Computer Modeling in Engineering & Sciences

Tech Science Press

DOI: 10.32604/cmes.2023.024742

check for updates

ARTICLE

DuFNet: Dual Flow Network of Real-Time Semantic Segmentation for Unmanned Driving Application of Internet of Things

Tao Duan[1], Yue Liu[1], Jingze Li[1], Zhichao Lian[2,*] and Qianmu Li[2]

[1]School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, 210094, China

[2]School of Cyberspace Security, Nanjing University of Science and Technology, Wuxi, 320200, China

*Corresponding Author: Zhichao Lian. Email: lzcts@163.com

Received: 07 June 2022 Accepted: 14 September 2022

ABSTRACT

The application of unmanned driving in the Internet of Things is one of the concrete manifestations of the application of artificial intelligence technology. Image semantic segmentation can help the unmanned driving system by achieving road accessibility analysis. Semantic segmentation is also a challenging technology for image understanding and scene parsing. We focused on the challenging task of real-time semantic segmentation in this paper. In this paper, we proposed a novel fast architecture for real-time semantic segmentation named DuFNet. Starting from the existing work of Bilateral Segmentation Network (BiSeNet), DuFNet proposes a novel Semantic Information Flow (SIF) structure for context information and a novel Fringe Information Flow (FIF) structure for spatial information. We also proposed two kinds of SIF with cascaded and paralleled structures, respectively. The SIF encodes the input stage by stage in the ResNet18 backbone and provides context information for the feature fusion module. Features from previous stages usually contain rich low-level details but high-level semantics for later stages. The multiple convolutions embed in Parallel SIF aggregate the corresponding features among different stages and generate a powerful global context representation with less computational cost. The FIF consists of a pooling layer and an upsampling operator followed by projection convolution layer. The concise component provides more spatial details for the network. Compared with BiSeNet, our work achieved faster speed and comparable performance with 72.34% mIoU accuracy and 78 FPS on Cityscapes Dataset based on the ResNet18 backbone.

KEYWORDS

Real-time semantic segmentation; convolutional neural network; feature fusion; unmanned driving; fringe information flow

1 Introduction

The Internet of Things (IoT) has brought unprecedented convenience to people's lives and can realize real-time tracking of the location and status of "things" [1,2]. The application of artificial intelligence technology in the Internet of Things has achieved many results including smart transportation, smart home [3,4] and smart wear [5–7]. Unmanned driving is one of the hottest fields in intelligent transportation applications, and great development has been achieved by utilizing the computer vision algorithm.

As an important topic in computer vision, image semantic segmentation aims to assign each pixel of the image a pre-defined class label [8], which can help the unmanned driving system recognize different entities near the current road, and then guide corresponding driving choices to avoid traffic jams or traffic accidents [9,10]. In addition, in order to reduce latency, existing studies often combine IoT with edge computing [11]. Edge computing has emerged as a promising solution to address the limitations of cloud computing in supporting delay-sensitive and context-aware services in the IoT era [12–14]. This allows the unmanned driving system to focus on its surroundings while moving at high speeds. Apart from autonomous driving [15,16] mentioned above, semantic segmentation also has wide-ranging applications, such as scene parsing [17] and medical segmentation [18]. Many previous approaches focus on research high-quality semantic segmentation but do not pay enough attention to real-time image segmentation [19,20]. Especially in the unmanned driving scene [21], the development of urban roads and the increase in car speed put higher demand for the faster response for real-time segmentation algorithms. Thus, we are committed to improving the speed of the model without sacrificing segmentation accuracy.

Towards more accurate prediction, many approaches rely on novel architectures and strong modules. Certain well-known backbones have achieved impressive results as great prior feature representations, e.g., AlexNet [22], ImageNet [23,24], VGGNet [25] and ResNet [26]. ResNet is widely known due to Residual Block and its identity skip connections. The goal of the module is clear the current layer can learn features that are different from the information encoded from previous layers already.

As a matter of fact, the internal idea behind this kind of method is that good performance depends on the fusion of semantic features and spatial features (or boundary details) [27–29]. In the residual block, the deeper layers extract semantic information through bigger receptive fields and the lower layers retain the spatial details. In BiSeNet [30], the Feature Fusion Module fused features from the Context Path and Spatial Path which provide encoded context information and spatial information respectively. To realize this theory, many methods employ two- or multi-branch architecture. They build strong semantic representations based on a state-of-the-art backbone with deeper structure and generate a spatial prediction map which has rich detailed information with a lightweight network.

The Boundary-Aware Network (BANet) [31] proposed a two-stream framework to achieve semantic segmentation which is not completely separated based on BiSeNet's feature fusion module. The Dual Stream Segmentation Network (DSSNet) [32] introduced its attention module and pyramid pooling module based on BiSeNet. Inspired by BiSeNet, TB-Net [33] proposed a three-stream boundary aware network which changes the context path with a context-aware attention module and adds Boundary Stream to enhance the segmentation performance, particularly for thin and small objects.

In addition, a few approaches try to capture multi-scale context information at the same stage instead of different levels. They proposed pyramid pooling modules with atrous convolution [34,35] or pooling layers to capture different fine-grained information. The pyramid Pooling Module (PPM) proposed by the Pyramid Scene Parsing Network (PSPNet) [36] attempted to merge coarse features and fine features for better inference.

For the real-time semantic segmentation task, how to predict fast without sacrificing too much accuracy is the main research direction [37–39]. As previously stated, many works focus on purchasing better accuracy with complicated structures and heavyweight components that do not perform well in real-time segmentation tasks. Therefore, many methods use speedup strategies such as downsampling input size, shrinking network channels, and model compression [40]. On the one hand, these strategies

take advantage of the most critical factors such as image resolution, number of features, and parameters of the network to build the models. On the other hand, their drawbacks are clear they cannot maintain the efficiency of high-resolution images without scarifying accuracy.

BiSeNet proposed the fusion of semantic features and spatial features with two custom paths and many powerful modules for pursuing better performance under two evaluation metrics of accuracy and speed. We believe the theory of two-branch architecture based on semantic and spatial information benefits real-time segmentation task on high-quality prediction tasks. However, some structures can still be cropped for faster speed without losing too much accuracy. Based on Semantic Information Flow and Fringe Information Flow, we constructed a novel end-to-end trainable network called Dual Flow Network (DuFNet). In addition, we also introduced our cascaded Semantic Information Flow (SIF) with Global Context Module to aggregate semantic information and spatial information. Therefore, through the reasonable combination of multiple components, we designed a two-stream segmentation framework that balances the accuracy and the speed, and thus the proposed method is more suitable for the Unmanned driving Application of the Internet of Things because of the lower computational burden.

Our main contributions are listed as follows:

1. We proposed a powerful architecture, DuFNet, for real-time semantic segmentation task. We introduced a classical two-branch model with unique Semantic Information Flow and compact Fringe Information Flow to capture information of different receptive fields. Feature fusion module borrowed from BiSeNet aggregates these features to generate excellent performance.
2. We also designed parallel SIF and cascaded SIF with different components. The parallel structure makes the network suitable for inference efficient implementation with slight accuracy loss. The cascaded structure is benefit for speed.
3. We validated the performance of our DuFNet on the Cityscapes dataset with ResNet18 backbone for real-time semantic segmentation. Our method achieves 72.34% mIoU with the speed of 78 FPS at high resolution image with, outperforming existing state-of-the-art methods.

2 Related Work

Real-time semantic segmentation has been studied for many years and lots of methods based on deep learning networks have achieved state-of-the-art performance. There are mainly two patterns for real-time semantic segmentation with high-resolution semantic map prediction. The critical difference between them is whether the network is a single-pipeline structure or a multi-branch structure.

One of these solutions is the Encoder-Decoder structure [41] with a single-pipeline model that has been successfully applied to semantic segmentation task. The encoder part gradually extracts contextual information by layer-by-layer convolution and generates high-dimensional feature representation. The decoder part gradually recovers the spatial information. SegNet [42] is a clear example of this structure. The encoder stage of SegNet is composed of a set of pooling and convolution layers and the decoder is composed of a set of upsampling and convolution layers. Deeplabv3 [43], an encoder variant, employs the spatial pyramid pooling module with atrous convolution to encode rich semantic information. Multi-scale feature [44,45] ensembles achieved impressive results in terms of accuracy without speed advantage.

Another solution to semantic segmentation is the two- or multi-branch architecture. They construct different networks to adapt to different tasks on multi branches and combine all subresults generated by each branch to further improve the performance with acceptable cost [40,46–49].

This kind of strategy overcomes weakness that single-pipeline structure cannot take full advantage of information from original image. For real-time semantic segmentation, ICNet [40] proposed a novel three-branch model for real-time semantic segmentation based on in-depth analysis of time budget in common frameworks and extensive experiments. The strategy of feeding low-resolution images into the full CNN and feeding medium-and high-resolution images into light-weighted CNNs is very clever. This architecture not only utilizes semantic information in low resolution along with details from high resolution images efficiently but also achieves faster prediction even on high resolution images. ICNet avoids the insufficiency of intuitive speedup strategies including downsampling input size, shrinking the channels of the network and model compression.

Two-branch architecture is dedicated to alleviating the problem in encoder-decoder architecture because partial information is lost during the downsampling process. GCN [47] proposed a multiresolution architecture that jointly exploits fine and coarse features and the two branches with partially shared weights achieve better speed performance. Context Aggregation Network [49] proposed a dual branch convolutional neural network with significantly lower computational cost while maintaining competitive accuracy.

RGPNet [38] proposed encoder-decoder structures for four level outputs from the backbone, respectively. The decoder reconstructs the lost spatial information from multi stages with feature fusion. Inspired by RGPNet, we also propose a parallel structure for context information which extracts semantic representation on four stages of the backbone independently.

BiSeNet is also an excellent approach of the two-branch. It constructs Semantic Path based on the ResNet18 or Xception backbone and an Attention Refinement Module to obtain large receptive field. While the Semantic Path encodes rich semantic contextual information, the Spatial Path utilizes the lightweight model to provide spatial information. BiSeNet uses Feature Fusion Module (FFM) to fuse the two paths which are different in level of feature representation. We use the same FFM of BiSeNet in our network. Actually, the weaknesses of BiSeNet are obvious that the light-weighted model of Spatial Path still causes certain computation time budget in order to make the features reach a similar level. We try to take full advantage of features which are generated in the process of backbone and utilize easy operators to form the so-called Spatial Path for efficiency. The U-shape cascaded structure used in Context Path does not fully exploit the potential of the network model (like ResNet18 backbone). We also conduct experiments with our cascaded model for extracting context information to demonstrate our speedup strategies.

3 Approach

3.1 Structure of DuFNet

Inspired by the Bilateral Segmentation Architecture of BiSeNet, we propose our DuFNet with two separate paths: Semantic Information Flow (SIF) and Fringe Information Flow (FIF). It comprises two components: an encoder based ResNet18 which takes full advantage of multi-scale contexts in parallel and a tiny lightweight network that reconstructs lost detailed information. The encoder extracts high-level features and generates semantic information with different levels of abstraction at different stages in SIF. The lightweight network shares features from the encoder and estimates low-level spatial information in FIF.

As shown in Fig. 1, we illustrate our proposed DuFNet with many modules or layers to be detailed later. Furthermore, we elaborate on how the semantic information and fringe information contained in the features are calculated and transmitted in each module and layer. It is clear that DuFNet just keeps

Bassinet's Bilateral architecture and Feature Fusion Module but designs totally different patterns for aggregating multi-scale contexts in our bilateral structure.

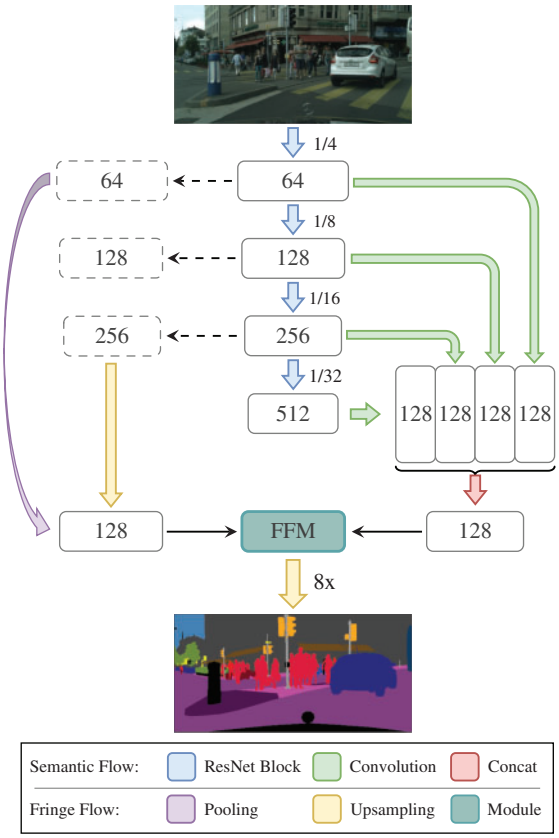


Figure 1: Overview of our proposed DuFNet. We first use CNN (ResNet18) to extract features of the input image stage by stage, then use multiple convolution layers to concatenate different fine-grained features in parallel. In this process, we aggregate information with different levels of abstraction and the concatenation layer merges all features. Secondly, the pooling layer and convolution layer with upsampling reconstruct spatial information flow from different stages of CNN. Finally, the two flows are fed into the feature fusion module to get a final prediction

3.2 Semantic Information Flow

Semantic Information Flow, the encoder, relies on changed ResNet18 as the backbone. It is divided into four stages named with numbers according to whether the resolution of features has changed. In the given diagram Fig. 2, features of four stages (stage1, stage2, stage3, and stage4) from ResNet18 backbone are extracted at different spatial resolutions 1/4, 1/8, 1/16, and 1/32 and with 64, 128, 256, and 512 channels mentioned in the scheme, respectively.

With the consideration of multi-scale context aggregation and computation demand simultaneously, we choose ResNet18 as an effective global prior representation for producing better performance and not just gaining accuracy. The residual blocks which address vanishing/exploding gradients also allow the network to easily enjoy the context information flowing through the model.

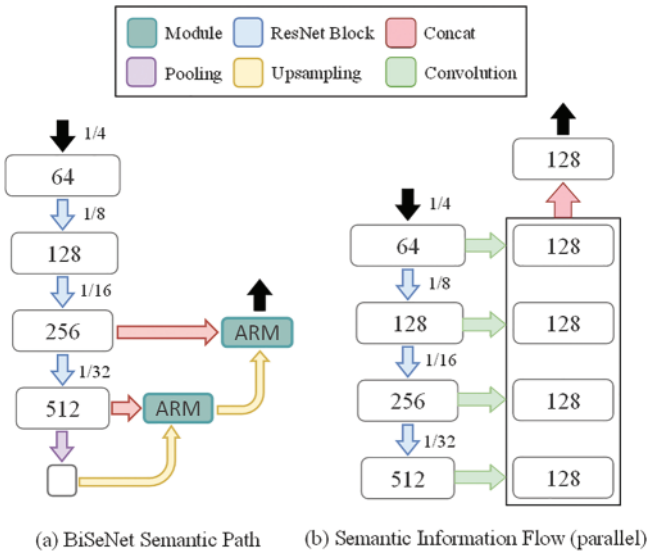


Figure 2: The schema of our Parallel Semantic Information Flow. Different from the original BiSeNet, we use paralleled structure and multiple convolution layers

In order to use context information generated at different stages, we propose a parallel convolution structure containing four convolutions with different receptive fields. For feature maps of stage1, we use a dilated convolution layer with kernel = 5, dilation = 2 and stride = 4 followed by batch normalization and ReLU to refine high-level and low-level information. Dilated convolution, a powerful feature extraction tool, can explicitly adjust field-of-view without extra parameters and computation burden. Feature maps extracted by stage 1 contain more spatial information and less semantic information due to the small receptive field. For stage2, we just use a normal convolution layer with kernel = 3 and stride = 2 to resize features. For features of stage3 and stage4, we use projection convolution with kernel size to project them so that they have the same number of channels, but we add an upsampling operation followed by the output of stage4.

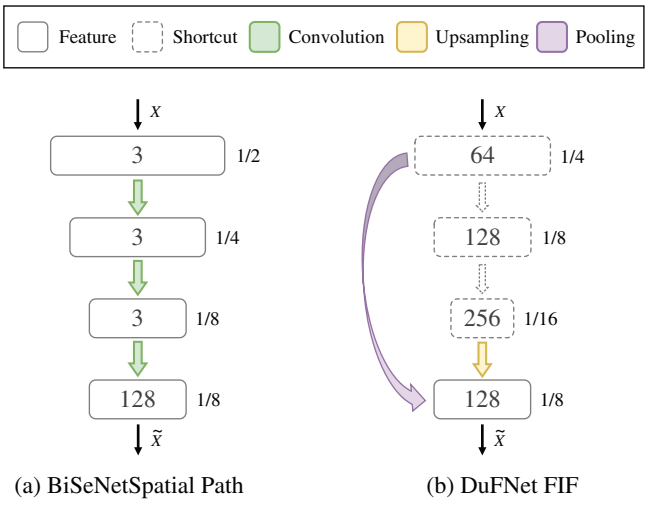
Given the different stages of the features, we transform them into feature maps with the same channels (e.g., 128) and the same spatial resolution, i.e., 1/8 of the original input. After that, we concatenate feature maps in all stages for aggregating different fine-grained context information. The concatenated features are fed into Feature Fusion Module to provide sufficient semantic information.

3.3 Fringe Information Flow

The goal of the Semantic Segmentation task is to split images into several regions or to confirm the boundaries of objects. Lots of methods encode rich contextual information into coarse prediction maps while missing object boundary details. However, extra spatial information or boundary details are crucial when the model retrieves resolution from prediction maps in the decode stage.

One of the most interesting approaches, BiSeNet, tries to use a lightweight network with three layers for preserving spatial information. It is no doubt that this idea produces better accuracy without the high computational overhead. However, it could be faster by using an operator with fewer parameters instead of multiple convolution layers.

Based on experiments, we propose the Fringe Information Flow architecture to preserve the boundary details and encode rich spatial information. Fig. 3 shows the details of this design. Different from BiSeNet's Spatial Path with three convolution layers, we use one convolution layer and an optional pooling layer to recover the missing spatial information in the process of downsampling. The pooling layer focuses on rich object boundary details in the features from stage1 of the backbone. The convolution layer regulates the channels of features after scaling up its spatial size. The FIF model just contains one convolution layer and one optional pooling layer. The feature maps in stage3 usually are rich in contextual information without losing many spatial details. Therefore, some details can be restored on these feature maps through some upsampling operations. We propose upsampling the feature maps of stage3 to produce a powerful representation whose size is 1/8 of the original image only. After that, we use a projection convolution with kernel = 1, to just squeeze the channels of feature maps to fit input requirements of Feature Fusion Module. The pooling layer works at feature maps of stage1, and the spatial size of output feature maps is also 1/8 of the original image. The meaning of the pooling feature of stage1 is to supplement the boundary details lost in the deeper stage.



(a) BiSeNetSpatial Path (b) DuFNet FIF

Figure 3: The schema of BiSeNet (a) and our Fringe Information Flow (b)

4 Experiments

We evaluate our proposed model on the publicly available Cityscapes dataset. In addition, we compare the performance with state-of-the-art works and perform ablation analysis for our DuFNet with BiSeNet under multiple evaluation metrics. Experimental results are as follows.

4.1 Datasets and Implementation Details

We introduce the public semantic segmentation datasets and show the details of our experimentation. Table 1 shows the settings of the training set, test set, retrieval set and label number of two data sets. During the training, the size of all images is uniformly changed to 224×224.

Table 1: Speed comparison of our proposed DuFNet against other state-of-the-art methods. "-" indicates that the methods did not give the corresponding speed results

Model	Base Model	Parameters	mIoU (%)	allAcc (%)	FPS
FCN-8 s	VGG16	-	65.3	-	2
ENet	no	-	58.3	-	76.9
SQ	SqueezeNet	-	59.8	-	16.7
ICNet	PSPNet50	-	69.5	-	30.3
DFANet A	Xception A	-	71.3	-	100
PSPNet	ResNet50	-	71.9	95.59	0.8
BiSeNet	ResNet18	101 MB	73.0	95.42	65.1
Ours	ResNet18	91 MB	72.3	95.22	77.6

4.1.1 Cityscapes

Cityscapes [50] is a large-scale urban street scene understanding dataset to train and test approaches for pixel-level, instance-level and panoptic semantic labeling task. It contains 2975 training and 500 validation images with high quality pixel-level annotations and involves 30 classes (e.g., road, person, car, wall and sky).

Implementation Details

Our architecture is started from the ResNet18 backbone and implemented with PyTorch [51]. We use poly strategy that learning rate varies according to $\text{base\_lr} \times \left(1 - \frac{\text{iter}}{\text{total\_iter}}\right)^{\text{power}}$ with $\text{base\_lr} = 1e-2$ and $\text{power} = 0.9$ in train process. The weight-decay and momentum are set as $5e-4$ and 0.9, respectively. For the structure, we make use of cross-entropy loss function. For Cityscapes dataset, we train our model with SGD optimizer with total 19 categories.

For data augmentation, the image is randomly scaled from 0.75 to 2.0 and randomly rotated between -10 to 10 degrees. Due to the cityscapes dataset having high quality images with a resolution of 1024×2048 , we crop the images into 768×1536 for train and valuation process.

The Cityscapes dataset is divided into 2975/500/1525 images for training, validation and testing respectively with 19 classes. Train and valuation are measured on a server with a single RTX 2080Ti. Limited by the performance of GPU card, the batch size is set as 8 to avoid Out of Memory (OOM) and some strategies which can make experiments efficiently, including multithreading training and Synchronized Multi-GPU Batch Normalization (SyncBN) are disabled.

Evaluation Metrics

For evaluation, we use mean intersection over union (mIoU), mIoU with no background (mIoU_no_back) and all pixel accuracy (allAcc) for accurate comparison. In addition, we still care for the speed and use frames per second (FPS) for speed comparison [52].

$$mAcc = \frac{1}{k+1}\sum_{i=0}^{k}\frac{p_{ii}}{p_{ii}+\sum_{j=0}^{k}p_{ji}} \tag{1}$$

$$allAcc = \frac{\sum_{i=0}^{k}p_{ii}}{\sum_{i=0}^{k}\sum_{j=0}^{k}p_{ji}} \tag{2}$$

where k represents the total number of predicted categories, P_{ii} denotes the number of pixels predicted as class i and the true class is also i, namely true positive. P_{ij} represents the number of pixels predicted as class i but actually class j, namely false positive, P_{ji} represents the number of pixels predicted as class j but actually class i, namely false negative.

4.2 Experiments

We compare our method with other state-of-the-art methods on Cityscapes dataset. We evaluate the accuracy and speed on NVIDIA GTX 1080Ti card with high resolution input (except DFANet A). Among them, all approaches are real-time semantic segmentation models except FCN-8 s and PSPNet which focus on high-quality image segmentation. Results are shown in Table 1 [53].

We choose the best version RN − S*P as the network for efficiency comparison with BiSeNet. For be fair, we use a common train strategy and report average speed from running through 5000 times with a single NVIDIA RTX 2080Ti card. All models are based on ResNet18 backbone with different image resolutions of {360 × 640, 512 × 1024, 720 × 1280, 1080 × 1920}.

As shown in Table 2, our model reaches better performance than BiSeNet at different image resolutions during speed test. DuFNet gets much faster inference speed 117.2 FPS with image resolution 720 × 1280 as input. With BiSeNet, our method yields 61.1 FPS while running at high resolution images with 1080 × 1920 and can gain about 43.8% speed improvement. It can be seen that the cascade structure of DuFNet is beneficial to improve the speed without losing the accuracy of the operation, making the network more suitable for inference efficient implementation and more streamlined to operate. As shown in Fig. 4, both methods achieve semantic segmentation, which are not different from the ground truth. But our method pays more attention to small targets, such as the car's logo. In addition, our method has better robustness compared to BiSeNet, with no incomplete regions.

Table 2: Speed comparison of our proposed DuFNet with different resolutions

Model	NVIDIA RTX 2080Ti			
	360 × 640	512 × 1024	720 × 1280	1080 × 1920
	FPS	FPS	FPS	FPS
BiSeNet	175.1	150.6	87.9	42.5
Ours	211.6	195.7	117.2	61.1

4.3 Ablation Analysis

In this section, we evaluate the performance of each component independently and compare the overall efficiency of our model in different situations with BiSeNet and other well-known real-time semantic segmentation models. We conduct the ablation analysis on Cityscapes dataset and train all models with the same strategy for fair.

Figure 4: Results of BiSeNet and our DuFNet on cityscapes validation dataset. (a) Image. (b) Ground Truth. (c) BiSeNet. (d) Ours

4.3.1 Ablation for FIF

The light-weighted network has been shown in famous works to be beneficial in providing rich spatial information with a low computation cost. In pursuit of faster performance, we propose a novel model with less computation and conduct experiments while preserving the Context Path of BiSeNet for comparison [54].

One form of FIF is combining features from stage1 of the backbone after pooling layer and stage3 of backbone after upsampling as illustrated in Fig. 3b, called RN − S* . As shown in Table 3, our FIF achieves 72.79% mIoU and 95.27% allAcc with the same setting. The accuracy of FIF drops about 0.2% in accuracy metrics, but it reaches 178.3 FPS with 27.7 absolute improvement compared with BiSeNet (150.6 FPS).

Table 3: Detailed performance comparison of Fringe Information Flow in our proposed DuFNet

Model	mIoU (%)	mIoU_noback (%)	allAcc (%)	FPS
BiSeNet	73.01	71.63	95.42	150.6
$RN - S$	72.88	71.50	95.17	183.4
$RN - S^*$	72.79	71.41	95.27	178.3

Other novel form of FIF is only using features from stage3 of backbone with upsampling and project convolution, called RN − S. As shown in Table 3, BiSeNet achieves 73.01% mIoU and 95.42% allAcc and 150.6 FPS. However, using our light FIF exceeds BiSeNet by 32.8 FPS and reaches 183.4 FPS. And the mIoU reaches 72.88% which is almost close to BiSeNet and only 0.13% drop. Compared with RN − S*, it yields fast inference because of the lack of computation produced by pooling layer.

The reason for using features of the backbone is twofold. First, the model does not need to use a lightweight network to extract features for spatial information, and this avoids a certain time budget. Second, the stage3 of the backbone can recover a certain extent of context and spatial information at the same time and rich object details for stage1.

4.3.2 Ablation for SIF

Convolution is a very flexible operator and can be easily deployed in any situation which needs to decrease or increase feature maps in spatial. Building on top of this idea, we propose our Parallel Semantic Information Flow with many types of convolutions including Atrous Convolution to yield stronger semantic representation. We concatenate all the feature maps from each stage of the ResNet18 after multi types of the convolution layer to generate global context representation. Some details are shown in Table 4.

Table 4: Detailed performance comparison of parallel

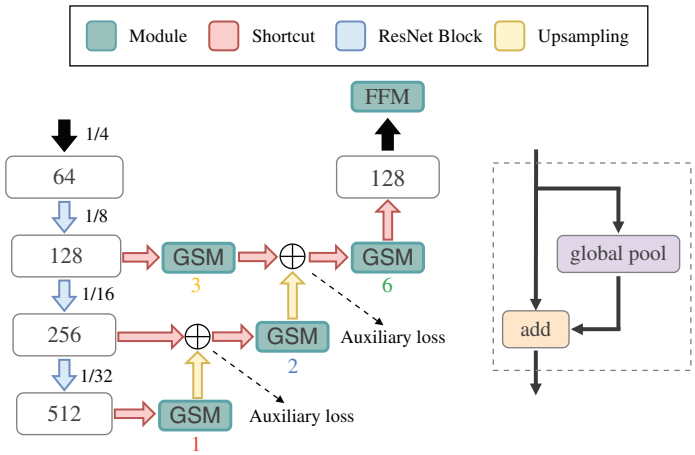
Model	mIoU (%)	mIoU_noback (%)	allAcc (%)	FPS
BiSeNet	73.01	71.63	95.42	150.6
$RN - C^{1236}$	71.62	70.17	95.02	145.8
$RN - C^{123}$	70.69	69.18	94.99	161.8
$RN - C^{124}$	72.39	70.98	95.05	161.5
$RN - C^{126}$	71.14	69.66	94.96	162.3

In order to independently evaluate our parallel SIF, we migrate it to BiSeNet and keep the original Spatial Path, called RN–P. Table 4 shows that our SIF can yield faster inference at the cost of sacrificing prediction accuracy and get score 164.9 in FPS. In addition, we also unite our FIF with parallel SIF to investigate the performance, called RN–SP and RN−S*P. As shown in Table 4, the accuracy of them is close. RN–SP achieves 71.36/95.00 in terms of mIoU and allAcc (%) and 189.6 in FPS which outperforming BiSeNet by 39 FPS. The most important is that the mIoU of RN−S*P drops to 72.34%, but the speed is still good enough to outperform the BiSeNet with an absolute improvement of 45.1 in FPS. Different from the parallel SIF mentioned above, we also proposed SIF with a cascaded structure. Fig. 5a shows the details of this design. In this part, we deploy U-shape structure to encode the features of different stages.

Global Semantic Module: Inspired by the context module of BiSeNet, we also utilize projection convolution and average pooling to construct our module for capturing more global context information [55]. As Fig. 5b shows, a global average pooling abstracts global context information by adopting varying-size pooling sizes of 1×1, 2×2, 3×3 and 6×6, respectively. After that, we recover the features automatically by upsampling on spatial size, and thus features can be added into input features by pixel-wise summation.

Global average pooling is a useful model as the global contextual prior, which is commonly applied in semantic segmentation tasks [36,56]. The strategy which we use to put pooling before convolution reduces the number of parameters and computational cost in our model. However, the most important significance is expected to extract more useful context information and makes a powerful global prior representation.

Inspired by the multi-scale parallel spatial pooling structure of PSPNet, we perform multiple cascade pooling layers with different downsampling sizes on the sub-decode stages. It is a classical U-shape structure with shortcut layer. For deeper feature maps, we use a GSM module with a smaller pooling size, and use a GSM module with larger pooling size for lower feature maps. For example, the pooling layer inside the GSM module marked with 1 will pool the feature of stage4 into a single bin feature map. And other pooling layers inside GSMs will generate different scale outputs as marked with different colors.



(a) Semantic Information Flow (Cascade) (b) Global Semantic Module

Figure 5: The schema of our Cascade Semantic Information Flow (a). Different from original BiSeNet, we use totally different structure and our custom Global Semantic Module (GSM). The GSM module contains only one global pooling layer which generates different sizes bin output of {1, 2, 3, 6}, as illustrated in part (b)

Auxiliary Branch: The auxiliary branches are added after the fusion of features that come from the adjacent stages. The auxiliary branches help to optimize the learning process at different levels. We still use BiSeNet's Segmentation Heads to deal with auxiliary branch outputs with tiny adjustments. We also add weights to balance the auxiliary loss for accuracy. In addition, we abandon all the auxiliary branches and use the master branch for the final prediction during the testing stage.

In the experiment, we also introduce the Cascade Semantic Information Flow (CSIF) to investigate the performance of the structure like U-shape. Where we use a lightweight model, ResNet18, as the backbone and keep the Spatial Path of BiSeNet for comparison. In addition, we propose the Global Semantic Module with four different factors (noted as G1, G2, G3 and G6) to combine the feature outputs from each stage in ResNet18 network, called $RN-C^{1236}$. The numbers represent the pooling size of pooling layer inside the GSM module. To further evaluate the performance of the component, we also conduct experiments with several settings, including removing G3 and replacing G6 with G3 or G4, called $RN-C^{126}$, $RN-C^{123}$ and $RN-C^{124}$, respectively. As listed in Table 5, $RN-C^{1236}$ works worse than BiSeNet on all metrics. Too many GSMs increase the computational burden and reduce the efficiency and parameters of the components are not set reasonably. However, $RN-C^{124}$ yields good result 161.49 in FPS with a little bit of accuracy loss. And compared to BiSeNet, it achieves 72.39%/95.05% in terms of mIoU and all Acc and is about 0.6% lower than BiSeNet.

Table 5: Detailed performance comparison of Cascade Semantic Information Flow in our proposed DuFNet

Model	mIoU (%)	mIoU_noback (%)	allAcc (%)	FPS
BiSeNet	73.01	71.63	95.42	150.6
$RN-S^{*}C^{1236}$	72.12	70.70	95.04	168.0
$RN-S^{*}C^{124}$	72.70	70.26	95.02	175.2

Based on the FIF we mentioned before, we also conducted experiments to verify the improvement of the unit. As listed in Table 6, RN−S*C[1236] yields 72.12/95.04 in terms of mIoU and allAcc and 168.0 in FPS. RN − S*C[124] has a better performance with 72.70/175.2 in terms of mIoU and FPS.

Table 6: Detailed performance comparison of Cascade SIF and FIF in our proposed DuFNet

Model	mIoU (%)	mIoU_noback (%)	allAcc (%)	FPS
BiSeNet	73.01	71.63	95.42	150.6
$RN - S^{*}C^{1236}$	72.12	70.70	95.04	168.0
$RN - S^{*}C^{124}$	72.70	70.26	95.02	175.2

The CSIF which is inspired by Pyramid Pooling Module of PSPNet [57] shows its efficiency on speed and also drawbacks of cascaded structure. Therefore, we propose a better Parallel Semantic Segmentation Flow.

4.4 Results

As our proposed architecture, RN−S*P, achieves comparable performance on multiple evaluation metrics. We evaluate the segmentation results on the server with a single RTX 2080Ti.

4.4.1 DuFNets vs. BiSeNet

Compared to the well-known real-time semantic segmentation model BiSeNet, DuFNet achieves absolute inference speed improvements with a slight sacrifice of accuracy. In Fig. 5, the accuracy of DuFNet is close to BiSeNet, which is shown in the object classification mission and the accurate prediction of boundary details. The quantitative results are summarized in Table 4.

In terms of inference speed, our DuFNets achieve a significant increase in speed with high-quality image segmentation results. Exemplary results are shown in Fig. 6.

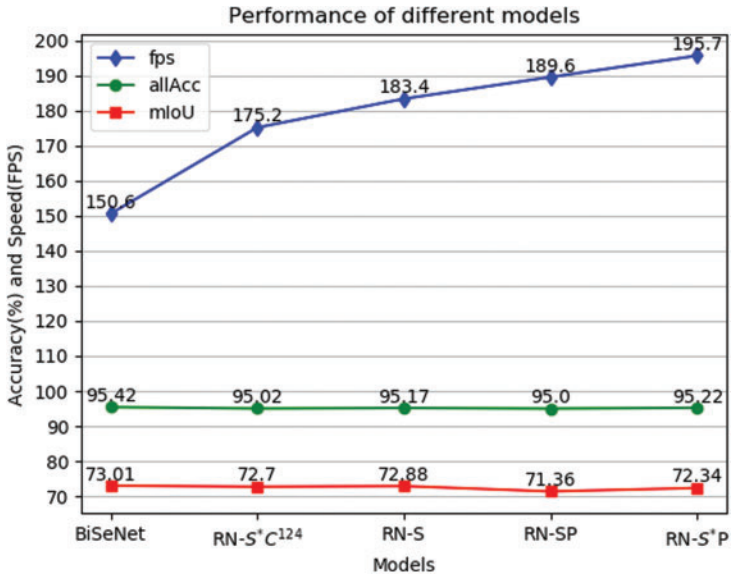


Figure 6: Accuracy and inference speed of BiSeNet and our DuFNets

5 Conclusion

We have proposed a novel architecture called DuFNet for real-time semantic segmentation tasks. The Fringe Information Flow takes advantage of the features of the backbone and reconstructs the spatial information with a light-weighted structure. The Cascade Semantic Information Flow enhances the quality of context encodings throughout its feature hierarchy with custom modules. The Parallel Semantic Information Flow enables the network to have better representational power by fusing spatial and context features more efficiently and contributes to the prior global representation generated by the feature fusion module. A wide range of experiments show the effectiveness of DuFNet, which achieves comparable performance on the Cityscape dataset. In the real world, autonomous driving and other applications still have a high demand for speed and accuracy of real-time semantic segmentation. The great trade-off between segmentation accuracy and inference speed will foster further research in this field.

Funding Statement: This work is supported in part by the National Key RD Program of China (2021YFF0602104–2, 2020YFB1804604), in part by the 2020 Industrial Internet Innovation and Development Project from Ministry of Industry and Information Technology of China, and in part by the Fundamental Research Fund for the Central Universities (30918012204, 30920041112).

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

1. Ren, J., Li, J., Liu, H., Qin, T. (2021). Task offloading strategy with emergency handling and blockchain security in SDN-empowered and fog-assisted healthcare IoT. *Tsinghua Science and Technology, 27(4),* 760–776. DOI 10.26599/TST.2021.9010046.
2. Xu, X., Li, H., Xu, W., Liu, Z., Yao, L. et al. (2021). Artificial intelligence for edge service optimization in Internet of vehicles: A survey. *Tsinghua Science and Technology, 27(2),* 270–287. DOI 10.26599/TST.2020.9010025.
3. Wei, D., Ning, H., Shi, F., Wan, Y., Xu, J. et al. (2021). Dataflow management in the Internet of Things: Sensing, control, and security. *Tsinghua Science and Technology, 26(6),* 918–930. DOI 10.26599/TST.2021.9010029.
4. Huo, Y., Fan, J., Wen, Y., Li, R. (2021). A cross-layer cooperative jamming scheme for social Internet of Things. *Tsinghua Science and Technology, 26(4),* 523–535. DOI 10.1109/TST.5971803.
5. Mabrouki, J., Azrour, M., Dhiba, D., Farhaoui, Y., El Hajjaji, S. (2021). IoT-based data logger for weather monitoring using arduino-based wireless sensor networks with remote graphical application and alerts. *Big Data Mining and Analytics, 4(1),* 25–32. DOI 10.26599/BDMA.2020.9020018.
6. Azrour, M., Mabrouki, J., Guezzaz, A., Farhaoui, Y. (2021). New enhanced authentication protocol for Internet of Things. *Big Data Mining and Analytics, 4(1),* 1–9. DOI 10.26599/BDMA.2020.9020010.
7. Zhong, G., Xiong, K., Zhong, Z., Ai, B. (2021). Internet of Things for high-speed railways. *Intelligent and Converged Networks, 2(2),* 115–132. DOI 10.23919/ICN.2021.0005.
8. Mottaghi, R., Chen, X., Liu, X., Cho, N. G., Lee, S. W. et al. (2014). The role of context for object detection and semantic segmentation in the wild. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 891–898. Columbus, OH, USA.
9. Malek, Y. N., Najib, M., Bakhouya, M., Essaaidi, M. (2021). Multivariate deep learning approach for electric vehicle speed forecasting. *Big Data Mining and Analytics, 4(1),* 56–64. DOI 10.26599/BDMA.2020.9020027.

10. Li, T., Li, C., Luo, J., Song, L. (2020). Wireless recommendations for Internet of vehicles: Recent advances, challenges, and opportunities. *Intelligent and Converged Networks, 1(1),* 1–17. DOI 10.23919/TUP-ICN.9195266.
11. Catlett, C., Beckman, P., Ferrier, N., Nusbaum, H., Papka, M. E. et al. (2020). Measuring cities with software-defined sensors. *Journal of Social Computing, 1(1),* 14–27. DOI 10.23919/JSCTUP.8964404.
12. Mabrouki, J., Azrour, M., Fattah, G., Dhiba, D., El Hajjaji, S. (2021). Intelligent monitoring system for biogas detection based on the Internet of Things: Mohammedia, Morocco city landfill case. *Big Data Mining and Analytics, 4(1),* 10–17. DOI 10.26599/BDMA.2020.9020017.
13. Su, Y. S., Ruan, Y., Sun, S., Chang, Y. T. (2020). A pattern recognition framework for detecting changes in Chinese Internet management system. *Journal of Social Computing, 1(1),* 28–39. DOI 10.23919/JSC.2020.0004.
14. Gao, X., Luo, J. D., Yang, K., Fu, X., Liu, L. et al. (2020). Predicting tie strength of Chinese Guanxi by using big data of social networks. *Journal of Social Computing, 1(1),* 40–52. DOI 10.23919/JSCTUP.8964404.
15. Büttner, S., Márton, Z. C., Hertkorn, K. (2016). Automatic scene parsing for generic object descriptions using shape primitives. *Robotics and Autonomous Systems, 76,* 93–112. DOI 10.1016/j.robot.2015.11.003.
16. Dong, J., Wu, W., Gao, Y., Wang, X., Si, P. (2020). Deep reinforcement learning based worker selection for distributed machine learning enhanced edge intelligence in internet of vehicles. *Intelligent and Converged Networks, 1(3),* 234–242. DOI 10.23919/TUP-ICN.9195266.
17. Treml, M., Arjona-Medina, J., Unterthiner, T., Durgesh, R., Friedmann, F. et al. (2016). Speeding up semantic segmentation for autonomous driving. *NIPS Workshop*, Barcelona, Spain.
18. Zhou, Z., Rahman Siddiquee, M. M., Tajbakhsh, N., Liang, J. (2018). UNet++: A nested U-Net architecture for medical image segmentation. In: *Deep learning in medical image analysis and multimodal learning for clinical decision support*, pp. 3–11. Cham: Springer.
19. Long, J., Shelhamer, E., Darrell, T. (2015). Fully convolutional networks for semantic segmentation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3431–3440. Boston, MA, USA.
20. Paszke, A., Chaurasia, A., Kim, S., Culurciello, E. (2016). ENet: A deep neural network architecture for real-time semantic segmentation. arXiv preprint arXiv:1606.02147.
21. Niu, Z., Shen, X. S., Zhang, Q., Tang, Y. (2020). Space-air-ground integrated vehicular network for connected and automated vehicles: Challenges and solutions. *Intelligent and Converged Networks, 1(2),* 142–169. DOI 10.23919/TUP-ICN.9195266.
22. Krizhevsky, A., Sutskever, I., Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. *Communications of the ACM, 60(6),* 84–90. DOI 10.1145/3065386.
23. Deng, J., Dong, W., Socher, R., Li, L. J., Li, K. et al. (2009). ImageNet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255. Miami Beach, FL, USA.
24. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S. et al. (2015). Imagenet large scale visual recognition challenge. *International Journal of Computer Vision, 115(3),* 211–252. DOI 10.1007/s11263-015-0816-y.
25. Simonyan, K., Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.
26. He, K., Zhang, X., Ren, S., Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778. Las Vegas, Nevada.
27. Wang, W., Zhou, T., Yu, F., Dai, J., Konukoglu, E. et al. (2021). Exploring cross-image pixel contrast for semantic segmentation. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7303–7313. Montreal, Canada.

28. Strudel, R., Garcia, R., Laptev, I., Schmid, C. (2021). Segmenter: Transformer for semantic segmentation. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7262–7272. Montreal, Canada.
29. Zhao, Y., Li, J., Zhang, Y., Tian, Y. (2019). Multi-class part parsing with joint boundary-semantic awareness. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9177–9186. Seoul, Korea (South).
30. Yu, C., Wang, J., Peng, C., Gao, C., Yu, G. et al. (2018). Bisenet: Bilateral segmentation network for real-time semantic segmentation. *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 325–341. Munich, Germany.
31. Chen, X., Qi, D., Shen, J. (2019). Boundary-aware network for fast and high-accuracy portrait segmentation. arXiv preprint arXiv:1901.03814.
32. Zhong, C., Hu, Z., Li, M., Li, H., Yang, X. et al. (2020). Dual stream segmentation network for real-time semantic segmentation. *2020 IEEE 5th International Conference on Image, Vision and Computing (ICIVC)*, pp. 144–149. Beijing, China, IEEE.
33. Zhang, Y., Li, Q., Zhao, X., Tan, M. (2021). TB-Net: A three-stream boundary-aware network for fine-grained pavement disease segmentation. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3655–3664. Waikoloa, Hawaii.
34. Yu, F., Koltun, V. (2015). Multi-scale context aggregation by dilated convolutions. arXiv preprint arXiv:1511.07122.
35. Chen, L. C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H. (2018). Encoder-decoder with atrous separable convolution for semantic image segmentation. *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 801–818. Munich, Germany.
36. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J. (2017). Pyramid scene parsing network. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2881–2890. Honolulu, HI, USA.
37. Zhuang, J., Yang, J., Gu, L., Dvornek, N. (2019). ShelfNet for fast semantic segmentation. *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, Seoul, Korea (South).
38. Arani, E., Marzban, S., Pata, A., Zonooz, B. (2021). RGPNet: A real-time general purpose semantic segmentation. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3009–3018. Waikoloa, Hawaii.
39. Yu, C., Gao, C., Wang, J., Yu, G., Shen, C. et al. (2021). BiSeNet V2: Bilateral network with guided aggregation for real-time semantic segmentation. *International Journal of Computer Vision, 129(11),* 3051–3068. DOI 10.1007/s11263-021-01515-2.
40. Zhao, H., Qi, X., Shen, X., Shi, J., Jia, J. (2018). ICNet for real-time semantic segmentation on high-resolution images. *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 405–420. Munich, Germany. DOI 10.1007/978-3-030-01219-9.
41. Noh, H., Hong, S., Han, B. (2015). Learning deconvolution network for semantic segmentation. *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1520–1528. Santiago, Chile.
42. Badrinarayanan, V., Kendall, A., Cipolla, R. (2017). SegNet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(12),* 2481–2495. DOI 10.1109/TPAMI.2016.2644615.
43. Chen, L. C., Papandreou, G., Schroff, F., Adam, H. (2017). Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587.
44. Hariharan, B., Arbeláez, P., Girshick, R., Malik, J. (2015). Hypercolumns for object segmentation and fine-grained localization. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 447–456. Boston, MA, USA.
45. Chen, L. C., Yang, Y., Wang, J., Xu, W., Yuille, A. L. (2016). Attention to scale: Scale-aware semantic image segmentation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3640–3649. Las Vegas, NV, USA.

46. Lin, G., Milan, A., Shen, C., Reid, I. (2017). RefineNet: Multi-path refinement networks for high-resolution semantic segmentation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1925–1934. Honolulu, HI, USA.
47. Mazzini, D. (2018). Guided upsampling network for real-time semantic segmentation. arXiv preprint arXiv:1807.07466.
48. Poudel, R. P., Bonde, U., Liwicki, S., Zach, C. (2018). ContextNet: Exploring context and detail for semantic segmentation in real-time. arXiv preprint arXiv:1805.04554.
49. Yang, M. Y., Kumaar, S., Lyu, Y., Nex, F. (2021). Real-time semantic segmentation with context aggregation network. *ISPRS Journal of Photogrammetry and Remote Sensing, 178,* 124–134. DOI 10.1016/j.isprsjprs.2021.06.006.
50. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M. et al. (2016). The cityscapes dataset for semantic urban scene understanding. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3213–3223. Las Vegas, NV, USA.
51. Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E. et al. (2017). Automatic differentiation in PyTorch. *NIPS Autodiff Workshop*, Long Beach, California, USA.
52. Li, H., Xiong, P., Fan, H., Sun, J. (2019). DFANet: Deep feature aggregation for real-time semantic segmentation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9522–9531. Long Beach, CA, USA.
53. Fan, M., Lai, S., Huang, J., Wei, X., Chai, Z. et al. (2021). Rethinking BiSeNet for real-time semantic segmentation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9716–9725.
54. Yuan, X., Shi, J., Gu, L. (2021). A review of deep learning methods for semantic segmentation of remote sensing imagery. *Expert Systems with Applications, 169,* 114417. DOI 10.1016/j.eswa.2020.114417.
55. Gao, G., Xu, G., Li, J., Yu, Y., Lu, H. et al. (2022). FBSNet: A fast bilateral symmetrical network for real-time semantic segmentation. *IEEE Transactions on Multimedia*. DOI 10.1109/TMM.2022.3157995.
56. He, K., Zhang, X., Ren, S., Sun, J. (2015). Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 37(9),* 1904–1916. DOI 10.1109/TPAMI.2015.2389824.
57. Pentland, A. (2020). Diversity of idea flows and economic growth. *Journal of Social Computing, 1(1),* 71–81. DOI 10.23919/JSCTUP.8964404.