Copyright ⓒ 2023 KSII

Saliency-Assisted Collaborative Learning Network for Road Scene Semantic Segmentation

Haifeng Sima[1], Yushuang Xu[1], Minmin Du[1], Meng Gao[1], and Jing Wang[1] *
[1] Department of Software Egineering,
Henan Polytechnic University, Jiaozuo [e-mail: smhf@hpu.edu.cm]
[e-mail: wjasmine@hpu.edu.cn]
*Corresponding author: Jing Wang

*Received November 4, 2022; revised December 26, 2022; accepted March 1, 2023;
published March 31, 2023*

Abstract

Semantic segmentation of road scene is the key technology of autonomous driving, and the improvement of convolutional neural network architecture promotes the improvement of model segmentation performance. The existing convolutional neural network has the simplification of learning knowledge and the complexity of the model. To address this issue, we proposed a road scene semantic segmentation algorithm based on multi-task collaborative learning. Firstly, a depthwise separable convolution atrous spatial pyramid pooling is proposed to reduce model complexity. Secondly, a collaborative learning framework is proposed involved with saliency detection, and the joint loss function is defined using homoscedastic uncertainty to meet the new learning model. Experiments are conducted on the road and nature scenes datasets. The proposed method achieves 70.94% and 64.90% mIoU on Cityscapes and PASCAL VOC 2012 datasets, respectively. Qualitatively, Compared to methods with excellent performance, the method proposed in this paper has significant advantages in the segmentation of fine targets and boundaries.

Keywords: Road scene semantic segmentation, collaborative learning, saliency detection, homoscedastic uncertainty

http://doi.org/10.3837/tiis.2023.03.010 **ISSN：1976-7277**

1. Introduction

Semantic segmentation, as the basis of image understanding, has been applied to imaging analysis [1], autonomous driving [2], augmented reality [3] and many other fields. Its purpose is to partition an image into several coherent and semantically meaningful parts by different colors [4]. Since the 1970s, image segmentation has been a classic challenge in the field of image processing, attracting many researchers to make efforts for it, for it is challenging in both traditional segmentation methods [5-8] and deep learning-based segmentation algorithms [9-14]. In the last decade, road image semantic segmentation, as a key technology for automatic driving, can provide important road condition information and ensure a safe ride. Consequently, it has important theoretical research significance and practical application needs.

The Full Convolutional Neural Network (FCN) [9] achieves pixel-level semantic segmentation for the first time, which can accept input images of arbitrary size and achieve pixel-level classification. The CNN architecture and the idea of pixel-by-pixel classification adopted by the FCN laid the foundation for the development of semantic segmentation, then, Researchers have proposed UNet [10], SegNet [11], and DeepLab [12] and other FCN-based deep neural networks. Among many computer-vision tasks, saliency detection aims to identify the compelling areas in an image, which can provide support for other image understanding tasks. Since the emergence of FCN, visual saliency detection has also been greatly improved. Among them, ResNet [13] and VGG [14] are the most widely used in visual saliency detection tasks. Both above two tasks require precise pixel-level annotation, and the requirements for the model are relatively high, so there is a certain correlation between the two tasks, and many existing semantic segmentation methods have benefited a lot from visual saliency detection [15-16]. However, there is no discussion of the interaction between the two tasks, but only use the results of visual saliency detection as a preprocessing operation for semantic segmentation. To make full use of the segmentation cues from visual saliency detection, the two tasks can be trained together through collaborative learning to improve each other.

At present, road scene semantic segmentation faces many challenges, such as, excavating the dissimilarity between different objects and the similarity between congener objects, to deal with the complexity of the road environment, and the changes in the relationship and position of objects caused by illumination, weather, shooting environment, etc. To adapt the model to environmental influences and to achieve accurate segmentation. First, to address fine-grained segmentation, Gao et al., [17] proposed an active and contrastive learning-based method. Which performs representation learning through comparison between image patchesThen, Chen et al., [18] introduced a Class-Guided Asymmetric Non-Local Network (CGAN-Net) Which emphasize class information in feature maps while reducing model complexity. However, feature fusion will introduce a certain amount of noise into the final semantic feature map, which will affect the accuracy of semantic segmentation. Finally, researchers propose to use parallel structures to capture richer contextual information. To approach this problem, Multi-Feature Fusion Network (MFNet) [19] is proposed, Which adopts parallel attention branch, semantic information acquisition branch and spatial information processing branch to process shallow and deep features. Meanwhile, asymmetric factorized (AF) blocks is used to process shallow features and deep features to obtain local and global information.

At present, the disadvantages of road scene segmentation mainly contains: insufficient context information, loss of some spatial details and slow segmentation speed, which cannot meet the requirements of real-time segmentation. Inspired by the above work, we propose a road scene semantic segmentation network based on the collaborative learning in this paper. The so-called collaborative learning refers to using the same network for information or knowledge sharing, and then processing different classification tasks through different classifiers. Compared with a single task, collaborative learning can better explore the relationship between tasks, obtain additional useful information, and enhancethe robustness of the model. Therefore, we consider saliency detection as a cascaded task, which shares the encoded feature information with the semantic segmentation task. Through collaborative learning, cascading tasks provide spatial information of salient objects, further enhancing the ability of knowledge acquisition. Meanwhile, inspired by the encoder-decoder structure, DeepLabV3+ is used as the backbone, Which performs multi-level processing of encoding features in a parallel manner to make up for some of the details lost in the down-sampling process and reduce the time for the decoder to learn features and the number of parameters. In order to improve the training efficiency of the atrous spatial pyramid pooling, we uses depthwise separable convolution instead of standard convolution of the atrous spatial pyramid pooling to reduce model complexity. Finally, the loss function of the multi-task model is improved. In multi-task learning, the weight of the loss function of each task affects the performance of the model to a certain extent, the traditional method of manually adjusting the task weight is too time-consuming and labor-intensive. This article uses the same variance uncertainty to set the weights of different tasks, and then select the optimal weights to improve the performance of the model.

The main contributions of this paper are as follows:

(1) We propose an improved Depthwise Separable Convolution-Atrous Spatial Pyramid Pooling (DSC-ASPP), which uses depthwise separable convolution to reduce model complexity.

(2) We propose A collaborative learning Network for Saliency Detection and Semantic Segmentation. By sharing the features of the convolutional layer, saliency detection as a cascaded task to provide relevant information for semantic segmentation tasks and further improve the feature learning ability.

(3) The homovariance uncertainty is derived and utilized to measure the joint loss function in the collaborative learning tasks. The uncertainty of classification tasks can capture the relative confidence between tasks, so as to learn the optimal weight of each task

(4) We choose DeepLabV3+ based on the encoding-decoding structure as the backbone network to effectively compensate for the loss of context and spatial information and prove the effectiveness of this model on the Cityscapes and PASCAL VOC 2012 datasets.

2. Related Work

2.1 Image Semantic Segmentation

Image Semantic Segmentation processes images using a pixel-by-pixel classification method. The development of deep learning has brought semantic segmentation into a new era, and the segmentation speed and accuracy have been greatly improved. As a pioneering work, FCN uses full convolution for image semantic segmentation. By using a convolutional layer instead of a fully connected layer, it can accept inputs of arbitrary size, which greatly promotes the development of semantic segmentation. Since then, Researchers have proposed many FCN-based semantic segmentation networks. Chen et al., [20] and Fisher Yu et al., [21] introduced dilated convolution on the basis of FCN and used atrous convolution with different dilation rates instead of ordinary convolution to improve the receptive field of features. Wang et al., [22] proposes to use a hybrid dilated convolution module to replace the dilation convolution module to solve the grid effect caused by using dilation convolution, which can further expand the receptive field and avoid the complete loss of local information. Chen et al., [23] combined the pyramid pooling model on the basis of [12] and proposed the Atrous Spatial Pyramid Pooling (ASPP), which uses different dilation rates of atrous convolution to obtain features of different scales, perform feature fusion, and better handle multi-scale problems. Segnet [11] and UNet [10], on the other hand, use an encoder-decoder structure to establish the correlation between shallow features and deep features. Among them, the encoder performs feature extraction, and then the decoder gradually restores the resolution of the image. In recent years, researchers have focused their work on how to improve segmentation accuracy based on the encoder-decoder structure and have achieved remarkable results. For example, the ENet [24] and LEDNet [25] models use an asymmetric encoder-decoder structure to reduce the number of parameters and effectively speed up semantic segmentation. Therefore, this paper chooses the DeepLabv3+ network [26] with an encoder-decoder structure as the backbone network.

2.2 Collaborative Learning

At present, knowledge distillation [27], multi-task learning [28] and collaborative learning can improve neural network performance without increasing complexity. Knowledge distillation adopts the idea of “teacher-student network” and transfers the knowledge of the trained teacher network to another small student network through two homogeneous or heterogeneous networks, so that the performance of the student network can reach the best. However, it is computationally intensive due to the presence of two complex networks. Multi-task learning is to train several different models simultaneously by learning several related tasks together and exploring the correlation between tasks. Song et al., [29] proposed a collaborative learning method that uses different classifiers to train the same data on the same network. Sogaard [30] proved that the degree of correlation between tasks determines the performance of multi-task learning. Compared with a single-task model, collaborative learning can coordinate the complementarity between different features. Combining the advantages of multi-task learning and knowledge distillation can improve the generalization ability of the model without increasing the complexity of the network, so that the network can achieve more accurate results. In recent years, researchers have gradually increased their exploration of collaborative learning. For example: Zhou [31] proposed a collaborative learning network for lesion segmentation and disease classification of medical images. The lesion mask is applied to the classification model to improve the classification accuracy, and the lesion attention model using specific category labels also benefit the segmentation result. Luo [32] proposed a single-stage collaborative learning network for the first time, which simultaneously solves the two tasks of referring expression comprehension and referring expression segmentation. Multiple interactions between the two task branches ensure that they can promote each other during the training process and improve speed and accuracy of real-time target detection. Wang [33] proposes a new cross-dataset collaborative learning segmentation network, which combines different datasets as a new input for training. It can learn the homogeneous representations and heterogeneous statistics of different datasets and add accuracy of segmentation.

This paper uses a collaborative learning method, chooses saliency detection similar to image segmentation as an auxiliary task, and builds a road scene semantic segmentation model based on collaborative learning.

2.3 Homoscedastic Uncertainty

In the field of artificial intelligence, the problem of uncertainty has always been the focus of academic study. Bayesian network can be utilized to represent the potential dependencies between variables and solve many uncertainty problems. It is excellent and convincing on revealing many types of probabilistic dependencies [34].

In the Bayesian model, epistemic uncertainty and aleatoric uncertainty can be modeled. Epistemic uncertainty is also called model uncertainty because it mainly reflects the uncertainty of model parameters, usually caused by ignorance of the collected training data. Because of insufficient training data, the model has a lower confidence in the data that has not been seen. If the amount of data is increased, this uncertainty can be reduced. Aleatoric uncertainty is the uncertainty caused by the inability of the training data to explain the information, and these uncertainties can only be explained by improving the accuracy of observing all explanatory variables [34].

Specifically, Aleatoric uncertainty includes data-dependent or heteroscedastic uncertainty and task-dependent or homoscedastic uncertainty. The former depends on the input data and is predicted as the model output, and the latter is a constant that remains the same for the input data and varies from task to task.

In the multi-task model of road image semantic segmentation, the uncertainty of tasks can capture the relative confidence between tasks, so as to learn the optimal weight of each task. Therefore, we use homoscedastic uncertainty to optimize the weights in multi-task learning.

2.4 Saliency Detection

Visual saliency refers to the extraction of salient areas in images by simulating human visual characteristics through algorithms. Saliency detection is used in fields such as computer vision, graphics, and robotics.

The first visual saliency detection model is proposed by Itti et al., [36] based on gray contrast. Since then, saliency detection has received extensive attention from researchers. Traditional saliency detection algorithms mostly use hand-made features. In [37], a region-based saliency is proposed and it takes the sum of the product of the contrast value and the weight value of the target region and all other regions as the saliency of the region. Achanta et al., [38] proposed a method of frequency adjustment to calculate the saliency map. With the application of convolutional neural networks in vision tasks, saliency detection algorithms based on deep learning methods have been developed. In [39], two networks are employed for local and global information estimation to overcome inaccurate boundary detection and complex texture objects. Zhang et al., [40] proposed progressive attention guided recurrent network PAGR, which adds an attention mechanism on the basis of recursion and multi-resolution, thereby improving the saliency detection performance. Later, researchers discovered that saliency detection is related to category semantic information, and category semantic information and saliency detection can be combined to improve model accuracy. In [41], a collaborative saliency detection method is designed to combine high-level semantic information of the category with depth vision features. First, a group of images with the same semantic information of the same category were used for supervised training, and then saliency detection maps were derived based on high-level in-group semantic information and depth vision features. Zhang et al., [42] proposed a collaborative aggregation and distribution network. First, the group semantic information between images is obtained, and then the group semantic information is adaptively assigned to different individuals and the collaborative saliency target prediction is performed through a decoder. In [43], the CoEGNet uses a collaborative attention network and a basic saliency detection network to extract the characterization and semantic features of the image at the same time, which improves the scalability and stability of the model. Inspired by the above views, the semantic segmentation model proposed in this paper also requires pixel-level semantic information, which has a strong correlation with saliency detection.

3. Methodology

Compared with single-task model, we introduce the saliency detection for collaborative learning. Through the encoder-decoder structure, feature sharing and collaborative learning are carried out in the encoding stage, thereby improving the complementarity of features. Then use decoders of different structures for the two tasks is used to generate unique segmentation results, and finally homoscedastic uncertainty is designed to learn the weight of the loss function automatically to improve network performance.

3.1 Network Architecture

In view of the problem of insufficient context information and loss of partial spatial information in road image segmentation, based on the idea of collaborative learning, we combined the saliency detection task, and the proposed road image semantic segmentation network is shown in **Fig. 1**.

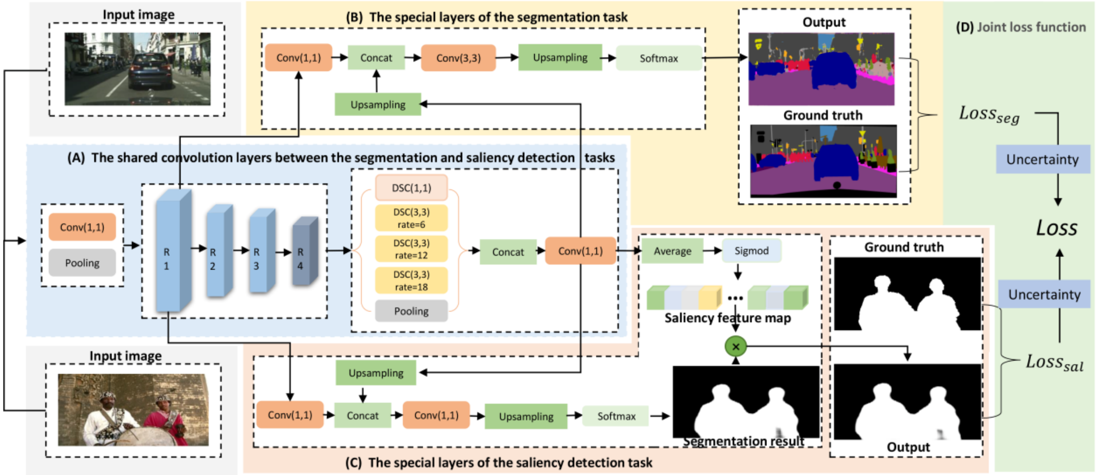


Fig. 1. Semantic segmentation network for road image based on collaborative learning. (A)The shared convolution layers between the segmentation and the saliency detection task. (B) The special layers of the segmentation task. (C) The special layers of the saliency detection task. (D)Joint loss function.

The network is composed of four parts: task-sharing module, semantic segmentation-specific module, saliency detection-specific module and joint loss function module. Among them, the task sharing module is responsible for the extraction of road features. The shared information of two different tasks is extracted by sharing the convolutional layer, thereby improving the feature extraction ability of each task. The semantic segmentation-specific module up-samples the extracted features through a decoding network, restores to the original image size, and obtains the segmentation map with class information. Compared to the semantic segmentation-specific module, the saliency detection-specific module has an additional branch for generating saliency feature maps. The joint loss function module is to model the importance of the two tasks and assign different weights through the uncertainty of the task, and finally obtains the total loss of the network.

3.2 Task- Sharing Module

The task-sharing module performs feature extraction through pooling layers and stacked convolutional layers, and the two tasks share features through a shared coding structure to strengthen the feature extraction capabilities, the structure is shown in **Fig. 2**. This module is designed as a network with powerful feature extraction capabilities, such as DenseNet [44], VGG [14], DeepLab [12], etc. We chose DeepLabV3+ [25] as the backbone network and made improvements. For the input image, this module first performs dimensionality reduction through an initial module while reducing the image resolution; then connect to the Resnet-101 network for down-sampling, where the last residual block uses a dilation convolution with an expansion rate of 2 instead of ordinary convolution to control the receptive field and extract richer feature information while ensuring the feature size remains unchanged. After down-sampling, the features of 1/16 size of the original image are obtained and sent to the DSC-ASPP module, which is improved from ASPP by replacing the parallel convolutional layers with parallel depth-separable convolutional layers. The information of different scales of the image is extracted while effectively reducing the number of parameters. Finally, the extracted multi-scale features are stitched together and the final extracted feature x_{mid} is output after passing through a convolution layer.

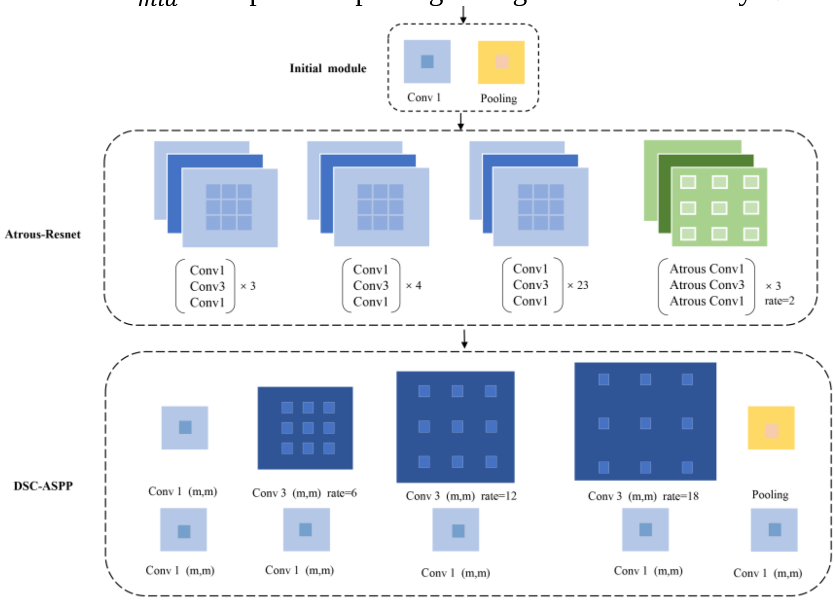


Fig. 2. Task- Sharing Module.From top to bottom: initial module, atrous-resnet, DSP-ASPP.

3.3 Task-specific modules

The semantic segmentation-specific module is essentially a decoder, as shown in **Fig. 1(C)**, which generates semantic segmentation results through a series of up-sampling and convolution operations. First, the x_{mid} is up-sampled 4 times, and the feature x_{R1}

obtained by the R1 module is channel-fused to obtain a 1/4-size feature, and then this feature is up-sampled 4 times to obtain the segmentation result y_{seg} of the same size as the input image:

$$y_{seg} = up^4(x_{R1} + up^4(x_{mid})) \tag{1}$$

Where, $up^4(\cdot)$ denotes the feature map is up-sampled by a factor of 4 , $x_{mid} \in \boldsymbol{R}^{c_1\times\frac{1}{16}h\times\frac{1}{16}w}$,$x_{R1} \in \boldsymbol{R}^{c_2\times\frac{1}{4}h\times\frac{1}{4}w}$$y_{seg} \in \boldsymbol{R}^{n\times h\times w}$, and n denotes the total number of classes, h and w respectively denote the height and width of the input image.

The saliency detection-specific module consists of two parts, as shown in **Fig. 1(B)**, For the feature x_{mid} output from the shared layer, the saliency feature map s_{fea} and the segmentation map s_{seg} are obtained through different operations. The specific operations are as follows:

First, average the pixels of each class of feature x_{mid} to obtain the weight that can characterize the importance of the feature, and then use the sigmoid function to normalize the weight to [0,1] to obtain the saliency feature map s_{fea}, which is defined as follows:

$$s_{fea} = \delta\big(\vartheta(x_{mid})\big) \tag{2}$$

Where, ϑ denotes the average value of each class of pixels, δ denotes the sigmoid function, $s_{fea} \in R^{n\times1\times1}$, and n denotes the total number of classes.

Second, similar to the semantic segmentation-specific module, an up-sampling operation is performed for x_{mid}, then feature fusion is performed with x_{R1}, and finally the fused features are upsampled to obtain a saliency segmentation map s_{seg} of the same size as the input image:

$$s_{seg} = up^4\big(x_{R1} + up^4(X_{mid})\big) \tag{3}$$

Similarly, $s_{seg} \in \boldsymbol{R}^{n\times h\times w}$. Finally, the saliency feature map and the saliency segmentation map are multiplied together to obtain the saliency detection map y_{sal}:

$$y_{sal} = \sum_{i=1}^{n} s_{fea}^{i} \cdot s_{seg}^{i} \tag{4}$$

where s_{fea}^i denotes the feature weights of the *i*-th class, and s_{seg}^i denotes the *i*-th class of the segmentation results.

3.4 Joint Loss Function

In [34], an automatically learning method of loss weights is proposed. It exploited homoscedastic uncertainty to learn multiple targets at the same time, and derive a multi-task loss function applicable to regression and classification tasks. Inspired by the above ideas, we derive a joint loss function suitable for the model proposed in this paper.

For semantic segmentation and saliency detection tasks, their essence belongs to classification tasks, and the final output is processed by the softmax function, so the probability model can be defined as follows:

$$p(y|f^W(x), \sigma) = softmax\left(\frac{1}{\sigma^2} f^W(x)\right) \tag{5}$$

$f^W(x)$ denotes the predicted result of the task with input x and weight w, y denotes the ground truth, and σ is a parameter learned by the neural network itself, which depends on the uncertainty of the task. Then the maximum likelihood estimation can be expressed as follows:

$$\log p(y=c|f^{w}(x),\sigma)=\frac{1}{\sigma^2}f_c^{w}(x)-\log\sum\nolimits_{c'}e^{\frac{1}{\sigma^2}f^{w}(x)} \tag{6}$$

Where $f_c^{w}(x)$ denotes the *c*-th output of $f^{W}(x)$.

For the proposed network in this paper, the outputs of the semantic segmentation and the saliency detection are y_{seg} and y_{sal}, respectively, and the joint probability model can be defined as follows：

$$\begin{gathered} p\left(y_{seg}=c_1,y_{sal}=c_2\middle|f^{w}(x)\right)\&=p\left(y_{seg}=c_1\middle|f^{w}(x)\right)\cdot p\left(y_{sal}=c_2\middle|f^{w}(x)\right) \\ =softmax\left(\frac{1}{{\sigma_1}^2}f^{w}(x)\right)\cdot softmax \\ \left(\frac{1}{{\sigma_2}^2}f^{w}(x)\right) \end{gathered} \tag{7}$$

If we want to maximize the likelihood estimation, we must minimize the negative log-likelihood function, that is, the joint loss function L can be defined as:

$$\begin{aligned} L(w,\sigma_1,\sigma_2) &= -\,log\,p(y_{seg}=c_1,y_{sal}=c_2|f^{W}(x) \\ &= -\,log\,softmax(\frac{1}{\sigma_1^2}f^{w}(x)\cdot softmax(\frac{1}{\sigma_2^2}f^{w}(x)) \\ &= -\,log\,softmax(\frac{1}{\sigma_1^2}f^{w}(x))-log\,softmax(\frac{1}{\sigma_2^2}f^{w}(x)) \\ &= -\frac{1}{\sigma_1^2}f_{c_1}^{w}(x)-log\sum\nolimits_{c\prime}exp(\frac{1}{\sigma_1^2}f_{c\prime}^{w}(x))-\frac{1}{\sigma_2^2}f_{c_2}^{w}(x) \\ &\quad +log\sum\nolimits_{c\prime\prime}exp(\frac{1}{\sigma_2^2}f_{c\prime\prime}^{w}(x)) \\ &= \frac{1}{\sigma_1^2}L_1(W)+\frac{1}{\sigma_2^2}L_2(W) \\ &\quad +log\frac{\sum_{c\prime}exp(\frac{1}{\sigma_1^2}f_{c\prime}^{w}(x))}{\sum_{c\prime}exp(f_{c\prime}^{w}(x))^{\frac{1}{\sigma_1^2}}}+log\frac{\sum_{c\prime\prime}exp(\frac{1}{\sigma_2^2}f_{c\prime\prime}^{w}(x))}{\sum_{c\prime\prime}exp(f_{c\prime\prime}^{w}(x))^{\frac{1}{\sigma_2^2}}} \\ &\approx \frac{1}{\sigma_1^2}L_1(W)+\frac{1}{\sigma_2^2}L_2(W)+log\,\sigma_1+log\,\sigma_2 \end{aligned} \tag{8}$$

where $L_1(W)=-\log softmax\left(y_{seg},f^{w}(x)\right)$ is the cross entropy loss of semantic segmentation, while $L_2(W)=-\log softmax\left(y_{sal},f^{w}(x)\right)$ is the cross entropy loss of saliency detection. The last step uses the following approximation: when $\sigma_1\rightarrow 1$, $\frac{1}{\sigma_1^2}\sum_{c'}exp\left(\frac{1}{\sigma_1^2}f_{c'}^{w}(x)\right)\approx\sum_{c'}exp\left(f_{c'}^{w}(x)\right)^{\frac{1}{\sigma_1^2}}$.

When σ increases, the task noise is relatively large, so the weight coefficient corresponding to the task will decrease. Compared with the previous direct loss weighting, the joint loss function can learn the relative weight of each task well. Experiments show that this method can improve the performance of the model well.

4. Experiments and Results

In this section, we evaluate our method on two benchmark semantic segmentation datasets: the Cityscapes dataset [45] and the PASCAL VOC 2012 dataset [46]. And we use DUTS [47], ECSSD [48] as the saliency detection dataset for collaborative learning. In the following experiments, all ablation experiments were performed on the Cityscapes dataset.

4.1 Datasets and Implementation Details

The Cityscapes Dataset[45] This dataset consists of 5000 finely annotated images, specifically, it is divided into three parts: 2975 images as training set, 500 images as validation set and 1525 images as test set. At the same time, in this experiment, we used 19 classes from this dataset for model training and evaluation.

The PASCAL VOC 2012 dataset [46] It contains 20 classes, of which 1464 iamges are used as training sets, 1449 images are used for evaluation, and 1456 images are used for testing. It has been widely used in image classification, target detection, and image segmentation. Same as [12], we add 10582 images to the training set to expand the amount of training data to improve model performance.

The DUTS dataset [47] It consists of the training set and the test set of the ImageNet DET total of 10553 images. At present, the DUTS dataset is the largest saliency detection benchmark with explicit training, including very challenging saliency detection scenarios, and all true labels are manually labeled.

The ECSSD dataset [48] This dataset is extended from the Complex Scene Saliency Dataset (CSSD) [49] and contains 1000 images with complex scenes and their true labels, presenting the texture and structure common to real-world images.

The sample images of the above datasets is shown in the **Fig. 3**.

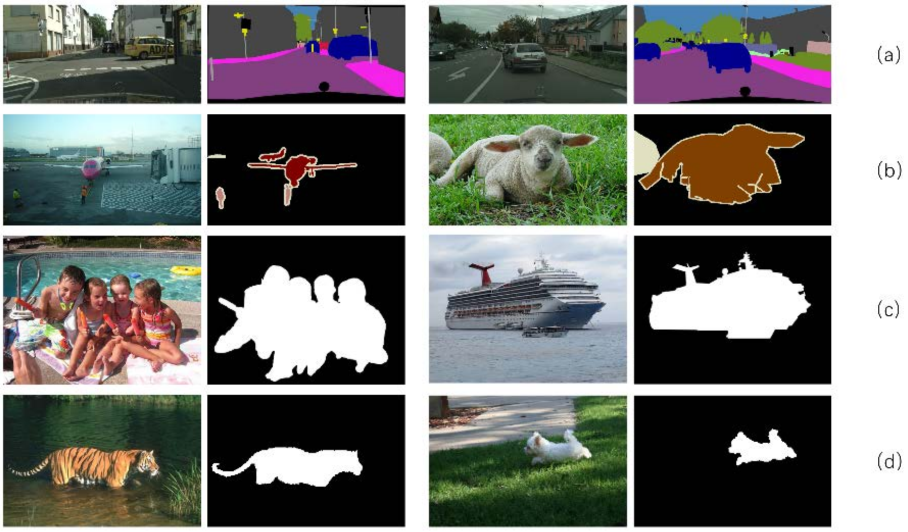


Fig. 3. The sample graph of the data set.The left side is the image and the right side is the ground truth.(a) Cityscapes. (b) PASCAL VOC 2012. (c) DUTS. (d) ECSSD.

Implementation Details This experiment is based on the pytorch framework and is trained on an NVIDIA Quadro RTX 8000 GPU with 48GB of memory. We use Adam as the optimizer and use cross-entropy loss to optimize all models. Due to the imbalance between classes, we adopt the custom class weighting scheme proposed by [23]: $w_{class} =$

$\frac{1}{ln(1.02+p_{class})}$. For the Cityscapes dataset, The size of the image is 1024*512, and trained with a batch size of 8, the PASCAL VOC 2012 dataset The size of the image is 256*256, and and trained with a batch size of 16.

We use mean Intersection over Union (mIoU) as our performance metric, which is used to calculate the mean intersection ratio of two collections. In the segmentation task, these two collections are the predicted value and the true value respectively.

4.2 Results on Cityscapes

In this section, we will investigate the effectiveness of our proposed method. In the following experiment, we choose the DeepLabV3+ model structure and evaluate it on the Cityscapes dataset.

4.2.1 Ablation study

Ablation of DSC-ASPP: In order to reduce the complexity of the model and the amount of calculation, we use the depthwise separable convolution to replace the normal convolution in ASPP. We choose 1024*512 as the resolution of the image, and in **Table 1** we have counted the number of parameters and computation required to process the resolution. The computing power is represented by the floating-point operations (FLOPs) required for forward propagation. Obviously, our method can effectively reduce the number of parameters and the computing power required for forward propagation without affecting the performance of the model.

Table 1. Comparison on the basis of operations

	DSC-ASPP	**GFLOPs**	**Parameters**	**mIoU (%)**
DeepLabv3+ （Resnet 101）		177.78	59.34M	75.52
	√	152.13	46.81M(↓21%)	75.66

Ablation of collaborative learning: Compared with the single-task semantic segmentation model, we propose a semantic segmentation model based on collaborative learning, which uses saliency detection as a cascade task to provide semantic segmentation with relevant information such as the spatial location of salient objects to improve the overall segmentation effect of the model. In order to verify the effectiveness of the model, we conducted experiments in accordance with **Table 2**.

Table 2. Ablation study for collaborative learning (SSM denotes Semantic Segmentation Module, SDM denotes Saliency Detection Module)

Method	**Backbone**	**SSM**	**SDM**	**mIoU (%)**
DenseNet	—	√		65.90
		√	√	69.53
DeepLabv3+	**ResNet101**	√		75.66
		√	√	**77.23**

As shown in Table 2, collaborative learning significantly improves performance. Compared with the DenseNet network that only performs semantic segmentation, the collaborative learning network with the saliency detection module has the mIoU of 69.53%, an increase of 3.63%. At the same time, for the DeepLabv3+ network we selected, by adding the saliency detection module, the mIoU increased by 1.57%. The results show that the collaborative learning framework combined with saliency detection brings significant benefits to road scene segmentation.

Ablation of loss function: On the basis of the collaborative learning framework, we optimized the weight with the homoscedastic uncertainty, and the results are shown in Table 3. Compared with manual adjustment of the loss weight, when using uniform weighting, the loss weight is average and consistent, and the performance of the model is increased. However, the manual assignment method shows poor performance. Because the optimal weight is uncertain, it needs to be tried continuously. There are thousands of possible combinations. This process requires a lot of time and resources. However, the performance of using the homoscedastic uncertainty is better, which is an increase of 3.28% compared with the semantic segmentation single task, and an increase of 1.71% compared with the using of uniform weights.

Table 3. Ablation for loss function

Loss	**Task Weight**		**mIoU (%)**
	Seg.	**Sal.**	
Segmentation Only	1	0	75.66
Unweighted sum of losses	0.5	0.5	77.23
Manually adjusted weight	0.8	0.2	71.34
	0.2	0.8	69.12
Multi-task uncertainty weighting	√	√	**78.94**

4.2.2 Compare with other methods

Experiments were conducted to compare our proposed method with other methods with equal training parameters for all experiments, and we compare the performance of the proposed algorithm with existing algorithm on the Cityscapes validation set, the results are shown in Table 4. It can be seen that our method outperforms all of them in mIoU index. In addition, as shown in Table 5, we used different methods to calculate the IoU for each category. Although our method cannot guarantee the best segmentation results for each category, our proposed method performs best in most categories.

Table 4. Comparison with some existing methods on the Cityscapes validation set.

Method	**Backbone**	**Mean IoU (%)**
DenseNet[44]	—	65.90
SqueezeNAS[55]	—	68.02
Edgenet[54]	—	71.00
Dilated ResNet[50]	ResNet101	73.15
MaskFormer[52]	ResNet101	74.63

DeepLabV3 [26]	ResNet101	75.66
VSANet [53]	DeepLabV3+	76.19
Cgan-Net[54]	ResNet34	76.80
PSPNet[19]	ResNet101	76.81
RepVGG[56]	—	77.15
Our method	ResNet101	**78.94**

Table 5. Per-class results on the Cityscapes validation set. The numbers marked in yellow indicate the best results for each category.

Model	**DenseNet**	**Dilated ResNet**	**Mask Former**	**DeepLab V3+**	**PSPNet**	**RepVGG**	**Ours**
Road	96.47	96.61	97.81	98.74	98.01	97.99	97.86
Sidewalk	79.58	78.09	83.05	82.56	82.69	80.14	83.77
Building	89.53	91.49	91.02	95.54	96.13	93.89	96.03
Wall	48.67	48.93	46.54	56.88	57.03	57.98	58.47
Fence	42.41	45.37	48.43	53.85	54.91	56.91	59.54
Pole	49.08	54.41	61.83	57.66	59.22	61.29	65.23
Traffic Light	26.57	65.04	67.72	58.91	66.91	67.12	68.03
Traffic Sign	61.65	71.09	75.14	70.93	73.24	72.41	74.12
Vegetation	89.96	91.88	91.51	90.19	91.09	91.76	92.02
Terrain	52.38	58.00	61.05	59.46	60.31	62.34	63.21
Sky	92.15	95.09	93.71	95.67	95.86	96.23	97.39
Person	69.62	78.80	79.25	85.45	80.86	85.24	87.38
Rider	48.53	58.26	59.53	56.32	57.91	54.32	58.23
Car	93.83	92.91	94.43	97.98	96.14	95.37	97.46
Truck	56.87	76.95	75.46	79.96	82.16	81.56	83.21
Bus	73.36	86.05	86.40	93.24	93.19	92.84	94.51
Train	62.06	70.55	69.91	66.45	67.98	68.43	70.81
Motorcycle	47.21	56.79	59.15	63.95	64.85	66.94	69.91
Bicycle	72.22	73.50	76.03	83.42	80.92	83.24	84.67
mIoU	64.90	73.15	74.63	75.66	76.81	77.15	**78.94**

For better comparison, we also counted the average accuracy ofadvanced models. As shown in **Fig. 4**, it can be seen that the accuracy of the proposed method has improved compared with other networks. We performed a visual comparison between different models in **Fig. 5**, and it can be seen from sensory intuition that our model can obtain a better segmentation effect.

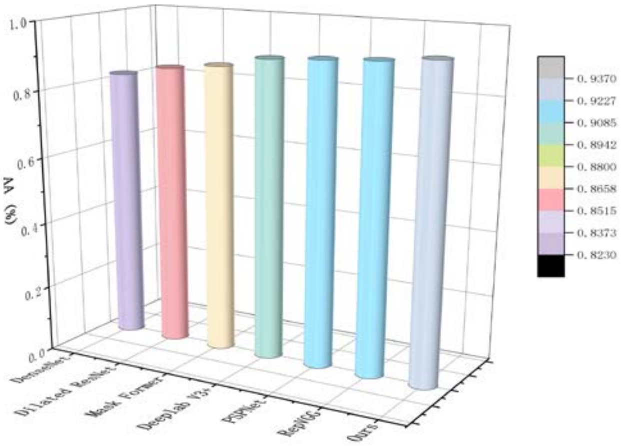


Fig. 4. the average accuracy of different models

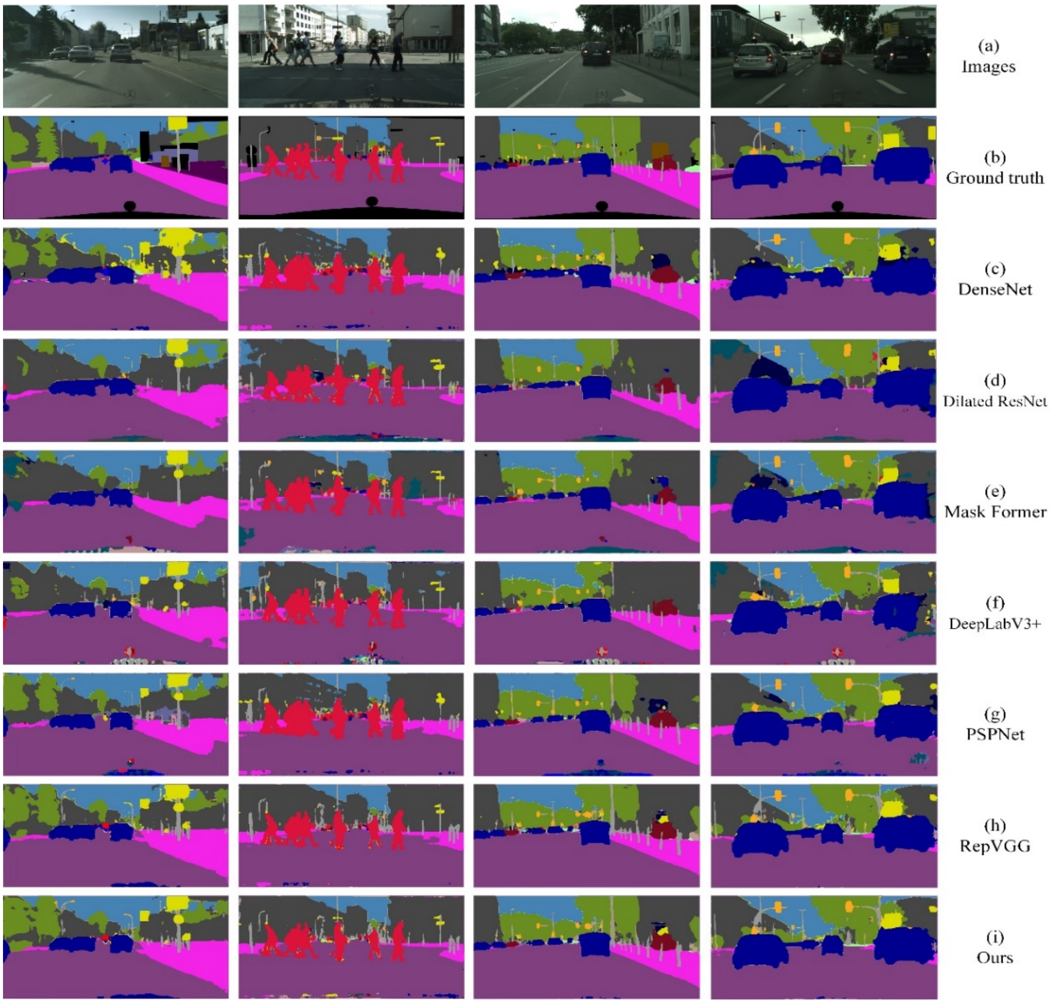


Fig. 5. Visual comparison on Cityscapes Dataset. (a) Image. (b) Ground Truth. (c) DenseNet. (d) Dilated ResNet. (e) MaskFormer. (f) DeepLabV3+ (g) PSPNet. (h) RepVGG. (i) Ours.

4.3 Results on PASCAL VOC 2012

We conducted experiments on another representative PASCAL VOC 2012 dataset, and verify the effectiveness of the model through the segmentation effect of the model on different datasets. As shown in **Table 6**, our proposed method also performed well on PASCAL VOC 2012. We achieved an mIoU of 60.90%, which is 10.76% better than DenseNet. In addition, we perform a visual comparison of different methods in **Fig. 6**.

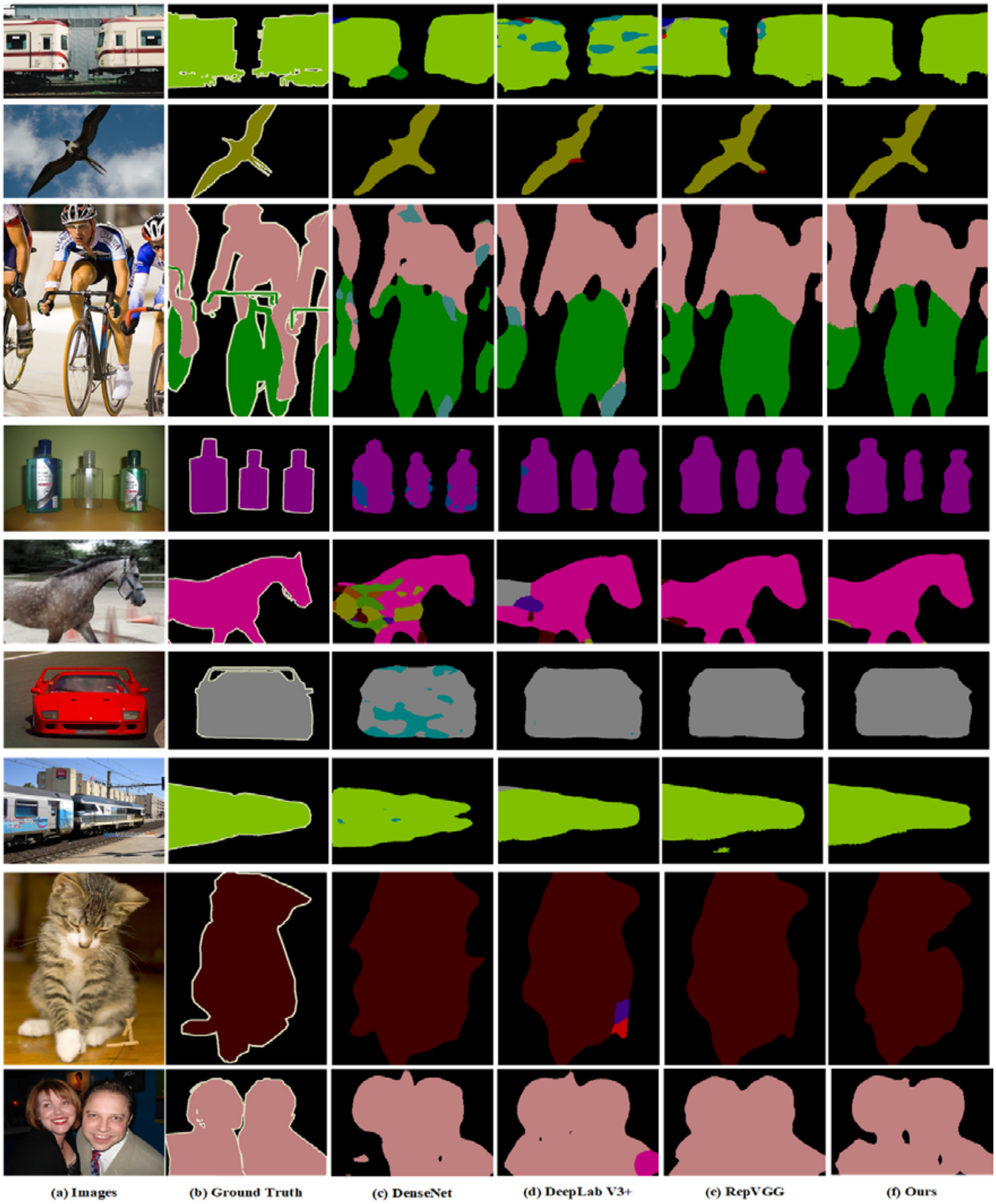


Fig. 6. Visual comparison on PASCAL VOC 2012 validation Dataset. (a) Image. (b) Ground Truth. (c)DenseNet. (d) DeepLabV3+. (e) RepVGG. (f) Ours.

Table 6. Comparison with some existing methods on PASCAL VOC 2012 validation set.

Method	Backbone	mIoU (%)
DenseNet	—	50.14
DeepLab V1	ResNet18	56.76
Dilated ResNet	ResNet101	57.66
DeepLabV3+	ResNet101	59.23
RepVGG	—	60.18
Our method	ResNet101	**60.90**

5. Conclusion

In this paper, combined with the saliency detection algorithm, a road image semantic segmentation algorithm based on collaborative learning is proposed. Specifically, we use deep separable convolution in the ASPP module to reduce the computational effort and number of parameters. In addition, the saliency detection module provides relative spatial information for semantic segmentation and improve the ability of the model to learn features. Experiments on the Cityscapes dataset show that the improved algorithm improves the mIoU in most categories to varying degrees and can better capture small-scale targets and segment object boundary regions. In addition, it has also achieved outstanding performance on the PASCAL VOC 2012 dataset.

References

[1] Jha D, Riegler M A, Johansen D, et al., “Doubleu-net: A deep convolutional neural network for medical image segmentation,” in *Proc. of 2020 IEEE 33rd International symposium on computer-based medical systems (CBMS)*, IEEE, 558-564, 2020. Article (CrossRef Link).

[2] Mozaffari S, Al-Jarrah O Y, Dianati M, et al., “Deep learning-based vehicle behavior prediction for autonomous driving applications: A review,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 1, pp. 33-47, 2022. Article (CrossRef Link).

[3] van Doormaal J A M, Fick T, Ali M, et al., “Fully automatic adaptive meshing based segmentation of the ventricular system for augmented reality visualization and navigation,” *World Neurosurgery*, vol. 156, pp. e9-e24, 2021. Article (CrossRef Link).

[4] Taghanaki S A, Abhishek K, Cohen J P, et al., “Deep semantic segmentation of natural and medical images: a review,” *Artificial Intelligence Review*, 54(1), 137-178, 2021. Article (CrossRef Link).

[5] Andrea H, Aranguren I, Oliva D, et al., “Efficient image segmentation through 2D histograms and an improved owl search algorithm,” *International Journal of Machine Learning and Cybernetics*, 12(1), 131-150, 2021. Article (CrossRef Link).

[6] Penaud-Polge V, Gavet Y., “Robustness Study of Edge Segmentation and Completion by CNN Based Method for Tessellation Images,” in *Proc. of 2020 10th International Symposium on Signal, Image, Video and Communications (ISIVC)*, IEEE, 1-6, 2021. Article (CrossRef Link).

[7] Fan H, Meng F, Liu Y, et al., “A novel breast ultrasound image automated segmentation algorithm based on seeded region growing integrating gradual equipartition threshold,” *Multimedia Tools and Applications*, 78(19), 27915-27932, 2019. Article (CrossRef Link).

[8] S SHI J, MALIK J., “Normalized cuts and image segmentation,” *IEEE Transactions on pattern analysis and machine intelligence*, 22(8), 888-905, 2000. Article (CrossRef Link)

[9] Long J, Shelhamer E, Darrell T., “Fully convolutional networks for semantic segmentation,” in *Proc. of the IEEE conference on computer vision and pattern recognition*, 3431-3440, 2015. Article (CrossRef Link).

[10] Ronneberger O, Fischer P, Brox T., “U-net: Convolutional networks for biomedical image segmentation,” in *Proc. of International Conference on Medical image computing and computer-assisted intervention*, Springer, Cham, 234-241, 2015. Article (CrossRef Link).

[11] Badrinarayanan V, Kendall A, Cipolla R., “Segnet: A deep convolutional encoder-decoder architecture for image segmentation,” *IEEE transactions on pattern analysis and machine intelligence*, 39(12), 2481-2495, 2017. Article (CrossRef Link).

[12] Chen L C, Papandreou G, Kokkinos I, et al., “Semantic image segmentation with deep convolutional nets and fully connected crfs,” *arXiv preprint arXiv:1412.7062*, 2014. Article (CrossRef Link).

[13] He K, Zhang X, Ren S, et al., “Deep residual learning for image recognition,” in *Proc. of the IEEE conference on computer vision and pattern recognition*, 770-778, 2016. Article (CrossRef Link).

[14] Simonyan K, Zisserman A., “Very deep convolutional networks for large-scale image recognition,” *arXiv preprint arXiv:1409.1556*, 2014. Article (CrossRef Link).

[15] S. J. Oh, R. Benenson, A. Khoreva, Z. Akata, M. Fritz, B. Schiele, “Exploiting saliency for object segmentation from image level labels,” in *Proc. of IEEE conference on computer vision and pattern recognition (CVPR)*, IEEE, pp. 5038–5047, 2017. Article (CrossRef Link)

[16] Wang, S. You, X. Li, H. Ma, “Weakly-supervised semantic segmentation by iteratively mining common object features,” in *Proc. of the IEEE conference on computer vision and pattern recognition*, pp. 1354–1362, 2018. Article (CrossRef Link)

[17] B. Gao, X. Zhao and H. Zhao, “An Active and Contrastive Learning Framework for Fine-Grained Off-Road Semantic Segmentation,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 1, pp. 564-579, 2023. Article (CrossRef Link)

[18] H. Chen, Q. Hu, J. Yang, J. Wu, Y. Guo, “Cgan-net: Class-guided asymmetric non-local network for real-time semantic segmentation,” in *Proc. of IEEE International Conference on Acoustics, Speech and SignalProcessing (ICASSP)*, pp. 2325–2329, 2021. Article (CrossRef Link)

[19] M. Lu, Z. Chen, C. Liu, S. Ma, L. Cai and H. Qin, “MFNet: Multi-Feature Fusion Network for Real-Time Semantic Segmentation in Road Scenes,” *IEEE Transactions on Intelligent Transportation Systems*, 23(11), 20991-21003, 2022. Article (CrossRef Link).

[20] Chen L C, Papandreou G, Kokkinos I, et al., “Semantic image segmentation with deep convolutional nets and fully connected crfs,” *arXiv preprint arXiv:1412.7062*, 2014. Article (CrossRef Link).

[21] Yu F, Koltun V, Funkhouser T, “Dilated residual networks,” in *Proc. of the IEEE conference on computer vision and pattern recognition*, 472-480, 2017. Article (CrossRef Link).

[22] Wang P, Chen P, Yuan Y, et al., “Understanding convolution for semantic segmentation,” in *Proc. of 2018 IEEE winter conference on applications of computer vision (WACV)*, 1451-1460, 2018. Article (CrossRef Link).

[23] Chen L C, Papandreou G, Kokkinos I, et al., “Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs,” *IEEE transactions on pattern analysis and machine intelligence*, 40(4), 834-848, 2018. Article (CrossRef Link).

[24] Wicker D, Rizki M M, Tamburino L A., “E-Net: Evolutionary neural network synthesis,” *Neurocomputing*, 42(1-4), 171-196, 2002. Article (CrossRef Link).

[25] Wang Y, Zhou Q, Liu J, et al., “Lednet: A lightweight encoder-decoder network for real-time semantic segmentation,” in *Proc. of 2019 IEEE International Conference on Image Processing (ICIP)*, IEEE, 1860-1864, 2019. Article (CrossRef Link).

[26] Chen L C, Zhu Y, Papandreou G, et al., “Encoder-decoder with atrous separable convolution for semantic image segmentation,” in *Proc. of the European conference on computer vision (ECCV)*, 801-818, 2018. Article (CrossRef Link).

[27] Hinton G, Vinyals O, Dean J., “Distilling the knowledge in a neural network,” *arXiv preprint arXiv:1503.02531*, 2015. RICH C., “Multitask learning,” *Machine learning*, 28.1, 41-75, 1997.

[28] Ruder S., "An overview of multi-task learning in deep neural networks," *arXiv preprint arXiv:1706.05098*, 2017. Article (CrossRef Link).

[29] Song G, Chai W., "Collaborative learning for deep neural networks," in *Proc. of 32nd Conference on Neural Information Processing Systems (NeurIPS 2018)*, 2018. Article (CrossRef Link).

[30] Søgaard A, Goldberg Y., "Deep multi-task learning with low level tasks supervised at lower layers," in *Proc. of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 231-235, 2016. Article (CrossRef Link).

[31] Zhou Y, He X, Huang L, et al., "Collaborative learning of semi-supervised segmentation and classification for medical images," in *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2079-2088, 2019. Article (CrossRef Link).

[32] Luo G, Zhou Y, Sun X, et al., "Multi-task collaborative network for joint referring expression comprehension and segmentation," in *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10034-10043, 2020. Article (CrossRef Link).

[33] Wang L, Li D, Zhu Y, et al., "Cross-Dataset Collaborative Learning for Semantic Segmentation in Autonomous Driving," *arXiv preprint arXiv:2103.11351*, 2021. Article (CrossRef Link).

[34] Cheng J, Bell D A, Liu W, "An algorithm for Bayesian network construction from data," in *Proc. of Sixth International Workshop on Artificial Intelligence and Statistics, PMLR*, 83-90, 1997. Article (CrossRef Link).

[35] Kendall A, Gal Y, Cipolla R, "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics," in *Proc. of the IEEE conference on computer vision and pattern recognition*, 7482-7491, 2018. Article (CrossRef Link).

[36] Itti L, Koch C, Niebur E, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on pattern analysis and machine intelligence*, 20(11), 1254-1259, 1998. Article (CrossRef Link).

[37] Vidal R, Ma Y, Sastry S, "Generalized principal component analysis (GPCA)," *IEEE transactions on pattern analysis and machine intelligence*, 27(12), 1945-1959, 2005. Article (CrossRef Link).

[38] Achanta R, Hemami S, Estrada F, et al., "Frequency-tuned salient region detection," in Pr*oc. of 2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009. Article (CrossRef Link).

[39] M. Ge, R. Ji and Y. Wu, "Saliency Detection Based on Local and Global Information Fusion," in Pr*oc. of 2019 IEEE 4th International Conference on Image, Vision and Computing (ICIVC)*, pp. 612-616, Xiamen, China, 2019. Article (CrossRef Link).

[40] Zhang X, Wang T, Qi J, et al., "Progressive attention guided recurrent network for salient object detection," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, 714-722, 2018. Article (CrossRef Link).

[41] L. Wang, H. Lu, X. Ruan, M.-H. Yang, "Deep networks for saliency detection via local estimation and global search," in *Proc. of the IEEE conference on computer vision and pattern recognition*, pp. 3183-3192, 2015. Article (CrossRef Link).

[42] Z.-J. Zha, C. Wang, D. Liu, H. Xie, Y. Zhang, "Robust deep co-saliency detection with group semantic and pyramid attention," *IEEE transactions on neural networks and learning systems*, 31 (7), 2398–2408, 2020. Article (CrossRef Link).

[43] Q. Zhang, R. Cong, J. Hou, C. Li, Y. Zhao, "Coadnet: Collaborative aggregation-and-distribution networks for co-salient object detection," in *Proc. of 34th Conference on Neural Information Processing Systems (NeurIPS 2020)*, 2020. Article (CrossRef Link).

[44] Huang G, Liu Z, Van Der Maaten L, et al., "Densely connected convolutional networks," in *Proc. of the IEEE conference on computer vision and pattern recognition*, 4700-4708, 2017. Article (CrossRef Link).

[45] Cordts M, Omran M, Ramos S, et al., "The cityscapes dataset for semantic urban scene understanding," in *Proc. of the IEEE conference on computer vision and pattern recognition*, 3213-3223, 2016. Article (CrossRef Link).

[46] Everingham M, Eslami S M A, Van Gool L, et al., “The pascal visual object classes challenge: A retrospective,” *International journal of computer vision*, 111(1), 98-136, 2015. Article (CrossRef Link).

[47] Wang L, Lu H, Wang Y, et al., “Learning to detect salient objects with image-level supervision,” in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, 136-145, 2017. Article (CrossRef Link).

[48] Shi J, Yan Q, Xu L, et al., “Hierarchical image saliency detection on extended CSSD,” *IEEE transactions on pattern analysis and machine intelligence*, 38(4), 717-729, 2016. Article (CrossRef Link).

[49] Yan Q, Xu L, Shi J, et al., “Hierarchical saliency detection,” in *Proc. of the IEEE conference on computer vision and pattern recognition*, 1155-1162, 2013. Article (CrossRef Link).

[50] Yu F, Koltun V, Funkhouser T. Dilated residual networks,” in *Proc. of the IEEE conference on computer vision and pattern recognition*, 472-480, 2017. Article (CrossRef Link).

[51] Ding X, Zhang X, Ma N, et al., “Repvgg: Making vgg-style convnets great again,” in *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13733-13742, 2021. Article (CrossRef Link).

[52] Cheng B, Schwing A G, Kirillov A., “Per-pixel classification is not all you need for semantic segmentation,” in *Proc. of 35th Conference on Neural Information Processing Systems (NeurIPS 2021)*, 2021. Article (CrossRef Link).

[53] R. Liu and D. He, "Semantic Segmentation Based on Deeplabv3+ and Attention Mechanism," in *Proc. of 2021 IEEE 4th Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC)*, pp. 255-259, 2021. Article (CrossRef Link).

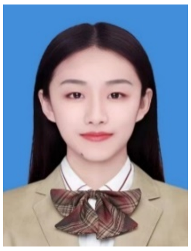
[54] H. Chen, Q. Hu, J. Yang, J. Wu and Y. Guo, "Cgan-Net: Class-Guided Asymmetric Non-Local Network for Real-Time Semantic Segmentation," in *Proc. of ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2325-2329, 2021. Article (CrossRef Link).

[55] Shaw A, Hunter D, Landola F, et al., “SqueezeNAS: Fast neural architecture search for faster semantic segmentation,” in *Proc. of the IEEE/CVF International Conference on Computer Vision*, 2019. Article (CrossRef Link).

[56] Han H Y, Chen Y C, Hsiao P Y, et al., “Using channel-wise attention for deep CNN based real-time semantic segmentation with class-aware edge information,” *IEEE Transactions on Intelligent Transportation Systems*, 22(2), 1041-1051, 2021. Article (CrossRef Link).

Haifeng Sima received the B.E. and M.E. degrees in computer science from Zhengzhou University, Zhengzhou, China, in 2004 and 2007, respectively, and the Ph.D. degree in software and theory from the Beijing Institute of Technology, Beijing, China, in 2015. Since 2007, he has been with the Faculty of Henan Polytechnic University, Jiaozuo, China, where he is currently a Lecturer with the School of Computer Science and Technology. His current research interests include pattern recognition, image processing, image segmentation, and image classification.

Yushuang Xu received the B.S. degree in computer science and technology from Luoyang Institute of Science and Technology, Luoyang, China, in 2019. He is a M.S. candidate in computer technology with the School of Computer Science and Technology, Henan Polytechnic University, Jiaozuo, China. His research interests include image processing, image segmentation.



Minmin Du received the B.S. degree in software engineering from Liaocheng University, Liaocheng, China, in 2019. He is a M.S. candidate in software engineering with the School of Computer Science and Technology, Henan Polytechnic University, Jiaozuo, China. His research interests include image processing, image segmentation, pattern recognition.

Meng Gao received the B.S. degree in data science and big data technology from Huanghe Jiaotong University, Jiaozuo, China, in 2022. He is a M.S. candidate in software engineering with the School of Computer Science and Technology, Henan Polytechnic University, Jiaozuo, China. His research interests include computer vision, deep learning and artificial intelligence.

Jing Wang received the B.Sc. degree in computer science and technology from the Henan University of Science and Technology, Luoyang, China, in 2006, and the Ph.D. degree in computer application technology from the College of Computing and Communication Engineering, Graduate University of Chinese Academy of Science, Beijing, China, in 2012. She is currently an Associate Professor with the College of Computer Science and Technology, Henan Polytechnic University, Jiaozuo, China. Her research interests include image processing computer vision and machine