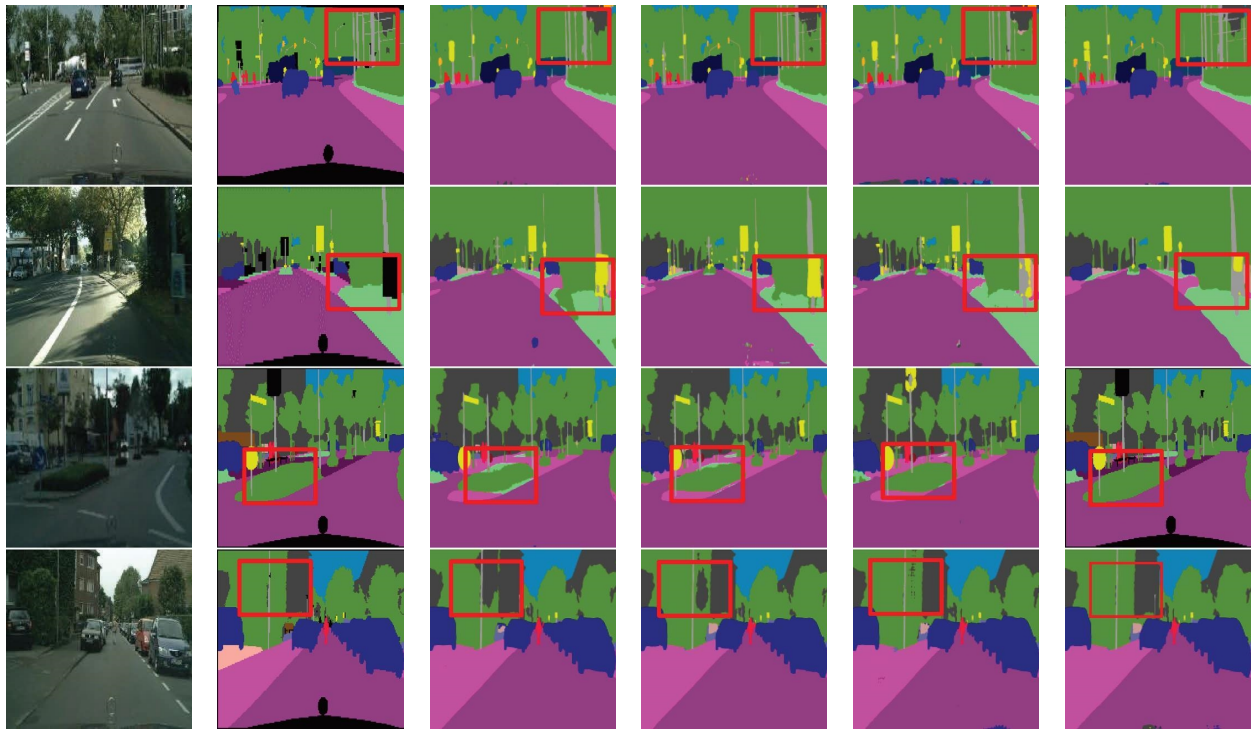
Figure 5: Visual comparison of segmentation graphs on the PASCAL VOC 2012 validation set; from left to right: (a) image, (b) ground truth, (c) OCRNet's object contextual representation (OCR) module, and (d) transformer based on convolution (CTrans) module.

For further verification, we visualized the segmentation maps of the dilated FCN, RDF, and CTrans modules. As shown in Figure 6, the pole in the fourth column of the first row has a limited number of texture features, so the RDF module is introduced to improve feature differentiation; then, the CTrans module is used to improve pixel-region representations. Pavement and grass sections of the second and third rows of the fourth column are divided into other categories in the RDF block. In the CTrans module, the same problem exists as in the RDF module. This is known as intraclass inconsistency. This paper combines RDF and CTrans modules to solve this problem. The RDF block can extract important information and create information interactions between the channels, whereas the CTrans module improves the pixel representation and creates global pixel relationships based on similar feature maps, which avoids the problem of intraclass inconsistency.

Table 6: Ablation experiments on the Cityscapes validation dataset.

Method	Backbone	RDF	CTrans	mIoU
Dilated FCN	ResNet50			71.56
RFT	ResNet50	✓		77.92
RFT	ResNet50		✓	78.23
RFT	ResNet50	✓	✓	78.88

The RDF module represents refined division features, and the CTrans module represents transformer based on depthwise convolution.



(a) (b) (c) (d) (e) (f)

Figure 6: Visual comparison of segmentation graphs; from left to right: (a) image, (b) actual label, (c) dilated FCN, (d) RDF, (e) CTrans, and (f) RDF combined with CTrans.

Table 7: In terms of mIoU, comparison with some existing methods on the cityscapes dataset.

Method	Backbone	Val	MIoU (%)
SA-FFNet	ResNet101	✓	73.1
RELAXNet	ResNet101		74.8
DFNet	ResNet101	✓	79.3
Axial-DeepLab-XL [56]	Axial-ResNet-XL	✓	79.9
PSANet	ResNet101	✓	80.1
SETR-PUP (100k) [57]	T-large	✓	81.08
SETR	ViT-large	✓	81.1
ANNet	ResNet101	✓	81.3
CCNet	ResNet101	✓	81.4
DANet	ResNet101	✓	81.5
OCRNet	ResNet101		81.8
SFNet [58]	ResNet101	✓	81.8
Ours	ResNet101		81.9

The val column indicates whether finely annotated validation set data containing cityscapes was used to train the model.

4.4. Comparison with Classical Semantic Segmentation Networks on Cityscapes Data. To demonstrate the effectiveness of the model, we conducted comparison experiments with other studies on the Cityscapes dataset. We trained our model using the finely labeled Cityscapes dataset, including the RDF module, which improved the distinguishability of

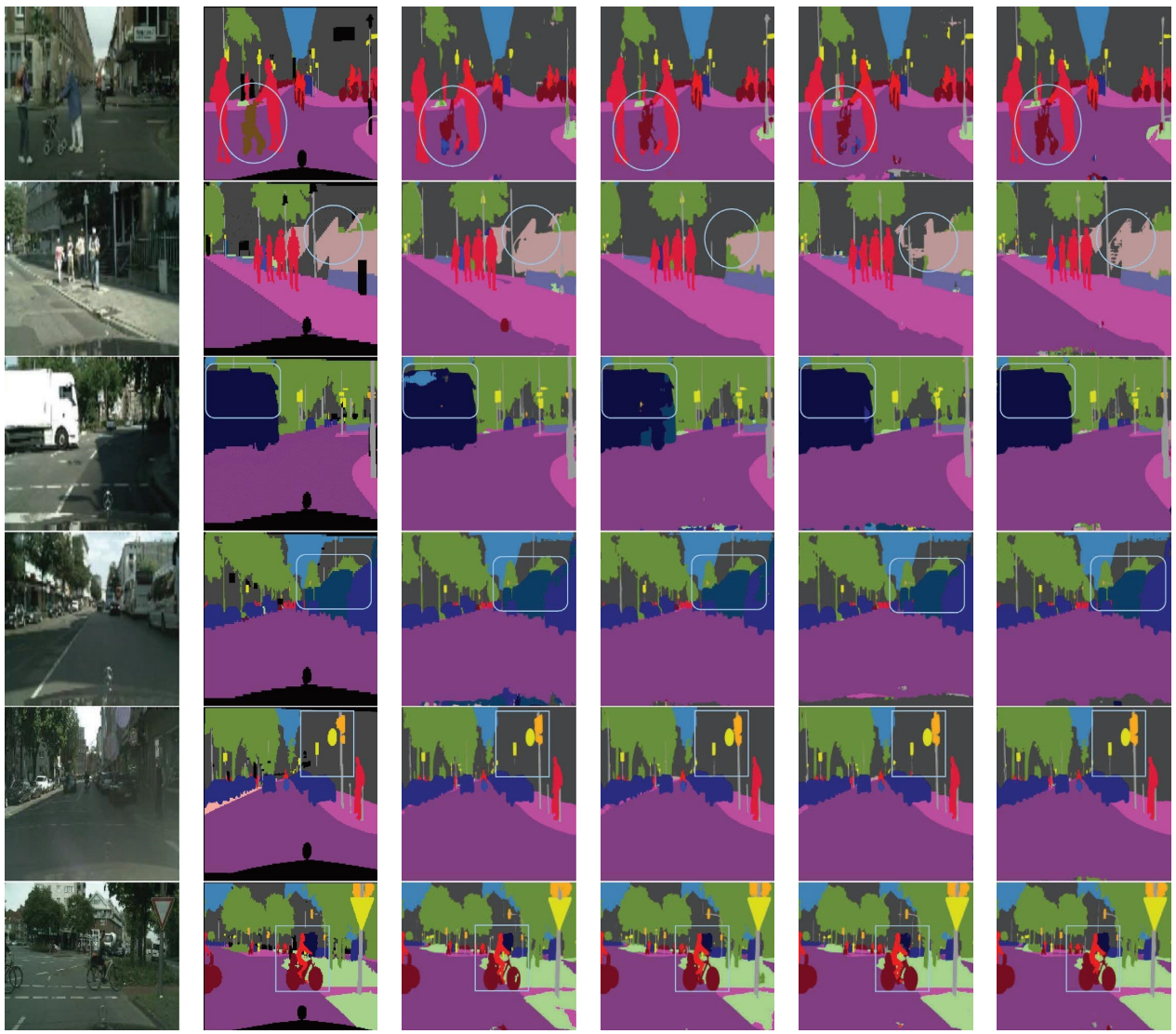


FIGURE 7: Segmentation plots of the cityscapes validation set are shown from left to right for (a) image, (b) ground truth, (c) PSPNet, (d) DANet, (e) OCRNet, and (f) RFTNet.

the feature maps, and the CTrans module, which builds long-distance pixel dependencies to improve pixel representation and construct global information relations.

To improve the segmentation results, we combined the RDF and CTrans modules into RFTNet. As shown in Figure 1, we compared RFTNet with some classical networks, including ANNet, PSPNet, GCNet, CCNet [55], DANet, and OCRNet. Table 7 shows that our network's segmentation mIoU is 81.9%, which is significantly better than that of other methods.

As shown in Figure 1, we visualize the data and find that our mIoU metric is 81.9%, which is 0.1% higher than that of OCRNet. In addition, our method has more network parameters than GCNet, but it has a 5% higher mIoU value than GCNet. Therefore, the parameters of our network will be a topic for future studies.

In Figure 7, we visualize the segmentation result graphs of the above networks, where the other methods segment bicycle wheels into different classes, while our network segments bicycle completely. Similarly, the wall, truck, and car in rows 2, 3, and 4 are partially segmented into other classes. We combine the RDF and CTrans modules to enhance feature differentiation first and then enhance pixel-region representation, which improves these problems greatly. In the fourth and fifth rows, the bars and bicycle front ends have few texture features, making them difficult to segment out. We can, however, segment out the subtle speed bars and bicycle front ends with the help of our method.

5. Conclusion

This paper enhances the category region representation and pixel representation from micro and macro aspects, respectively. The image-level context information is easily affected by the outside world, and other categories of context information are introduced into the pixel representation, resulting in network misclassification. Inspired by this problem, from a macro point of view, this paper proposes RDF module to enhance the representation of channel category region in the feature graph. To further enhance the performance of semantic segmentation, we design the CTrans module from the micro point of view. First, it compacts and enriches the feature map to reduce the computational load of CTrans module. Then, the similarity between pixels is used to enhance the pixel representation. Finally, the global relationship between the pixels is established. The method in this paper can accurately segment object categories under the conditions of illumination changes, similar colors, background, and so on. Compared with other methods, our segmentation index and segmentation effect are optimal. However, the method in this paper has some limitations. On the one hand, it is still necessary to improve the ability of the model to segment the boundary of small objects with fuzzy edges. On the other hand, our model was trained, validated, and tested on Cityscapes and PASCAL VOC 2012 data sets, which are commonly used. They are both refined and annotated data sets. The generalization ability of the model is not strong enough in the face of images that differ greatly from the two data sets. Therefore, for the method proposed in this paper, we need to test and modify our model on more data sets to maximize our model generalization ability.

Data Availability

The data that support the findings of this study are available from the author Tianping Li but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data are, however, available from the author on reasonable request and with permission of the author Tianping Li.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Authors' Contributions

Tianping Li performed formal analysis, Tianping Li and Yanjun Wei proposed the methodology, Jun Du supervised the study, and Tianping Li and Yanjun Wei wrote the original draft.

Acknowledgments

The study was supported by the National Natural Science Foundation of China, grant/award numbers: 61472220 and 61572286.

References

[1] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 4, pp. 640–651, 2017.

[2] Y. Chen, H. Zhang, L. Liu et al., "Research on image inpainting algorithm of improved total variation minimization method," *Journal of Ambient Intelligence and Humanized Computing*, vol. 14, no. 5, pp. 5555–5564, 2021.

[3] Y. Chen, R. Xia, K. Zou, and K. Yang, "FFTI: image inpainting algorithm via features fusion and two-steps inpainting," *Journal of Visual Communication and Image Representation*, vol. 91, Article ID 103776, 2023.

[4] R. Xia, Y. Chen, and B. Ren, "Improved anti-occlusion object tracking algorithm using Unscented Rauch-Tung-Striebel smoother and kernel correlation filter," *Journal of King Saud University-Computer and Information Sciences*, vol. 34, no. 8, pp. 6008–6018, 2022.

[5] Y. Chen, L. Liu, V. Phonevilay et al., "Image super-resolution reconstruction based on feature map attention mechanism," *Applied Intelligence*, vol. 51, no. 7, pp. 4367–4380, 2021.

[6] T. Y. Lin, P. Dollár, and R. Girshick, "Feature pyramid networks for object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2117–2125, Honolulu, HI, USA, July 2017.

[7] V. Badrinarayanan, A. Kendall, and R. C. SegNet, "A deep convolutional encoder-decoder architecture for image segmentation," *arXiv Preprint ArXiv:1511.00561*, vol. 5, 2015.

[8] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 1520–1528, Venice, Italy, 2015.

[9] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, https://arxiv.org/abs/1706.05587.

[10] 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)*, pp. 833–851, Glasgow, United Kingdom, September 2018.

[11] 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*, pp. 2881–2890, Honolulu, Hawaii, USA, July 2017.

[12] Z. Zhou, Y. Zhou, D. Wang, J. Mu, and H. Zhou, "Self-attention feature fusion network for semantic segmentation," *Neurocomputing*, vol. 453, pp. 50–59, 2021.

[13] J. Liu, X. Xu, Y. Shi, C. Deng, and M. Shi, "Relaxnet: residual efficient learning and attention expected fusion network for real-time semantic segmentation," *Neurocomputing*, vol. 474, pp. 115–127, 2022.

[14] H. Zhang, K. Dana, and J. Shi, "Context encoding for semantic segmentation," in *Proceedings of the-IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 7151–7160, Seattle, WA, USA, 2018.

[15] M. Teichmann, M. Weber, M. Zoellner, R. Cipolla, and R. Urtasun, "Multinet: real-time joint semantic reasoning for autonomous driving," in *Proceedings of the IEEE Intelligent Vehicles Symposium*, pp. 1013–1020, Changshu, China, June 2018.

[16] H. Huang, L. Lin, and R. Tong, "A full-scale connected unet for medical image segmentation," 2004, https://arxiv.org/abs/2004.08790.

[17] F. Yang, C. Lu, Y. Guo, L. J. Latecki, and H. Ling, "Dually supervised feature pyramid for object detection and segmentation," 2019, https://arxiv.org/abs/1912.03730.

[18] M. Salman and S. E. Yüksel, "Fusion of hyperspectral image and lidar data and classification using deep convolutional neural networks," in *Proceedings of the 26th Signal Processing and Communications Appl Conference (SIU)*, pp. 1–4, Izmir, Turkey, May 2018.

[19] H. Zhao, Y. Zhang, and S. Liu, "Psanet: point-wise spatial attention network for scene parsing," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 270–286, Glasgow, United Kingdom, September 2018.

[20] Z. Zhu, M. Xu, S. Bai, T. Huang, and X. Bai, "Asymmetric non-local neural networks for semantic segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 593–602, Montreal, BC, Canada, October 2019.

[21] Y. Yuan, X. Chen, and J. Wang, "Object-contextual representations for semantic segmentation," in *Proceedings of the Lecture Notes in Computer Science Eur Conference on Computer Vision*, pp. 173–190, Glasgow, UK, August 2020.

[22] J. Fu, J. Liu, and H. Tian, "Dual attention network for scene segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3146–3154, New Orleans, Louisiana, USA, June 2019.

[23] Y. Chen, H. Gan, and Z. Zeng, "DADCNet: dual attention densely connected network for more accurate real iris region segmentation," *International Journal of Intelligent Systems*, vol. 37, no. 1, pp. 829–858, 2022.

[24] Z. Liu, J. Ou, W. Huo, Y. Yan, and T. Li, "Multiple feature fusion-based video face tracking for IoT big data," *International Journal of Intelligent Systems*, vol. 37, 2021.

[25] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, "Learning a discriminative feature network for semantic segmentation," in *Proceedings-IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 1857–1866, Salt Lake City, UT, USA, June 2018.

[26] F. Lin, T. Wu, S. Wu, S. Tian, and G. Guo, "Feature selective transformer for semantic image segmentation," 2022, https://arxiv.org/abs/2203.14124.

[27] A. Dosovitskiy, L. Beyer, and A. Kolesnikov, "An image is worth 16x16 words: transformers for image recognition at scale," 2020, https://arxiv.org/abs/2010.11929.

[28] J. Guo, K. Han, and H. Wu, "Cmt: convolutional neural networks meet vision transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12175–12185, New Orleans, Louisiana, USA, June 2022.

[29] K. Li, Y. Wang, and J. Zhang, "Uniformer: unifying convolution and self-attention for visual recognition," 2022, https://arxiv.org/abs/2201.09450.

[30] Q. Liu, Y. Dong, and X. Li, "Multi-stage context refinement network for semantic segmentation," *Neurocomputing*, vol. 535, pp. 53–63, 2023.

[31] Q. Zhou, X. Wu, and S. Zhang, "Contextual ensemble network for semantic segmentation," *Pattern Recognition*, vol. 122, Article ID 108290, 2022.

[32] X. Zhang, Q. Li, and Z. Quan, "Pyramid geometric consistency learning for semantic segmentation," *Pattern Recognition*, vol. 133, Article ID 109020, 2023.

[33] S. Yi, J. Li, and G. Jiang, "CCTseg: a cascade composite transformer semantic segmentation network for UAV visual perception," *Measurement*, vol. 211, Article ID 112612, 2023.

[34] G. Lin, A. Milan, C. Shen, and I. Reid, "Refinenet: multipath refinement networks for high-resolution semantic segmentation," in *Proceedings of the IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit*, pp. 1925–1934, Honolulu, Hawaii, USA, July 2017.

[35] C. Peng, X. Zhang, G. Yu, G. Luo, and J. Sun, "Large kernel matters–improve semantic segmentation by global convolutional network," in *Proceedingsof the-IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 4353–4361, Honolulu, Hawaii, USA, July 2017.

[36] Y. Yuan and J. O. Wang, "Object context network for scene parsing," 2018, https://arxiv.org/abs/1809.00916.

[37] Q. Yang, S. Hu, W. Zhang, and J. Zhang, "Attention mechanism andadaptive convolution actuated fusion network for next POI recommendation," *International Journal of Intelligent Systems*, vol. 37, 2022.

[38] J. Chen, Y. Lu, and Q. Yu, "Transunet:transformers make strong encoders for medical image segmentation," 2021, https://arxiv.org/abs/2102.04306.

[39] W. Wang, C. Chen, and M. Ding, "Transbts:Multimodal brain tumor segmentation using transformer," in *Proceedings of the International Conference on Medical Image Computing and Computer-AssistedIntervention*, pp. 109–119, Strasbourg, France, October 2021.

[40] O. Ronneberger, P. Fischer, and T. Brox, "U-net:Convolutional networks for biomedical image segmentation," in *Proceedings of the International Conference on Medical Image Computing and Computer-AssistedIntervention*, pp. 234–241, Munich, Germany, October 2015.

[41] X. Li, S. Ma, and J. Tang, "Transiam: fusing multimodal visual features using transformer for medical image segmentation," 2022, https://arxiv.org/abs/2204.12185.

[42] Z. Dai, H. Liu, and Q. V. Le, "Coatnet: marrying convolution and attention for all data sizes," *Advances in Neural Information Processing Systems*, vol. 34, pp. 3965–3977, 2021.

[43] Z. Peng, W. Huang, and S. Gu, "Conformer: local features coupling global representations for visual recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 367–376, Montreal, BC, Canada, October 2021.

[44] Y. Chen, X. Dai, and D. Chen, "Mobile-former: bridging mobilenet and transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5270–5279, New Orleans, Louisiana, USA, June 2022.

[45] T. Xiao, M. Singh, E. Mintun, T. Darrell, P. Dollár, and R. Girshick, "Early convolutions help transformers see better," *Advances in Neural Information Processing Systems*, vol. 34, pp. 30392–30400, 2021.

[46] L. C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected crfs," 2014, https://arxiv.org/abs/1412.7062.

[47] 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, 2018.

[48] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, Las Vegas, Nevada, USA, June 2016.

[49] H. Zhang, K. Zu, J. Lu, Y. Zou, and D. Meng, "Epsanet: an efficient pyramid squeeze attention block on convolutional neural network," 2021, https://arxiv.org/abs/2105.14447.

[50] A. Vaswani, N. Shazeer, and N. Parmar, "Attention is all you need," in *Proceedings of the Advances in NeuralInformation Processing Systems*, pp. 5998–6008, Long Beach, CA, USA, December 2017.

[51] M. Everingham and J. Winn, "The pascal visual object classes challenge; 2012 1–45:(voc2012) development kit," *Pattern Anal Stat Model Comput Learn Tech Rep*, vol. 2007, 2012.

[52] M. Cordts, M. Omran, and S. Ramos, "The cityscapes dataset for semantic urban scene understanding," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3213–3223, Las Vegas, NV, USA, June 2016.

[53] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[54] J. Deng, "A large-scale hierarchical image database," in *Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition*, Miami, FL, USA, June 2009.

[55] 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*, pp. 603–612, Montreal, BC, Canada, October 2019.

[56] H. Wang, Y. Zhu, and B. Green, "Axial-deeplab: stand-alone axial-attention for panoptic segmentation," in *Proceedings of the Computer Vision–ECCV 2020: 16th European Conference*, pp. 108–126, Springer International Publishing, Glasgow, UK, August 2020.

[57] S. Zheng, J. Lu, and H. Zhao, "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6881–6890, Nashville, TN, USA, June 2021.

[58] X. Li, A. You, and Z. Zhu, "Semantic flow for fast and accurate scene parsing," in *Proceedings of the Computer Vision–ECCV 2020: 16th European Conference*, pp. 775–793, Springer International Publishing, Glasgow, UK, August 2020.