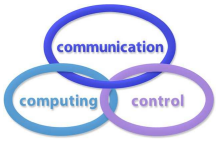
INTERNATIONAL JOURNAL OF COMPUTERS COMMUNICATIONS & CONTROL
Online ISSN 1841-9844, ISSN-L 1841-9836, Volume: 18, Issue: 6, Month: December, Year: 2023
Article Number: 5879, https://doi.org/10.15837/ijccc.2023.6.5879



CCC Publications

An Improved Deeplabv3+ Model for Semantic Segmentation of Urban Environments Targeting Autonomous Driving

Wang Wang,Hua He, Changsong Ma

Wang Wang
Office of Development and Planning
Geely University of China,China
Chengdu , Sichuan, 641423, China
wangwang@guc.edu.cn

Hua He
1. Chongqing Technology and Business University
Chongqing, 400067,China
2. International College
Krirk University
Bangkok, 10220,Thailand
huahe@guc.edu.cn

Changsong Ma
1. Geely University of China,China
Chengdu , Sichuan, 641423, China
2. International College
Krirk University
Bangkok, 10220,Thailand
changsongma@guc.edu.cn

Abstract

This paper proposes an improved Deeplabv3+ model for semantic segmentation of urban scenes targeting autonomous driving applications. A high-quality semantic segmentation dataset is constructed from 2,967 manually labeled aerial images captured at 200m height with a 5-eye camera. The images contain 5 classes - buildings, vegetation, ground, lake and playgrounds. The improved Deeplabv3+ network enriches high-level semantics by replacing max pooling with depthwise separable convolutions. Dilated convolutions extract multi-scale features to avoid overfitting. Experiments demonstrate that the model achieves an overall mean IoU of 0.87 on the test set, with IoU scores of 0.90, 0.92 and 0.94 on buildings, vegetation and water respectively. The model shows promising results for extracting semantic information from complex urban environments to support navigation for autonomous vehicles.

Keywords: Autonomous Driving; Semantic Segmentation; Urban Environments; Improved Deeplabv3+ Model.

1 Introduction

In recent years, the driverless industry has developed rapidly, bringing great convenience to people's travel. Research and application of related aspects have advanced by leaps and bounds. Driverless technology covers many technical fields, such as machinery, control, positioning, mapping, navigation and so on. With the rise of driverless technology and the development of computer vision, the 3D scene map containing rich information provides more possibilities for applications such as autonomous navigation and environmental exploration. Traditional maps only contain spatial geometric information, which limits the ability of robots to perform advanced tasks or better understand the meaning of surrounding scenes. To this end, researchers have carried out a lot of research work, among which the research on semantic maps is an important field in various robot applications such as autonomous robots[1] and driverless cars. Semantic maps means reconstructing the real environment into space and embedding semantic information in the map. Semantic segmentation is an essential step in generating semantic maps, and it is based on deep learning techniques[2, 3]. According to the spatial domain of the algorithm, the research on semantic segmentation can be divided into two-dimensional (2D) semantic segmentation and three-dimensional (3D) semantic segmentation. 2D semantic segmentation assigns a semantic label to each pixel in an image. The deep learning is also widely used in image recognition[4], image classification[5, 6, 7, 8],object detection[9, 10, 11, 12, 13]. In literature [14, 15],probability models such as Markov Random Field (MRF) and Conditional Random Field (CRF) are used for semantic segmentation.In addition, with the development of the Convolutional Neural Network (CNN), several studies have used CNN to solve semantic segmentation problems and achieved significant performance improvements. Long et al [16] ,proposed a representative Fully Convolutional Network (FCN), which can perform pixel-by-pixel classification while preserving the spatial information of the original input image by introducing an upsampling layer of transposed convolutional layers. On this basis, several excellent 2D semantic segmentation network architectures such as ENET [17],SegNet [18], DilatedNet[19] and RefineNet [20]have also been developed . The Pascal VOC2012 [21] and Cityscapes [22]datasets are widely recognized benchmarks for training and evaluating the performance metrics of deep learning models. Ibrahem et al[23],introduced an innovative method employing the efficient sub-pixel convolutional neural network, achieving an mIOU of 91.1% on the PASCAL VOC2012 dataset. Similarly, Wang et al[24], developed a novel, large-scale CNN-based foundational model named InternImage. This model harnesses the benefits of increasing parameters and training data, like ViTs, and elevated the mIoU to 86.1% on the Cityscapes dataset.Qin et al[25],proposed a cost-efficient localization solution that leverages low-cost cameras and compact visual semantic maps rather than relying on expensive sensors and high-resolution maps. Despite challenges with segmentation noise, a strategic application of statistics further refines this light-weight semantic mapping technique, showcasing a practical localization solution for autonomous driving with room for continuous evolution and improvement [26] . Presents a novel approach of utilizing autopilot drones for asphalt surface monitoring to effectively maintain road serviceability, however, practical challenges associated with drone-based road detection, segmentation, and following are yet to be addressed.

Road discontinuity caused by noise and occlusion in the automatic extraction of road information from remote sensing images, were proposed in [27]. Yet, there might be potential limitations or unaddressed complications in handling more complex interferences beyond noise and occlusion. Ref [28],proposed a novel method incorporating atrous convolution into deep learning models for improved semantic image segmentation [25]. The method exhibits significant practical benefits; however, potential limitations are not explicitly addressed in the paper.

The authors of [29] presented OccGAN, a novel technique for generating plausible occluded images with annotation to counteract the shortcomings of long-tail distribution faced during training of intricate driving scenes datasets. The method strives to address the challenges faced by prevalent algorithms when dealing with images encumbered by complicated environments and occlusions. However, there's the potential risk of these synthetically generated samples overwhelming rare real-world instances.

The authors of [30] provided a comprehensive review of advancements in deep learning that have significantly improved the accuracy of semantic image segmentation for autonomous driving, a task that requires high effectiveness and efficiency. Ref [31],proposed a method for enhancing autonomous

driving applications through semantic image segmentation. The research aims to provide advanced AI features, such as automatic brakes and park assist, more affordably in vehicles.Ref [32],presented an innovative framework for reliable image segmentation in low-light scenarios, increasing the safety and practicality of autonomous vehicles. The novel nighttime segmentation framework catapults the applicability of autonomous driving technologies under unfavorable weather conditions via an efficient synthetic data collection and style transfer mechanism. Notwithstanding, the study's approach may encounter limitations stemming from a lack of access to a large-scale, labeled nighttime dataset.

Reviewed current multimodal road-scene segmentation approaches [33], focusing on imaging modalities and datasets. These methods combine inputs to enhance performance, typically employing separate network branches for each modality. Challenges include limited labeled data for training and the need for more diverse datasets, particularly in thermal imaging.Ref [34], presented an effective approach for semantic segmentation for self-driving automobiles. They combined deep learning architectures like convolutional neural networks and autoencoders, as well as cutting-edge approaches like feature pyramid networks and bottleneck residual blocks, to develop the model. Sun et al [35], focused on utilizing semantic segmentation techniques, employing the neural network-based Fully Convolutional Network (FCN) on the cityscapes dataset, to process road condition information. Both safety and real-time processing are considered, aiming for accuracy and processing speed in semantic segmentation.

When researchers in the field of autonomous driving develop a new algorithm, considering safety and cost issues, they will not directly drive a car for testing. However, all scenarios must be tested to ensure that it is stable enough, which takes a lot of time. Finding and reproducing problems through software simulation does not require a real environment and hardware, and can greatly save costs and time. With the rise of deep learning, simulation has new uses in the field of autonomous driving. The autonomous driving platform collects data through simulation, which can greatly increase the training time, far exceeding the time of road tests, and speed up the model iteration speed. Firstly, the cluster training model is used, and then it is tested in the actual road test, and the data-driven method is adopted to carry out the autonomous driving research. However, the quality of the simulation is closely related to the similarity between the simulation model and the real-world scene. The model in this paper can realize the semantic segmentation of two-dimensional pictures through the method of deep learning. At present, there are generally three implementation methods for constructing semantic maps: voxel-based point cloud segmentation [36], direct point cloud segmentation [37] and multi-visual image-based point cloud segmentation [38].

In this work, our primary contributions can be distilled into three pivotal advancements:

We create a high-quality semantic segmentation dataset which is constructed from 2,967 manually labeled aerial images captured at 200m height with a 5-eye camera.

We adopt depthwise separable convolutions, replacing all max-pooling operations, leading to richer high-level semantics.

We utilize dilated convolutions to extract feature maps at arbitrary resolutions, thereby facilitating the capture of multi-scale information. We also add a Batch Normalization (BN) layer and a ReLU layer after each deep convolution operation to mitigate the risk of overfitting.

2 Materials and Methods

Deep learning consists of many layers of neural networks, such as an input layer, an intermediate layer (also known as a hidden layer), an output layer, and so on. A deep neural network refers to a feedforward neural network (also called a multi-layer perceptron) that contains multiple hidden layers, and the nodes of two adjacent layers are fully connected. Generally, the unsupervised pre-training method is used to initialize the network weights. And a classifier is constructed between the last hidden layer and the output layer, such as Logistic Regression (LR), Support Vector Machine (SVM) or SoftMax network, etc. Finally, the weight of the entire network is adjusted through supervised training. The figure below shows a multi-layer perceptron that includes 1 input layer, 1 hidden layer and 1 output layer.

Each layer of the neural network is composed of many neurons. A neuron is actually a computing

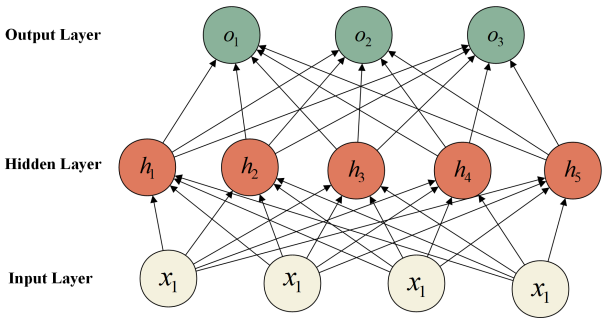


Figure 1: **Multi-Layer Perceptron Model**

unit, which needs to receive N input signals to start computing. These signals are passed to the neuron through a weighted connection, and the neuron starts to calculate a weighted sum of these input signals to obtain a value, and then the neuron processes this value through the activation function to generate the final output. The function of the activation function is to perform nonlinear calculations and squeeze the input values that may vary in a large range into the range of (0, 1) output values. The activation function used below is relu.

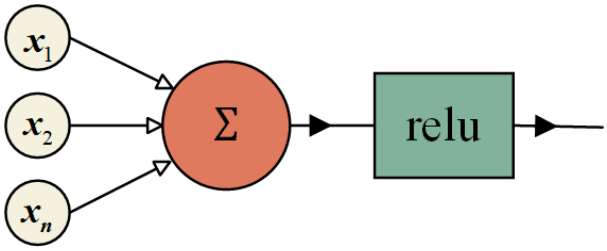


Figure 2: **Structural diagram of a single neuron**

Then the output formula function of the neuron is:

$$h_{\mathbf{W},b}(\mathbf{x}) = f\left(\mathbf{W}^T\mathbf{x}\right) = f\left(\sum_{i=1}^{n} W_i x_i + b\right) \tag{1}$$

For the above neural network, assuming there are m training samples, then for any sample, its loss function is defined as:

$$J\left(W, b; x, y\right) = \frac{1}{2} h_{w,b}\left(x\right) - y^2 \tag{2}$$

In order to prevent the overfitting of the model, it is generally necessary to add a regular term to the loss function. At this time, the loss function is:

$$J = loss + R \tag{3}$$

Among them, the loss represents the loss function, and R represents the regular term. For the above training set containing m samples, using L2 regularization, then the expression of the loss function is:

$$J(\mathbf{W}, \mathbf{b}) = \left[\frac{1}{m}\sum_{i=1}^{m} J\left(\mathbf{W}, \mathbf{b}; \mathbf{x}^{(i)}, \boldsymbol{y}^{(i)}\right)\right] + \frac{\lambda}{2}\sum_{l=1}^{n_l-1}\sum_{i=1}^{s_l}\sum_{j+1}^{s_{l+1}}\left(\boldsymbol{W}_{ij}^{(l)}\right)^2 \tag{4}$$

Our optimization goal is to find the parameters W and b so that the loss function J (W, b) reaches the minimum value. The most commonly used method is the backpropagation algorithm, and the weight update adopts the form of gradient descent.

Autoencoder: The essence of deep learning is the feature learning. The sample features are transformed from the original space to a new feature space through nonlinear transformation. Autoencoder

is an important neural network for feature learning. Autoencoder is an unsupervised deep learning method by classification. It is mainly composed of two parts: an encoder and a decoder. The structure is shown in the figure below. The original input feature vector x is transformed into the encoded output through the mapping of the hidden layer h, and then the reverse transformation through the hidden layer is reproduced to the original input data as much as possible. The connection weight of the network is adjusted by reconstructing the error between the data and the input data. Its simple model is shown in the figure below, where x represents the input, r represents the reconstructed output, h represents the hidden layer, f refers to the encoding process, and g represents the decoding process.

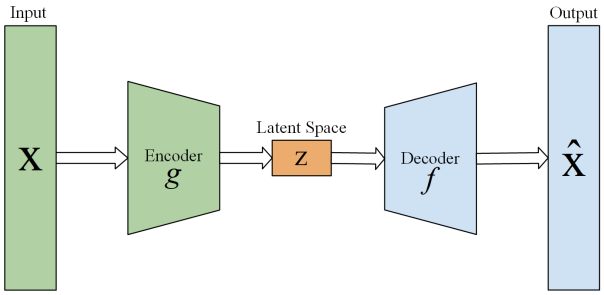


Figure 3: **Autoencoder structure diagram**

The convolutional Neural Network (CNN) is a special neural network structure. The nodes in two adjacent layers are not fully connected. It is generally composed of convolutional layers and pooling layers alternately cascaded. Among them, the convolutional layer generally contains multiple filter factors, and multiple feature subgraphs are obtained by performing convolution operations with the input of the previous layer, so as to realize the feature extraction of the data. The pooling layer realizes feature dimensionality reduction by downsampling the input data of the convolutional layer. Since CNN adopts perceptual field, weight sharing and pooling technology, its network size is greatly reduced compared with the deep neural network. Moreover, the features extracted in this way have the advantage of invariance, so it is widely used in fields such as computer vision and image recognition. At present, the medium is divided into several categories from the perspective of geographical elements: buildings, vegetation, undulating terrain, roads, and playgrounds.

The development process of deep learning always requires a large amount of data to support it. The quality of deep learning algorithms in the field of computer vision has a lot to do with the size and quality of the dataset. High-quality, large-scale datasets work well with algorithms for training and validation. To achieve accurate segmentation of different media in complex urban scenes, a high-quality media semantic segmentation dataset is essential.

At present, open-source image segmentation datasets in the world, such as COCO, VOC, and Cityscapes, cannot be used for media recognition of UAV detection pictures. Forcible use will lead to a decrease in accuracy or even failure to correctly complete the identification of the medium. Therefore, in order to maximize the accuracy of media semantic segmentation, we can only label and establish a semantic recognition dataset that can realize object segmentation.

After referring to the establishment specifications of many semantic segmentation datasets, we proposed the following media semantic segmentation dataset process, which specifically includes: UAV detection image data collection, data preprocessing and artificial media semantic annotation, as shown in the figure below.

2.1 The Improved DeepLabv3+ Algorithm

DeepLabV3+ is gradually developed from DeepLab and is the fourth version of this series of algorithms. The DeepLab semantic segmentation algorithm uses the VGG network to extract image features, uses dilated convolution to expand the receptive field, and uses the conditional random field commonly used in traditional segmentation algorithms to optimize the final segmentation results. Based on DeepLab and inspired by Spatial Pyramid Pooling (SPP), DeepLabV2 proposes Atrous Spatial Pyramid Pooling (ASPP) module. This module extracts rich image contextual information

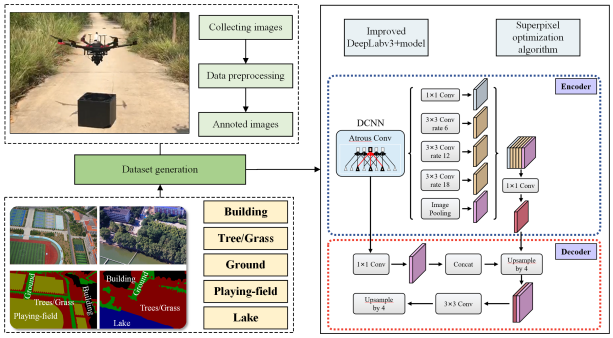


Figure 4: **The process of semantic segmentation algorithm**

at multiple scales by parallel sampling using dilated convolutions with different dilation rates. In addition, DeepLabV2 uses the ResNet-101 network with stronger feature extraction capabilities to replace the VGG-16 network used by DeepLab, and obtains more excellent semantic segmentation performance. Subsequently, DeepLabV3 further optimized the model on the basis of DeepLabV2, improved the ASPP structure of DeepLabV2, and obtained better semantic segmentation results by improving the training strategy. On the basis of the above, DeepLabV3+ was proposed in March 2018. The algorithm is based on DeepLabV3, uses the DeepLabV3 structure as an encoder, and replaces ResNet-101 with the Xception model in the feature extraction stage, further improving the accuracy and speed of operation of the image semantic segmentation algorithm. In addition, DeepLabV3+ upsamples the output features of the ASPP module by four times, and then concatenates the features with the low-level features in the feature extractor with the same resolution, and then the cascaded features are quadruple upsampled again in the way of bilinear upsampling to get the final semantic segmentation result. DeepLabV3+ is used to solve the problem of loss of a large amount of detailed information caused by directly upsampling the feature map to restore the original image resolution size in the DeepLabV3 model by adding a decoding module, so that the model has achieved top results in semantic segmentation tasks on multiple datasets. The network structure diagram of DeepLabV3+ is shown in Figure 5.

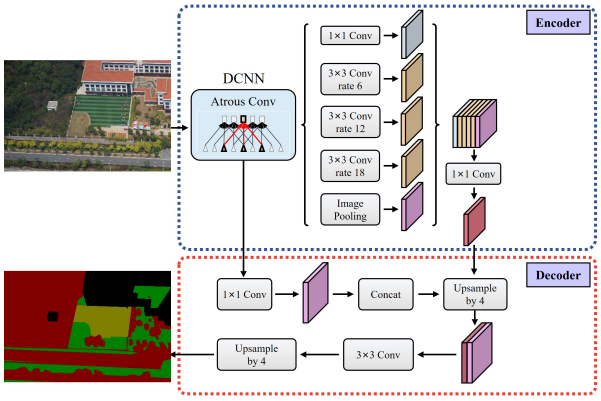


Figure 5: **DeepLabv3+ network structure d**

The DeepLabV3+ semantic segmentation algorithm uses the Xception network pre-trained by ImageNet for feature extraction, and applies the Depthwise Separable Convolution structure to the ASPP module and decoding module. The Xception network is another improvement to Inception-V3 proposed after Inception and Inception-V2. It mainly introduces Depthwise Separable Convolution on the basis of Inception-V3, and improves the performance of the model without increasing the complexity of the network. Among them, Depthwise Separable Convolution is also called separable convolution. For the correlation between channels, the model first performs depthwise convolution, that is, depth convolution, performs a 3*3 convolution operation on each input channel, and concatenates the results; then for spatial correlation, the model performs pointwise convolution, that is, point-by-point convo-

lution, and perform a 1*1 convolution operation on the concatenated results in the depth convolution. The structure of Xception is based on ResNet, and the entire network is divided into three parts, namely Entry flow, Middle flow and Exit flow. The Xception network structure and its improvements in DeepLabV3+ are shown in Figure 6.

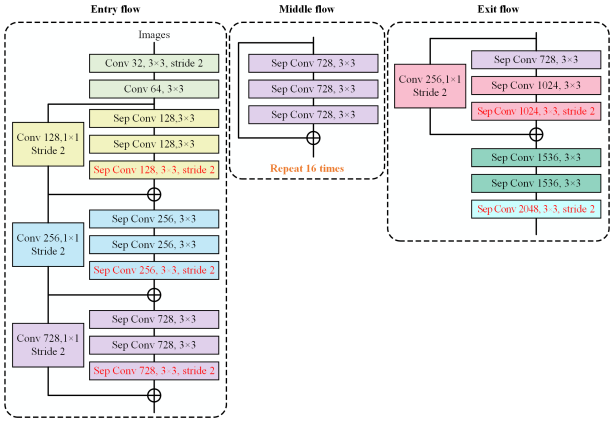


Figure 6: **Schematic diagram of Xception and its improvements in Deeplabv3+**

2.2 Dilated convolution

The DeepLabV3+ semantic segmentation algorithm uses dilated convolution to control the resolution of the output feature map and expand the receptive field of the convolution kernel, reducing the downsampling rate while keeping the training parameters unchanged. Dilated convolution, also known as atrous convolution, extends the standard convolution operation by inserting a value of 0 in the convolution kernel. Taking the two-dimensional feature map as an example, assuming that the convolution kernel is w, when the dilated convolution is applied to the input feature map x, for each position i in the output feature map y, there are:

$$y\left(i\right)=\sum_{k}x\left[i+r*k\right]w\left[k\right] \tag{5}$$

Among them, r represents the atrous rate. A schematic diagram of dilated convolution is shown in Figure 2.7. Among them, Figure 2.7(a) corresponds to the dilated convolution with a convolution kernel of 3*3 and an atrous rate of 1. Like ordinary convolution, the size of the receptive field is 3*3; Figure 2.7(b) corresponds to the dilated convolution with an atrous rate of 2 of 3*3, and a value of 0 is inserted between ordinary convolutions, and the size of the receptive field is 5*5; Figure 2.7(c) corresponds to a dilated convolution with a 3*3 atrous rate of 4, and the receptive field size is 9*9.

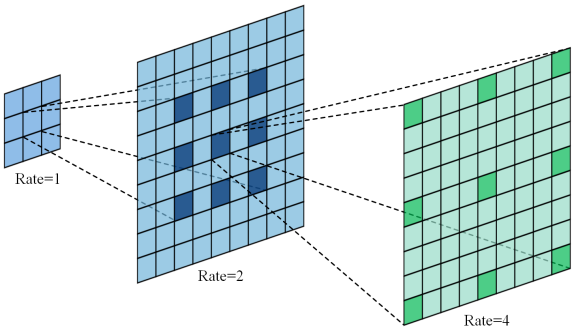


Figure 7: **Schematic diagram of dilated convolution**

2.3 ASPP Module

The DeepLabv3+ semantic segmentation algorithm uses the dilated spatial pyramid pooling ASPP module. It uses parallel sampling of dilated convolutions with different atrous rates and combines

image-level features to mine rich media semantic context information at different scales to solve the problem of different scales of the media to be segmented.

2.4 Deficiencies of DeepLabv3+

Although the DeepLabV3+ semantic segmentation algorithm has achieved great success in various image semantic segmentation datasets, there are still the following deficiencies: DeepLabV3+ and other semantic segmentation algorithms based on deep learning usually increase the receptive field by continuously stacking pooling layers or downsampling layers in the feature extraction stage, so many details such as object edges are lost in the convolution process. The ratio between the spatial resolution of the input image and the resolution of the final output feature map is defined as the Output Stride. Even though the DeepLabV3+ algorithm proposes to generate intensive pixel prediction by using dilated convolution instead of a pooling layer, the resolution of the encoded output feature map is still reduced by 16 times compared with that of the input image. That is, the Output Stride=16. Under this premise, DeepLabV3+ first performs bilinear upsampling of the encoded features with a factor of 4 and then connects the feature layers output by the low-level network with the same spatial resolution. Finally, a 3*3 convolution kernel and bilinear upsampling with a factor of 4 are used to restore the output feature map to the spatial resolution of the input image, thus completing the semantic segmentation. Nevertheless, a lot of detailed information is still lost in this process. Only relying on two consecutive bilinear upsampling with a factor of 4 is not enough to fully restore the detailed information of the image, resulting in an unsatisfactory semantic segmentation effect in detail parts such as object edges. Models with multiple pooling layers that are deeper in architecture enhance translational invariance and expand the receptive field of the top-layer nodes. However, such structures may lead to the loss of high-frequency details, which are crucial for the precise localization of object edges. This loss of edge details is fatal both for point cloud classification for media reconstruction and for improving the correct semantics for subsequent calculations. Therefore, we must improve it on the basis of the DeepLabV3+ model to fully obtain the edge information of the medium.

Based on the above, we adopt a method that directly performs semantic segmentation on multi-visual images of UAVs. The neural network structure of Deeplab V3+ (combination of codec and Xception) is the method we use to implement semantic segmentation in this paper [39]. We use an improved DeepLabv3+ deep learning network to carry out semantic recognition on UAV aerial survey pictures and use UAV with five-eye camera to obtain aerial survey pictures. It is a key step to build a semantic map by manually labelling it to give it semantic information, and continuously training the algorithm to obtain the media category to which each pixel of each aerial survey image belongs. Traditional pooling operations, while reducing spatial dimensions, can also lead to the loss of fine-grained spatial details. Depthwise separable convolutions, on the other hand, allow for a reduction in computational complexity without significantly downsampling the feature maps. This ensures that more spatial details are retained, which is crucial for tasks that require precise localization or where fine-grained patterns are important. The improved deep learning network model enriches the high-level semantics of the original network. Multi-scale information is easier to be extracted and overfitting is avoided. Finally, the semantic segmentation task of the medium is completed accurately, and the semantic recognition of urban scenes suitable for automatic driving is realized.

3 Experiment and Results

The images in our dataset were sourced from two universities in China, namely Wuhan University and Xiamen University, during the summer of 2020. Throughout this period, we accumulated a collection of over 5,000 images. The quality of the media semantic segmentation dataset established by surveying and mapping images will be affected by many factors: the height, angle, resolution, content, and quantity of the image when acquiring the image. These factors will have different and non-negligible effects on the subsequent semantic segmentation. Considering the above factors comprehensively, it is determined that the collection height of the dataset is 200m; the angle is the shooting angle of the five-eye camera; and the resolution is . The dataset has a total of 2,967 pictures, including 2,067 training sets, 500 validation sets, and 400 test sets. There are five categories:

buildings, vegetation, ground, lakes, and playgrounds. The following is a detailed introduction to the consideration angle of each category.

(1) Height

The height of the image acquisition is an important factor affecting the quality of surveying and mapping images. If the height is too high, the proportion of pixels occupied by small objects such as branches of trees, vehicles, and doors and windows on buildings will be too low to even be recognized, which will cause great errors in the subsequent semantic segmentation of the medium. And if it is too low, it will not be able to detect the whole picture of taller buildings, and even cause the drone to crash into trees or buildings. Here the height is chosen as 200m.

(2) Angle

The angle at which images are taken also affects the quality of the dataset. If you shoot vertically, then a lot of side information will be ignored, such as doors and windows of buildings, etc. In order to obtain this side information, the angles of the surveying and mapping pictures should be varied. Here, we can use the five-eye camera to take pictures to obtain pictures containing media data. The five-eye camera can collect images from one vertical angle and four side-view angles at the same time, which more truly reflects the actual situation of ground objects and makes up for the deficiency of orthophoto images.

(3) Resolution

The resolution of surveying and mapping pictures is also an important factor that cannot be ignored. If the resolution is too low, small objects will occupy too few pixels, making recognition extremely difficult. If it is too high, it will take too much time to process the data, and it may even fail to train smoothly due to insufficient memory. Therefore, is formulated here as the standard for dataset resolution.

(4) Quantity

The size of the dataset is also an issue worth discussing. If the dataset is too small, that is, the number of pictures contained is too small, which will lead to overfitting in the subsequent training. Since manual labelling is a very time-consuming process, and the dataset is too large, that is, it contains too many pictures, it will also lead to too long training time and cannot build our map in real time. In the end, a total of 2,567 pictures were selected in the training set of the dataset to train the convergence of the model. A total of 400 pictures were selected in the test set to verify and test the accuracy of the model.

(5) Category

In order to avoid the imbalance of media categories affecting the training of semantic segmentation, the proportion of pixels occupied by all categories should be roughly the same, and there should be no too many or too few pixels occupied by a certain category. After the data is successfully collected, we need to preprocess the collected data before using the labelme software to manually label the media.

3.1 Data Preprocessing

Before deep learning performs image classification, target detection, and image semantic segmentation, image data needs to be preprocessed. The two commonly used preprocessing methods are image standardization processing and image normalization processing. The two image preprocessing methods are described in detail below.

(1) Image standardization processing

The image standardization process uses the theoretical knowledge of two parts, convex optimization theory and data probability division. The data in the image is de-averaged to complete the centralization process. The data centralization satisfies the data distribution law, and it is easier to obtain the generalization effect after training. The image normalization processing formula is as follows:

$$adjusted_stddev = \max(\delta, \frac{1.0}{\sqrt{N}}) \tag{6}$$

$$image\_standardization = \frac{x - \mu}{adjusted\_stddev} \tag{7}$$

Table 1: **Proportion of pixels occupied by categories**

Category	Buildings	Vegetation	Ground	Lake	Playground
The percentage of pixels occupied	24.03%	49.24%	16.11%	6.24%	4.28%

In the formula, represents the standard deviation, is the standard deviation, N is the number of pixels in the image x, is the mean value of the image, x is the image matrix, and is the standardized processing result of the image.

(2) Image normalization processing

The commonly used method for normalization processing is the maximum and minimum value normalization method, and the formula is as follows:

$$norm = \frac{x_i - min(x)}{max(x) - min(x)} \tag{8}$$

In this formula, represents the image pixel point value, min(x) and max(x) represent the minimum value and maximum value of the image pixel, respectively. The normalized image result is exactly the same as the original image. Normalization does not change the information of the original image, but only changes the value range, changing the value range of the pixel value of the original image from 255 to 0-1. This is helpful for training the data later using the deep learning network. Considering that the task is based on the medium segmentation of UAV detection images, the normalization processing of images is suitable for image classification and target detection tasks. However, the semantic segmentation of images needs to process the overall information of the images, so the normalization processing method of the image is used to preprocess the dataset.

3.2 Dataset Labelling

After the data image is acquired, it is necessary to manually label the data image. Since image semantic segmentation is a pixel-level segmentation task, manual data labelling is a time-consuming and laborious process. It took 26 days to complete the labelling task of the media semantic recognition datasets with a total of 2,967 pictures.

The statistics include a total of 2,967 pictures in the training set and the test set, and the ratio of pixels occupied by each category is obtained, as shown in the table.

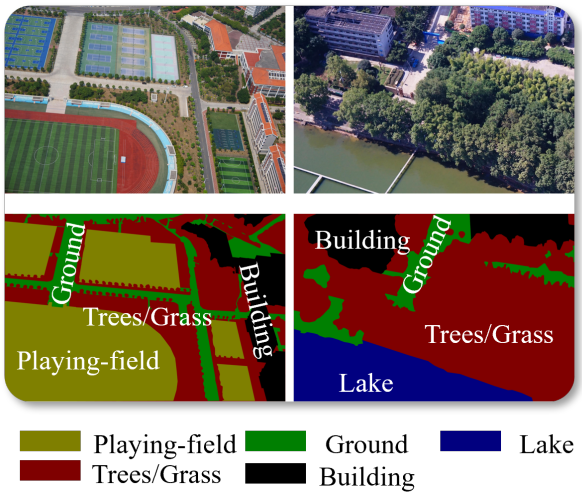


Figure 8: **Examples of manually labelled images**

After the labeling is completed using labelme annotation software, the labled information is stored in the form of.json file. These json files are then sorted together and labled diagrams are generated. The color of each category in the generated label image is inconsistent, and a unified label map will be generated after processing. An example label diagram is shown in Figure 8.

The Softmax classifier in the last layer of the medium semantic segmentation network used will classify the model with one-dimensional values based on probability. Therefore, before the sample is input into the network, it is necessary to convert the data format of the sample image whose colour space is RGB, and convert the RGB grayscale of 3 channels into the grayscale value of 1 channel.

Since the gray value of each category is very low, no change can be seen directly from the grayscale image, so it is not shown here. However, during training, the network model will learn according to the gray value of each category in the figure to understand the category to which the pixel belongs. After the manual labelling is completed, our dataset can be divided into three folders: image, mask, and index, as shown in Figure 9.

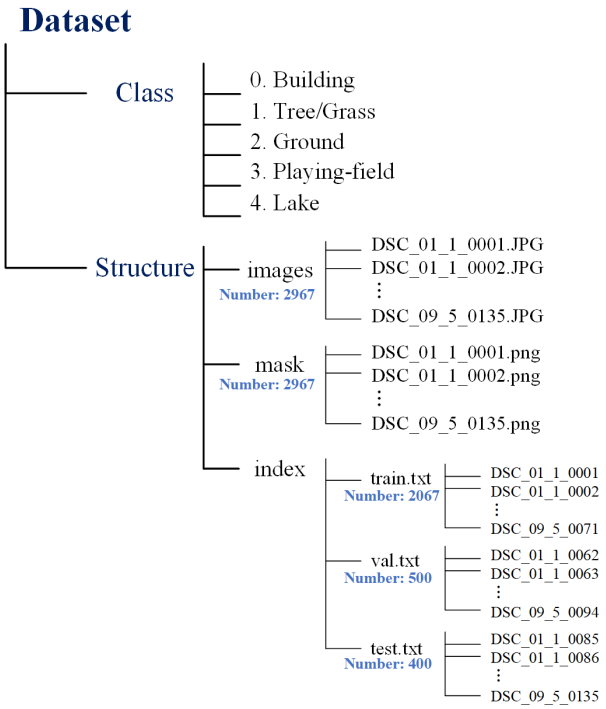


Figure 9: **The directory display of the generated media semantic recognition dataset**

Among them, the image stores all input pictures, including training, testing, and validation set pictures, as shown in Figure 10. All the label pictures are stored in the mask, and there is a one-to-one correspondence with the input pictures (that is, the pictures in the image folder), as shown in Figure 11. The index directory contains three .txt files: train.txt, val.txt, and test.txt, which contain the file names of all training sets, validation sets, and test sets, respectively, and are used to distinguish training sets, validation sets, and test sets.

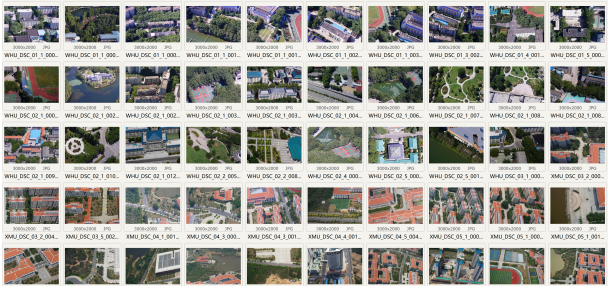


Figure 10: **The original images in our annotated dataset**

The final improved network model is shown below.

The parameters are shown in Table 2, and the loss function of the training process is shown in Figure 12.

.

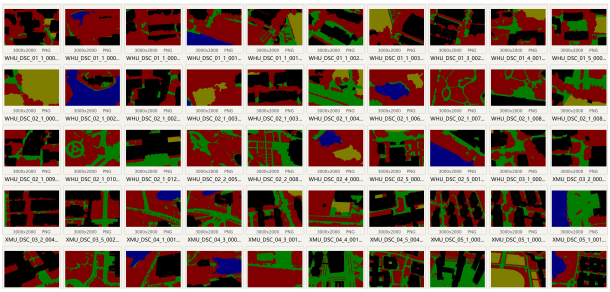


Figure 11: **The segmented image in our annotated datase**

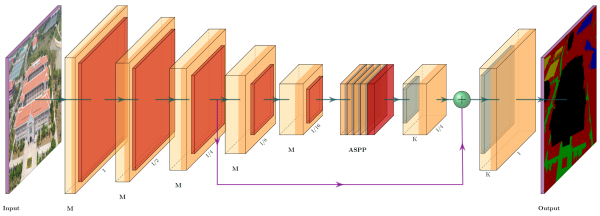


Figure 12: **The final improved network model**

It can be seen from Figure 13 that after 100,000 iterations, the loss value of the model stabilizes around 0.4, and the model converges successfully. To provide a clearer perspective on the dataset's quality and the precision of our annotations, Figure 14 presents a side-by-side comparison of an original image, ground truth and prediction generated by our improved deepLabv3+ model.

The trained model was used to verify the dataset data, and the average pixel ratio (m_IOU) of each category in the test set data was obtained, as shown in Table 3.

It can be seen that IOU of lakes, buildings and vegetation is as high as 0.9, and the accuracy of segmentation is very high. As per the literature, particularly cited in [25], the leading metric for urban scene segmentation pertinent to autonomous driving is an mIoU of 83.4%, achieved and validated on the Cityscapes dataset. Our proposed method has demonstrated promising results, with an mIoU of 86.7% on our assembled dataset. However, we acknowledge that direct comparisons should be interpreted cautiously due to the intrinsic differences in dataset characteristics, evaluation conditions, and protocols. Although our dataset is meticulously curated and relevant to the tasks at hand, the disparities between datasets may lead to variations in performance metrics. Future endeavors in our research will aim at conducting a comprehensive comparative analysis with uniform datasets and evaluation criteria to facilitate a more accurate and fair assessment against the current state-of-the-art methods in urban scene segmentation for autonomous driving applications.

Visualize the segmentation results of our model, as shown in Figure 15.

In acknowledging the limitations of the present model, it is crucial to note that it yields optimal results predominantly with images of higher resolution, as these provide finer details imperative for accurate segmentation. Performance may decline with lower resolution inputs due to the consequential loss of these intricate details. Furthermore, while our model is adept at handling various capture angles, having been trained with a diverse dataset, it might exhibit inconsistencies under extreme or unanticipated weather and lighting scenarios due to limited training exposure to such conditions.

Addressing the model's limitations presents substantial avenues for future work. Engaging with simulated images during the training step is a viable strategy. The use of simulated images would allow

Table 2: **Improved DeepLabv3+ network parameters**

Parameters	Size
Input size	2000×3000
Output size	2000×3000
Total layers of the network	69
Total number of nodes	1.7×10^{10}

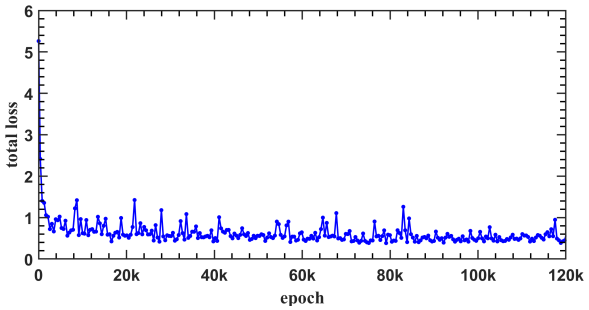


Figure 13: **Loss function change curve during model training**

Figure 14: **Sample results of ground truth and predicted**

the model to encounter and learn from a broader set of conditions, enhancing its adaptability and performance across diverse scenarios not present in the current dataset. Moreover, the incorporation of 3D point cloud data is another promising direction. By integrating depth information provided by 3D data, our model could achieve a more nuanced understanding of the scenes, facilitating improved accuracy in image segmentation tasks, particularly in environments characterized by their complexity and dynamism.

4 Conclusion

In conclusion, this paper presents an improved DeepLabv3+ model for semantic segmentation tailored for urban driving scenes. A high-quality dataset of aerial images captured specifically for this application is constructed. The model enriches semantics and extracts multi-scale features through several enhancements to the DeepLabv3+ architecture. Extensive experiments demonstrate state-of-the-art performance with mean IoU of 0.87, showing potential to provide the semantic maps needed for environment perception in autonomous vehicles. Future work includes expanding the dataset diversity, investigating multi-modal inputs, and evaluating on additional urban scene benchmarks.

Funding

This research was funded by Chengdu Key Research Base of Philosophy and Social Sciences (Project No. CXZL202307); Humanities and Social Science Fund of Ministry of Education, China (Project No. 21XJA630004); Supported by Sichuan Science and Technology Program, China (Project No. 2023JDR0194).

Author contributions

The authors contributed equally to this work.

Table 3: **Validation results on val and test dataset**

Category	IoU on val set	IoU on test set
Building	91.5%	90.3%
Tree/Grass	92.2%	92.0%
Ground	70.0%	69.2%
Playground	88.3%	87.2%
Lake	95.7%	94.4%
Over_all	87.5%	86.6%

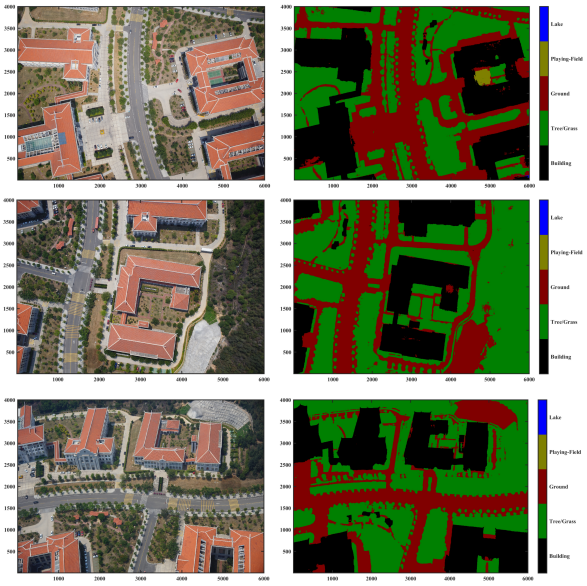


Figure 15: **Schematic diagram of segmentation results**

Conflict of interest

The authors declare no conflict of interest.

References

[1] A. Nüchter and J. Hertzberg, "Towards semantic maps for mobile robots," *Robotics and Autonomous Systems*, vol. 56, pp. 915–926, Nov. 2008.

[2] L. Teng and Y. Qiao, "BiSeNet-oriented context attention model for image semantic segmentation," *ComSIS*, vol. 19, no. 3, pp. 1409–1426, 2022.

[3] F. Zeng, B. Yang, M. Zhao, Y. Xing, and Y. Ma, "MASANet: Multi-Angle Self-Attention Network for Semantic Segmentation of Remote Sensing Images," *Tehnički Vjesnik*, vol. 29, pp. 1567–1575, Jan. 2022. Publisher: Faculty of Mechanical Engineering in Slavonski Brod, Faculty of Electrical Engineering in Osijek, Faculty of Civil Engineering in Osijek.

[4] J. Zhang, X. Yu, X. Lei, and C. Wu, "A novel deep LeNet-5 convolutional neural network model for image recognition," *Computer Science and Information Systems*, vol. 19, pp. 36–36, Jan. 2022.

[5] X. Ma, Z. Li, and L. Zhang, "An Improved ResNet-50 for Garbage Image Classification," *Tehnički Vjesnik*, vol. 29, pp. 1552–1559, Jan. 2022. Publisher: Faculty of Mechanical Engineering in Slavonski Brod, Faculty of Electrical Engineering in Osijek, Faculty of Civil Engineering in Osijek.

[6] M. Ozkahraman and M. Ozkahraman, "Artificial Intelligence in Foreign Object Classification in Fenceless Robotic Work Cells Using 2-D Safety Cameras | Request PDF."

[7] L. Teng and Y. Q. , "Classification of Beef by Using Artificial Intelligence."

[8] D. M. Asriny and R. Jayadi, "Transfer Learning VGG16 for Classification Orange Fruit Images," vol. 13, no. 1, 2023.

[9] S. Sarp and M. Kuzlu, "[PDF] A comparison of deep learning algorithms on image data for detecting floodwater on roadways | Semantic Scholar."

[10] Y. Kim, H. Song, J. Han, and Konyang, "A Deepfake-Based Deep Learning Algorithm for Medical Data Manipulation Detection," 2022.

[11] Y. Lee, "A Study on Abnormal Behavior Detection in CCTV Images through the Supervised Learning Model of Deep Learning," vol. 9, no. 2, 2022.

[12] M. M. Gomaa, E. R. Mohamed, A. M. Zaki, and A. Elnashar, "Deep Learning to Detect Image Forgery Based on Image Classification," vol. 12, no. 6, 2022.

[13] Y.-H. Cho and M.-D. Shahbe, "Vision-based In-room Fall Detection Application," vol. 9, no. 4, 2022.

[14] B. L. G. Floros, "[PDF] Joint 2D-3D temporally consistent semantic segmentation of street scenes | Semantic Scholar."

[15] D. L. and F. Jurie, "Combining appearance models and Markov Random Fields for category level object segmentation | IEEE Conference Publication | IEEE Xplore."

[16] J. Long, E. Shelhamer, and T. Darrell, "Fully Convolutional Networks for Semantic Segmentation," Mar. 2015. arXiv:1411.4038 [cs].

[17] A. Paszke, A. Chaurasia, S. Kim, and E. Culurciello, "ENet: A Deep Neural Network Architecture for Real-Time Semantic Segmentation," June 2016. arXiv:1606.02147 [cs].

[18] V. B. A. K. R. Cipolla, "SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation | IEEE Journals & Magazine | IEEE Xplore."

[19] F. Yu and V. Koltun, "Multi-Scale Context Aggregation by Dilated Convolutions," Apr. 2016. arXiv:1511.07122 [cs].

[20] G. L. A. M. C. S. I. Reid, "RefineNet: Multi-path Refinement Networks for High-Resolution Semantic Segmentation | IEEE Conference Publication | IEEE Xplore."

[21] L. V. G. C. K. I. W. J. W. A. Z. Mark Everingham, S. M. Ali Eslami, "The PASCAL Visual Object Classes Challenge: A Retrospective — University of Edinburgh Research Explorer."

[22] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The Cityscapes Dataset for Semantic Urban Scene Understanding," Apr. 2016. arXiv:1604.01685 [cs].

[23] H. Ibrahem, A. Salem, and H.-S. Kang, "DTS-Net: Depth-to-Space Networks for Fast and Accurate Semantic Object Segmentation," *Sensors*, vol. 22, Jan. 2022.

[24] W. Wang, J. Dai, Z. Chen, Z. Huang, Z. Li, X. Zhu, X. Hu, T. Lu, L. Lu, H. Li, X. Wang, and Y. Qiao, "InternImage: Exploring Large-Scale Vision Foundation Models with Deformable Convolutions," Nov. 2022.

[25] T. Qin, Y. Zheng, T. Chen, Y. Chen, and Q. Su, "RoadMap: A Light-Weight Semantic Map for Visual Localization towards Autonomous Driving," June 2021. arXiv:2106.02527 [cs].

[26] H. Ranjbar, P. Forsythe, A. A. F. Fini, M. Maghrebi, and T. S. Waller, "Addressing practical challenge of using autopilot drone for asphalt surface monitoring: Road detection, segmentation, and following," *Results in Engineering*, vol. 18, p. 101130, June 2023.

[27] H. T. , H. Xu, and J. Dai, "BSIRNet: A Road Extraction Network with Bidirectional Spatial Information Reasoning."

[28] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs," *IEEE Trans Pattern Anal Mach Intell*, vol. 40, pp. 834–848, Apr. 2018.

[29] Y. Wang, L. Mo, H. Ma, and J. Yuan, "OccGAN: Semantic image augmentation for driving scenes," *Pattern Recognition Letters*, vol. 136, pp. 257–263, Aug. 2020.

[30] I. Papadeas, L. Tsochatzidis, A. Amanatiadis, and I. Pratikakis, "Real-Time Semantic Image Segmentation with Deep Learning for Autonomous Driving: A Survey," *Applied Sciences*, vol. 11, p. 8802, Sept. 2021.

[31] G. L. A. M. C. S. I. Reid, "SegFast-V2: Semantic image segmentation with less parameters in deep learning for autonomous driving — Manipal Academy of Higher Education, Manipal, India."

[32] H. Wang, Y. Chen, Y. Cai, L. Chen, Y. Li, M. A. Sotelo, and Z. Li, "SFNet-N: An Improved SFNet Algorithm for Semantic Segmentation of Low-Light Autonomous Driving Road Scenes," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, pp. 21405–21417, Nov. 2022. Conference Name: IEEE Transactions on Intelligent Transportation Systems.

[33] G. Rizzoli, F. Barbato, and P. Zanuttigh, "Multimodal Semantic Segmentation in Autonomous Driving: A Review of Current Approaches and Future Perspectives," *Technologies*, vol. 10, p. 90, July 2022.

[34] Q. S. , R. Priyadarshini, and A. Vidyarthi, "Intelligent Semantic Segmentation for Self-Driving Vehicles Using Deep Learning."

[35] H. S. and T. Wang, "Semantic segmentation in autonomous driving—an example of FCN | Proceedings of the 2023 7th International Conference on Innovation in Artificial Intelligence."

[36] L. P. Tchapmi, C. B. Choy, I. Armeni, J. Gwak, and S. Savarese, "SEGCloud: Semantic Segmentation of 3D Point Clouds," Oct. 2017. arXiv:1710.07563 [cs].

[37] R. Q. C. H. S. M. K. L. J. Guibas, "PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation | IEEE Conference Publication | IEEE Xplore."

[38] A. Boulch, J. Guerry, B. Le Saux, and N. Audebert, "SnapNet: 3D point cloud semantic labeling with 2D deep segmentation networks," *Computers & Graphics*, vol. 71, pp. 189–198, Apr. 2018.

[39] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation," in *Computer Vision – ECCV 2018* (V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, eds.), Lecture Notes in Computer Science, (Cham), pp. 833–851, Springer International Publishing, 2018.

Journal's webpage: http://univagora.ro/jour/index.php/ijccc/

This journal is a member of, and subscribes to the principles of,
the Committee on Publication Ethics (COPE).
https://publicationethics.org/members/international-journal-computers-communications-and-control

Cite this paper as:

Wang Wang ; Hua He; Changsong Ma (2023). An Improved Deeplabv3+ Model for Semantic Segmentation of Urban Environments Targeting Autonomous Driving, *International Journal of Computers Communications & Control*, 18(6), 5879, 2023.

https://doi.org/10.15837/ijccc.2023.6.5879