Research Article

Honghuan Chen* and Xiaoke Lan

Real-time semantic segmentation based on BiSeNetV2 for wild road

https://doi.org/10.1515/jisys-2023-0205
received October 23, 2023; accepted November 28, 2023

Abstract: State-of-the-art segmentation models have shown great performance in structured road segmentation. However, these models are not suitable for the wild roads, which are highly unstructured. To tackle the problem of real-time semantic segmentation of wild roads, we propose a Multi-Information Concatenate Network based on BiSeNetV2 and construct a segmentation dataset Dalle Molle institute for artificial intelligence feature segmentation (IDSIAFS) based on Dalle Molle institute for artificial intelligence. The proposed model removes structural redundancy and optimizes the semantic branch based on BiSeNetV2. Moreover, the Dual-Path Semantic Inference Layer (TPSIL) reduces computation by designing the channel dimension of the semantic branch feature map and aggregates feature maps of different depths. Finally, the segmentation results are achieved by fusing both shallow detail information and deep semantic information. Experiments on the IDSIAFS dataset demonstrate that our proposed model achieves an 89.5% Intersection over Union. The comparative experiments on Cityscapes and India driving dataset benchmarks show that proposed model achieves good inference accuracy and faster inference speed.

Keywords: semantic segmentation, real-time, unstructured road, multi-information concatenation

MSC 2020: 68T45

1 Introduction

The development of big data and advancements in deep learning technology have catalyzed significant enhancements across various fields [1]. A prime example is in the realm of environmental monitoring [2,3], where Remote Sensing and Satellite Imagery have seen substantial progress [4,5], coupled with notable strides in satellite technology and applications [6,7]. These technological advancements find echoes in other industries, notably the automotive sector. Here, the evolution of autonomous driving technology underscores the growing importance of accurate and effective road detection. This capability is essential, not merely as an advanced feature but as a fundamental prerequisite for the practical engineering and application of autonomous vehicles. Roads can be broadly classified into two categories: structured roads with clear road markings, such as city streets, and unstructured roads without distinct boundary lines, such as those found in wilderness areas [8]. Compared to structured roads, the roads in wilderness areas are characterized by ambiguous road boundaries, complex and ever-changing road surfaces, and are highly susceptible to environmental interference as shown in Figure 1. Although state-of-the-art segmentation models for structured roads have shown remarkable performance, they are typically not directly applicable to the segmentation task of roads in the wilderness.

Traditional road detection techniques can be categorized into two main types: model-based detection methods and feature-based detection methods. Model-based road detection methods [9–13] are designed to

* **Corresponding author: Honghuan Chen,** College of Internet of Things Technology, Hangzhou Polytechnic, Hangzhou, 311402, China, e-mail: chh@mail.hzpt.edu.cn
Xiaoke Lan: College of Internet of Things Technology, Hangzhou Polytechnic, Hangzhou, 311402, China, e-mail: lxk@mail.hzpt.edu.cn

(a) (b)

Figure 1: (a) Structured road and (b) unstructured road.

handle different shapes of lane lines by creating various road image models. These methods work well for structured roads with clear lane lines or distinct road boundary features. However, they are not suitable for wild roads that lack clear boundary information. Feature-based road detection methods [14–16] rely on detecting gradient differences between road and non-road areas and the consistency of local image texture features like rut marks. These methods estimate the vanishing point of unstructured roads and use it, along with local texture information, to estimate an idealized triangular road shape. However, determining the vanishing point can be challenging, especially for sparsely traveled wild roads where the boundary between the blurred road region and non-road region may not be clearly defined, and there might be few vehicle driving traces. Previously published studies [17–19] use features such as color and texture and combine them with methods such as clustering or region growing to distinguish road and non-road areas. They can adapt well to the shape of the road surface, but are susceptible to interference from obstacles (e.g., pedestrians and vehicles) and noise on the road surface (e.g., puddles and shadows). Moreover, the aforementioned road recognition methods are constructed by human prior knowledge and often suffer from problems such as complicated model construction and weak generalization ability.

The full convolutional network (FCN) [20] is capable of autonomously summarizing picture information and learning deeper semantic features. It addresses the challenge of adapting models to complex and variable environments, which is often hindered by human knowledge limitations. Subsequently, several semantic segmentation models based on convolutional neural networks have introduced new network structures. For example, the U-Net architecture [21] and the Atrous Spatial Pyramid Pooling in DeepLabv3+ [22] have achieved excellent performance. Other FCN-based methods [23–28] are proposed to alleviate the weak generalization of traditional detection methods and reduce the reliance of models on human priori knowledge. Among them, SegNet [23] employs an encoding network and a symmetric decoding network, uses jump connection and pool indexing strategies to speed up the inference of the network, and finds a good balance between the number of parameters and accuracy. DeepLabv3 [24] uses ResNet [29,30] as the backbone and employs a Atrous Spatial Pyramid Pooling to obtain multi-scale context. Graph convolutional network [25] finds a balance between classification and localization problems from a pair of contradictions in semantic segmentation and alleviates the problem of large convolutional kernel with many parameters. Deep feature network [26] uses Smooth Network to solve the problem of intra-class inconsistency in the same class and uses Border Network to make the features learned by the model have stronger inter-class inconsistency. Another study [27] introduces a Context Contrasted Local Model to obtain multi-scale and multi-level context-contrasted local features and uses Gated Sum to selectively fuse multi-scale features at each location. Dynamic module net [28] uses Dynamic Convolutional Module to adaptively capture multi-scale information and eventually integrate it to obtain segmentation maps. With the improvement of accuracy, semantic segmentation network has been widely used in autonomous driving [31], medical image processing [32], defect detection [33], and other fields.

Although there are datasets available for structured road images, such as Cityscapes [34], Camvid [35], and India driving dataset (IDD) [36], there is currently a lack of datasets that contain unstructured wild road

images. While segmentation models trained on these structured datasets perform well in road detection due to their distinct structural features, they may not be suitable for wild road segmentation. To bridge this gap, we constructed Dalle Molle institute for artificial intelligence feature segmentation (IDSIAFS), a segmentation dataset based on Dalle Molle institute for artificial intelligence (IDSIA) [37]. IDSIAFS comprises a substantial collection of images featuring wild roads and rural trails. This dataset aims to provide the necessary data for training and evaluating segmentation models in the context of unstructured road scenes.

Furthermore, segmentation models designed for wild road detection often need to run on the mobile platforms, which requires higher inference efficiency and lower hardware resource overhead. One way to achieve real-time inference is to appropriately limit the input image's resolution, although this may result in some loss of boundary detail information. Additionally, some approaches, such as BiSeNetV1 [38] and SwiftNet [39], use lightweight backbone networks extracted from classification tasks as backbone networks for semantic segmentation. However, classification-based backbone networks may not always be suitable for participating in semantic segmentation tasks [40].

In this study, we propose a novel real-time semantic segmentation network Multi-Information Concatenate Network (MICNet) based on BisenetV2 for wild road segmentation. The proposed network has higher inference efficiency and lower number of parameters compared with the original BisenetV2 architecture. In MICNet, we introduce a new component called the Multi-Information Concatenate Module (MICM). This module is designed to extract deep semantic features while ensuring the fusion of shallow and deep information at the module level. Importantly, it achieves this fusion with reduced computational effort. This addresses the challenge of losing shallow detail information when extracting deeper semantic features. Additionally, we incorporate a Detail Guide Module (DGM), which guides the shallow learning of spatial detail information. It is important to note that this module consumes computational resources only during the training phase and is not involved in prediction, which helps maintain efficient inference. Extensive experiments are carried out to demonstrate the effectiveness of the proposed model on the IDSIAFS dataset. The results demonstrate that our model performs exceptionally well in wild road segmentation, achieving an impressive 89.5% Intersection over Union (IoU) with a rapid inference speed of 704.9 frames per second on an RTX3090 GPU. It also achieves good accuracy and fast inference speed on Cityscapes and IDD benchmarks, demonstrating its adaptability to both structured and unstructured road environments.

The contributions are summarized in the following aspects:

(1) A novel unstructured wild road-based segmentation dataset named IDSIAFS is developed to improve the accuracy of recognizing complex road scenes outside urban areas.
(2) A MICM is proposed to integrate multi-scale sensory fields and multi-depth semantic information effectively, which improves segmentation accuracy while reducing computational overhead.
(3) DGM is introduced to enhance the learning of spatial detail information in the shallow network, which in turn enhances segmentation performance without adding computational burden.

2 Related work

2.1 Real-time semantic segmentation

Methods such as high-resolution feature and complex network structures are usually used to improve segmentation accuracy. But deep convolutional neural networks with high-resolution feature or complex structures have the disadvantage of requiring large float operations, and their real-time performance is greatly challenged. Due to the limitation of hardware resources and the requirement for real-time performance, wild road segmentation models need to increase the inference speed while maintaining high segmentation accuracy. Therefore, it is crucial to improve the speed of semantic segmentation models on mobile devices.

DFANet [41] uses a lightweight backbone network to improve the prediction speed. The core idea of its sub-network aggregation is to optimize the front network with the back network. The goal of its sub-stage aggregation is to achieve the exchange of spatial and semantic information between different network layers

to alleviate the problem of information loss due to increasing network depth. Faste segmentation convolutional neural network (Fast-SCNN) [42] uses deeper branches to obtain global context for low-resolution features and applies shallower networks to learn detailed information for high-resolution features to reduce the computational effort of the network. ShelfNet [43] improves inference speed by reducing the number of channels, but its multiple codec branch pairs extract detailed information on different levels of the backbone network, which may lead to duplicate feature extraction and thus structural redundancy problems. ICNet [44] uses low-resolution branches to obtain semantic information and medium and high-resolution branches to recover and refine the coarse prediction results, respectively. The three branches are cascaded to achieve progressive fusion. BiSeNetV2 [45] uses two independent branches of different depths to obtain low-level detailed information and high-level semantic information, respectively, thus achieving a balance between inference speeds. However, the independence of the two branches at the shallow level not only results in a lack of communication between the detailed and semantic information, but also may generate unnecessary computational effort because the deep semantic information usually depends on the shallow detailed information.

2.2 Real-time semantic segmentation for the wild road

There have been some segmentation networks, such as LSPANet [46] and WFDCNet [47], that take real time into account, but most of them are built and trained based on open datasets containing structured roads and simple unstructured roads. Although they achieve good results, the performance of these models may not be optimal on complex wild roads. In this study, we build an IDSIAFS dataset mainly containing a large number of wild road images and use it as a benchmark to train a network that is better adapted to the wild road environment.

3 Network architecture

3.1 Overview

The wild road real-time semantic segmentation model MICNet proposed in this study adopts an encoder–decoder framework and is an improvement based on BiSeNetV2 [45], which consists of four main components: Information Sharing Layer (ISL), TPSIL, DGM, and Aggregation Layer (AL) [45]. The ISL and the TPSIL form the encoder, and the decoder consists of the AL.

The structure of MICNet is shown in Figure 2. First, the original image is input into ISL (a) to obtain shallow spatial information. Second, the TPSIL (b) converts the shallow detail information into deeper semantic information. Finally, the AL (d) is used to selectively aggregate the spatial and semantic information to obtain the final inference results. In addition, a DGM (c) is added at the end of the ISL, and used two identical auxiliary loss modules in the TPSIL to improve the model training process.

Compared with BiSeNetV2, the AL and CE modules are retained, and substantial changes are made to the overall network structure. The shallow spatial information and deep semantic information is incorporated, which improves the segmentation accuracy of wild roads without sacrificing low-level detail information. A MICM is also designed, which combines different scale receptive fields and multi-depth semantic information to improve accuracy while alleviating the waste of computational resources caused by the repeated computation of similar semantic information. The TPSIL includes both the CE and MICM modules.

3.2 ISL

The two branches of BiSeNetv2 have no information exchange in the shallow layer. Considering that the semantic information in the deep layer always comes from the spatial detail information in the shallow layer,

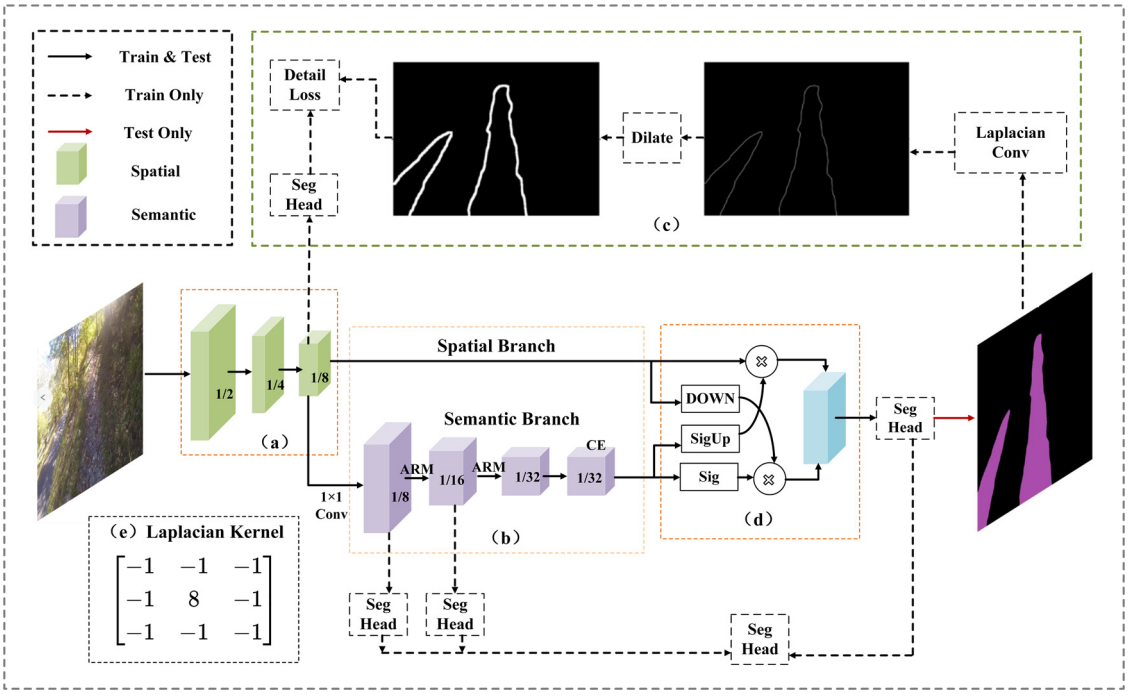

Figure 2: General overview of MICNet: (a) ISL, (b) TPSIL, (c) DGM, (d) AL – CE is the contextual embedding, and (e) Laplacian convolution kernel.

we do not adopt its completely independent two-way backbone network structure but only use its detail branch. This is equivalent to leaving the extraction of semantic information at the shallow level in the semantic branch to the ISL, which enables the exchange of semantic and spatial information at the shallow level of the network and can reduce the computational overhead of the model. The first half of the ISL as the encoder part mainly consists of 3 × 3 standard convolution, batch normalization, and rectified linear unit (ReLu). We divide the ISL into three stages (S1, S2, and S3) by adjusting the number of output channels of the convolution kernel, the step size, and the number of repetitions. The details are illustrated in Table 1.

Table 1: Structure of ISL

Stage	Opr	*k*	*c*	*s*	*r*	Output size
Input						3*480*640
S1	Conv2d	3	32	2	1	32*240*320
	Conv2d	3	32	1	1	32*240*320
S2	Conv2d	3	64	2	1	64*120*160
	Conv2d	3	64	1	2	64*120*160
S3	Conv2d	3	128	2	1	128*60*80
	Conv2d	3	128	1	2	128*60*80

S1, S2, and S3 constitute three stages, and each stage contains one or more operations. The parameters k, c, s, and r, respectively, represent the kernel size, output channels, stride, and repeat times.

3.3 TPSIL

The TPSIL, as the second half of the encoder part, mainly consists of the spatial branch and the semantic branch, respectively. Specifically, the spatial branch does not perform any operation on the put shallow spatial information, and the semantic branch is responsible for converting the shallow spatial information into more

advanced semantic information. In the semantic branch, we first use 1X1 convolution to convert the spatial information into shallow semantic information and compress the number of channels to reduce the computational effort. Subsequently, instead of using the GE [45], MICM and the Attention Refinement Module (ARM) in BiSeNetV1 [38] are used to iteratively extract the higher-level semantic information. Table 2 shows its detailed construction process. In order to reduce the complexity of the model structure, ARM is directly concatenated into semantic branches to make the model more focused on useful feature information, and it is found that this also yields better results. Finally, the semantic features are embedded with the contextual information containing the strongest semantics through global average pooling and residual concatenation in CE.

Table 2: Structure of the TPSIL

Branch	Opr	*k*	*c*	*s*	*r*	Output size
Semantic	Conv2d	1	32	1	1	32*60*80
	MICM	3	64	2	1	64*30*40
	ARM	3	64	1	1	64*30*40
	MICM	3	128	2	1	128*15*20
	MICM	3	128	1	2	128*15*20
	ARM	3	128	1	1	128*15*20
	CE	3	128	1	1	128*15*20
Spatial	—	—	—	—	—	128*60*80

Compared with the GE module of BiSeNetV2, the core design idea of MICM is to emphasize more on the fusion of multi-scale receptive field and multi-depth semantics and to reduce the computational overhead of deep semantic information due to the excessive number of channels.

To reduce the computational effort of MICM, we first use a 3 × 3 convolution to decrease the feature map's size rapidly. As shown in Figure 3, four paths are extended from different depths of the backbone, namely, Path1, Path2, Path3, and Path4. Except for the first two paths, the later paths have increasing perceptual field size as the backbone network deepens. Table 3 shows the perceptual field size of each path. The multi-scale

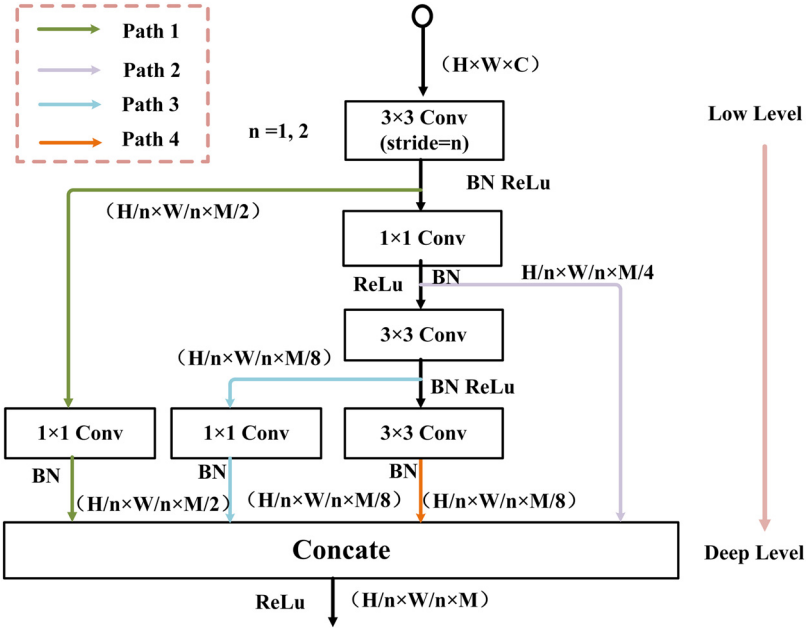

Figure 3: Description of the information splicing module.

Table 3: Receptive field of paths in MICM

MICM	Path 1	Path 2	Path 3	Path 4	Fusion
RF (S = 1)	3 × 3	3 × 3	5 × 5	7 × 7	3 × 3, 5 × 5, 7 × 7
RF (S = 2)	3 × 3	3 × 3	7 × 7	11 × 11	3 × 3, 7 × 7, 11 × 11

RF denotes the receptive field, and S denotes the stride.

perceptual field facilitates the model to learn the wild road features with different sizes, thus alleviating the problem of pixel-level classification errors due to the variation of road size. Moreover, as the path labeling increases, each path has more advanced semantic information due to its extension from different depths of the backbone. The fusion of multi-depth semantics by splicing all paths can alleviate the problem of shallow semantic information loss due to the deepening of the network within the module. In order to alleviate the problem of poor fusion performance due to possible excessive information differences in features of different depths, a 1 × 1 convolution at the end of each path except path4 is used to moderate the information differences. The shallow layer inside the module contains more semantic information, while the deep semantic information is often a generalization of the shallow semantic information, so assigning too many channels to the deeper network may result in a waste of computational resources. The deeper the network, the more detailed information is lost, and the deep network has too many channels without exchanging information with the shallow network, which may also cause the model to stick to learning useless detailed information that is not lost, thus affecting the training effect. In the channel number setting of MICM, the deeper the depth, the smaller the channel number, and the final output of MICM module is.

$$X_{\text{out}}^{M} = F\left(P_1^{\frac{M}{2}}, P_2^{\frac{M}{4}}, P_3^{\frac{M}{8}}, P_4^{\frac{M}{8}}\right) X_{\text{out}}^{M} = F\left(P_1^{\frac{M}{2}}, P_2^{\frac{M}{4}}, P_3^{\frac{M}{8}}, P_4^{\frac{M}{8}}\right), \tag{1}$$

where X_{out} denotes the output of MICM; M denotes the number of channels; F denotes the fusion operation; and P_1, P_2, P_3, and P_4 are the output feature maps of each path.

3.4 DGM and auxiliary loss

The DGM is introduced to guide the shallow layer (Stage 3) to learn more detailed information, and its specific structure is shown in Figure 2(c). It mainly consists of a 2D Laplacian convolution kernel (e) as well as a maximum pooling layer. First, we perform a Laplacian convolution operation (stirde = 1) on the semantic segmentation truth map to obtain the detailed feature map. Subsequently, the detailed feature map is dilated using a maximum pooling layer with a kernel size of 5 to obtain enhanced detailed information. Finally, the detailed information is transformed into a binary detail ground truth with rich road boundary information by using a threshold of 0.1.

Although a dilation operation is used to enrich the boundary information, the number of pixels containing road boundary information is still much less than the number of pixels at the non-road boundary, which leads to the class imbalance problem. In the study by Deng et al. [48], dice loss can be used to calculate the similarity between two samples and is applicable to the case of class imbalance. However, using dice loss alone can have a detrimental effect on backpropagation, which leads to unstable network training. For this reason, a combination of binary cross-entropy and dice loss is used to optimize the detailed learning. The final detailed loss formula is as follows:

$$L_{\text{detialloss}}(P, G) = L_{\text{dice}}(P, G) + L_{\text{bce}}(P, G), \tag{2}$$

where P denotes the prediction result, g denotes the detailed ground-truth, L_{dice} denotes the dice loss, and L_{bce} denotes the binary cross-entropy loss.

In this study, the Segment Head (Seg Head, in Figure 4) is set at different positions of the DGM and the semantic branch for training assistance. The Seg Head is also used for the final prediction acquisition. It contains 3 × 3 convolutions, batch normalization, ReLU, and a standard 1 × 1 convolution. All Seg Heads used for detail bootstrapping and auxiliary loss in the inference phase do not consume computational resources, and these parts are only useful in the training phase. In our approach, the Semantic Segmentation Head's C_m is set to 1024, which is used to obtain the segmentation results, and the other auxiliary prediction heads have their C_m set to 128.

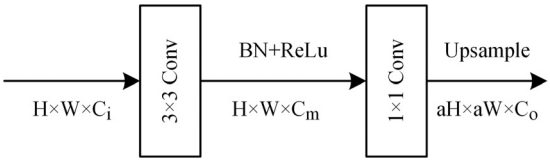

Figure 4: Structure of the Segment Head: a is used to control the size of the split header output.

4 Experimental results

4.1 Benchmarks

To enhance the generalization ability of the model, it is necessary to have a wild road dataset that accounts for diverse factors such as climate changes and variations in lighting conditions throughout the day. In this study, we have used three benchmark datasets, namely, IDSIA, Cityscapes, and IDD, as the foundational datasets.

The IDSIA dataset [37] is used as the foundation for our wild road semantic segmentation dataset. Although the original IDSIA was collected from forest trails in the Swiss Alps and accounted for seasonal and lighting factors during the collection process, it was primarily created for image classification tasks as shown in Figure 5. Therefore, we manually re-tuned the annotation process of the dataset for segmentation tasks using the semi-automated annotation tool PaddleSeg [49], resulting in the creation of the IDSIAFS (IDSIA For Segmentation) dataset. The images of IDSIA were captured together by three cameras in three different directions, allowing them to be divided into three classes (middle, left 30°, and right 30°) and used to train the classification networks that control the flight direction of the unmanned aerial vehicle along the road. However, when considering semantic segmentation as a pixel-level classification task requiring higher input image

Figure 5: Original samples of IDSIAF.

resolution, the actual images captured by the IDSIA (which reach up to 1,280 × 720) are often blurry due to the quality of the cameras. Directly inputting such images into the semantic segmentation network may cause a waste of computational resources and introduce unnecessary noise to the network, ultimately degrading training effectiveness. Therefore, the input image resolution was adjusted to 640 × 480 to remove unnecessary interference information during model training. Additionally, only images captured by the middle camera were used, and those with excessive light intensity or blur were removed. The annotation process of the dataset involved categorizing objects in each image into two categories (road areas and non-road areas).

To ensure that the IDSIAFS can also be used to train models for structured and normal unstructured roads, we do not exclude the images of non-extreme structured roads (asphalt, concrete, etc.) from the IDSIA, which account for about 30% of the entire dataset. Finally, the IDSIAFS is randomly divided into three parts: training, validation, and testing set. The number of images in the training set is 6,049, and the corresponding number in both the validation and testing sets is 981. Some manually annotated samples of IDSIAFS are shown in Figure 6.

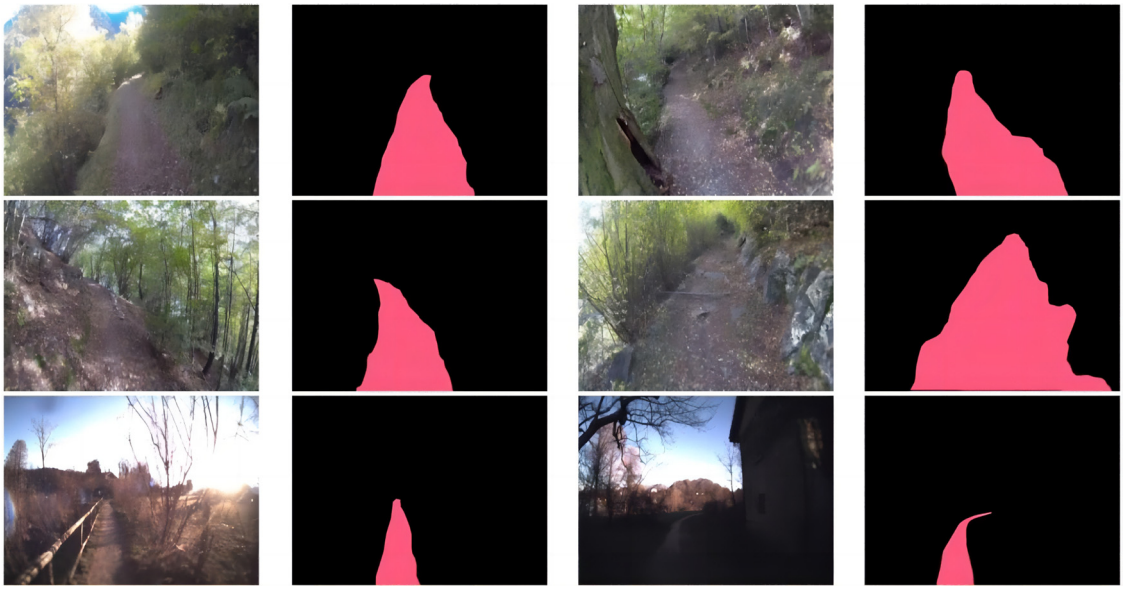

Figure 6: Manually annotated samples of IDSIAFS.

The Cityscapes dataset is a semantic parsing dataset for urban road scenes, which was captured from a car and comprises 2,975, 500, and 1,525 images in the training, validation, and testing sets, respectively. For the semantic segmentation task, annotations for 19 image classes are provided, but in this study, only the road class is used, with all other classes considered as non-road areas during training.

Compared with Cityscapes, IDD dataset contains more images of unstructured roads in urban and rural areas with less developed infrastructure. These roads often have blurred boundaries, but their pavement features are still relatively simple and obvious, and there are very few images of extreme unstructured roads. The IDD dataset has three classes for chosen, with higher class indicating a higher number of category labels. In this study, we use class 1, and all objects in the dataset are divided into seven classes. Six classes except the Drivable class are classified as off-road areas during training, while the Drivable class is classified as drivable areas. The dataset consists of 20,000 images and is divided into three sets. The number of images in the training, validation, and testing sets is 14,027, 2,036, and 4,038, respectively.

4.2 Implementation details

To ensure fairness, the same training code is used for different models on the same dataset. All models are initialized using "Kaiming Normal" [50] and trained from scratch. Stochastic gradient descent is employed for training with momentum of 0.9 and weight decay of 0.001. The batch size for IDSIAFS, IDD, and Cityscapes is set at 32, 24, and 16, respectively. A poly learning rate strategy is adopted, as shown in the following equation:

$$\mathrm{LR} = \mathrm{LR}_{\text{initial value}} \times \left(1 - \frac{\text{current iteration}}{\text{maximum iteration}}\right)^{\text{power}}, \tag{3}$$

where power is 0.9 and the initial learning rate is 0.02. The iterations for training IDSIAFS, IDD, and Cityscapes are set at 40k, 60k, and 100k, respectively.

Additionally, a warm-up strategy is adopted during the first 1,000 iterations. Data augmentation techniques, such as color jittering, random horizontal flip, random crop, and random scaling, are utilized. The resolution of the random crop is set at 1,024 × 512 for both Cityscapes and IDD datasets, and 640 × 480 for IDSIAFS. The random scaling sizes include 0.5, 0.75, 1.0, 1.25, 1.5, and 1.75.

4.3 Ablation study

To better showcase the advantages of the ISL, TPSIL, and detail guidance module in the proposed MICNet model, this study conducts three different ablation experiments to prove the effectiveness of each layer or module. All ablation experiments for each module are trained on a self-built road dataset and the training sets of Cityscapes and IDD.

In Section 3.2, we have described how the ISL merges shallow semantic branches with spatial branches. To demonstrate its effectiveness, ablation experiments are carried out based on BiSeNetV2, and the results are shown in Table 4. When ISL is used on IDSIAFS, it achieves a 0.2% improvement in accuracy for wild roads while reducing the number of parameters and increasing the inference speed by 101 FPS. Additionally, we conduct experiments on both Cityscapes and IDD datasets to verify the effectiveness of the ISL on other datasets. The experiments show that the proposed method is effective on IDSIAFS. Furthermore, the accuracy also improves by 0.4 and 0.2% on Cityscapes and IDD datasets, respectively.

Table 4: Ablation experiments of ISL based on BiSeNetV2

Dataset	Initial TIB	Down-sample ISL	IoU (%)	Params (M)	GFLOPs	FPS
IDSIAFS	✓	—	88.4	3.19	14.42	442
	—	✓	88.6	3.16	14.12	543
Cityscapes	✓	—	92.9	3.19	24.61	307
	—	✓	93.3	3.16	24.1	348
IDD	✓	—	95.1	3.19	24.61	307
	—	✓	95.3	3.16	24.1	348

TIB under initial down-sample indicates the use of two independent branches in BiSeNetV2, and ISL indicates the merging of shallow semantic branches onto spatial branches.

The IDD dataset has more unstructured roads but does not have extremely unstructured wild road images compared to IDSIAFS.

The ablation experiments of ISL have also been carried out based on the proposed MICNet, and the results are shown in Table 5. The segmentation accuracy of the proposed model is improved on both IDSIAFS and Cityscapes after using ISL. Although the same IoU results are obtained on the IDD dataset, the inference speed

Table 5: Ablation experiments of ISL based on MICNet

Data set	Initial down-sample		IoU (%)	Params (M)	FLOPs (G)	FPS
	TIB	ISL				
IDSIAFS	✓	—	88.7	2.63	11.46	604
	—	✓	89.3	2.61	11.27	704
Cityscapes	✓	—	92.6	2.63	19.55	414
	—	✓	93.1	2.61	19.24	463
IDD	✓	—	95.2	2.63	19.55	414
	—	✓	95.2	2.61	19.24	463

is improved by ISL, which also reflects the role of the ISL in facilitating information exchange and reducing computation in shallow networks.

To compare the effectiveness of our MICM to the GE module on the IDSIAFS dataset, we replace all MICMs in the proposed method with GE to perform comparative experiments. By comparing the results shown in the first row with the second row in Table 5, the inference accuracy decreases by 0.2%, the number of parameters increases by 0.4 M, and the inference speed decreases by 66.6 FPS after replacing MICM with GE. This indicates that the multi-scale sensory field fusion and the multi-depth semantic information splicing of MICM are beneficial for wild road segmentation. Meanwhile, reducing the number of channels of deep feature maps within the module does not significantly impact MICM but reduces the computational effort. The experimental results in the first and last rows of Table 6 demonstrate the effectiveness of ARM. ARM can achieve a 0.6% improvement in prediction accuracy with only a slight increase in computational overhead.

Table 6: Ablation experiments of semantic reasoning layer on IDSIAFS

GE	MICM	ARM	IoU (%)	Params (M)	GFLOPs	FPS
—	✓	✓	89.3	2.6	11.27	704
✓	—	✓	89.1	3	11.48	638
—	✓	—	88.7	2.4	11.19	726

GE denotes the Gather-and-Expansion Layer, MICM denotes the Multi-Information Concatenate Module, and ARM denotes the Attention Refinement Module.

The DGM is involved in the computation during the training phase, to guide the shallow network to learn more detailed information. In order to determine the size of the dilated convolution kernel that makes our model optimal, ablation experiments of the DGM are performed on the validation set of IDSIAFS by varying the kernel size. According to the experimental results as shown in Table 7, we finally determined the kernel size to be 5. Furthermore, ablation experiments of the proposed model with and without the DGM are also performed.

Table 7: DGM ablation experiments on IDSIAFS

DGM				IoU (%)	Params (M)	GFLOPs	FPS
No DGM	$K = 3$	$K = 5$	$K = 7$				
✓				88.4	2.6	11.27	704
	✓			88.9			
		✓		89.3			
			✓	88.9			

DGM denotes the Detail Guide Module and k denotes the kernel size of dilated convolution.

The experimental results in the first and third rows of Table 6 show that the DGM can improve the accuracy by 0.9% on the IDSIAFS dataset without increasing the computational resource overhead in the inference phase.

4.4 Compared with state-of-the-arts

As shown in Table 8, as MICNet is significantly improved on the basis of BiSeNetV2, its performance metrics have been improved. Specifically, the IoU accuracy of MICNet has increased by 0.9% on the validation set and 1.2% on the test set. In terms of parameter volume, MICNet has decreased the number of parameters by 0.6M. In terms of inference speed, MICNet has increased by 262 frames per second, i.e., an increase of 59.2% in speed. Compared with Fast-SCNN, BiSeNetV1, and STDC, the IoU accuracy of MICNet is 2.3, 0.7, and 0.7% higher on the validation set and 1.5, 0.8, and 1% higher on the test set, respectively. Although Fast-SCNN has the lowest number of parameters, too few parameters also lead to its worst performance among all networks. It is worth mentioning that BiSeNetV2 is an upgrade and improvement of BiSeNetV1, but its accuracy on wild roads in this study is lower than that of BiSeNetV1. This reflects that the improvement direction of the model is constrained by the dataset: a model that performs well on urban road data sets does not necessarily perform optimally on wild roads. To demonstrate the effectiveness of MICNet, the visual segmentation results of the proposed model and other benchmark models on IDSIAFS are illustrated. In Figure 7, it can be observed from the first row of

Table 8: Comparison with some state-of-the-arts on the IDSIAFS

Method	Backbone	IoU (%)		FPS	Params (M)	FLOPs (G)
		val	test			
BiSeNetV1 [38]	ResNet18	88.6	88.7	418	12.6	17.39
BiseNetV2 [45]	No	88.4	88.3	442	3.19	14.42
Fast-SCNN [42]	No	87.0	88.0	441	1.08	1.02
STDC [40]	STDC1	88.6	88.5	589	8.5	13.57
Ours	No	89.3	89.5	704	2.6	11.27

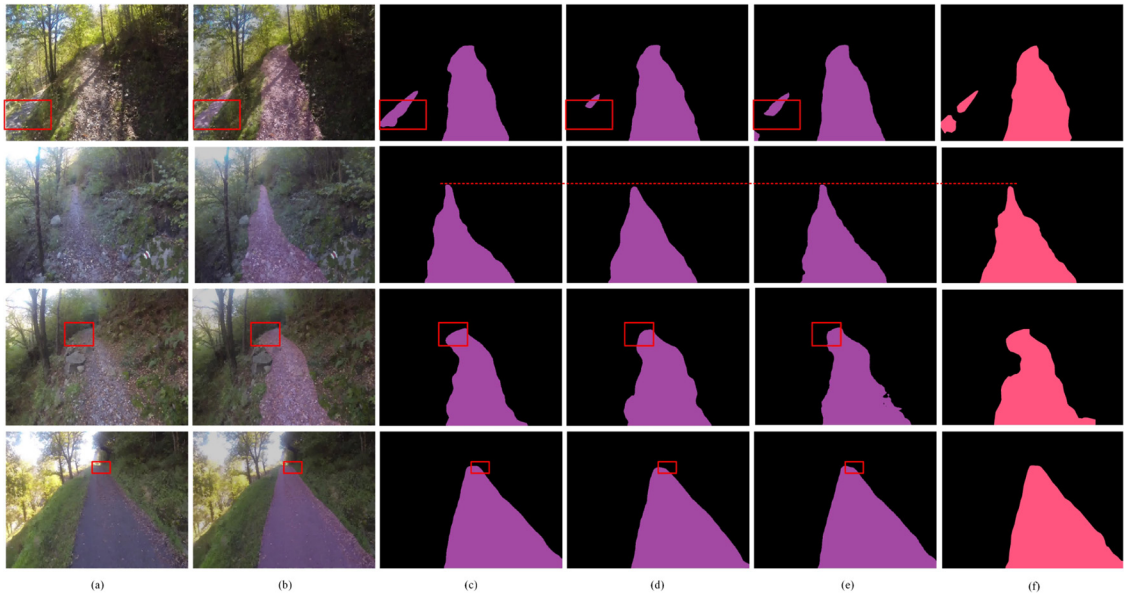

Figure 7: Visual segmentation results on IDSIAFS: (a) is the input image, (b) is the projection of the predicted output on the input image, (c) is the result of MICNet, (d) and (e) are the results of BiSeNetV2 and STDC, respectively, and (f) is the ground truth of the input image.

images that even though the roads in the ground truth are not fully labeled due to the roads being obscured by tree branches during manual labeling, the model is still able to have predictions for the obscured roads by learning information from the whole dataset. It shows that MICM can prevent the model from sticking to useless details. The remaining three rows of images show that our model is more sensitive to edge detailed information of the distant wild roads, and the segmentation results fit the distant road boundaries better.

To evaluate the performance of the proposed model on structured roads, the experiments of our model and some other state-of-the-arts are carried out on Cityscapes and IDD. Table 9 shows that the IoU accuracy and calculation time consumption of MICNet in Cityscapes are optimal. Compared with BiSeNetV2, although the IoU is only increased by 0.2%, the FPS improvement rate reaches 50.8%. Compared with all networks, the lightweight STDC's overall performance is medium. Although Fast-SCNN has the lowest number of parameters, it has the most insufficient segmentation accuracy compared with other models and even a lower inference speed than MICNet. This shows that pursuing low parameter quantities without considering the rational design of the model structure will lead to poor performance on key performance indicators.

Table 9: Comparison with some state-of-the-arts on Cityscapes and IDD datasets with resolution 512 × 1,024

Data set	Method	Backbone	IoU (%)	FPS
Cityscapes	BiSeNetV1 [38]	ResNet18	91.9	280
	BiSeNetV2 [45]	No	92.9	307
	Fast-SCNN [42]	No	85.7	450
	STDC [40]	No	92.2	408
	Ours	No	93.1	463
IDD	BiSeNetV1 [38]	ResNet18	95.5	280
	BiSeNetV2 [45]	No	95.1	307
	Fast-SCNN [42]	No	94.1	450
	STDC [40]	No	95.3	408
	Ours	No	95.2	463

On the IDD dataset, although MICNet is 0.3% worse than the BiSeNetV1 in terms of IoU, the inference speed is increased by 65.3%. The maximum difference in IoU accuracy for all models except Fast-SCNN is only 0.4%. This is because the IDD dataset is collected from Indian streets with more unstructured roads, but its pavement features are simpler and more uniform than Cityscapes and the wild road dataset in this study. The data are easy to distinguish, which leads to the fact that all models are almost optimal during the training process. Therefore, it is difficult to further improve the segmentation accuracy, even if the models are well designed. In general, the higher the resolution of the image input to the model, the more time-consuming its computation will be. However, comparing with Table 8, it can be observed that regardless of the resolution of the input images, the inference speed of Fast-SCNN is almost the same. This is because although the frequent use of depthwise separable convolution can reduce the number of parameters, its structural characteristics will occupy a large amount of memory bandwidth.

4.5 Deployment experiments

To further verify the practicality of IDSIAFS and MICNet, we deployed MICNet trained on IDSIAFS directly to the Jetson Xavier NX, then tested it on a wild road, and further converted the segmentation result into the directional angle that could guide the car to move along the center-line of the road. In order to ensure stable operation of the Nvidia. Jetson Xavier NX in outdoor environments, it was mounted on a modified robot, as shown in Figure 8.

After acquiring the real-time image, the camera first uses the distortion correction method to correct the barrel distortion in the image, extracts the centerline of the road from the segmentation result obtained by

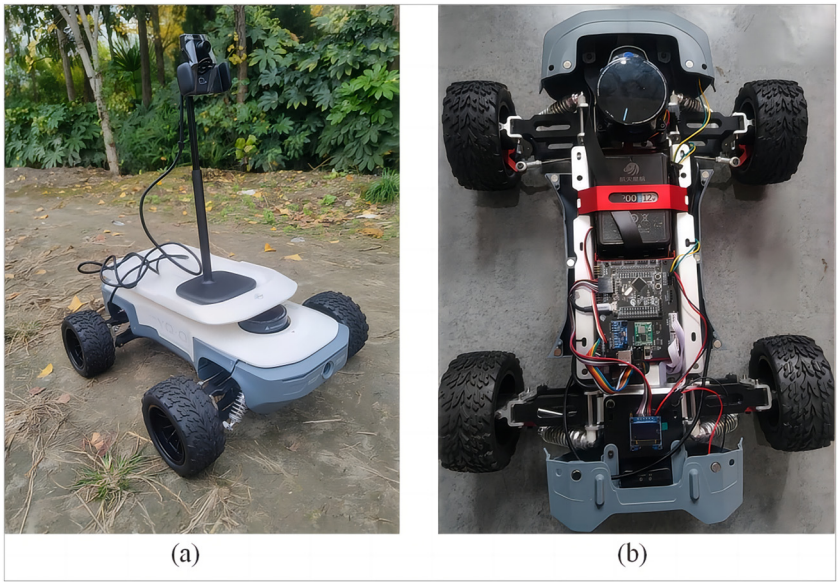

Figure 8: Outdoor testing robot: (a) modified robot and (b) Jetson Xavier NX-mounted module.

predicting the corrected image, and selects the appropriate pixel points on the centerline as the target points for the car to move forward. Then, through perspective transformation, the pixel coordinates of the target point in the image are converted into coordinate points with high accuracy under the overhead view and finally converted into the directional angle required for the car to move forward. The reason for using

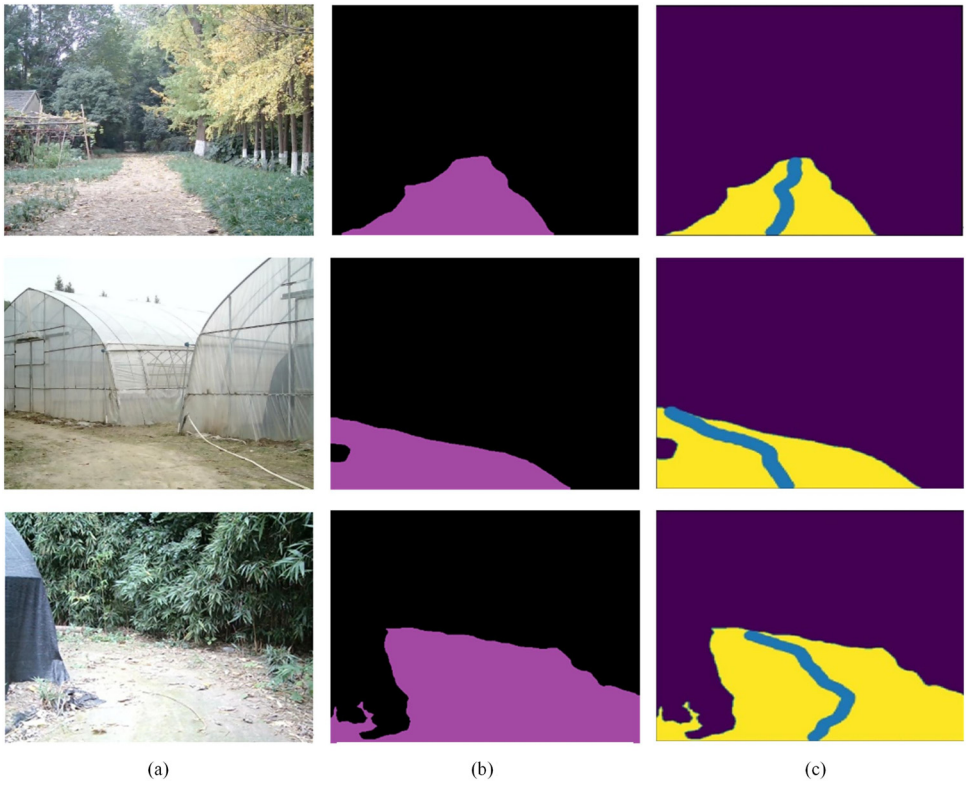

Figure 9: Visualization of the calculation results: (a) shows some examples of the results of the wild road image after camera distortion correction, (b) shows the wild road segmentation results, and (c) shows the road centerlines extracted from the segmented wild road.

perspective transformation is that because the number of pixel points occupied by objects close to the camera is much larger than that of objects far away from the camera view, there will be a significant error when the car's driving direction angle is obtained directly from the pixel coordinates. Converting the pixel coordinates to the coordinates under the camera's overhead view through perspective transformation can improve the accuracy of the directional angle calculation. The perspective transformation formula is as follows:

$$\begin{bmatrix} x'' \\ y'' \\ 1 \end{bmatrix} = \frac{1}{b_{31}x + b_{32}y + 1} \begin{bmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b\mathrm{b}_{23} \\ b_{31} & b_{32} & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = \frac{1}{b_{31}x + b_{32}y + 1} H \begin{bmatrix} x \\ y \\ 1 \end{bmatrix}, \tag{4}$$

where H denotes the perspective transformation matrix, (x'', y'') denotes the pixel coordinates, and (x'', y'') denotes the coordinates under the overhead view.

On the Jetson Xavier NX, it takes 0.04 s to calculate the directional angle for each picture frame, but it only takes 0.015 s for the model to predict each frame. This shows that the method in this study can meet the real-time requirements in the embedded system. Figure 9 shows the visualization of part of the calculation process of calculating the directional angle on the actual wild road. The segmentation results of the model on the actual wild road in Figure 9(b) demonstrate two aspects: the MICNet for wild road segmentation in this study has practical application ability, and the model trained based on the IDSIAFS also has good generalization ability in unfamiliar wild environments. In the third row of images, the model appears to be mis-segmented because the black shed in the lower left corner use dirt for fixation. However, in the later stage of extracting the road centerline, the method of taking the midpoint of the segment with the most consecutive pixels belonging to the road class as the road centerline point in a row can well compensate for the abnormal calculation of the deflection angle caused by this type of mis-segmentation. The robustness of the overall algorithm is further improved.

5 Conclusion

In this study, we build the dataset IDSIAFS and propose the network MICNet for wild road segmentation. In the proposed method, we remove the redundant part of BiSeNetV2 to build an ISL and propose a Muti-Information Concatenate Module to obtain multi-scale perceptions of the wild road and alleviate the problem of information loss in the shallow layer as the network deepens. We also add a DGM to enable the shallow network to learn more detailed information. Extensive experiments demonstrate that the proposed network achieves excellent results on IDSIAFS. Furthermore, the proposed network has good segmentation capability for structured roads on Cityscapes and IDD as well. Finally, we have carried out actual deployment on the embedded system, which proves that the IDS IAFS and MICNet built in this study are practical.

This study has achieved results in wild road segmentation but faces certain limitations. First, the IDSIAFS dataset may not completely cover the diverse and variable unstructured road conditions worldwide. Second, the performance of MICNet, especially under extreme conditions, remains to be validated in real-world scenarios. Additionally, the dependence on high-end hardware limits its application in resource-constrained environments. Future research could focus on expanding the dataset, optimizing the network structure to suit different hardware resources, and conducting long-term testing in more varied real-world environments. These efforts will enhance the network's generalizability, better meeting the needs of rapidly evolving autonomous driving technologies.

Acknowledgement: The author would like to thank the editor and anonymous reviewers for their contributions toward improving the quality of this manuscript.

Funding information: This work was supported by the Scientific Research Fund of Zhejiang Provincial Education Department of China (Grant Number: Y202146001).

Author contributions: Honghuan Chen: Responsible for Conceptualization, Methodology, and Writing–review & editing. Xiaoke Lan: Responsible for Visualization and Writing–original draft.

Conflict of interest: The authors declare that there is no conflict of interests.

Data availability statement: All data generated or analyzed during this study are included in this published article.

References

[1] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521(7553):436–44.
[2] Sharifi A. Flood mapping using relevance vector machine and SAR data: A case study from Aqqala, Iran. J Indian Soc Remote Sens. 2020;48(9):1289–96.
[3] Sharifi A, Amini J, Tateishi R. Estimation of forest biomass using multivariate relevance vector regression. Photogramm Eng Remote Sens. 2016;82(1):41–9.
[4] Ghaderizadeh S, Abbasi-Moghadam D, Sharifi A, Tariq A, Qin S. Multiscale dual-branch residual spectral–spatial network with attention for hyperspectral image classification. IEEE J Sel Top Appl Earth Obs Remote Sens. 2022;15:5455–67.
[5] Esmaeili M, Abbasi-Moghadam D, Sharifi A, Tariq A, Li Q. Hyperspectral image band selection based on CNN embedded GA (CNNeGA). IEEE J Sel Top Appl Earth Obs Remote Sens. 2023;16:1927–50.
[6] Kosari A, Sharifi A, Ahmadi A, Khoshsima M. Remote sensing satellite's attitude control system: rapid performance sizing for passive scan imaging mode. Aircr Eng Aerosp Technol. 2020;92(7):1073–83.
[7] Sharifi A, Amini J, Sri Sumantyo JT, Tateishi R. Speckle reduction of PolSAR images in forest regions using fast ICA algorithm. J Indian Soc Remote Sens. 2015;43:339–46.
[8] Huang JG, Kong B, Li BC, Zheng F. A new method of unstructured road detection based on hsv color space and road features. In 2007 International Conference on Information Acquisition. IEEE; 2007. p. 596–601.
[9] Wang Y, Teoh EK, Shen D. Lane detection and tracking using b-snake. Image Vis Comput. 2004;22(4):269–80.
[10] Tan C, Hong T, Chang T, Shneier M. Color model-based real-time learning for road following. In 2006 IEEE Intelligent Transportation Systems Conference. IEEE; 2006. p. 939–44.
[11] Alvarez JM, Lopez A, Baldrich R. Illuminant-invariant model-based road segmentation. In 2008 IEEE Intelligent Vehicles Symposium. IEEE; 2008. p. 1175–80.
[12] Li Z, Dai B, He HG. A novel fast segmentation method of unstructured roads. In 2006 IEEE International Conference on Vehicular Electronics and Safety. IEEE; 2006. p. 53–6.
[13] Alvarez JM, Gevers T, Diego F, Lopez AM. Road geometry classification by adaptive shape models. IEEE Trans Intell Transp Syst. 2012;14(1):459–68.
[14] Rasmussen C. Grouping dominant orientations for ill-structured road following. In Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004. Vol. 1. IEEE; 2004. p. I.
[15] Chang CK, Siagian C, Itti L. Mobile robot monocular vision navigation based on road region and boundary estimation. In 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems. IEEE; 2012. p. 1043–50.
[16] Kong H, Audibert JY, Ponce J. General road detection from a single image. IEEE Trans Image Process. 2010;19(8):2211–20.
[17] Hu MH, Yang WJ, Ren MW, Yang JY. A vision based road detection algorithm. In IEEE Conference on Robotics, Automation and Mechatronics. Vol. 2. IEEE; 2004. p. 846–50.
[18] Wang YQ, Chen D, Shi CX. Vision-based road detection by adaptive region segmentation and edge constraint. In 2008 Second International Symposium on Intelligent Information Technology Application. Vol. 1. IEEE; 2008. p. 342–6.
[19] Chern MY. Knowledge-based region classification for rural road area in the color scene image. In IEEE International Conference on Networking, Sensing and Control. Vol. 2. IEEE; 2004. p. 891–6.
[20] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision And Pattern Recognition. 2015. p. 3431–40.
[21] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. In International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer; 2015. p. 234–41.
[22] Chen LC, Zhu YK, Papandreou G, Schroff F, Adam H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In Proceedings of the European Conference on Computer Vision (ECCV). 2018. p. 801–18.
[23] Badrinarayanan V, Kendall A, Cipolla R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans Pattern Anal Mach Intell. 2017;39(12):2481–95.
[24] Chen LC, Papandreou G, Schroff F, Adam H. Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587. 2017.

[25] Peng C, Zhang XY, Yu G, Luo GM, Sun J. Large kernel matters–improve semantic segmentation by global convolutional network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2017. p. 4353–61.
[26] Yu, CQ, Wang JB, Peng C, Gao CX, Yu G, Sang N. Learning a discriminative feature network for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2018. p. 1857–66.
[27] Ding HH, Jiang XD, Shuai B, Liu AQ, Wang G. Context contrasted feature and gated multi-scale aggregation for scene segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2018. p. 2393–402.
[28] He J, Deng ZY, Qiao Y. Dynamic multi-scale filters for semantic segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision; 2019. p. 3562–72.
[29] He KM, Zhang XY, Ren SQ, Sun J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2016. p. 770–8.
[30] He KM, Zhang XY, Ren SQ, Sun J. Identity mappings in deep residual networks. In European Conference on Computer Vision. Springer; 2016. p. 630–45.
[31] Zou Q, Jiang HW, Dai QY, Yue YH, Chen L, Wang Q. Robust lane detection from continuous driving scenes using deep neural networks. IEEE Trans Veh Technol. 2019;69(1):41–54.
[32] Chen X, Zhao Y, Liu CC. Medical image segmentation using scalable functional variational bayesian neural networks with gaussian processes. Neurocomputing. 2022;500:58–72.
[33] Zhang HB, Pan D, Liu JH, Jiang ZH. A novel mas-gan-based data synthesis method for object surface defect detection. Neurocomputing. 2022;499:106–14.
[34] Cordts M, Omran M, Ramos S, Rehfeld T, Enzweiler M, Benenson R, et al. The cityscapes dataset for semantic urban scene understanding. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2016. p. 3213–23.
[35] Brostow GJ, Fauqueur J, Cipolla R. Semantic object classes in video: A high-definition ground truth database. Pattern Recognit Lett. 2009;30(2):88–97.
[36] Varma G, Subramanian A, Namboodiri A, Chandraker M, Jawahar CV. Idd: A dataset for exploring problems of autonomous navigation in unconstrained environments. In 2019 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE; 2019. p. 1743–51.
[37] Giusti A, Guzzi J, Cireşan DC, He FL, Rodríguez JP, Fontana F, et al. A machine learning approach to visual perception of forest trails for mobile robots. IEEE Robot Autom Lett. 2015;1(2):661–7.
[38] Yu CQ, Wang JB, Peng C, Gao CX, Yu G, Sang N. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In Proceedings of the European Conference on Computer Vision (ECCV); 2018. p. 325–41.
[39] Orsic M, Kreso I, Bevandic P, Segvic S. In defense of pre-trained imagenet architectures for real-time semantic segmentation of road-driving images. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2019. p. 12607–16.
[40] Fan MY, Lai SQ, Huang JS, Wei XM, Chai ZH, Luo JF, et al. Rethinking bisenet for real-time semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2021. p. 9716–25.
[41] Li H, Xiong PF, Fan HQ, Sun J. Dfanet: Deep feature aggregation for real-time semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2019. p. 9522–31.
[42] Poudel RPK, Liwicki S, Cipolla R. Fast-scnn: Fast semantic segmentation network. arXiv preprint arXiv:1902.04502, 2019.
[43] Zhuang JT, Yang JL. Shelfnet for real-time semantic segmentation. arXiv preprint arXiv:1811.11254, 2018.
[44] Zhao HS, Qi XJ, Shen XY, Shi JP, Jia JY. Icnet for real-time semantic segmentation on high-resolution images. In Proceedings of the European Conference on Computer Vision (ECCV); 2018. p. 405–20.
[45] Yu CQ, Gao CX, Wang JB, Yu G, Shen CH, Sang N. Bisenet v2: Bilateral network with guided aggregation for real-time semantic segmentation. Int J Comput Vis. 2021;129(11):3051–68.
[46] Xiao CJ, Hao XJ, Li HB, Li YQ, Zhang WM. Real-time semantic segmentation with local spatial pixel adjustment. Image Vis Comput. 2022;123:104470.
[47] Hao XC, Hao XJ, Zhang YR, Li YY, Wu C. Real-time semantic segmentation with weighted factorized-depthwise convolution. Image Vis Comput. 2021;114:104269.
[48] Deng RX, Shen CH, Liu SJ, Wang HB, Liu XR. Learning to predict crisp boundaries. In Proceedings of the European Conference on Computer Vision (ECCV); 2018. p. 562–78.
[49] Liu Y, Chu LT, Chen GW, Wu ZW, Chen ZY, Lai BH, et al. Paddleseg: A high-efficient development toolkit for image segmentation. arXiv preprint arXiv:2101.06175, 2021.
[50] He KM, Zhang XY, Ren SQ, Sun J. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In Proceedings of the IEEE International Conference on Computer Vision; 2015. p. 1026–34.