Hindawi
Mathematical Problems in Engineering
Volume 2021, Article ID 9963974, 13 pages
https://doi.org/10.1155/2021/9963974

*Research Article*

# A Real-Time Image Semantic Segmentation Method Based on Multilabel Classification

**Ran Jin**,[1,2] **Xiaozhen Han**,[1] **and Tongrui Yu**[1]

[1]*College of Big Data and Software Engineering, Zhejiang Wanli University, Ningbo 315100, China*
[2]*College of Computer Science and Technology, Zhejiang University, Hangzhou 310058, China*

Correspondence should be addressed to Ran Jin; ran.jin@163.com

Received 30 March 2021; Revised 5 May 2021; Accepted 23 May 2021; Published 1 June 2021

Academic Editor: Jude Hemanth



Image semantic segmentation as a kind of technology has been playing a crucial part in intelligent driving, medical image analysis, video surveillance, and AR. However, since the scene needs to infer more semantics from video and audio clips and the request for real-time performance becomes stricter, whetherthe single-label classification method that was usually used before or the regular manual labeling cannot meet this end. Given the excellent performance of deep learning algorithms in extensive applications, the image semantic segmentation algorithm based on deep learning framework has been brought under the spotlight of development. This paper attempts to improve the ESPNet (Efficient Spatial Pyramid of Dilated Convolutions for Semantic Segmentation) based on the multilabel classification method by the following steps. First, the standard convolution is replaced by applying Receptive Field in Deep Convolutional Neural Network in the convolution layer, to the extent that every pixel in the covered area would facilitate the ultimate feature response. Second, the ASPP (Atrous Spatial Pyramid Pooling) module is improved based on the atrous convolution, and the DB-ASPP (Delate Batch Normalization-ASPP) is proposed as a way to reducing gridding artifacts due to the multilayer atrous convolution, acquiring multiscale information, and integrating the feature information in relation to the image set. Finally, the proposed model and regular models are subject to extensive tests and comparisons on a plurality of multiple data sets. Results show that the proposed model demonstrates a good accuracy of segmentation, the smallest network parameter at 0.3 M and the fastest speed of segmentation at 25 FPS.

## 1. Introduction

Multilabel classification evolved as the single-label classification method is gradually away from having the present needs satisfied. At first, it mainly took the form of text classification. The development of deep learning and the updates on computer vision have boosted image semantic segmentation, target recognition, and detection. Many kinds of deep learning-based methods for image semantic segmentation have been reported, including Fully Convolutional Network (FCN), Convolution and Graphics Model, Encoder-Decoder Model, Multiscale Pyramid Model, Region-Based Convolutional Neural Network (R-CNN) Model, Dilated Convolution and DeepLab Family Model, Recurrent Neural Network (RNN) Model, Attention Mechanism Model, Generate Adversarial Network (GAN), and Active Contour Model [1, 2]. Given the characteristics, these methods can be roughly categorized to the method based on region classification and that based on pixel classification. The method based on region classification refers to an alternative of dividing image into several blocks, extracting image feature by Convolutional Neural Network (CNN) and classifying the image blocks. This alternative can be subdivided into the method based on candidate region and that based on segmentation mask. In general, the category to which a pixel belongs may be marked according to the highest score region. This alternative may regard Visual Geometry Group Network (VGGNet), GoogLeNet, ResNet (Residual Neural Network), and other networks as the backbone network of the model for classification of image blocks. Since these methods contain Fully Connected Layer (FCL) in the classification network, the size of input image is required to be fixed and the model with generally a higher cost of memory, resulting in computational inefficiency and

unsatisfactory segmentation effect. At present, there are also some extensions on this basis, such as the composite segmentation method based on encoder-decoder, combined with Dense Residual Block (DRB) and FCN [3–5]. Given this, FCN was put forward in 2014, which has been one of the most popular pixel-based classification methods. The space size of feature image as extracted by the CNN structure can be adjusted by upsampling until it matches the original image. For image segmentation task, FCN appears to be superior over conventional CNN because the input image in the model doesn't have to be fixed size, and the network has an even higher computational efficiency. The stricter demand for real-time performance and huge computational power has put lightweight semantic segmentation under the spotlight. In 2017, Andrew Howard et al. proposed MobileNets (Efficient Convolutional Neural Networks for Mobile Vision) [6] and, later in 2018, proposed MobileNetV2. The underlying idea of MobileNets is to reduce the number of model parameters by means of separable convolution, leading to faster running speed of the model. In 2017, Zhang et al. from Megvill proposed ShuffleNet (an Extremely Efficient Convolutional Neural Network for Mobile) [7], and Ma et al. proposed ShuffleNetV2 in 2018. The underlying idea of ShuffleNet is to reduce the computational workload by using the convolution channel shuffle. For real-time semantic segmentation models, Adam Paszke et al. proposed ENet (a deep neural network architecture for real-time semantic segmentation) [8] in 2016, improved the pooling operation and output pooling mask at the time of downsampling, and improved recognition accuracy at the time of upsampling. In 2020, Tan [9] et al. from Google proposed ESPNet, which can capitalize on the two-way weighted feature pyramid structure for feature fusion and use the composite size method to uniformly scale down the resolution, depth, and width of backbone network, feature network, and predictive network.

## 2. Related Works

*2.1. Multilabel Classification.* Multilabel classification is considered as an issue in relation to classification, where a sample may be assigned with multiple target labels concurrently. For example, an image may contain urban buildings, vehicles, and people; a song is both lyrical and sentimental. Accordingly, a data sample (picture or music) may contain a plurality of different labels concurrently, which are used to characterize data attributes. What makes it hard to carry out multilabel learning is the explosive growth of output space. For example, if there are 10 labels available, the output space would be 210 in size. An effective mining of the label-to-label correlation is the only way to reduce the huge amount of output, which underpins the success of multilabel learning. Multilabel algorithms can be divided into three categories if we consider the intensity of correlation mining. First-order strategy: the correlation between one label and other labels is neglected. Second-order strategy: the pairwise correlation among labels is considered. High-order strategy: the correlation among a plurality of labels is considered. It should be noted that multilabel classification can be solved in three options. One alternative is issue transform options, including label transform-based options and instance transform-based options, e.g., binary relevance (BR) [10]. The second alternative is adaptive algorithm, that is, to modify some available learning algorithms, to the extent that the multilabel learning capability can be satisfied, e.g., Multilabel K-Nearest Neighbor (ML-KNN) [11]. The third alternative is integration method, an option evolved from regular issue transform or adaptive algorithm. The most famous ensemble of issue transform can be illustrated by RAKEL system [12], Ensemble of Pruned Sets (EPS) [13], and Ensemble of Classifier Chains (ECC) [14] proposed by Tsoumakas et al. Further details about these options are available in Figure 1.

*2.2. ESPNet.* The ESPNet was introduced by Mehta et al. [15], where a semantic segmentation network architecture featuring fast calculation and excellent effect of segmentation is presented in details. ESPNet can process data at 112 FPS on GPU in an ideal state or up to 9 FPS on edge device at a level even faster than the well-known lightweight networks—MobileNet [6], ENet [8], and ShuffleNet [7]. Provided that the control model only losses 8% of the classification accuracy, the ESPNet has the model parameters only 1/180 of PSPNet, known as the most excellent architecture at that time, but its processing speed is 22 times faster than PSPNet. In this published paper, a convolution module which is referred to as "Effective Spatial Pyramid" was introduced as a part of ESPNet. Consequently, such network architecture is characterized by fast speed, low power consumption, and low latency, which in turn makes it more suitable to deploy in some edge devices subject to more resource limits.

Figure 2 is the basic network architecture of ESPNet. In this model, point-by-point convolution is used to reduce the number of channels and sent to the hollow convolution pyramid. The greater receptive field is obtained from different scales of dilated convolution, alongside with feature fusion, so the amount of parameters is quite few. Following the reduced number of channels, the amount of parameters with respect to each dilated convolution is quite few. Figure 3 presents the number of channels, ratio, and merging strategy. The feature fusion method for merging strategy is sharply contrasted with that for the regular dilated convolution. The stepwise addition strategy is used as a way to avoiding gridding artifacts.

ESPNetv2 was introduced by Mehta et al. in 2019. With the increased network depth in the EESP module, each convolution layer is improved by using the PRelu activation function, and the activation function is removed from the final group level convolution layer. The dilated convolution is used to the extent that the receptive field is dilated, the number of network parameters is reduced, and the running speed is increased. Figure 4(a) provides an overview of the performance level of individual models by comparing the accuracy rate attainable by the respective model under different FLOPs, where the floating-point operations per second (FLOPs) is used as a reference. Figure 4(b) is the loss under the respective model, where the time is used as a reference.

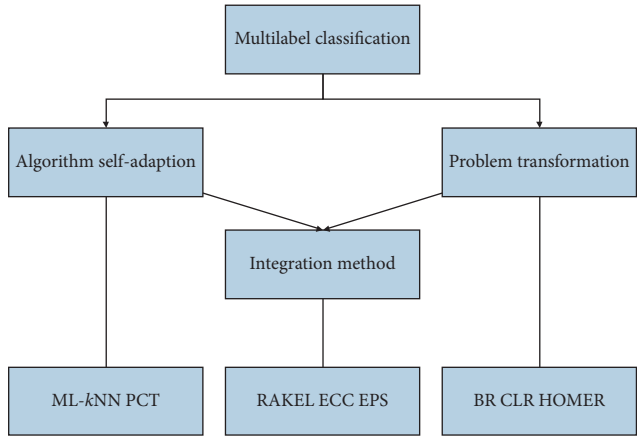


Figure 1: Method for multilabel classification.

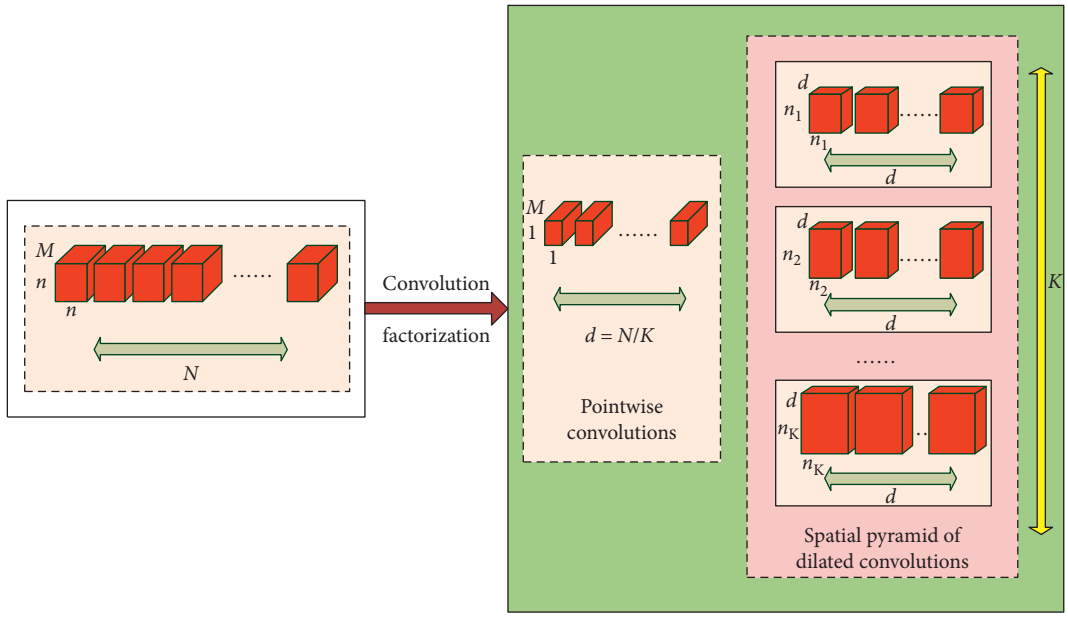


Figure 2: Basic network architecture of ESPNet.

In the past two years, some scholars have carried out useful research based on ESPNet [16–19]. Kim [16] proposed ESCNet based on ESPNet architecture which is one of the state-of-the-art real-time semantic segmentation network that can be easily deployed on edge devices. Nuechterlein [17] extended ESPNet, a fast and efficient network designed for vanilla 2D semantic segmentation, to challenging 3D data in the medical imaging domain.

## 3. The Proposed Algorithm

ESPNet is evolved from the Efficient Spatial Pyramid (ESP) module, where the point convolution maps high-dimensional features to low-dimensional space by $1 \times 1$ convolution. In this section, ESPNet is improved based on integration and tuning of a plurality of technical methods as mentioned earlier, and its core constituent modules are described here. Figure 5 is the process flow with respect to the improved model. The spatial pyramid of dilated convolution exploits $K$ and $N \times N$ dilated convolution kernels, while resampling these low-dimensional feature images. The dilation rate of each convolution kernel is $2K-1(K=F1)$. This decomposition sharply reduces the number of parameters and memory required for the ESP module and retains a large effective receiving field (n−1) $2K-1$. This sort of pyramid convolution operation is also referred to as "Spatial Dilation Convolution Pyramid." Each dilated convolution kernel learns the weight of the respective receptive field, so it appears to be similar to spatial pyramid. Since ESPNet is superior to all high-efficiency CNN networks that are currently available, this model is designed and improved. Figure 6 is the ESPNet improvement based on the convolution factor decomposition as the first step.

Provided that the parameters are constantly the same, a greater receptive field can be assured by atrous convolution, but it may be unfriendly to the recognition effect of some tiny objects. Finally, the improved model generates segmented images by exploiting the deconvolution principle of the

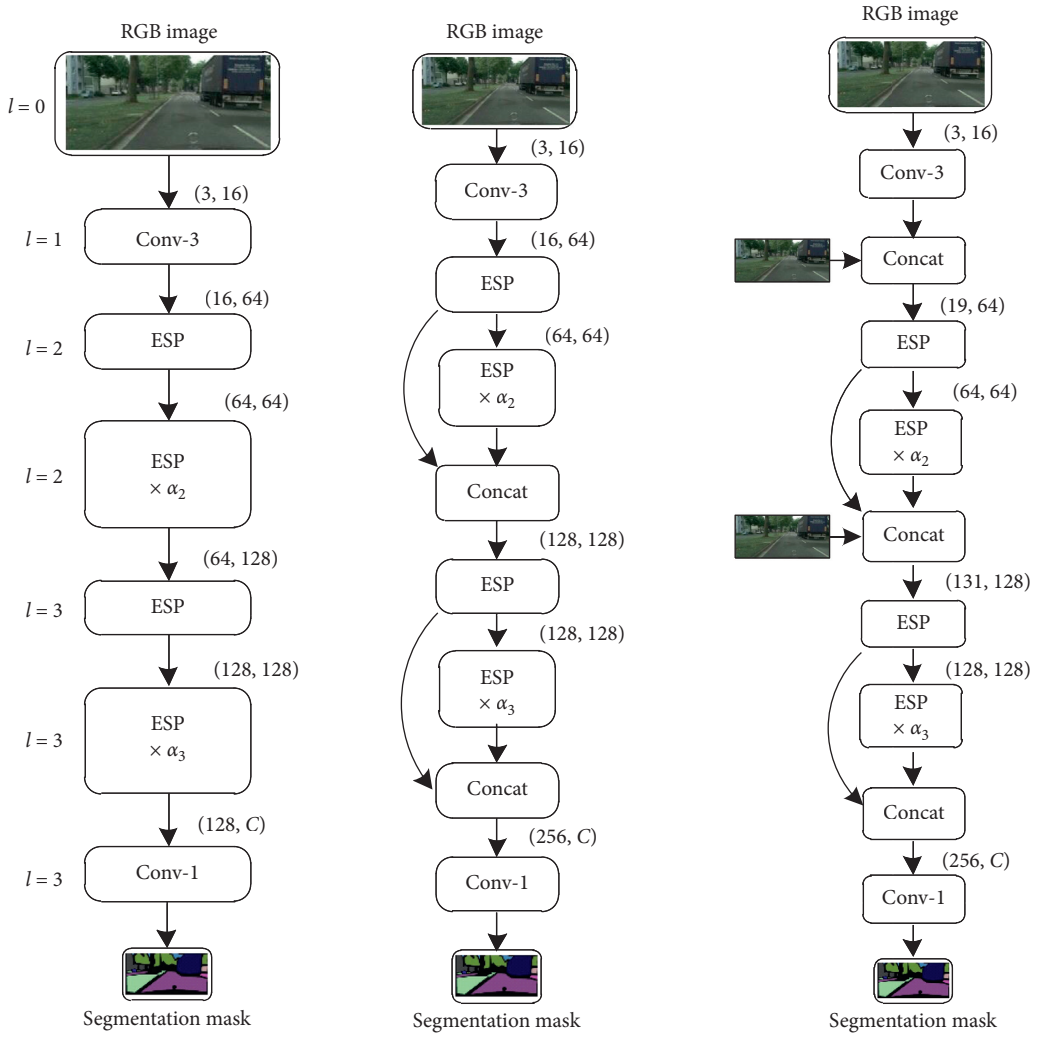


Figure 3: Main architecture of ESPNet.

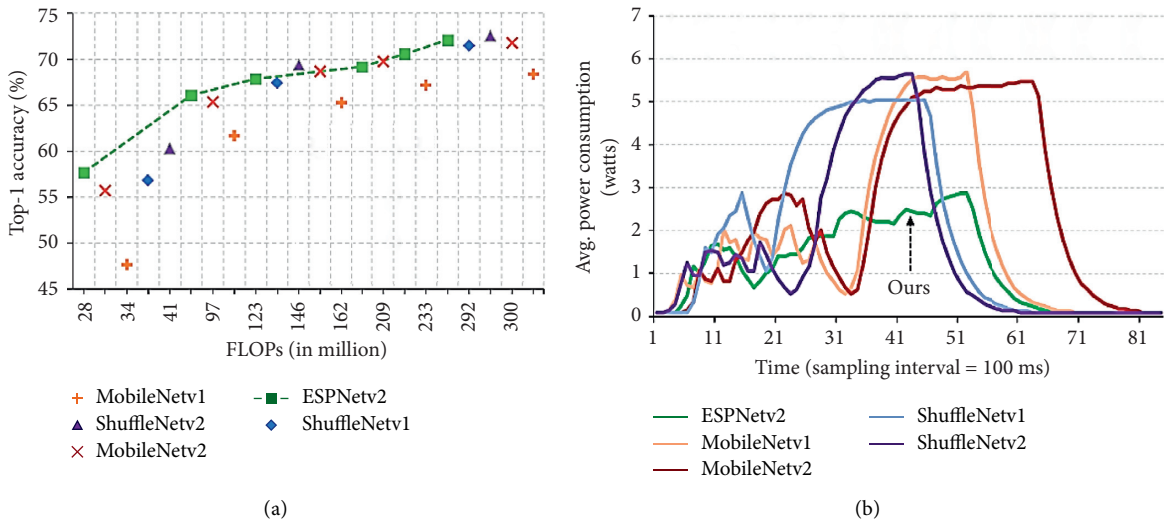


Figure 4: (a) Accuracy rate variation as a function of FLOPs. (b) Loss variation as a function of time.

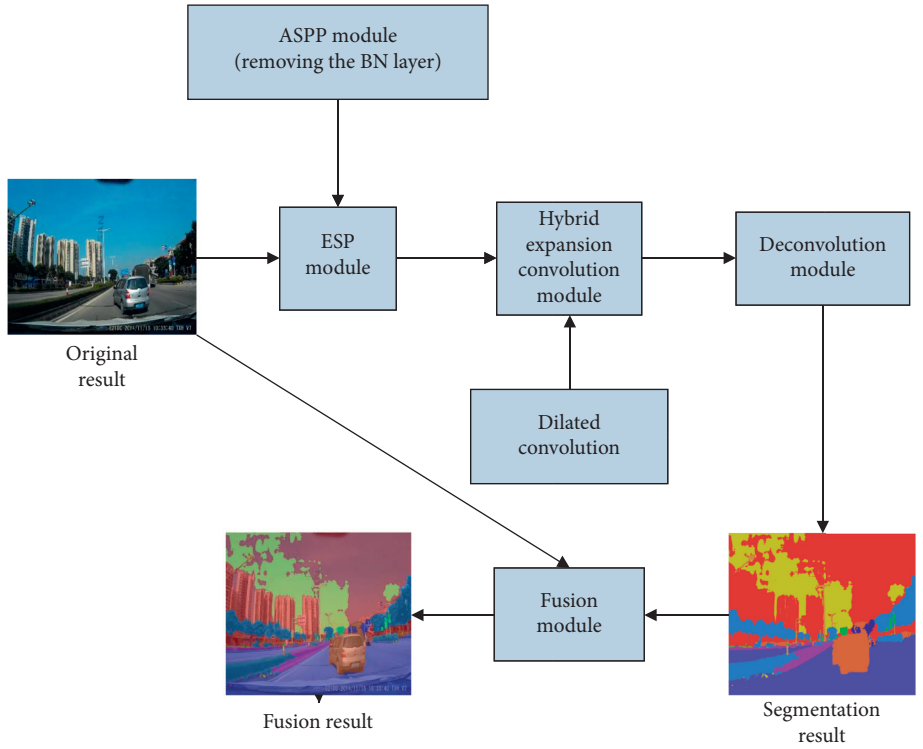


Figure 5: Functional flow figure of the improved model.

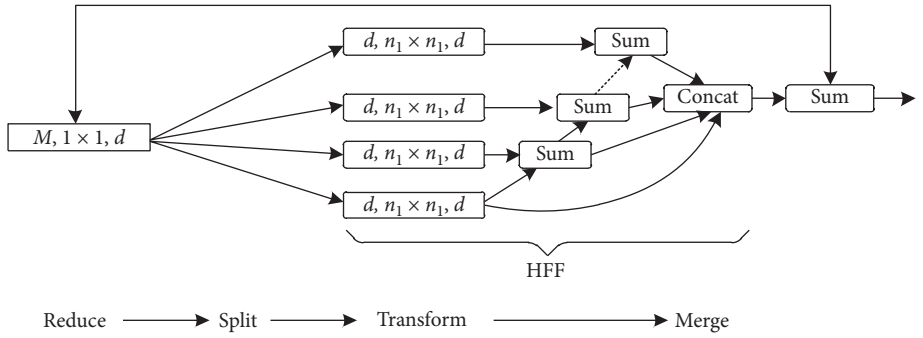


Figure 6: Schematic diagram of convolution factorization.

decoding part in the similar encoding-decoding structure. The segmented images are fused with original images on the merging module, which provides an intuitive feeling of the accuracy of the model segmentation. Figure 7 is the working principle figure of the improved model, and the algorithm used in the proposed model will use image pyramid during training, as expressed in equation (1).where $P_n^{\text{out}}$ is the feature prediction output of *nth* layer and $P_n^{\text{in}}$ is the feature input of *nth* layer. Resize(.) is used to adjust the image size.

$$\begin{aligned} P_n^{\text{out}} &= \text{Conv}\left(P_n^{in}\right) \\ P_{n-1}^{\text{out}} &= \text{Conv}\left(P_{n-1}^{in} + \text{Resize}\left(P_n^{\text{out}}\right)\right) \\ &\cdots \\ P_1^{\text{out}} &= \text{Conv}\left(P_1^{in} + \text{Resize}\left(P_2^{\text{out}}\right)\right), \end{aligned} \tag{1}$$

*3.1. Depthwise Separable Convolution.* It can be inferred from the semantic segmentation analysis of CNN and decoding-encoding that the convolution layer stands as the core part. A matching convolution method should be available for adapting to different kinds of environments; otherwise, gridding, gridding artifacts, and other unfriendly phenomena would be aggravated. As a consequence, the model may not lead to a good effect of semantic segmentation. Given this, the convolution layer is improved by treating depthwise separable convolution as its core part, using a set of dilation rates and joining them by the segmentation method in ResNet. Figure 8 describes how it works.

As seen from Figure 8, the input in the layer-by-layer convolution is $M$ channel feature images, which are, respectively, convolved with $M$ filters until $M$ feature images are output. In contrast with the conventional convolution

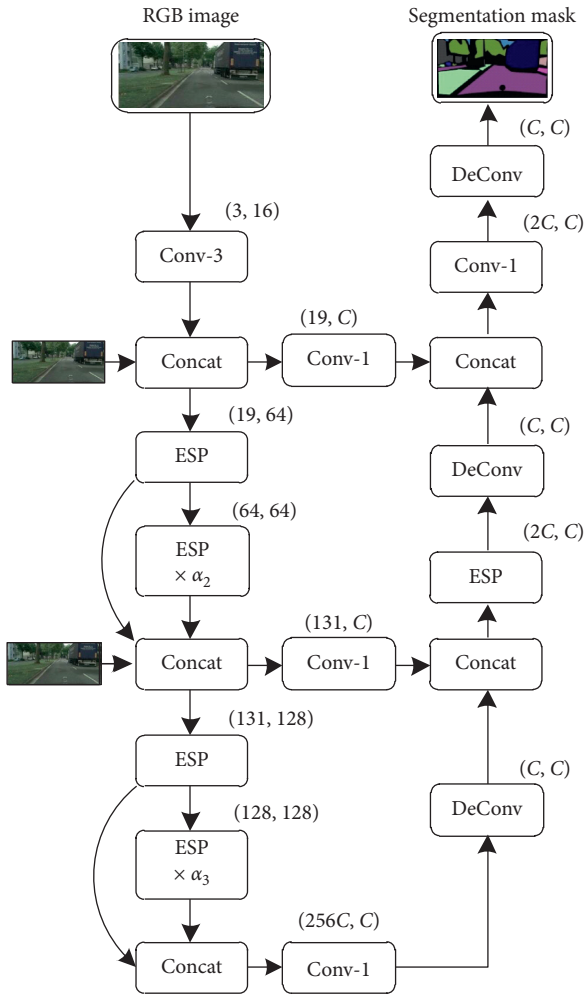


Figure 7: Working principle of the improved model.

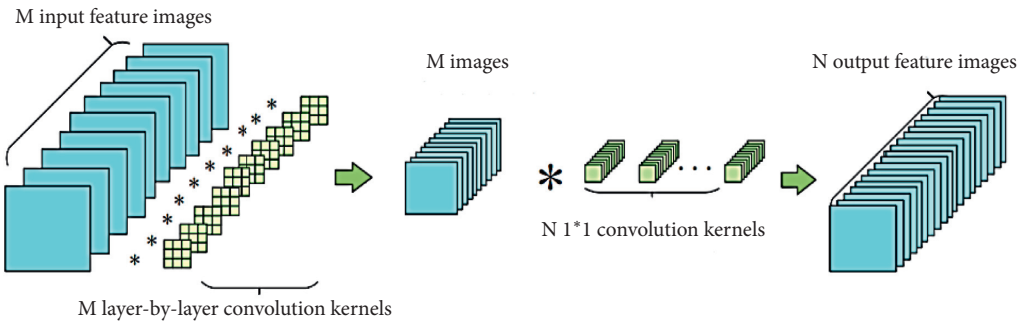


Figure 8: Schematic structure diagram of depthwise separable convolution.

method, what makes this convolution method significantly different is that the learning process with respect to the channel-space correlation is asynchronous. To put it in other way, it will not follow synchronous learning, just as conventional convolution method does. Comparing the regular convolution equation (2) and the depthwise separable convolution equation (3), this can increase the speed of network training and widen up the network. As a result, the network can accommodate and transmit more available feature information, leading to improved working efficiency.

$$\mathrm{Conv}(W, y)_{(i,y)} = \sum_{k,l,m}^{K,l,M} W_{(k,l,m)} \cdot y_{(i+k,j+l,m)}. \quad (2)$$

In the first step, the depth separable convolution is subject to channel convolution by equation (3) (in this paper, Θdenotes the multiplication of the corresponding elements) and then the pointwise convolution is performed by equation (4). Substituting equation (3) into equation (4), equation (5) with respect to the depthwise separable convolution can be obtained.

$$\text{Depthwiseconv}(W,Y)_{(i,j)} = \sum_{k,l}^{K,L} W_{(k,l)} \Theta y_{(i+k,j+l)}, \tag{3}$$

$$\text{Pointwiseconv}(W,Y)_{(i,j)} = \sum_{m}^{M} W_m \cdot y_{(i,j,m)}, \tag{4}$$

$$\text{Sepconv}\left(W_p, W_d, y\right)_{(i,j)} = \text{Pointwiseconv}_{(i,j)}\left(W_p, \text{Depthwiseconv}_{(i,y)}\left(W_d, y\right)\right), \tag{5}$$

where $W$ is convolution kernel, $y$ is input feature image, both $i$ and $j$ are the resolution of input feature image, both $k$ and $l$ are the resolution of output feature image, and $m$ is the number of channels.

*3.2. DB-ASPP.* In this paper, the ASPP module is introduced as a part of HDC to collect multiscale information, and the image-level feature information is integrated in available ASPP module. Considering the fusion needs, the batch normalization (BN) layer is filtered out. The course of ablation experiment can increase the number of BN layers and improve the accuracy of the activation function PRelu by approximately 1.4%, but the benefit of removal is that the parallel branch results directly disappear without post-processing. In other words, the network parameters are reduced, and the speed is increased. Accordingly, the improved DB-ASPP module based on ASPP is proposed here.

The atrous convolution has two functions: first, the receptive field is dilated; for example, when $r = 1$, subject to dilated convolution, it becomes $r = 2$. However, the deficiency is that the reduced resolution in spatial distribution, and if the compression level is high, it will add to the difficulty level of the subsequent upsampling or deconvolution to restore the original image size. Further, the continuous downsampling combination layer will cause a serious reduction in the spatial resolution of feature image. And more context information can be extracted by atrous convolution. Figure 9 is the schematic diagram of how atrous convolution works. When $r$–1, the receptive field is $3 \times 3$. Subject to atrous convolution, namely, when $r$–2 as shown below, the receptive field will be $5 \times 5$. It is apparent that as the atrous rate increases, the range of receptive field that is recognizable by original convolution kernel has been significantly increased.

Atrous convolution can increase the receptive field and control the resolution, but the current atrous convolution method is still vulnerable to an inherent issue—gridding issue. If the atrous convolution is continuously used and the atrous rate is improperly selected, certain pixels may not be always involved in the calculation process. For example, with respect to the pixel $p$ in a certain layer of the atrous convolution, its value is limited to the adjacent zone of the upper layer, and its size is $k\text{size} \times k\text{size}$ with $p$ as the center point. Assume that the atrous rate is $r = 1$ and $k\text{size} = 3$, the pixel $p$ is expressed by the red points as shown in Figure 10, the blue area denotes the range to be captured by the convolution, and then the lower layer image of Figure 10 can be obtained after two steps of operation ($r = 1$).

From the white spots as shown in Figure 10, many adjacent pixels are overlooked, and only a small part is used in the repetitive atrous convolution calculations. In addition, since the atrous convolution is constructed by zero value insertion among parameters in the convolution kernel, when the applicable atrous rate increases, the distance between non-zero values would also increase, and the relevance between local information will be destructed, leading to more serious loss of local information, aggravating the gridding effect in the generated feature image.

Accordingly, Wang et al. proposed the Hybrid Dilated Convolution (HDC), and the atrous convolutions with different atrous rates are used continuously and alternately to reduce the impact of gridding issue. At one dimension, HDC is defined in the following equation:

$$k[i] = \sum_{l=1}^{L} h[i + r \cdot l] g[l], \tag{6}$$

where $h[i]$ denotes input signal, $k[i]$ denotes output signal, $g[l]$ denotes the filter with length $L$, and $r$ is the dilation rate used for sampling process $h[i]$. In standard convolution, $r = 1$. Assume there are $N$ atrous convolutions, whose convolution kernel size is $ks \times ks$, $\{d_1, \ldots, d_i, \ldots, d_m\}$ is its atrous rate, and $M_i$ is the maximum distance between two non-zero points, which is computed using $d_i$ as in the following equation:

$$M_i = \text{MAX}\left[M_{i+1} - 2d_i, M_{i+1} - 2\left(M_{i+1} - di\right), di\right]. \tag{7}$$

Figure 11 is the schematic diagram of the receptive field with respect to $d = \{1, 2, 5\}$. It can be confirmed that all pixels are involved as a part of the convolution operation, which suggests that HDC can solve gridding issue well.

Based on the above HDC, the ASPP module as a part of ESPNet is improved here by introducing HDC and removing the BN layer. Figure 12 presents the functional architecture. The dilated convolution available with four dilation rates can capture multiscale information in parallel on the top-level feature response of the backbone network. The improved ASPP module confers a greater receptive field to neurons, and the Pyramid Pooling Module (PPM) is introduced to the proposed ESP. As a result, the contextual semantic information in different regions can be aggregated to attain a better effect of segmentation.

In addition, for control of the model size and prevention of over-sized network, $1 \times 1$ convolution layer is added in front of each atrous convolution layer in DB-ASPP with reference to DenseNet and DenseASPP in order to reduce

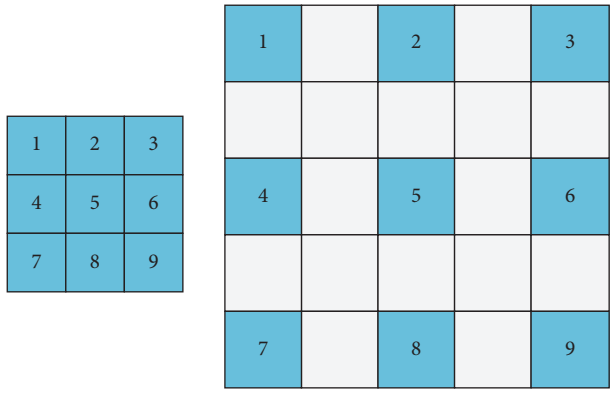


Figure 9: Schematic structure diagram of depthwise separable convolution.

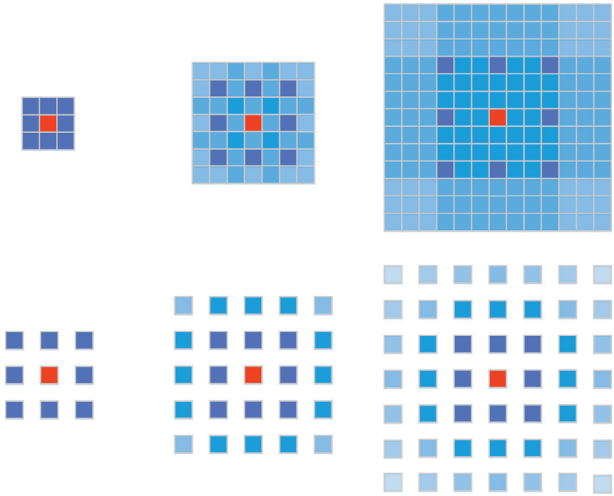

Figure 10: Schematic structure diagram of depthwise separable convolution.

the depth of feature image to the specified size and further control the output size. Assume that each atrous convolution layer output has $n$ feature images, DB-ASPP has $C_0$ feature images as input, and the $lth$ $1 \times 1$ convolution layer in front of the $lth$ convolution layer has $C_l$ input feature images. $C_l$ is computed using input $C_0$, $n$, and $l$ as in the following equation:

$$C_l = C_0 + n(l-1). \tag{8}$$

In DB-ASPP, each $1 \times 1$ convolution layer in front of the atrous convolution layer reduces the depth of the corresponding input feature image to $C_0/4$, and all atrous convolution layers output $C_0/4$. The parameters in DB-ASPP can be computed as written in the following equation:

$$\text{Params} = \sum_{l=1}^{L} \left(c_l \times 0.25c_0 + 0.25k^2 \times nc_0\right), \tag{9}$$

where $L$ is the number of atrous convolution layers in DB-ASPP and $k$ is the size of the convolution kernel to validate the effectiveness of DB-ASPP.

## 4. Experiment and Analysis

*4.1. Parameter Setting and Criteria for Evaluation.* The proposed network model is trained based on the SGD algorithm, and its parameters are given in Table 1. Following the experiment comparison as described above, the PReLU activation function and maximum pooling with proven best effect are selected. For assessment of the generalization ability in transfer learning, the loss function set with 4,200 iterations is used for testing so as to observe the numerical results in the optimization process.

The Mean Intersection over Union (MIoU), Params, and FPs are used to evaluate the performance of model. MIoU is one of the important evaluation indexes in the semantic segmentation model, which measures the advantages and disadvantages of the algorithm by calculating the intersection and union ratio (that is, calculating the ratio between TP and TP + FN + FP). The calculation method is shown in formula (10). Params is the parameter value, and the smaller value means the better lightweight feature of the model and the lower dependence on high-performance equipment. FPS is the number of frames transmitted and recognized per second in semantic segmentation. The higher the T values, the faster speed it means:

$$\text{MIoU} = \frac{1}{k+1} \sum_{i=0}^{k} \frac{p_{ii}}{\sum_{j=0}^{k} p_{ij} + \sum_{j=0}^{k} p_{ji} - p_{ii}}, \tag{10}$$

where $P_{ij}$ is the number of pixels misjudged as class $j$ in class $i$. $P_{ii}$ is the number of pixels predicted correctly.

*4.2. Self-Built Datasets.* The experimental data are modified on the basis of Pascal VOC dataset, adding the road images taken by the author around the campus and removing some small category images in Pascal VOC dataset, such as potted plant and chair. The classification of the self-built datasets is shown in Table 2.

*4.3. Experiment Results.* We conduct three kinds of comparative experiments in order to fully prove the performance of the proposed algorithm. The first is the ablation comparative experiment of DB-ASPP proposed in this paper. The second is the comparison of the experimental results between ESPNet and improved model. The third is the comparison of the improved model and other sever models such as SegNet (a Deep Convolutional Encoder-Decoder Architecture for Image Segmentation).

*4.3.1. Performance Comparison of DB-ASPP Ablation Experiment.* The Pascal VOC verification set is used to conduct ablation experiment. Provided that other parameters are the same, the performance of ASPP, DenseNet, DenseASPP, and DB-ASPP is compared. Table 3 lists the experiment results. Referring to the experiment results, the accuracy of DB-ASPP increases by 0.4% MIoU, 1.1% MIoU, and 2.3% MIoU, respectively, compared with DenseASPP, DenseNet, and ASPP.

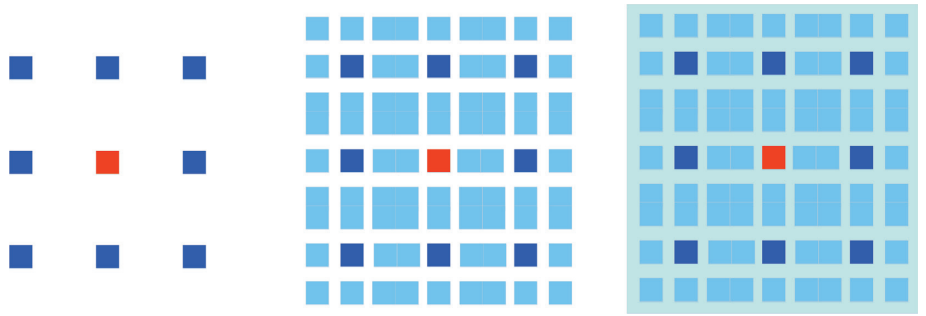

Figure 11: Schematic diagram of HDC.

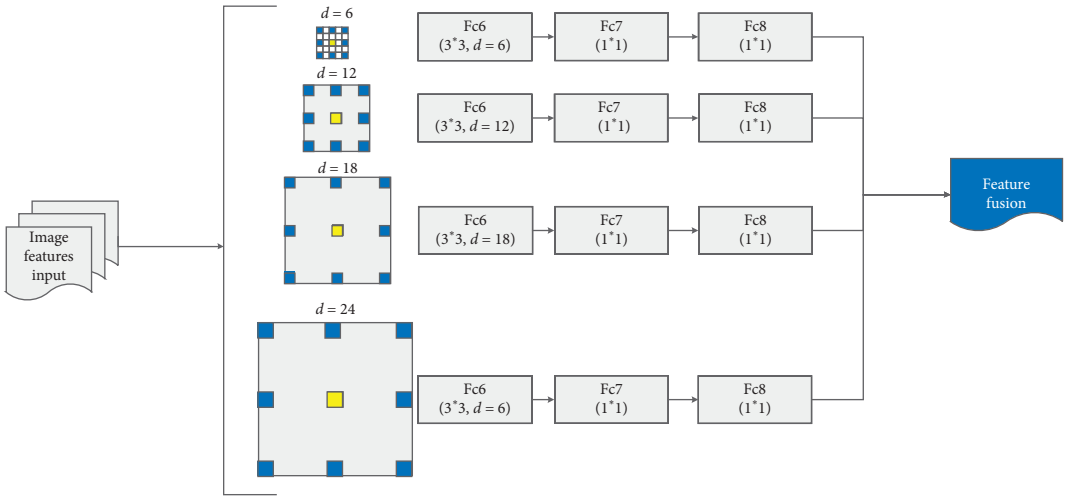


Figure 12: DB-ASPP schematic.

Table 1: SGD parameter setting.

| Description parameter | Parameter setting |
|---|---|
| Learning rate | 0.01 |
| Momentum | 0.8 |
| Weight decay | 4 |

Table 2: The self-built datasets.

| | Train set | | Train set | | Train set 1 | | Train set 1 | |
|---|---|---|---|---|---|---|---|---|
| | Images | Object | Images | Object | Images | Object | Images | Object |
| Aeroplane | 112 | 151 | 126 | 155 | 238 | 306 | 204 | 285 |
| Bicycle | 116 | 176 | 127 | 177 | 243 | 353 | 239 | 337 |
| Bus | 97 | 115 | 89 | 114 | 186 | 229 | 174 | 213 |
| Car | 376 | 625 | 337 | 625 | 713 | 1250 | 721 | 1201 |
| Horse | 139 | 182 | 148 | 180 | 287 | 362 | 274 | 348 |
| Motorbike | 120 | 167 | 125 | 172 | 245 | 339 | 222 | 325 |
| Person | 1025 | 2358 | 983 | 2332 | 2008 | 4690 | 2007 | 4528 |
| Total | 1985 | 3774 | 1935 | 3755 | 3920 | 7529 | 3841 | 7237 |

Table 3: Comparison of experiment results.

| ASPP | DenseASPP | DenseNet | DB-ASPP | MIoU (%) |
|---|---|---|---|---|
| √ | | | | 68.2 |
| | √ | | | 70.1 |
| | | √ | | 69.4 |

Figure 13: The segmentation of ESPNet on self-built datasets.

(a) (b) (c)

Figure 14: Results of the improved ESPNet model on self-built dataset: (a) original image; (b) segmented image; (c) fused image.

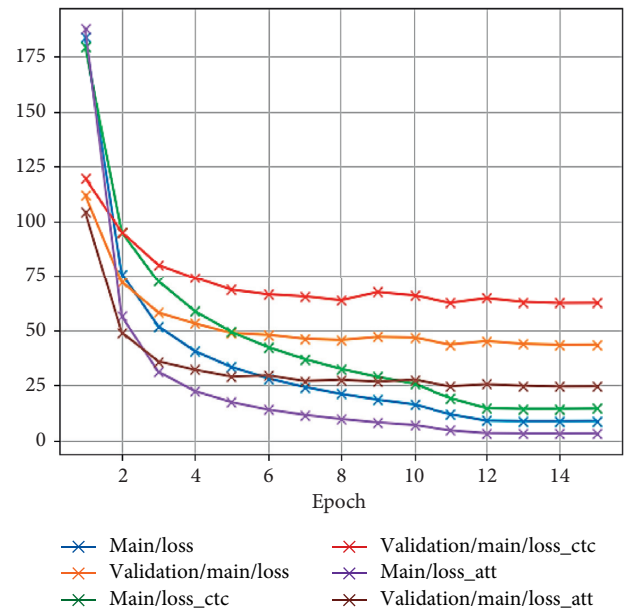


Figure 15: Results of the improved ESPNet model on self-built datasets.

Table 4: Comparison of the experiment results on self-built datasets by category.

| Network model | Buildings | Trees | Sky | Cars | Road | Pedestrians | Cyclists | MIoU (%) | Accuracy (%) |
|---|---|---|---|---|---|---|---|---|---|
| The proposed model | 81.1 | 85.9 | 82.0 | 83.7 | 95.2 | 42.6 | 50.3 | 74.4 | 89.1 |
| SegNet | 78.8 | 79.9 | 90.5 | 83.2 | 92.9 | 51.2 | 40.8 | 73.9 | 88.8 |
| RefineNet | 78.2 | 74.5 | 92.1 | 83.5 | 93.5 | 46.5 | 48.1 | 73.8 | 88.4 |
| DeepLab | 81.5 | 73.2 | 88.3 | 84.1 | 94.9 | 48.4 | 50.1 | 74.4 | 89.1 |
| PSPNet | 80 | 74.2 | 90.5 | 83.1 | 90.5 | 49.7 | 47.3 | 73.6 | 88.3 |
| LRR | 76.4 | 79.3 | 92.5 | 81.5 | 89.7 | 55.2 | 47.5 | 74.6 | 89.2 |
| Dilation-8 | 77.9 | 85.0 | 92.2 | 79.3 | 92.4 | 52.4 | 48.9 | 75.4 | 91.5 |
| FCN-8s | 79.9 | 81.5 | 88.6 | 76.8 | 92.2 | 41.6 | 40 | 71.5 | 84.3 |

Table 5: The self-built datasets.

| Model | The proposed model | SegNet | RefineNet | DeepLab | PSPNet | LRR | Dilation-8 | FCN-8 |
|---|---|---|---|---|---|---|---|---|
| Params | 3.0 K | 29.5 M | 42.6 M | 44.04 M | 65.7 M | 48M | 141.13 M | 134.5 M |
| FPs | 25 | 16 | 12 | 12 | 6 | 9 | 19 | 15 |

*4.3.2. Comparison of ESPNet and Improved ESPNet Model.* In this section, the segmentation results of ESPNet and improved ESPNet on self-result datasets are showed, as well as the loss function of the improved model in the numerical optimization process. As shown in Figures 13–15, respectively, it can be seen from Figures 13 and 14 that the output segmentation images of the improved model are almost consistent with the segmentation standard image, and the output segmentation images of the improved model are also well fused with the original images. It shows that the improved models have good accuracy segmentation and good semantic segmentation effect.

There are 6 loss curves in Figure 15, with an aim of analyzing the loss of different functions in a full scale, so that the experiment results can be accurately optimized.

In addition to the loss function of train sets and the validation set loss function, the attention mechanism loss curve (loss_att) and the time correlation loss function (loss_ctc) with respect to train sets and test sets are configured to detect the model capability to solve and generalize real-time issues.

*4.3.3. Comparison of the Proposed Model and Common Models.* The proposed model is validated on the self-built datasets. Provided with the same memory and calculation condition, its performance is superior to some efficient convolutional neural networks under the standard metrics and introduced performance metrics, with the test results given in Tables 4 and 5.

Table 4 provides a summary of the recognition ability of the proposed model and other seven models for different kinds of objects on the self-built dataset, where the bold figures denote the highest accuracy in the respective category. The MIoU refers to the mean value of the overlapping rates with respect to the target window generated by the

proposed model and the previously marked window. The higher value of this parameter means the higher recognition accuracy.

It can be seen from Table 4 that the recognition accuracy of the proposed model is high in most categories. However, in the Sky and Pedestrian, corresponding value is 82.0 and 42.6, respectively, and the ranking is the penultimate and the penultimate, respectively, which is the obvious shortcomings of the proposed model. The preliminary analysis is due to the fuzzy absence of boundary information in the ablation experiment and data training stage.

In addition, different models are compared for the amount of parameters and real-time performance, as given in Table 5.

Referring to Table 5, the amount of parameters involved in the paper is very small and the recognition and segmentation are fast. This suggests that provided with good accuracy, the proposed model has high real-time performance without the support of strong computing power.

## 5. Conclusion

In this paper, a real-time image semantic segmentation model based on multilabel classification is proposed. The ESPNet model is improved with reference to the characteristics of multilabel classification learning by the following steps: first, the standard convolution is replaced by applying Receptive Field in Deep Convolutional Neural Network in the convolution layer, to the extent that every pixel in the covered area would facilitate the ultimate feature response; second, the ASPP module is improved based on the atrous convolution, the DB-ASPP is proposed as a way to reducing gridding artifacts due to the multilayer atrous convolution, acquiring multiscale information, and integrating the feature information in relation to the image set; finally, subject to extensive tests and comparisons, the proposed model demonstrates smaller number of parameters, faster segmentation, and higher accuracy, compared with other models.

Although the proposed model has improved in real-time and accuracy, there is still a gap compared with the accuracy of non real-time image semantic segmentation model. The next work will focus on improving the accuracy, mainly considering the integration of shallow network in feature information and the optimization of boundary information collection and processing methods.

## Data Availability

The experimental datasets used in this work are publicly available, and the bundled data and code of this work are available from the corresponding author upon request.

## Conflicts of Interest

The authors declare that they have no conflicts of interest.

## Acknowledgments

This work was supported by the National Natural Science Foundation of China under grant nos. 61472348 and 61672455, the Humanities and Social Science Fund of the Ministry of Education of China under grant no.17YJCZH076, Zhejiang Science and Technology Project under grant nos. LGF18F020001 and LGF21F020022, and the Ningbo Natural Science Foundation under grant no. 202003N4324.

## References

[1] S. Minaee, Y. Boykov, F. Porikli, A. Plaza, N. Kehtarnavaz, and D. Terzopoulos, "Image segmentation using deep learning: a survey," *Computer Vision and Pattern Recognition*, vol. 4, no. 1, pp. 1–23, 2020.

[2] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proceedings of the European conference on computer vision*, pp. 818–833, Zurich, Switzerland, September 2014.

[3] T. Akilan, Q. J. Wu, A. Safaei, J. Huo, and Y. Yang, "A 3D CNN-LSTM-Based image-to-image foreground segmentation," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 3, pp. 959–971, 2020.

[4] P. W. Patil, K. M. Biradar, and A. Dudhane, "An end-to-end edge aggregation network for moving object segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8146–8155, Seattle, WA, USA, May 2020.

[5] A. Thangarajah, Q. M. J. Wu, and W. Zhang, "Video foreground extraction using multi-view receptive field and encoder-decoder DCNN for traffic and surveillance applications," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 10, pp. 9478–9493, 2019.

[6] A. G. Howard, M. Zhu, B. Chen et al., "Mobilenets: efficient convolutional neural networks for mobile vision applications," *Computer Vision and Pattern Recognition*, vol. 20, no. 4, pp. 1–9, 2017.

[7] X. Hang, X. Zhou, M. Lin, and J. Sun, "Shufflenet: an extremely efficient convolutional neural network for mobile devices," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6848–6856, Salt Lake City, UT, USA, June 2018.

[8] A. Paszke, A. Chaurasia, S. Kim, and E. Culurciello, *Enet: A Deep Neural Network Architecture for Real-Time Semantic Segmentation*, , pp. 1–10, ICLR, 2017.

[9] M. Tan, R. Pang, and V. L. Quoc, "EfficientDet: scalable and efficient object detection," in *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–10, Seattle, WA, USA, June 2020.

[10] M.-L. Zhang and Z.-H. Zhou, "A review on multi-label learning algorithms," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 8, pp. 1819–1837, 2013.

[11] M.-L. Zhang and Z.-H. Zhou, "ML-KNN: A lazy learning approach to multi-label learning," *Pattern Recognition*, vol. 40, no. 7, pp. 2018–2048, 2007.

[12] G. Tsoumakas and I. Vlahavas, "Random k-labelsets: an ensemble method for multilabel classification," in *Proceedings. of the 18th European conference on Machine Learning*, pp. 406–417, Warsaw, Poland, September 2007.

[13] J. Read, B. Pfahringer, and G. Holmes, "Multi-label classification using ensembles of pruned sets," in *Procceedings of the 8th IEEE International Conference on Data Mining*, pp. 995–1000, Pisa, Italy, December 2008.

[14] J. Read, B. Pfahringer, G. Holmes, and E. Frank, "Classifier chains for multi-label classification," in *Procceedings of the*

*20th European Conference on Machine Learning*, pp. 254–269, Bled, Slovenia, Febuaray 2009.

[15] S. Mehta, M. Rastegari, A. Caspi, L. Shapiro, and H. Hajishirzi, "ESPNet: efficient spatial pyramid of dilated convolutions for semantic segmentation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 561–580, Munich, Germany, September 2018.

[16] J. Kim and Y. S. Heo, "Efficient semantic segmentation using spatio-channel dilated convolutions," *IEEE Access*, vol. 7, pp. 154239–154252, 2019.

[17] N. Nuechterlein and S. Mehta, "3D-ESPNet with pyramidal refinement for volumetric brain tumor image segmentation," in *Proceedings of the 4th International Workshop, BrainLes 2018*, pp. 245–253, Granada, Spain, September 2018.

[18] X. Han and R. Jin, "A Small Sample Image Recognition Method Based on ResNet," in *Proceedings of the 5th International Conference on Computational Intelligence and Applications ICCIA 2020*, pp. 76–81, Beijing, China, June 2020.

[19] R. Jin, G. Chen, A. H. Tung et al., "DIM: a distributed air index based on MapReduce for spatial query processing in road networks," *EURASIP Journal on Wireless Communications and Networking*, vol. 280, pp. 1–15, 2018.