CRLNet: Cascaded Resolution Learning Network for Unstructured Semantic Segmentation in Natural Scenes

Wei Li[b,c], Shishun Tian[a,b,c], Guoguang Hua[b,c], Muxin Liao[a,c], Yuhang Zhang[b,c], Wenbin Zou[a,b,c,*]

[a]*Guangdong Key Laboratory of Intelligent Information Processing, Shenzhen University, China, 518060, Guangdong, shenzhen*
[b]*Shenzhen Key Laboratory of Advanced Machine Learning and Applications, Shenzhen University, China, 518060, Guangdong, shenzhen*
[c]*College of Electronics and Information Engineering, Shenzhen University, China, 518060, Guangdong, shenzhen*

Abstract

With the increasing demand for environmental perception in practical applications such as rescue robots, autonomous driving, and drone navigation, unstructured semantic segmentation has received widespread attention. Different from the structured street scene analysis with clear road topography, natural scenes contain irregular objects and different terrains with highly similar appearances, which pose significant challenges for unstructured semantic segmentation. Existing semantic segmentation methods typically employ feature fusion to improve segmentation performance. However, they cannot fully leverage feature information of different resolutions and their network receptive fields are limited, which leads to some degradation in segmentation accuracy. To tackle this issue, we propose a cascaded resolution learning network (CRLNet) to improve the segmentation performance through global textual embedding and multi-resolution feature learning. The proposed network constructs a multi-path segmentation system, which integrates multi-resolution feature information from different paths and progressively achieves fine learning of local features. CRLNet includes two vital modules:Partition-Fusion Channel Attention Module (PFCAM) and Features Learning Module (FLM). The PFCAM is a computationally efficient channel attention module to improve the segmentation confusion caused by similar objects, which assigns partition-fusion and channel attention on multi-path textual information. FLM is designed to learn resolution feature maps of different paths, which is help to refining object representation and improving the segmentation performance by integrating multi-path resolution feature maps. Extensive experiments on real natural scene datasets (RUGD and RELLIS) and street scene dataset (Cityscapes) demonstrate that the proposed CRLNet outperforms existing efficient semantic segmentation methods in terms of accuracy. The code will be available at https://github.com/lv881314/CRLNet.

Keywords: Unstructured semantic segmentation, deep neural networks, partition-fuse channel attention module, features learning module

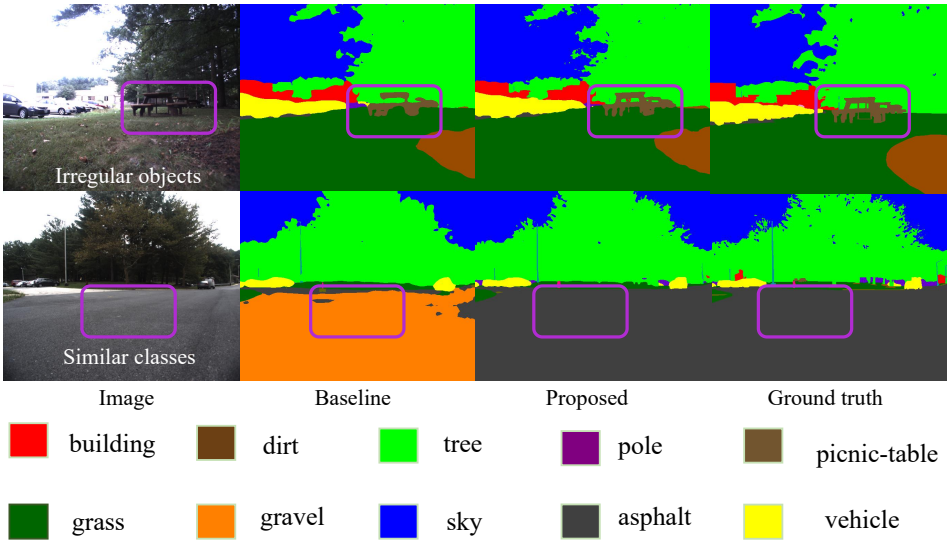

Figure 1: Typical deficiencies for unstructured semantic segmentation in natural scenes, dominant irregular objects segmentation and similar terrain classes.

1. Introduction

Semantic segmentation is a fundamental but important perception task, which predicts and assigns appropriate semantic labels to each pixel in an input image. Semantic segmentation provides valuable information about the types, contours, and geometric positions of terrain in visual scenes, it is significant to various practical application, such as rescue robots [1], autonomous driving [2], and drone navigation [3], etc. However, when images contain lighting changes, texture differences, and complex natural structures, semantic segmentation in natural environments becomes more challenging compared to regular street scene segmentation, as street scenes have a uniform environment and terrain structures that are easy to segment [4].

When using deep neural networks [5, 6, 7, 8, 9, 10, 11, 12, 13] to perform unstructured semantic segmentation tasks, there are two significant challenges that current deep neural networks will face in achieving unstructured semantic segmentation. These challenges arise due to the limited depth of the network architecture [5, 6, 11] and the insufficient overall receptive field of the network [7, 8, 9]. First, the down-sampling operation can lead to coarse segmentation labels, especially when segmenting some small unstructured objects, such as irregular "picnic-table" is shown in Fig. 1. Secondly, due to the lower feature extraction ability of shallow network structures compared to deep ones, it is difficult for a neural network to accurately assign semantic labels to some terrains that are similar in

⋆This document is the results of the research project funded by the National Science Foundation, Guangdong Provincial Natural Science Foundation, Shenzhen Natural Science Foundation.

*Corresponding author at: College of Electronics and Information Engineering, Shenzhen University, Shenzhen, 518060, China. e-mail: wzou@szu.edu.cn (Wenbin Zou)

[1]College of Electronics and Information Engineering, Shenzhen University, Shenzhen, 518060, China.e-mail: 2250432002@szu.edu.cn

appearance and geometry. For example, similar terrain classes "gravel" and "asphalt" are shown in Fig. 1, which poses significant challenges for applications such as rescue robots that require outdoor operations on terrains with different safety levels.

To address these issues, we propose a cascaded resolution learning network (CRLNet) that continuously aggregates multi-resolution features. In the multi-resolution features aggregation procedure, CRLNet continuously integrates global semantic information from multiple paths and collects edge details, thereby effectively alleviating segmentation errors in irregular objects and similar terrain categories. Concretely, CRLNet consists of two important feature learning modules: (1) The Partition-Fusion Channel Attention Module (PFCAM) is used to collect local fine-grained channel features. The use of PFCAM in multiple paths effectively ensures that rich local details are embedded in the multi-stage network, which mitigates the problem of rough semantic labels; (2) Features Learning Module (FLM) based on gated recurrent unit (GRU) [12, 14] is designed to integrate resolution feature maps from adjacent paths. FLM uses two different GRUs to address the problem of similar terrain classification, and the multi-resolution feature learning in different paths can improve the segmentation effect of different terrain categories. The main contributions of our work are summarized as follows.

- **Multi-stage unstructured semantic segmentation network.** We propose a natural environment-oriented unstructured semantic segmentation method based on multi-stage network structure to gradually extract high-resolution features and improve the representation ability of feature maps.
- **Partition-Fusion Channel Attention Module.** We propose a partition-fusion channel attention module (PFCAM), which utilizes the partition-fusion and channel attention mechanism to collect delicate local channel features.
- **Features Learning Module.** We propose a features learning module (FLM) to learn the significance of multi-path resolution feature maps during fusion, which achieves a multi-stage representation validly.

2. Related work

This section introduces some relevant works to our proposed approach, such as self-attention mechanism, multi-stage feature embedding and semantic segmentation.

2.1. Self-attention mechanism

Self-attention mechanism was first applied in natural language processing to represent the global independence of different words in a sentence [15]. In visual segmentation tasks, self-attention mechanisms are often used in deep

neural networks for denoising and restoring high-quality feature maps [16, 17, 18, 19]. Self-attention mechanisms in semantic segmentation can be classified into spatial self-attention mechanisms and channel self-attention mechanisms [12, 20]. The spatial information attention map from high-level feature indicates the key of per pixel, which attention on restoring the edges with spatial information and positioning the objects. However, due to the fact that spatial self-attention mechanisms are based on spatial pixel positions, the burden on the network increases when the scale of the input image is too large. For example, the adjacent position attention module proposed by MANet [21] repeatedly calculates the feature matrix between adjacent pixels many times, and the multi-view attention mechanism module proposed by CFNet [22] uses parallel multiple different kernel sizes of dilated convolution. Although these methods can improve the segmentation quality of images, they are limited by local resources. Different from the aforementioned methods, we use a simple and effectively partition-fusion operation to divide the input image into blocks, which greatly simplifying the computational complexity.

Moreover, the channel attention map obtained from multi-resolution represents the significance of per channel, which focuses on global context information [23, 24]. SENet [25] proposes a squeeze-and-excitation idea to model the correlation between feature channels, enhance the features of important channels, and weaken the features of unimportant channels, thereby strengthening the important features to improve the accuracy of image segmentation. Based on squeeze-and-excitation (SE), ECANet [26] propose a local cross-channel interaction strategy to reduce the complexity of the network and ensure the effectiveness of network. SE-CNN [27] construct a feature extractor by combining SE and convolutional neural network to improve the accuracy of image classification. RegSeg [28] designed an SE-ResNeXt backbone structure to achieve a large receptive field and retain local detailed information. Following the aforementioned work, we designed a partition-fusion channel attention module based on partition-fusion operation and SE in this paper to collect local detailed information.

2.2. Multi-stage feature embedding

Feature embedding is regularly utilized in semantic segmentation to integrate multi-stage representations [29, 30, 31]. For feature embedding, one approach is to apply multi-stage architecture. MCRNet [14] build a inverted residual pyramid block module to learn abundant context feature; MSFNet [32] propose two multi-stage transformers to fuse detail information from different stages. However, the aforementioned method ignores the representation gap between multi-stage features, which hinders friendly propagation of information between multi-stage features. Recently, recurrent neural network (RNN) [33] uses gated recurrent unit (GRU) [34] to learn high-dimensional features and progressively update these local embeddings into the network. It controls the information propagation between multiple dimensions well through gate units, but is limited by computation while ensuring the effectiveness of the network. In that regard, we propose features learning module (FLM) which uses two different activation functions

(sigmoid and tanh) to learn important information in multi-stage resolution feature maps. To enhance the effectiveness of information propagation, we also incorporate a simple self-attention mechanism into the module. Details will be introduced in Section 3 on features learning module.

2.3. semantic segmentation

Thanks to the advancement of convolutional neural networks, many deep neural networks such as Resnet[35] and ResNext [36] are often used as encoders in neural networks for feature extraction. To improve the performance of the network, many decoders have been proposed to generate accurate segmentation labels. Early semantic segmentation networks [37, 38, 39] often had poor segmentation performance due to their simple network structure. To improve the performance of the network, most current segmentation networks typically use multi-stage feature fusion [32, 33, 34] and local self-attention mechanisms [26, 27, 28] to enhance feature representation. The effectiveness of these methods has been demonstrated in the regularized street scene analysis, but only a few works have focused on unstructured semantic segmentation tasks in real natural scenes. Considering this point, we attempt to learn irregular object features in natural environments through network structure and feature optimization modules in this paper, thereby solving the classification confusion problem caused by similar terrain in unstructured environments.

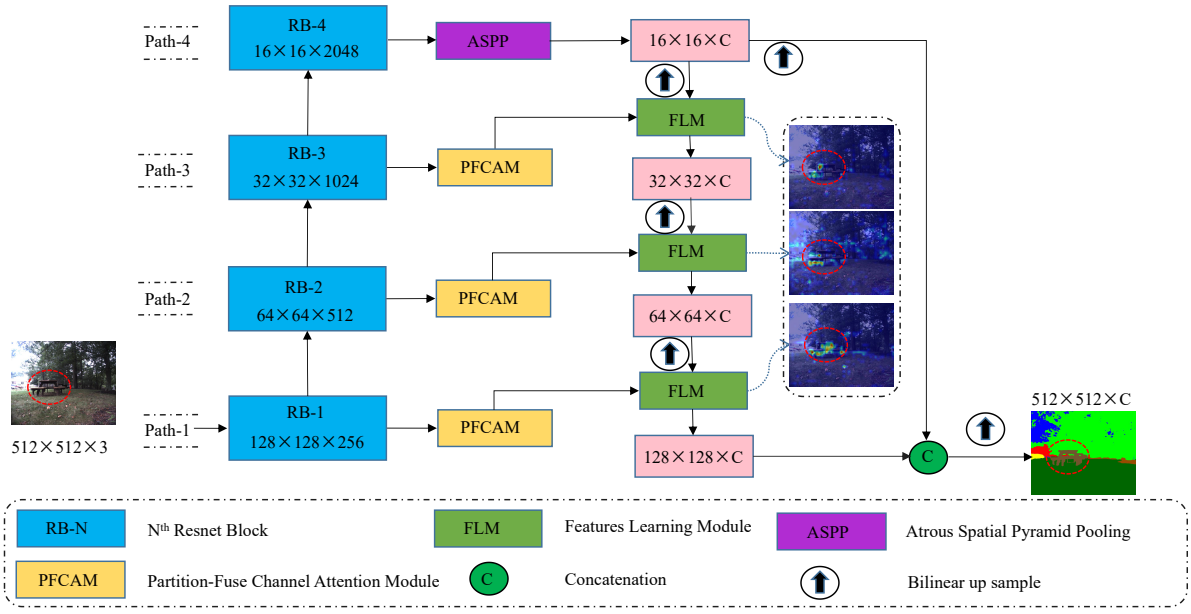

Figure 2: An overall of the framework of the proposed CRLNet. It includes four multi-resolution Resnet features. We set the value of C to 256 in all experiments. The blue dashed lines represent the heatmaps of features learned through multi-level cascaded learning in the FLM module. The irregular object marked in red circle demonstrates the effectiveness of CRLNet network in extracting features from natural scenes. Please zoom in to view more details.

3. Method

In this section, we provide a detailed introduction to the unstructured semantic segmentation network CRLNet that we proposed. First, we introduce the overview architecture of CRLNet, then we discuss two important modules in CRLNet: partition-fusion channel attention module (PFCAM) and features learning module (FLM), and we finally presented the lightweight and medium versions of CRLNet.

3.1. The overview architecture of CRLNet

As shown in Fig. 2, CRLNet has 4 Resnet paths for feature extraction. The resolution of feature on the same path is the same. Given a RGB image X$\in \mathbb{R}^{H_i \times W_i \times 3}$, where H_i and W_i represent the width and height of the image, and the number of channels is 3. In the CRLNet network, the RGB image is first passed through the Resnet blocks (RB-1 to RB-4) to generate four different feature maps. The spatial resolution of these four feature maps is 1/4, 1/8, 1/16, and 1/32 times of the original RGB image, and the number of channels increases to 256, 512, 1024, and 2048, respectively. For each Resnet block on paths 1-3, we utilize the PFCAM to collect local channel features of the Resnet blocks. For the Resnet block on path 4. As CRLNet is designed based on a multi-stage network structure starting from low-resolution images, it is necessary to appropriately expand the receptive field of the network to meet the needs of traversing all input images. Therefore, we choose to use the Atrous Spatial Pyramid Pooling (ASPP) [40] to expand the receptive field of the network and extract enough feature representations on path 4.

In addition to the Resnet backbone blocks on the four different paths, the low-resolution feature maps from higher paths are first 2× upsampled using bilinear interpolation to match the spatial resolution of the corresponding second highest path. Then they are embedded into the high-resolution feature maps from the adjacent second highest path. After multiple low-to-high feature embeddings in FLM, this allows the low-resolution feature maps from the higher paths can comprehensively learn different feature representations at multiple stages, and ultimately obtain rich high-resolution semantic information with a low number of channels. As described in [41, 42, 43], the combination of semantic local information and global spatial detail information can achieve good network performance. Therefore, when predicting the final segmentation label, we concatenate the low spatial resolution feature maps processed by ASPP on path-4 with the high-resolution feature maps obtained through multi-stage embedding, and then perform 4× upsampling to obtain segmentation label prediction.

3.2. Partition-fusion channel attention module

The partition-fusion channel attention modulet (PFCAM) is contrived to collect local channel features from path-1 to path-3. The detail of our proposed PFCAM is shown in Fig. 3. Given a feature map RB-N$\in \mathbb{R}^{H \times W \times C}$, where W

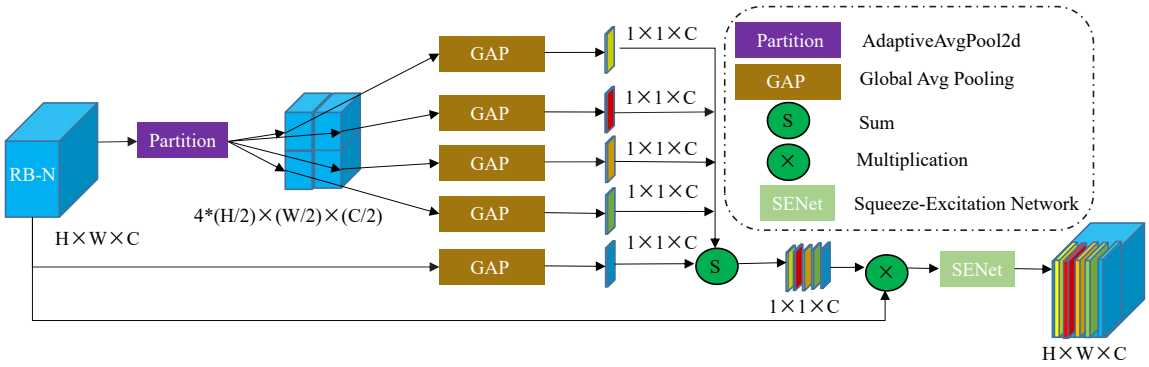

Figure 3: The details of Partition-Fusion Channel Attention Module.

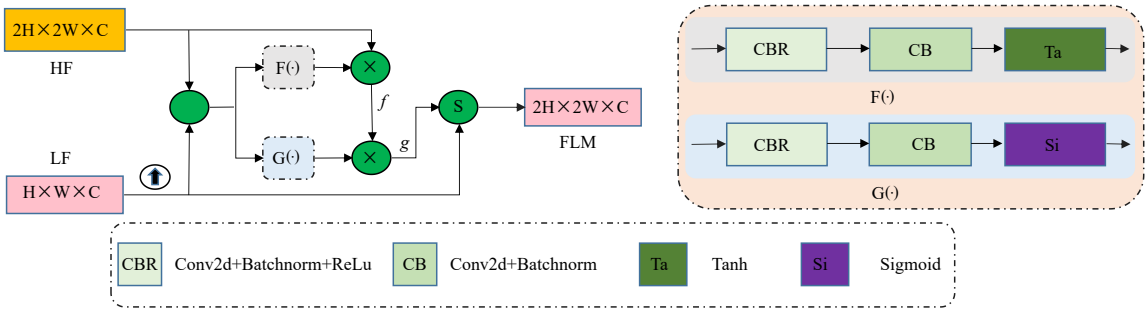

Figure 4: The details of Features Learning Module.

and H represent the width and height of the image, and the number of channels is C. The RB-N feature map is first split into four local feature maps ($\mathbb{R}^{H/2\times W/2\times C}$) with the same resolution and channel number through adaptive average pooling in the spatial domain. Then, global average pooling is used to capture the contextual information of the entire image. In order to establish contextual weight relationships between the RB-N feature map and the four local segmentation regions, we also use global average pooling to obtain the contextual information of the original feature map, and then the contextual information of the original feature map is superimposed with the textual information of the four local segmentation regions to strengthen the contextual weighting between different positions of the RB-N feature map. Afterwards, we multiply the dense contextual weighting with the original RB-N feature map to collect key information. Finally, we use the SENet [25] to extract important channel feature map and achieve local channel feature extraction for segmentation. It should be noted that although dividing the input feature map into more local regions can appropriately improve the overall segmentation performance, it will also increase the computational burden of the network. Therefore, in this paper, we only divide it into 2×2 local regions to extract local channel feature information of the RB-N feature map.

3.3. Features learning module

Low-resolution contextual information is a global summary of the input image, which has a significant impact on generating accurate segmentation labels. It contains little local detail information. Therefore, intermediate features from Resnet blocks can be extracted to compensate for this deficiency. Although many network models based on multi-branch and multi-scale architectures have been proposed to fill in the missing detail information, their segmentation performance is poor or requires complex computational overhead [14, 30, 31, 32]. To address this issue, we propose a feature learning module (FLM) based on the gate recurrent unit mechanism in this paper, which continuously learns high-resolution local detail features from the low pathway and fills in this part of local feature information into low-resolution images. The detail of our proposed FLM is shown in Fig. 4.

Assuming that there are two feature maps ($HF \in \mathbb{R}^{2H\times 2W\times C}$ and $LF \in \mathbb{R}^{H\times W\times C}$), one is the low-resolution feature map in the high pathway and the other is the high-resolution feature map in the sub-high pathway, where H, W, and C are the height, width, and channel of the feature maps respectively. First, the low-resolution feature map ($LF \in \mathbb{R}^{H\times W\times C}$) in the high pathway is bilinearly upsampled to the same size as the high-resolution feature map in the sub-high pathway. Then, the feature maps with the same resolution size in the high and low pathways are concatenated. Next, two different gate functions ($F(\cdot)$ and $G(\cdot)$) based on the gate recurrent unit mechanism are designed to determine which important local information can be passed into the next stage of feature patching. The weight function $F(\cdot)$ is constrained to the interval [-1, 1], while the weight function $G(\cdot)$ is constrained to the interval [0, 1]. Subsequently, the high-resolution feature map ($HF \in \mathbb{R}^{2H\times 2W\times C}$) in the sub-high pathway is multiplied by the two weight functions (f and g) separately to obtain the local detail information contained in the intermediate backbone. Finally, to facilitate friendly information dissemination, we have incorporated a simple self-attention mechanism into this module. Specifically, it is added to the low-resolution feature map in the high pathway to obtain complete semantic information. This module is applied to the fusion of high and low-resolution feature maps in different paths, which can continuously learn different high-resolution local detail information to compensate for the missing detail information in the low-resolution feature map. At the same time, the two different gate weight functions constrain the data value calculation to the range [-1, 1], which effectively reducing computational complexity.

3.4. Light and medium network

Generally speaking, calculations on high-resolution feature maps require more computation resources than those on low-resolution feature maps. From the network architecture perspective, the computational resources occupied by the low pathway are greater than those of the high pathway. For example, Path-1 occupies more computational resources than Path-2, and the computational resources of each path increase in accordance with this pattern. In the

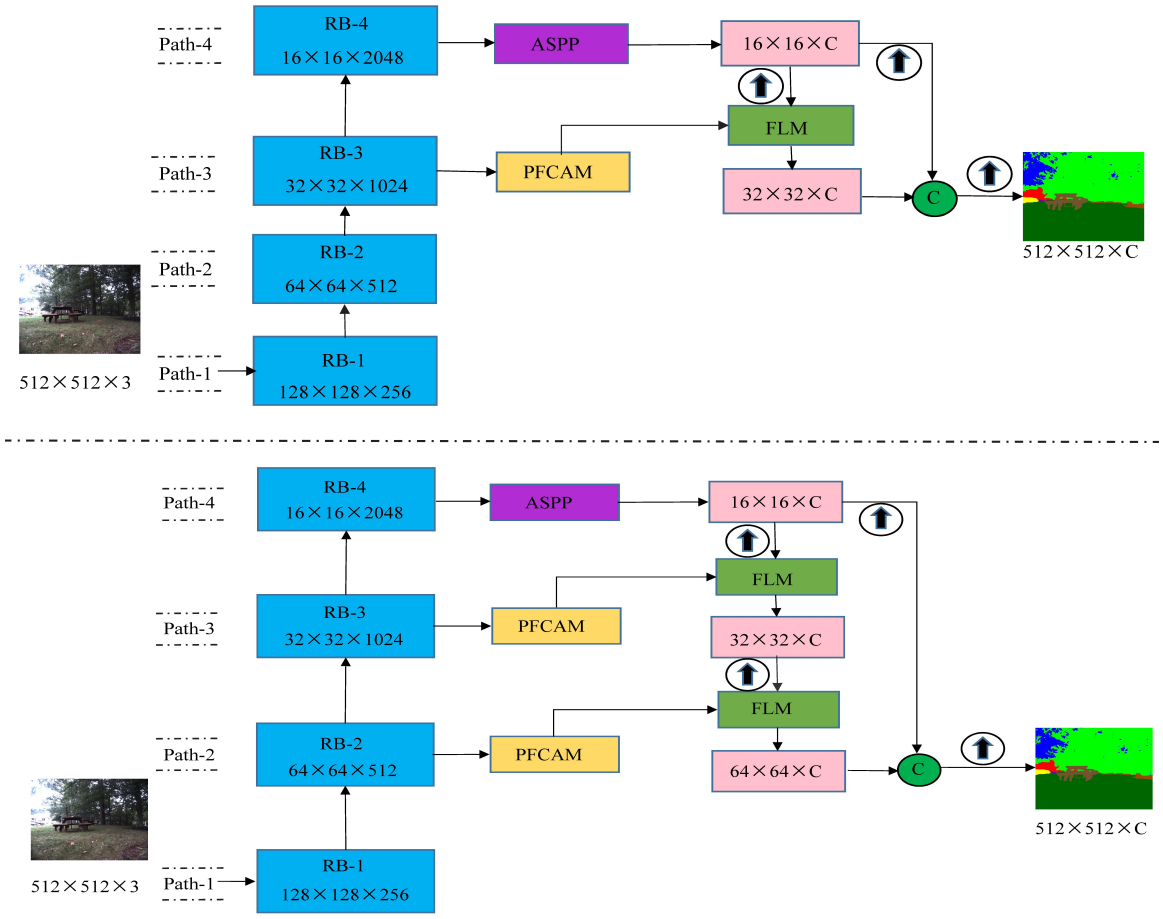

Figure 5: The framework of the proposed CRLNet-light (upper) and CRLNet-medium (bottom) network.

CRLNet, the feature maps on Path-1 to Path-4 are downsampled to 1/4, 1/8, 1/16, and 1/32 of the original image, respectively. Due to the high computational cost of processing high-resolution feature maps in the low paths, we selectively remove the high-low resolution feature learning modules in Path-1 to obtain a simplified network called CSRNet-medium (the bottom image in Fig. 5). Similarly, we can also remove the feature learning modules in Path-2 to obtain the simplest network called CSRNet-light (the upper image in Fig. 5). The medium and light versions of the CRLNet network avoid heavy computations on high-resolution feature maps, which further improves the computational speed of the network. The original network is referred to as CRLNet-heavy (as shown in Fig. 2). CRLNet-heavy continuously utilizes the Feature Learning Module (FLM) to learn global information from high-path feature maps and aggregates the global semantics into low-path feature maps that contain rich spatial information. Using FLM multiple times in multiple stages can embed rich contextual information and significantly expand the receptive field.

4. Experiments

In this section, we conducted a large number of experiments to validate our proposed method. Firstly, we introduced the unstructured natural datasets (RUGD [44] and RELLIS [45]) and the structured street scene dataset Cityscapes [46]. Additionally, the experimental details of the CRLNet network are also elaborated. Then, we compared the effectiveness of the CRLNet network with some advanced semantic segmentation methods. Next, a series of ablation experiments were conducted on the RUGD to test the roles of each element.

4.1. Experimental Settings

1) RUGD: The RUGD is a recent popular semantic segmentation dataset for scene parsing. It is collected from natural and unstructured off-road environments. Due to highly irregular scene semantic information, it poses great challenges to outdoor off-road driving. Additionally, the resolution of RUGD is 668×550. It contains 4759 images in the training set, 1964 images in the validation set and 733 images in the testing set. There are 24 classification labels, such as building, grass, picnic-table, and asphalt, etc.

2) RELLIS: The RELLIS is another prevalent unstructured out-door scene dataset. RELLIS was collected by Texas A&M University and raises challenges to current algorithm involved to terrain environmental. The resolution of RELLIS is 1920×1200. It includes 3302 images for training, 983 images for validating and 1672 images for testing. It has 19 classification labels, such as: tree, vehicle, bicycle and building, etc.

3) Cityscapes: The Cityscapes dataset is prevalent urban street scene dataset in the area of self-driving. It includes 2975 images for training, 500 images for validating and 1525 images for testing. The performance of Cityscapes is mainly assessed by pixel mean intersection over union (mIoU) across the 19 classes. Different from RUGD and RELLIS, the Cityscapes test set only have the original images for testing. It requires user to submit the segmentation results to the official website of Cityscapes and obtains the segmentation performance from the Cityscapes official assessment system. The segmentation data from the CRLNet on the Cityscapes official website can be found in **https://github.com/lv881314/CRLNet**.

4) Implementation protocol: All the training experiments are executed with one RTX 3090 GPU, CUDA 11.7 on the PyTorch platform. The mini-batch stochastic gradient descent ([40]) is applied for training with batch size of 8, momentum 0.9. Besides, we use the "policy" from ([47]) for learning rate policy and the initial learning rate of $1e^{-2}$ with power of 0.9 for RUGD, RELLIS and Cityscapes. The random scale and random mirror are employed during training, where the random parameters contain 0.75, 1.0, 1.25, 1.5, 1.75 and 2.0. Besides, data agumentation including horizontal flipping and rotation(-10° to 10°) is applied to avoid overfitting during training.

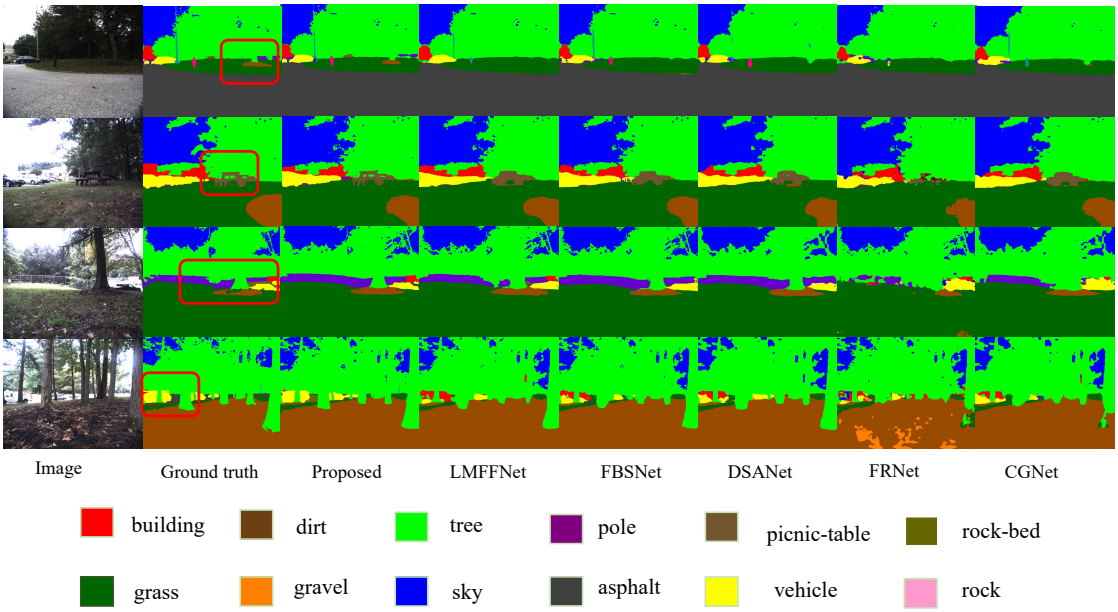

Figure 6: Comparison of visualization segmentation results on the RUGD dataset. Please zoom in to view more details.

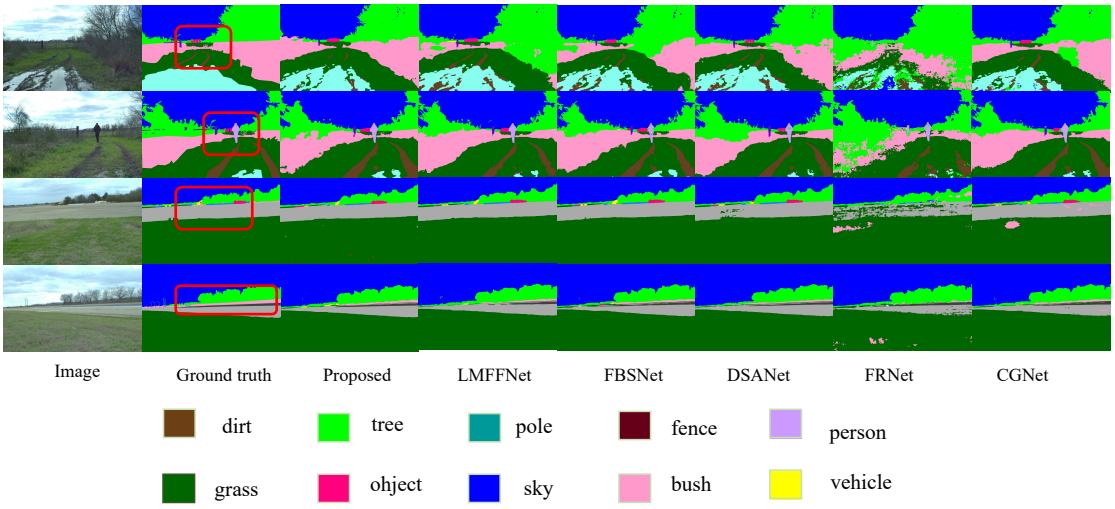

Figure 7: Comparison of visualization segmentation results on the RELLIS dataset. Please zoom in to view more details.

4.2. Results

In this subsection, we compare our proposed CRLNet with other state-of-the-art networks on two outdoor unstructured datasets (RUGD dataset and RELLIS dataset) and one urban street scene dataset (Cityscapes dataset).To evaluate the segmentation performance, we used the mean intersection over union (mIoU), Mean Accuracy (mAcc) and Overall Accuracy (OA) to evaluate the segmentation accuracy, where mAcc represents the average accuracy of all categories, and OA means the overall accuracy, which does not consider categories and only considers the classification of all

Table 1: Results on RUGD dataset for different SOTA methods. The best-performing data is highlighted in bold.

Methods	mIoU % ↑	mAcc % ↑	OA % ↑
Deeplabv3 [40]	38.10	<u>46.46</u>	91.02
MobileNet [48]	31.24	36.68	88.84
PSPNet [49]	37.86	46.03	90.24
HRNet [50]	32.88	39.87	87.27
DANet [51]	34.35	41.50	89.11
DABNet [52]	35.33	41.06	90.46
DSANet [43]	36.85	43.51	90.50
CGNet [53]	34.22	39.75	90.40
LMFFNet [54]	36.83	43.06	91.04
FBSNet [41]	37.45	44.43	<u>91.24</u>
S2FPNet [55]	<u>39.13</u>	45.47	90.97
FRNet [56]	26.14	30.34	89.10
CRLNet	**42.07**	**49.79**	**91.73**

Table 2: Results on RELLIS dataset for different SOTA methods. The best-performing data is highlighted in bold.

Methods	mIoU % ↑	mAcc % ↑	OA % ↑
Deeplabv3 [40]	41.13	48.58	92.19
MobileNet [48]	32.19	35.98	90.12
PSPNet [49]	36.60	43.24	92.00
DANet [51]	36.22	42.75	91.68
DABNet [52]	40.20	45.32	90.72
DSANet [43]	42.23	47.37	87.60
CGNet [53]	34.22	39.75	90.40
LMFFNet [54]	43.62	48.75	91.94
FBSNet [41]	43.32	50.04	<u>92.47</u>
S2FPNet [55]	<u>44.02</u>	<u>50.14</u>	91.39
FRNet [56]	27.69	33.74	79.84
CRLNet	**48.33**	**55.80**	**95.54**

samples.

4.2.1. Comparison with the state-of-the-art networks

We present the segmentation results of our proposed CRLNet network on the outdoor dataset. Table 1 reports the segmentation performance of CRLNet compared to some excellent segmentation networks on the RUGD dataset. Most of these excellent algorithms are based on models for semantic segmentation using multi-scale features or multi-level cascaded features, such as CGNet [53], LMFFNet [54], FBSNet[57], S2FPNet [55], and FRNet [56], etc. From Table 1, it can be seen that our segmentation performance in natural environments is superior to these multi-scale feature models. Specifically, compared to S2FPNet, our segmentation performance has improved by 2.94% in terms of mIoU, and our segmentation results have improved by 3.33% in terms of mAcc compared to Deeplabv3. Our model's segmentation performance outperforms FBSNet by 0.495% in terms of OA. In addition, we provide visual comparisons of the segmentation results between our network and five other state-of-the-art networks on the

Table 3: Results on RUGD test set for different SOTA methods. The best-performing data is highlighted in bold, while the second-best data is underlined.

Method	Di	Sa	Gras	Tr	Po	Wa	Sk	Ve	CG	As	Grav	Bu	Mu	RB	Lo	Bi	Pe	Fe	Bu	Si	Ro	Br	Co	PT
MoblileNet[43]	0	0	83.9	88.5	8.2	50.7	60.0	47.7	0.6	90.7	79.3	32.4	79.2	0	24.2	0	0	21.1	27.4	0	39.1	0	2.0	52.0
HRMet[50]	0	0	84.9	90.7	29.1	**60.1**	67.8	58.4	26.4	8.5	51.8	36.7	79.6	0	32.6	0	0	17.9	31.0	0	45.4	0	5.6	62.7
DANet[51]	0	0	84.4	88.6	9.5	55.4	59.0	55.2	12.9	82.3	75.3	31.9	79.1	0	27.9	0	0	27.1	26.0	0	46.2	0	9.7	53.9
DSANet[41]	0	0	85.7	89.0	23.4	54.0	65.9	56.8	28.1	**94.9**	81.0	33.7	80.4	0	31.5	0	0	23.1	31.3	0	43.3	0	2.2	59.30
CGNet[53]	0	0	85.7	89.5	23.1	50.8	65.6	51.7	1.78	94.6	**81.4**	35.0	80.8	0	29.1	0	0	13.2	29.8	0	37.8	0	0.2	51.3
LMFFNet[54]	0	0	86.5	90.5	26.4	52.9	69.9	55.3	8.2	94.3	81.1	36.1	80.8	0	32.9	0	0	22.3	33.6	0	46.0	0	6.21	60.4
S2FPNet[56]	0	0	86.3	90.6	28.0	55.4	68.0	66.9	27.2	87.4	78.6	34.3	82.1	0	38.6	0	0	29.0	**35.9**	0	51.6	0	10.3	64.2
FBSNet[57]	0	0	**86.7**	91.0	25.4	56.7	71.2	55.9	12.6	91.8	79.3	37.3	80.8	0	33.8	0	0	27.5	40.7	0	44.5	0	0.71	62.3
FRNet[56]	0	0	84.8	88.7	6.2	8.8	67.0	39.4	0.7	90.7	72.6	25.9	75.8	0	23.2	0	0	9.43	18.7	0	3.1	0	0.24	11.42
CRLNet	0	0	86.1	**92.0**	**34.9**	55.4	**73.7**	**75.5**	**44.9**	94.0	81.1	**41.6**	82.9	0	**40.6**	0	0	**32.3**	30.2	0	**55.3**	0	**13.9**	**75.2**

[1] Di:dirt; Sa:sand; Gras:grass; Tr:tree; Po:pole; Wa:water; Sk:sky; Ve:vehicle; CG:container/generic-object; As:asphalt; Grav:gravel; Bu:bilding; Mu:mulch; RB:rock-bed; Lo:log; Bi:bicycle; Pe:person; Fe:fence; Bu:bush; Si:sign; Ro:rock; Br:bridge; Co:concrete; PT:picnic-table.

Table 4: Results on RELLIS test set for segmentation models. The best results are highlighted in bold.

Method	Di	Gr	Tr	Po	Wa	Sk	Ve	Ob	As	Bu	Lo	Pe	Fe	Bu	Co	Ba	Pu	Mu	Ru
MobliIeNet[43]	0	85.6	72.4	1.9	0.7	95.9	11.4	24.6	0	0	0	57.6	9.6	68.8	67.5	20.7	66.3	26.3	2.4
PSPNet[49]	0	88.3	77.3	0.4	8.2	95.9	7.5	19.8	1.9	0	0	72.7	10.7	74.4	72.2	30.9	70.3	34.1	27.3
DANet[51]	0	88.3	76.0	1.0	11.2	96.1	5.9	20.3	0	0	0	72.8	14.7	74.5	72.2	21.0	**69.6**	34.6	30.1
DSANet[41]	0	84.5	**78.7**	2.0	0.1	86.5	18.4	50.4	37.5	0.4	0	83.8	27.0	69.4	80.6	**56.9**	63.3	32.8	29.9
CGNet[53]	0	89.3	76.9	1.2	0.4	96.7	16.7	43.1	27.9	0.8	0	77.8	17.5	73.2	72.1	45.5	55.9	36.9	47.3
LMFFNet[54]	0	89.2	77.6	2.3	2.2	**96.9**	28.9	44.5	24.5	**2**	0	83.4	19.7	74.5	82.9	47.9	65.9	34.8	51.4
S2FPNet[56]	0	86.1	79.5	1.1	5.3	**96.9**	**29.5**	**51.2**	6.9	**11.5**	0	**86.8**	27.7	71.3	83.1	53.6	54.7	38.1	52.3
FBSNet[57]	0	**89.7**	76.8	**2.9**	0	96.5	26.9	42.3	41.8	0.7	0	59.3	21.5	**75.3**	**84.1**	54.7	63.7	39.2	47.8
FRNet[56]	0	62.2	65.2	0	5.1	96.6	5.5	0.8	45.9	0.1	0	45.1	0.1	48.4	52.7	35.9	32.4	7.3	21.6
CRLNet	0	89.6	78.3	1.7	**46.6**	**96.9**	27.7	49.8	**51.9**	0	0	77.5	26.4	74.6	81.0	55.8	66.2	**40.0**	**54.2**

[1] Di:dirt; Gras:grass; Tr:tree; Po:pole; Wa:water; Sk:sky; Ve:vehicle; Ob:object; As:asphalt; Bu:bilding; Lo:log; Bi:bicycle; Pe:person; Fe:fence; Bu:bush; Co:concrete; Ba:barrier; Pu:puddle; Mu:mud; Ru:rubble.

Table 5: Results on Cityscapes test set for segmentation models. The best-performing data is highlighted in bold, while the second-best data is underlined.

Method	Roa	Sid	Bui	Wal	Fen	Pol	TLi	TSi	Veg	Ter	Sky	Per	Rid	Car	Tru	Bus	Tra	Mot	Bic	mIoU↑	Iou-C↑
CGNet[53]	95.5	78.7	88.1	40	43	54.1	59.8	63.9	89.6	67.6	92.9	74.9	54.9	90.2	44.1	59.5	25.2	47.3	60.2	64.8	85.7
EDANet[58]	97.8	80.6	89.5	42	46	52.3	59.8	65	91.4	68.7	93.6	75.7	54.3	92.4	40.9	58.7	56	50.2	64	67.3	85.8
ERFNet[59]	97.2	80	89.5	41.6	45.3	56.4	60.5	64.6	91.4	68.7	94.2	76.1	56.4	92.4	45.7	60.6	27	48.7	61.8	66.3	85.2
ICNet[60]	97.1	79.2	89.7	43.2	48.9	61.5	60.4	63.4	91.5	68.3	93.5	74.6	56.1	92.6	51.3	72.7	51.3	53.6	70.5	69.5	86.4
LEDNet[61]	**98.1**	79.5	91.6	47.7	49.9	**62.8**	61.3	72.8	92.6	61.2	94.9	76.2	53.7	90.9	64.4	64	52.7	44.4	71.6	70.6	87.1
ESNet[62]	**98.1**	80.4	**92.4**	48.3	49.2	61.5	62.5	72.3	92.5	61.5	94.4	76.6	53.2	94.4	62.5	74.3	52.4	45.4	71.4	70.7	87.4
AGLNet[63]	97.8	80.1	91.0	51.3	50.6	58.3	63.0	68.5	92.3	**71.3**	94.2	80.1	59.6	93.8	48.4	68.1	42.1	52.4	67.8	70.1	87.0
NDNet[64]	97.4	80.8	91.0	49.1	51.9	55.7	64.1	72.1	91.3	61.0	93.5	76.9	58.4	92.9	70.7	79.1	71.5	60.4	73.2	64	73.2
DSANet[41]	96.8	78.5	91.2	50.5	50.8	59.4	64.0	71.7	92.6	70.0	94.5	81.8	61.9	92.9	56.1	75.6	50.6	50.9	66.8	71.4	87.4
FDDWNet[65]	98.0	82.4	91.1	52.5	51.2	59.9	64.4	68.9	92.5	70.3	94.4	80.8	59.8	**94.0**	56.5	68.9	48.6	55.7	67.7	71.5	-
FPANet[66]	97.1	78.1	90.0	45.0	50.0	58.4	65.9	72.0	92.3	69.4	94.4	81.9	62.7	93.9	50.2	71.7	60.8	57.1	69.6	72.0	87.7
JPANet[67]	**98.1**	82.7	90.6	47.1	48.2	53.1	60.2	66.1	91.6	69.2	94.5	78.6	60.5	**94.0**	64.2	75.9	66.5	63.7	65.9	71.6	86.7
CRLNet	98.0	**82.8**	92.2	**53.8**	**58.7**	62.3	**72.4**	**75.5**	**92.9**	69.7	**95.0**	**83.9**	**67.9**	93.8	**64.9**	**79.5**	**72.2**	**64.9**	**72.9**	**76.5**	**89.4**

[1] Roa:road; Sid:sidewalk; Bui:building; Wa:wall; Fen:fence; Pol:pole; TLi:traffice light; Tsi:traffice sign; Veg:vegetation; Ter:terrain; Sky:sky; Per:person; Rid:rider; Car:car; Tru:truck; Bus:bus; Tra:train; Mot:motorcycle; Bic:bicycle; Iou-C: iIou Categories.

Table 6: Results on Cityscapes dataset for different SOTA methods. The best-performing data is highlighted in bold.

Methods	Resolution	mIoU %(test/val) ↑
CSRNet [12]	768×768	76.0/77.3
FRNet [56]	512×1024	70.4/-
GDN [68]	512× 1024	75.6/-
Relaxnet [69]	512×1024	74.8/-
LEANet [70]	512×1024	71.9/-
LMFFNet [54]	512×1024	75.1/74.9
FBSNet [41]	512×1024	70.9/-
DABNet [52]	1024× 2048	70.1/69.6
CGNet [53]	1024× 2048	64.8/63.5
LSPNet [71]	1024×2048	74.9/76.5
CRLNet	1024× 2048	**76.5**/**82.09**

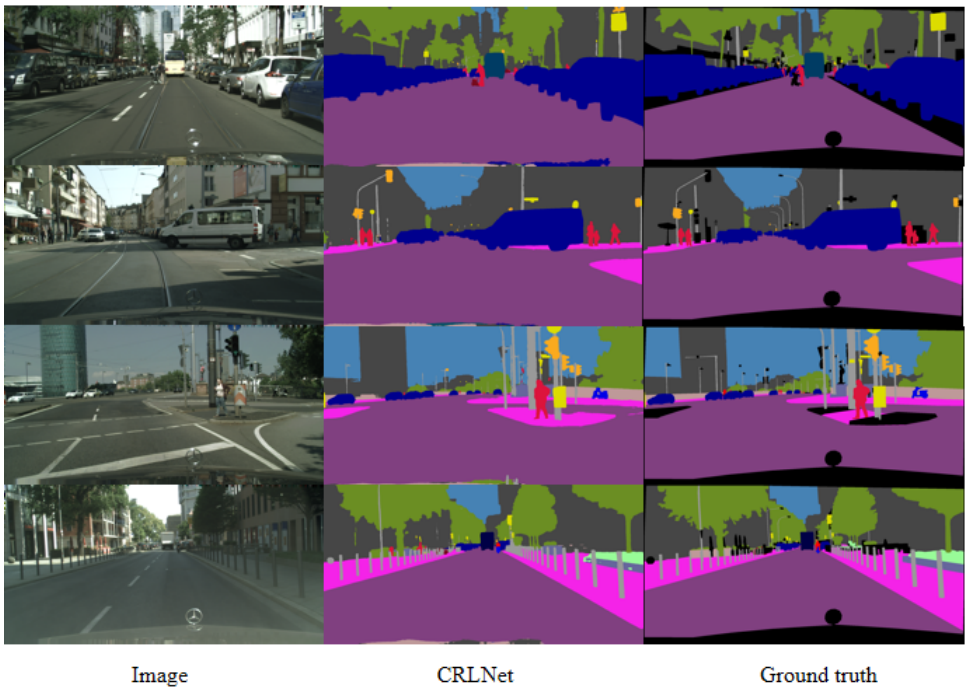

Figure 8: Qualitative examples of CRLNet on Citysacpes validation set. Please zoom in to view more details.

RUGD dataset, as shown in Fig. 6. We can observe that our proposed CRLNet network achieves comparable visual segmentation results to the other networks.

Furthermore, to validate the robustness of CRLNet, we conducted experiments comparing CRLNet with some advanced segmentation networks on the RELLIS unstructured outdoor dataset. The results are shown in Table 2. Compared to the S2FPNet network, CRLNet outperforms S2FPNet with improvements of 4.31% mIoU, 5.66% mAcc, and 4.15% OA and our segmentation performance has improved by 2.94% in term of mIoU, 5.66% in term of mAcc, and 3.07% in term of OA compared to FBSNet. This indicates that the CRLNet network outperforms the aforementioned

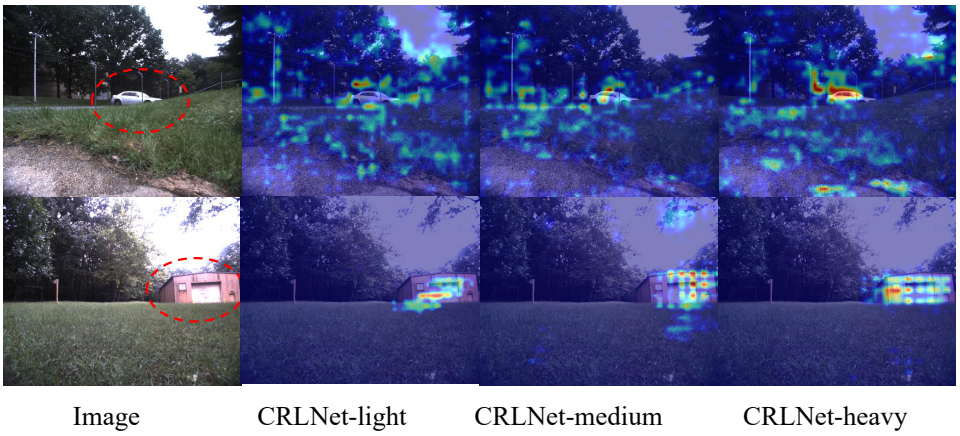

Figure 9: Visual feature extraction heatmaps of different versions of CRLNet models. From top to bottom, we selected car and building for analysis respectively.

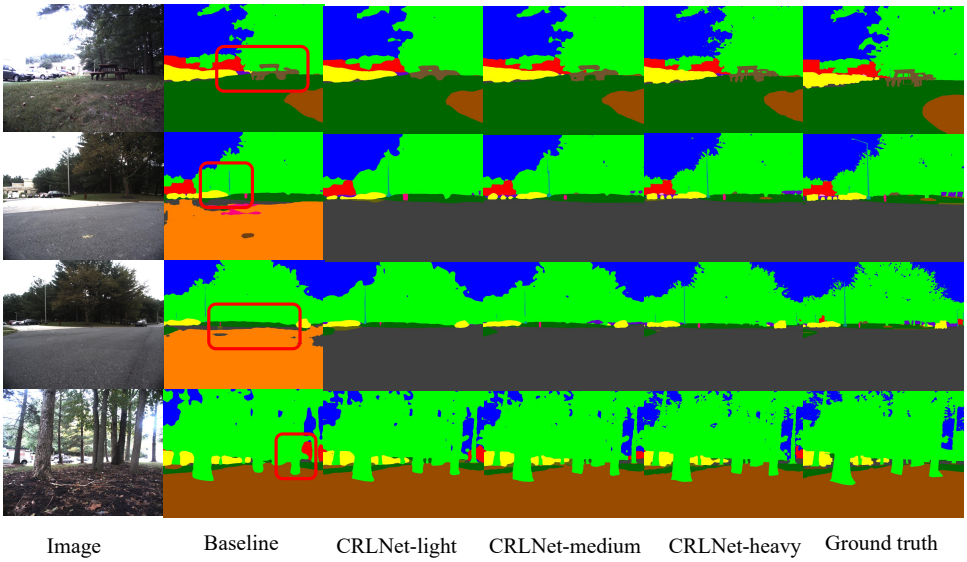

Figure 10: Performance on segmenting irregular objects or similar classes on the RUGD dataset. Please zoom in to view more details.

advanced segmentation networks in terms of unstructured semantic segmentation performance. Likewise, we provide the visual results of color segmentation on the RELLIS dataset (as shown in Fig. 7). From Fig. 7, it can be observed that our segmentation results on the RELLIS dataset are more detailed compared to several other segmentation methods. In addition, we have shown the segmentation results for a single category in Table 3 and Table 4. From Table 3 and Table 4, it can be observed that our proposed CRLNet network consistently maintains high category segmentation accuracy. Specifically, our CRLNet network demonstrates excellent segmentation performance for non-structured objects such as grass, water, tree, asphalt, and picnic tables. Additionally, it also shows decent segmentation results for regular objects like vehicles and buildings. Comparing with the state-of-the-art networks mentioned above, the

Table 7: Results on RUGD dataset for different CRLNet network. The best-performing data is highlighted in bold.

Model description	mIoU % ↑	mAcc % ↑	OA % ↑
Baseline	37.02	44.70	89.85
CRLNet-light	37.86	46.03	90.24
CRLNet-medium	40.88	48.35	91.65
CRLNet-heavy	**42.07**	**49.79**	**91.73**

segmentation results indicate that our proposed network is competitive in segmenting non-structured objects in natural scenes.

Moreover, we also validate the generalizability of CRLNet on another widely used outdoor street scene dataset, Cityscapes. We first demonstrate the segmentation performance of CRLNet compared to some recent state-of-the-art multi-scale multi-feature semantic segmentation algorithms in Table 6. From Table 6, it can be seen that our proposed CRLNet achieves a mIoU of 76.5% on the cityscapes test set at a resolution of 1024x2048, which is 0.9% mIoU higher than GDN and 1.6% mIou higher than LSPNet, both recently published methods. Additionally, in Table 5, we present the classification results for individual objects on the Cityscapes test set. From Table 5, it can be seen that our network CRLNet exhibits good segmentation performance on regular objects such as walls, fences, traffic lights, and traffic signs. It also performs well on non-structured objects such as vegetation and trucks. This indicates the strong generalizability of CRLNet to various outdoor scenes. Besides, some results on the Cityscapes validation set are shown in Fig. 8, which intuitively show that our proposed CRLNet network can segment the urban street well.

4.2.2. Performance on segmenting irregular objects or similar classes

As described in section 3, we removed some modules from the original network and proposed two low-complexity CRLNet networks (CSRNet-light, CSRNet-medium). The segmentation performance of the three different versions of the CRLNet network is reported in Table 7, where "Baseline" refers to the basic network without PFCAM and FLM. As shown in Table 7, with the incorporation of PFCAM and FLM modules, the CRLNet network learns more and more different resolution information from multiple pathways. CRLNet-heavy achieved the best segmentation performance with a mIoU of 42.07%, mAcc of 49.79%, and OA of 91.73%, which improved by 5.05%, 5.09%, and 1.88% respectively compared to "Baseline" in terms of mIoU, mAcc, and OA. This indicates that our proposed CRLNet network improves the segmentation performance of unstructured objects and similar classes through multi-stage feature learning.

To visually observe the segmentation performance of different CRLNet networks on unstructured objects and similar classes, we first use Grad-CAM [72] to generate visualized feature maps to demonstrate the feature extraction performance of various versions of CRLNet that integrate multi-level PFCAM and FLM modules. From Fig. 9, it can be observed that CRLNet can effectively extract feature information from natural environments through the proposed

PFCAM and FLM modules in a progressive manner. This indicates the rationality and effectiveness of the network structure. Secondly, we present the segmentation results of different versions of CRLNet in Fig. 10. As shown in Fig. 10, the Baseline network exhibits rough segmentation results for outdoor tables (as shown in the first row of Fig. 10). However, as CRLNet continuously learns high-resolution features from different paths, the segmentation performance is continually enhanced. Moreover, CRLNet achieves more accurate segmentation of unstructured objects like picnic-table. At the same time, CRLNet achieves more detailed segmentation results for structured objects like lamp posts (as shown in the second row of Fig. 10). Additionally, as depicted in the third and fourth rows of Fig. 10, CRLNet-heavy effectively addresses the segmentation errors caused by the complexity of outdoor terrain and similar class issues.

4.3. Ablation studies

We design several experiments to estimate and validate the effectiveness of different components in our proposed CRLNet.

4.3.1. Ablation study of ASPP

In the CRLNet network, ASPP is used to expand the network's receptive field and also for obtaining global spatial information. Then, CRLNet adds the global spatial information and local semantic information to improve the overall segmentation of the network [41, 42, 43]. As reported in Table 8, the segmentation results show that CRLNet achieved better segmentation performance compared to the CRLNet network without ASPP. By incorporating ASPP, CRLNet obtained gains of 1.25%, 1.12%, and 0.19% in mIoU, mAcc, and OA, respectively.

Table 8: Ablation study of ASPP. The best-performing data is highlighted in bold.

ASPP	mIoU % ↑	mAcc % ↑	OA ↑%
✗	40.82	48.67	91.54
✓	**42.07**	**49.79**	**91.73**

4.3.2. Ablation study of PFCAM

To verify the role of PFCAM in the CRLNet network, we conducted ablative experiments on the PFCAM module. The experimental results are shown in Table 9. It can be seen from Table 9 that the CRLNet network without the PFCAM module can achieve good segmentation results compared to the "Baseline" in Table 7, indicating that multi-stage feature learning can enhance the network's feature representation capability. Moreover, adding the PFCAM module to the CRLNet network significantly improves the performance in term of mIoU and mAcc, while there is a minimal decrease in OA. This demonstrates that the PFCAM module can improve the segmentation performance of the CRLNet network by further constraining local channel features while ensuring the overall segmentation performance

Table 9: Ablation study of PFCAM. The best-performing data is highlighted in bold.

PFCAM	mIoU % ↑	mAcc % ↑	OA %↑
✗	41.71	48.98	**91.75**
✓	**42.07**	**49.79**	91.73

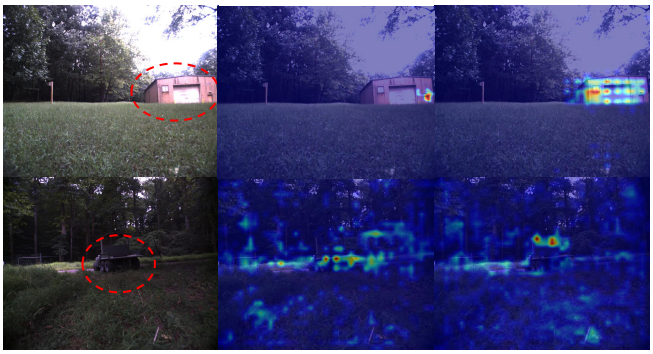

Image Feature w/o PFCAM Feature w/ PFCAM

Figure 11: The feature visualization heatmaps of PFCAM ablation experiments. we selected building and mobile car for analysis respectively. Please zoom in to view more details.

of the network. In addition, to visually validate the effectiveness of the FLM module, we have shown the feature visualization heatmaps of the PFCAM module ablation experiments in Fig. 11. From Fig. 11, we can see that using the PFCAM module proposed by us can effectively improve the feature extraction ability of CRLNet compared to not using the PFCAM module. This indicates that properly utilizing the PFCAM module to collect fine local channel information from various layers is beneficial for improving the model's semantic segmentation performance in complex outdoor environments.

4.3.3. Ablation study of FLM

To verify the contribution of FLM in the CRLNet network, we conducted ablative experiments on the FLM module. Firstly, we did not add the FLM module to the CRLNet network, and then we gradually increased different weight gating functions to validate the rationality of the FLM design. The segmentation results of the FLM ablative experiment are shown in Table 10, where "No FLM" refers to using a simple concatenation to connect different resolution images from the high and low paths. From Table 10, it can be observed that learning high-resolution feature maps from different paths using FLM significantly improves the unstructured segmentation performance. Specifically, using the $G(\cdot)$ weight gating function achieved gains of 0.77% mIoU, 0.21% mAcc, and 0.14% OA; using the $F(\cdot)$ weight gating function achieved gains of 1.1% mIoU, 0.48% mAcc, and 0.33% OA; finally, by jointly using the $G(\cdot)$ and $F(\cdot)$ gating functions to design the FLM module and incorporating it into the CRLNet network. It is evident from Table 10 that

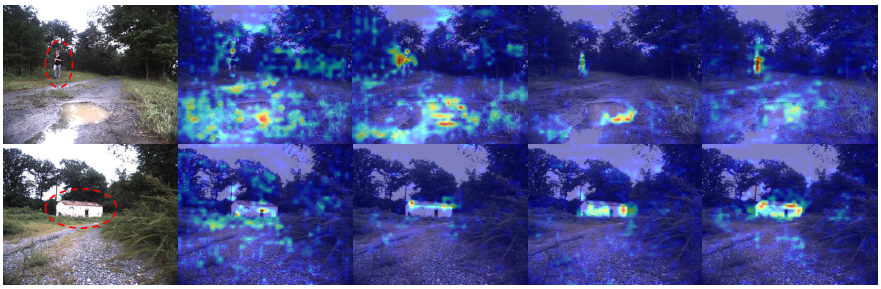

Figure 12: The feature visualization heatmaps of FLM ablation experiments. we selected people and suburban building for analysis respectively. Please zoom in to view more details.

Table 10: Ablation study of FLM. The best-performing data is highlighted in bold.

FLM	$G(\cdot)$	$F(\cdot)$	mIoU % ↑	mAcc % ↑	OA %↑
✗	✗	✗	40.86	48.68	91.54
✓	✓	✗	41.63	48.89	91.68
✓	✗	✓	41.96	49.16	**91.87**
✓	✓	✓	**42.07**	**49.79**	91.73

the CRLNet network with the FLM module can improve the overall segmentation accuracy compared to the CRLNet network without the FLM module, with gains of 1.21% mIoU, 1.11% mAcc, and 0.18% OA. In addition, to visually validate the effectiveness of the FLM module, we have shown the feature visualization heatmaps of the FLM module ablation experiments in Fig. 12. From Fig. 12, it can be observed that by using two different weight gating functions, the features can be gradually localized, thereby obtaining more complete feature information of outdoor objects. This effectively indicates that our proposed FLM module can efficiently learn important features from different branches, thus enhancing the network's representation ability for target features in natural environments.

5. Conclusion

To address the issue of rough segmentation of unstructured objects in complex natural environments, we propose a novel multi-path cascaded feature learning network called CRLNet in this paper. CRLNet utilizes gate weight functions to aggregate feature maps of different spatial resolutions and fine local channel information. Meanwhile, the multi-path local semantic embedding effectively extracts important feature information from different spatial resolutions. In each stage, it learns high spatial resolution features in the sub-high path and continuously embeds them into the low spatial resolution feature maps. The multi-stage network structure can be seen as a process of continuously refining features, and the information propagation between different stages is flexible. The feature maps generated at each stage have the same size (dimension) as the high spatial resolution feature maps in their respective paths, which enhances the discriminability of the feature maps. Extensive experiments have shown that the proposed CRLNet

effectively improves the performance of unstructured semantic segmentation in natural environments.

In this work, we have validated the generalizability of some advanced structured street scene parsing methods to the task of semantic segmentation in natural environments. We also hope that this work can further contribute to the research on unstructured semantic segmentation. In future work, we will attempt to use the proposed network to obtain semantic labels with different levels of safety driving, meeting the navigation requirements of rescue robots and other tasks in natural environments.

6. CRediT authorship contribution statement

Wei Li: Methodology, Writing–original draft. **Shishun Tian**: Supervision, Methodology, Writing–review & editing. **Guoguang Hua**: Methodology & editing. **Muxin Liao**: Methodology & editing. **Yuhang Zhang**: Methodology & editing. **Wenbin Zou**: Supervision, Methodology, Writing–review & editing.

7. Data availability

Data will be made available on request.

8. Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

9. Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under grants 62171294, 62101344, in part by the Natural Science Foundation of Guangdong Province, China under grants 2020A1515010959, in part by Natural Science Foundation of Shenzhen under grants JCYJ20200109105832261, JCYJ20190808122409660, in part by the Tencent "Phinoceros Birds" Scientific Research Foundation for Young Teachers of Shenzhen University and in part by the Interdisciplinary Innovation Team of Shenzhen University.

References

[1] Y. Xia, H. Wu, L. Zhu, W. Qi, S. Zhang, J. Zhu, A multi-sensor fusion framework with tight coupling for precise positioning and optimization, Signal Processing 217 (2024) 109343.

[2] D.-H. Lee, J.-L. Liu, End-to-end deep learning of lane detection and path prediction for real-time autonomous driving, Signal, Image and Video Processing 17 (2023) 199–205.

[3] M. Y. Arafat, M. M. Alam, S. Moh, Vision-based navigation techniques for unmanned aerial vehicles: Review and challenges, Drones 7 (2023) 89.

[4] Y. Jin, D. Han, H. Ko, Trseg: Transformer for semantic segmentation, Pattern recognition letters 148 (2021) 29–35.

[5] X. Zhang, Z. Zhao, L. Ran, Y. Xing, W. Wang, Z. Lan, H. Yin, H. He, Q. Liu, B. Zhang, et al., Fasticenet: A real-time and accurate semantic segmentation model for aerial remote sensing river ice image, Signal Processing (2023) 109150.

[6] Z. Song, Z. Zhang, F. Fang, Z. Fan, J. Lu, Deep semantic-aware remote sensing image deblurring, Signal Processing 211 (2023) 109108.

[7] Y. Liu, Y. Duan, T. Zeng, Learning multi-level structural information for small organ segmentation, Signal Processing 193 (2022) 108418.

[8] L. Zhang, Z. Yang, G. Zhou, C. Lu, A. Chen, Y. Ding, Y. Wang, L. Li, W. Cai, Mdmasnet: A dual-task interactive semi-supervised remote sensing image segmentation method, Signal Processing (2023) 109152.

[9] Y. Yang, T. Yan, X. Jiang, R. Xie, C. Li, T. Zhou, Mh-net: Model-data-driven hybrid-fusion network for medical image segmentation, Knowledge-Based Systems 248 (2022) 108795.

[10] X. Wu, L. Wang, C. Wu, C. Guo, H. Yan, Z. Qiao, Semantic segmentation of remote sensing images using multiway fusion network, Signal Processing 215 (2024) 109272.

[11] S. Yang, X. Zhang, Y. Chen, Y. Jiang, Q. Feng, L. Pu, F. Sun, Ucunet: A lightweight and precise medical image segmentation network based on efficient large kernel u-shaped convolutional module design, Knowledge-Based Systems 278 (2023) 110868.

[12] J. Xiong, L.-M. Po, W.-Y. Yu, C. Zhou, P. Xian, W. Ou, Csrnet: Cascaded selective resolution network for real-time semantic segmentation, Expert Systems with Applications 211 (2023) 118537.

[13] Y. Rao, J. Ni, H. Xie, Multi-semantic crf-based attention model for image forgery detection and localization, Signal Processing 183 (2021) 108051.

[14] H. Guo, X. Wu, W. Feng, Multi-stream deep networks for human action classification with sequential tensor decomposition, Signal Processing 140 (2017) 198–206.

[15] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems 30 (2017).

[16] M. Ding, Y. Zhou, Y. Chi, Self-attention generative adversarial network interpolating and denoising seismic signals simultaneously, Remote Sensing 16 (2024) 305.

[17] S. Liu, S. Tian, Y. Zhao, Q. Hu, B. Li, Y.-D. Zhang, Lg-dbnet: Local and global dual-branch network for sar image denoising, IEEE Transactions on Geoscience and Remote Sensing (2024).

[18] B. Zhang, J. Sun, F. Sun, F. Wang, B. Zhu, Image deblurring method based on self-attention and residual wavelet transform, Expert Systems with Applications 244 (2024) 123005.

[19] W. Huang, Y. Deng, S. Hui, Y. Wu, S. Zhou, J. Wang, Sparse self-attention transformer for image inpainting, Pattern Recognition 145 (2024) 109897.

[20] N. Lv, Z. Zhang, C. Li, J. Deng, T. Su, C. Chen, Y. Zhou, A hybrid-attention semantic segmentation network for remote sensing interpretation in land-use surveillance, International Journal of Machine Learning and Cybernetics 14 (2023) 395–406.

[21] D. Wang, S. Xiang, Y. Zhou, J. Mu, H. Zhou, R. Irampaye, Multiple-attention mechanism network for semantic segmentation, Sensors 22 (2022) 4477.

[22] B. Zhan, E. Song, H. Liu, Z. Gong, G. Ma, C.-C. Hung, Cfnet: A medical image segmentation method using the multi-view attention mechanism and adaptive fusion strategy, Biomedical Signal Processing and Control 79 (2023) 104112.

[23] X. Hu, P. Zhang, Q. Zhang, F. Yuan, Glsanet: Global-local self-attention network for remote sensing image semantic segmentation, IEEE Geoscience and Remote Sensing Letters (2023).

[24] Z. Wang, J. Wang, K. Yang, L. Wang, F. Su, X. Chen, Semantic segmentation of high-resolution remote sensing images based on a class feature attention mechanism fused with deeplabv3+, Computers & Geosciences 158 (2022) 104969.

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

[26] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, Q. Hu, Eca-net: Efficient channel attention for deep convolutional neural networks, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 11534–11542.

[27] P. Valsalan, C. P. Latha G, et al., Hyperspectral image classification model using squeeze and excitation network with deep learning, Computational Intelligence and Neuroscience 2022 (2022).

[28] R. Gao, Rethinking dilated convolution for real-time semantic segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 4674–4683.

[29] Y. Li, W. Chen, X. Huang, Z. Gao, S. Li, T. He, Y. Zhang, Mfvnet: a deep adaptive fusion network with multiple field-of-views for remote sensing image semantic segmentation, Science China Information Sciences 66 (2023) 140305.

[30] H. Yin, W. Xie, J. Zhang, Y. Zhang, W. Zhu, J. Gao, Y. Shao, Y. Li, Dual context network for real-time semantic segmentation, Machine Vision and Applications 34 (2023) 22.

[31] O. Frigo, L. Martin-Gaffe, C. Wacongne, Doodlenet: Double deeplab enhanced feature fusion for thermal-color semantic segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 3021–3029.

[32] X. Ma, X. Zhang, M.-O. Pun, M. Liu, Msfnet: Multi-stage fusion network for semantic segmentation of fine-resolution remote sensing data, in: IGARSS 2022-2022 IEEE International Geoscience and Remote Sensing Symposium, IEEE, 2022, pp. 2833–2836.

[33] R. Zhai, J. Zou, Y. He, L. Meng, Iagc: interactive attention graph convolution network for semantic segmentation of point clouds in building indoor environment, ISPRS International Journal of Geo-Information 11 (2022) 181.

[34] Y. Yu, X. Si, C. Hu, J. Zhang, A review of recurrent neural networks: Lstm cells and network architectures, Neural computation 31 (2019) 1235–1270.

[35] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[36] S. Xie, R. Girshick, P. Dollár, Z. Tu, K. He, Aggregated residual transformations for deep neural networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 1492–1500.

[37] H. Noh, S. Hong, B. Han, Learning deconvolution network for semantic segmentation, in: Proceedings of the IEEE international conference on computer vision, 2015, pp. 1520–1528.

[38] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, 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.

[39] O. Ronneberger, P. Fischer, 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.

[40] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, 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 40 (2017) 834–848.

[41] G. Gao, G. Xu, J. Li, Y. Yu, H. Lu, J. Yang, Fbsnet: A fast bilateral symmetrical network for real-time semantic segmentation, IEEE Transactions on Multimedia (2022).

[42] Q. Song, K. Mei, R. Huang, Attanet: Attention-augmented network for fast and accurate scene parsing, in: Proceedings of the AAAI conference on artificial intelligence, volume 35, 2021, pp. 2567–2575.

[43] M. A. Elhassan, C. Huang, C. Yang, T. L. Munea, Dsanet: Dilated spatial attention for real-time semantic segmentation in urban street scenes, Expert Systems with Applications 183 (2021) 115090.

[44] M. Wigness, S. Eum, J. G. Rogers, D. Han, H. Kwon, A rugd dataset for autonomous navigation and visual perception in unstructured outdoor environments, in: 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE, 2019, pp. 5000–5007.

[45] P. Jiang, P. Osteen, M. Wigness, S. Saripalli, Rellis-3d dataset: Data, benchmarks and analysis, in: 2021 IEEE international conference on robotics and automation (ICRA), IEEE, 2021, pp. 1110–1116.

[46] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, B. Schiele, The cityscapes dataset for semantic urban scene understanding, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 3213–3223.

[47] A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks, Communications of the ACM 60 (2017) 84–90.

[48] A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan, et al., Searching for mobilenetv3, in: Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 1314–1324.

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

[50] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang, et al., Deep high-resolution representation learning for visual recognition, IEEE transactions on pattern analysis and machine intelligence 43 (2020) 3349–3364.

[51] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, 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.

[52] G. Li, I. Yun, J. Kim, J. Kim, Dabnet: Depth-wise asymmetric bottleneck for real-time semantic segmentation, arXiv preprint arXiv:1907.11357 (2019).

[53] T. Wu, S. Tang, R. Zhang, J. Cao, Y. Zhang, Cgnet: A light-weight context guided network for semantic segmentation, IEEE Transactions on Image Processing 30 (2020) 1169–1179.

[54] M. Shi, J. Shen, Q. Yi, J. Weng, Z. Huang, A. Luo, Y. Zhou, Lmffnet: a well-balanced lightweight network for fast and accurate semantic segmentation, IEEE Transactions on Neural Networks and Learning Systems (2022).

[55] M. A. Elhassan, C. Yang, C. Huang, T. L. Munea, X. Hong, S2fpn: Scale-ware strip attention guided feature pyramid network for real-time semantic segmentation, arXiv preprint arXiv:2206.07298 (2022).

[56] M. Lu, Z. Chen, Q. J. Wu, N. Wang, X. Rong, X. Yan, Frnet: Factorized and regular blocks network for semantic segmentation in road scene, IEEE Transactions on Intelligent Transportation Systems 23 (2020) 3522–3530.

[57] F. Chollet, Xception: Deep learning with depthwise separable convolutions, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 1251–1258.

[58] S.-Y. Lo, H.-M. Hang, S.-W. Chan, J.-J. Lin, Efficient dense modules of asymmetric convolution for real-time semantic segmentation, in: Proceedings of the ACM Multimedia Asia, 2019, pp. 1–6.

[59] E. Romera, J. M. Alvarez, L. M. Bergasa, R. Arroyo, Erfnet: Efficient residual factorized convnet for real-time semantic segmentation, IEEE Transactions on Intelligent Transportation Systems 19 (2017) 263–272.

[60] H. Zhao, X. Qi, X. Shen, J. Shi, J. Jia, Icnet for real-time semantic segmentation on high-resolution images, in: Proceedings of the European conference on computer vision (ECCV), 2018, pp. 405–420.

[61] Y. Wang, Q. Zhou, J. Liu, J. Xiong, G. Gao, X. Wu, L. J. Latecki, Lednet: A lightweight encoder-decoder network for real-time semantic segmentation, in: 2019 IEEE international conference on image processing (ICIP), IEEE, 2019, pp. 1860–1864.

[62] W. Yu, Q. Zhou, J. Xiong, X. Wu, X. Jin, Esnet: An efficient symmetric network for real-time semantic segmentation, in: Pattern Recognition

and Computer Vision: Second Chinese Conference, PRCV 2019, Xi'an, China, November 8–11, 2019, Proceedings, Part II 2, Springer, 2019, pp. 41–52.

[63] Q. Zhou, Y. Wang, Y. Fan, X. Wu, S. Zhang, B. Kang, L. J. Latecki, Aglnet: Towards real-time semantic segmentation of self-driving images via attention-guided lightweight network, applied soft computing 96 (2020) 106682.

[64] Z. Yang, H. Yu, M. Feng, W. Sun, X. Lin, M. Sun, Z.-H. Mao, A. Mian, Small object augmentation of urban scenes for real-time semantic segmentation, IEEE Transactions on Image Processing 29 (2020) 5175–5190.

[65] J. Liu, Q. Zhou, Y. Qiang, B. Kang, X. Wu, B. Zheng, Fddwnet: a lightweight convolutional neural network for real-time semantic segmentation, in: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2020, pp. 2373–2377.

[66] H. Yin, W. Xie, J. Zhang, Y. Zhang, W. Zhu, J. Gao, Y. Shao, Y. Li, Dual context network for real-time semantic segmentation, Machine Vision and Applications 34 (2023) 22.

[67] X. Hu, L. Jing, U. Sehar, Joint pyramid attention network for real-time semantic segmentation of urban scenes, Applied Intelligence 52 (2022) 580–594.

[68] D. Luo, H. Kang, J. Long, J. Zhang, X. Liu, T. Quan, Gdn: Guided down-sampling network for real-time semantic segmentation, Neurocomputing 520 (2023) 205–215.

[69] J. Liu, X. Xu, Y. Shi, C. Deng, M. Shi, Relaxnet: Residual efficient learning and attention expected fusion network for real-time semantic segmentation, Neurocomputing 474 (2022) 115–127.

[70] X.-L. Zhang, B.-C. Du, Z.-C. Luo, K. Ma, Lightweight and efficient asymmetric network design for real-time semantic segmentation, Applied Intelligence 52 (2022) 564–579.

[71] Y. Zhang, T. Yao, Z. Qiu, T. Mei, Lightweight and progressively-scalable networks for semantic segmentation, International Journal of Computer Vision (2023) 1–19.

[72] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 618–626.